Inner Voice Experiences During Processing of Direct and Indirect Speech

Bo Yao and Christoph Scheepers

Abstract In this chapter, we review recent research concerned with "inner voice" experiences during silent reading of direct speech (e.g., *Mary said, "This dress is beautiful!"*) and indirect speech (e.g., *Mary said that the dress was beautiful*). Converging findings from speech analysis, brain imaging, and eye tracking indicate that readers spontaneously engage in mental simulations of audible-speech like representations during silent reading of direct speech, and to a much lesser extent during silent reading of indirect speech. This "simulated" implicit prosody is highly correlated with the overt prosody generated during actual speaking. We then compare this "simulated" implicit prosody with the sort of "default" implicit prosody that is commonly discussed in relation to syntactic ambiguity resolution. We hope our discussion will motivate new interdisciplinary research into prosodic processing during reading which could potentially unify the two phenomena within a single theoretical framework.

Keywords Implicit prosody · Inner voice experience · Direct speech · Indirect speech · Reading · Mental simulation · Embodied cognition · fMRI · Eye tracking

1 Overview

In this chapter, we review a new body of research on language processing, focussing particularly on the distinction between direct speech (e.g., *Mary said, "This dress is absolutely beautiful!"*) and indirect speech (e.g., *Mary said that the dress was absolutely beautiful*).

First, we will discuss an important pragmatic distinction between the two reporting styles and highlight the consequences of this distinction for prosodic processing.

B. Yao (✉)
School of Psychological Sciences, University of Manchester, Manchester, UK
e-mail: Bo.Yao@manchester.ac.uk

C. Scheepers
Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, UK
e-mail: Christoph.Scheepers@glasgow.ac.uk

L. Frazier, E. Gibson (eds.), *Explicit and Implicit Prosody in Sentence Processing,*
Studies in Theoretical Psycholinguistics 46, DOI 10.1007/978-3-319-12961-7_15

While direct speech provides vivid *demonstrations* of the reported speech act (informing recipients about *how* something was said by another speaker), indirect speech is more descriptive of *what* was said by the reported speaker. This is clearly reflected in differential prosodic contours for the two reporting styles during speaking: Direct speech is typically delivered with a more variable and expressive prosody, whereas indirect speech tends to be used in combination with a more neutral and less expressive prosody.

Next, we will introduce recent evidence in support of an "inner voice" during language comprehension, especially during silent reading of direct speech quotations. We present and discuss a coherent stream of research using a wide range of methods, including speech analysis, functional magnetic resonance imaging (fMRI), and eye tracking. The findings are discussed in relation to overt (or "explicit") prosodic characteristics that are likely to be observed when direct and indirect speech are used in spoken utterances (such as during oral reading). Indeed, the research we review here makes a convincing case for the hypothesis that recipients spontaneously activate voice-related mental representations during silent reading, and that such an "inner voice" is particularly pronounced when reading direct speech quotations (and much less so for indirect speech). The corresponding brain activation patterns, as well as correlations between silent and oral reading data, furthermore suggest that this "inner voice" during silent reading is related to the suprasegmental and temporal characteristics of actual speech. For ease of comparison, we shall dub this phenomenon of an "inner voice" (particularly during silent reading of direct speech) *simulated implicit prosody* (SIP) to distinguish it from *default implicit prosody* (DIP) that is commonly discussed in relation to syntactic ambiguity resolution.

In the final part of this chapter, we will attempt to specify the relation between SIP and DIP. Based on the existing empirical data and our own theoretical conclusions, we will discuss the similarities and discrepancies between the two not necessarily mutually exclusive terms. We hope that our discussion will motivate a new surge of interdisciplinary research that will not only extend our knowledge of prosodic processes during reading, but could potentially unify the two phenomena in a single theoretical framework.

2 Direct and Indirect Speech: Pragmatic and (Explicit) Prosodic Differences

In everyday language use, prosody carries rich information not only about the structure and pragmatic function of an utterance but also about the source of the utterance (e.g., the speaker and their emotional state). When reporting speech (as in quotations), prosody is a key feature that differentiates direct speech (1) from indirect speech (2).

(1) *Mary said, "This dress is absolutely beautiful!"*
(2) *Mary said that the dress was absolutely beautiful.*

Direct speech is often a literal quotation of what the original speaker said. Indirect speech, by contrast, involves more of a summary or paraphrase of what the original speaker said. The quoted utterance in direct speech is usually treated as an independent prosodic unit and is typically marked with a phonetic pitch reset (i.e., resetting vocal pitch to a higher level in order to continue speaking). In contrast, an indirect speech utterance is usually embedded in a complement clause and not prosodically distinguished from the matrix clause.

While there are semantic and syntactic differences between direct and indirect speech (e.g., Banfield 1973, 1982; Li 1986; Partee 1973; Wierzbicka 1974), linguists have also recognized the “theatrical” nature of direct speech, meaning that it tends to carry more vivid paralinguistic information than indirect speech during communication (Li 1986; Tannen 1986, 1989; Wierzbicka 1974). As first conceptualized by Clark and Gerrig (1990), an important pragmatic function of direct speech is to provide *demonstrations* of the reported speech act. Demonstrations enable others to directly *experience* the things depicted. For example, to demonstrate the action of taking a photograph, one may take an imaginary camera to one’s eyes and click the imaginary shutter. Direct speech is often used to demonstrate *how* something was said by another speaker. As Clark and Gerrig (1990) argue, direct speech is an important stylistic device for enlivening stories. It provides vivid demonstrations of the reported speech act, thereby enabling the addressee to experience what it would be like to see, hear, or feel what the original speaker did in saying something. Consider example (1): when the reporter quotes *Mary,* he/she may depict *Mary*’s voice (e.g., high-pitch, squeaky), her accent (e.g., southern, northern), her emotional state (e.g., excitement), and/or Mary’s supposed facial expressions and gestures while making the utterance, so as to demonstrate *how* Mary said those words. Indirect speech, on the other hand, typically provides a mere *description* of what was said, without depicting paralinguistic information surrounding the reported speech act. In terms of prosody, this pragmatic distinction might become manifest in more dramatized and expressive vocal modulations for direct speech as compared to indirect speech, with the latter being generally reported in a more neutral tone.

Indeed, our own research suggests that in an oral reading task, direct speech tends to be interpreted in a more vivid fashion than indirect speech (Yao 2011, experiment 3). In this exploratory study, we examined whether individuals would spontaneously adjust their voices to “act out” the contextually implied emotional state of the reported speaker when reading aloud *direct speech* or meaning-equivalent *indirect speech* text passages. It is well established that a speaker’s emotional arousal is reliably reflected in modulations of vocal pitch (fundamental frequency, F_0) during speaking (Banse and Scherer 1996). If direct speech reporting is associated with demonstrations of the reported speech act, it should display a pitch profile that represents the reported speaker’s emotional state. In contrast, indirect speech reporting is likely to be characterised by a pitch profile that is emotionally detached from the original source. To test this idea, we prepared short fictitious stories containing direct or indirect speech utterances. Critically, between-items we manipulated the emotional arousal level of the reported speaker (the main protagonist in the story) by using introductory contexts implying "high", "medium", or "low" arousal of the quoted speaker (see below for examples; the different arousal levels were verified in a separate rating study).

Examples from Yao (2011, experiment 3):

(3) [HIGH AROUSAL] *Millionaire Joseph was addicted to betting on horses. Tipped by a so-called 'insider', he recently placed an enormous bet, but shockingly, the horse had lost.*

[DIRECT SPEECH] *Angry with his informant, Joseph shouted furiously on the phone: "Where did your bloody information come from!? That was a huge amount of money—almost one million pounds!"*

[INDIRECT SPEECH] Angry with his informant, Joseph shouted furiously on the phone, asking where the information had come from, because that was a huge amount of money—almost one million pounds.

(4) [MEDIUM AROUSAL] *Britney is a student at the University of Glasgow. After a heavy snow in the afternoon, she was complaining to her boyfriend James about the weather on their way home.*

[DIRECT SPEECH] *Her voice sounded very grumpy and unpleasant: "I really hate the winter! It's always dark and the roads are too slippery."*

[INDIRECT SPEECH] *Her voice sounded very grumpy and unpleasant, saying that she really hated the winter because it's always dark and the roads are too slippery.*

(5) [LOW AROUSAL] *Smith was working in a small antiques shop down the local high street. Today, a middle-aged posh lady with thick glasses came into the shop.*

[DIRECT SPEECH] *She looked around and said in a nonchalant tone: "You may be surprised to learn that I'm a world-renowned collector of rare memorabilia of White-eared Pheasant."*

[INDIRECT SPEECH] *She looked around and said, in a nonchalant tone, that he might be surprised to learn that she was a world renowned collector of rare memorabilia of White-eared Pheasant.*

Participants were instructed to read these stories aloud as naturally and fluently as possible. Each participant read each story only once, and importantly, the instructions did not explicitly encourage participants to vocally "act out" the stories. We recorded and analysed pitch contours and other characteristics of participants' speech during reading. Overall, we observed significantly larger variation of F_0 during oral reading of direct speech as opposed to indirect speech (mean SD for F_0 over time:12.63 [direct speech] vs. 8.68 [indirect speech], paired-sample $t\text{s} > 11$, $p\text{s} < 0.001$). In line with their hypothesised *demonstration* pragmatics, direct speech quotations appeared to have been orally interpreted in more varied, fluctuating pitch profiles than indirect speech utterances. More importantly, when reading direct speech aloud, readers' mean F_0 increased as a function of the contextually implied emotional arousal of the quoted speaker, with more arousal leading to a steady increase in F_0. In contrast, no such linear trend was observed during oral reading of indirect speech (Fig. 1). The data confirmed that readers spontaneously adjust their voices in accordance with the contextually implied emotional arousal of the quoted speaker. This was the case particularly for oral reading of direct speech, but not (or considerably less so) for oral reading of indirect speech. These findings highlight the distinctive prosodic profiles of direct and indirect speech in speaking.

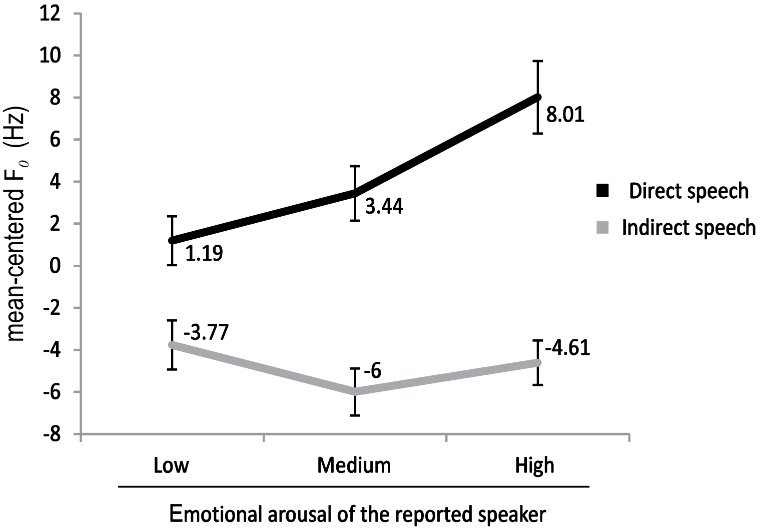

Fig. 1 Reporting style × emotional arousal interaction in oral reading (Yao 2011, experiment 3). The *numbers* indicate the condition means (in mean-centred F_0 to remove systematic gender differences in pitch). The error bars represent the 95 % confidence intervals for the means per condition

3 Direct and Indirect Speech During Silent Reading and "Prosodically Impoverished" Listening

Prosodic features of direct and indirect speech are easily measurable in *spoken* language. Here we are going to review evidence suggesting that perceivers also differentiate between the two reporting styles during *written* language processing. To illustrate this, Yao et al. (2011) explored how direct and indirect speech utterances are processed in the brain during silent reading of text where no auditory stimulation is present. Inspired by the common intuition of hearing an "inner voice" during silent reading of quotations, they speculated that the brain might take direct speech as a cue to activate "audible-speech"-like mental representations, even during silent reading of text. Recent embodied cognition theories (e.g., Barsalou 1999, 2008) lend theoretical support to this conjecture. Such theories propose that language processing is grounded in mental simulations (or re-enactments) of sensory and motor experiences that have been acquired through individuals' interaction with the environment and their internal states. Under such a premise, accumulated experiences with how direct versus indirect speech are typically reported in spoken language could form the basis for differential mental simulations during written-language processing. In other words, silent reading of direct speech would be grounded in mental simulations of vivid vocal depictions whereas silent reading of indirect speech would be grounded in simulations of voices that are more neutral. The brain may therefore be more prone to activate "audible-speech"-like representations during silent reading of direct speech than of indirect speech.

To test this hypothesis, Yao and colleagues combined fMRI and eye tracking to measure neural activity within the auditory cortex. The fMRI technique captures changes in oxygen consumption in local blood flow, which in turn estimates the degrees of neural activity within certain brain areas in vivo (Ogawa et al. 1990a, b; Ogawa and Lee 1990). Using this technique, neuroscientists have established that certain areas in the auditory cortex, i.e., those along the upper bank of the superior temporal sulcus (STS), are selectively sensitive to "bottom-up" auditory stimulation by human voices (Belin et al. 2000). These areas, labelled temporal voice areas (TVAs), provided Yao et al. (2011) with clearly defined functional hot spots for locating activations of voice-related representations during silent reading. In their experiment, participants' individual TVAs were identified in a voice localizer task in which audio clips of nonvocal sounds (e.g., telephone ringing) were compared to vocal sounds generated from speech (e.g., vowels) and nonspeech (e.g., laughing) utterances (Belin et al. 2000). Participants' TVAs were thus localizable via the contrast of their brain responses to vocal sounds versus nonvocal sounds. Before this functional voice-localizer task, Yao and colleagues measured neural activity (in the same participants) during silent reading of direct versus indirect speech text passages, as shown in the following example:

Examples from Yao et al. 2011:

(6) *PhD student Ella was summoned to her supervisor Jim's office to give a report on her current progress. Ella asked for an extension but Jim looked concerned.*

[DIRECT SPEECH] He said: "Hmm, we really need those data in by next month for that conference."

[INDIRECT SPEECH] *He said that they really needed those data in by next month for that conference.*

Importantly, the reported speech utterances in both conditions were kept equivalent in terms of linguistic content within each story (see underscored sentences in the above example); this was to rule out potential confounding factors between conditions. In the magnetic resonance imaging (MRI) scanner, these stories were visually presented to participants in a sentence-by-sentence fashion and for a fixed duration. Participants were instructed to silently read these stories for comprehension while their eye movements and brain activity were simultaneously monitored. Yao and colleagues observed that during silent reading of the critical speech utterances (determined via eye tracking), direct speech was associated with greater neural activity across multiple brain areas than indirect speech. The enhanced activity was distributed not only in the right auditory cortex but also in bilateral occipital lobes (associated with visual processing), superior parietal lobules, and precuneus (associated with visuo-spatial imagery, episodic memory retrieval, and self-processing). Such an activation pattern seemed to suggest an enriched multisensory mental simulation process for direct speech, which is consistent with Clark and Gerrig's (1990) hypothesis of direct speech as demonstration. Critically, reading of direct speech quotations (compared to meaning-equivalent indirect speech utterances) elicited significantly higher neural activity along the right STS (rSTS) areas which were clearly part of the TVAs identified in the voice-localizer task. This was the first direct indication that silent reading of direct speech is more strongly associated with "top-down" simulations of voice-related sensory experiences. Interestingly, compared to a baseline without linguistic stimulation, even indirect speech elicited some activation in those TVAs, but to a considerably lesser extent than direct speech.

Similar kinds of "inner voice" experiences were also observed during silent reading of direct speech in German (Brück et al. 2014). The authors' primary aim in that study was to investigate the neural correlates in processing emotional voice signals described in written texts (e.g., *Als sie sprach, klang ihre Stimme sanft und kehlig und mit einem italienischen Akzent behaftet—When she spoke, her voice sounded smooth and throaty and beset with an Italian accent*). Although not central to their research question, they also explored how direct speech reporting might modulate TVA activation during silent reading. This was possible because one third of their stimuli actually comprised direct speech quotations (e.g., *"Das ist nicht zu ertragen", sprach die Fürstin leise mit zitterender Stimme—"This is unbearable", said the baroness quietly with a quivering voice*). As expected, Brück et al. (2014) observed significantly higher activations of the right TVAs during silent reading of direct speech quotations as opposed to the other types of descriptions without quotations. Although this finding was established "post-hoc", it largely agrees with Yao et al.'s (2011) results, confirming that direct speech is likely to activate speech-(or voice-)related sensory experiences "top down", i.e., without acoustic stimulation.

One objection might be that the direct versus indirect speech materials used in Yao et al. (2011) sometimes differed in grammatical tense (present vs. past), syntactic structure (coordination vs. subordination), the use of pronouns (e.g., first vs. third person), or the use of emotion-signalling punctuation ("!" vs. "."). It is therefore conceivable that the observed differences between direct and indirect speech may be evoked by these extraneous differences, rather than the reporting styles "*per se*". However, Yao et al.'s (2011) additional reading performance analyses revealed no clear differences in either reading time (204 vs. 203 ms/word) or comprehension accuracy (83 vs. 82 %) between the direct and indirect speech conditions. More importantly, Yao et al. (2011) could show that the critical fMRI effect did not disappear when only a subset of items (34 out of 90) was considered, in which the direct and indirect speech conditions could be regarded as equivalent in terms of grammar and punctuation. With respect to the locus of the fMRI effect, the right-lateralized STS activation pattern hardly overlaps with activation patterns observed during processing of present versus past (D'Argembeau et al. 2008), syntax (e.g., Friederici et al.2000a, b), perspective (Vogeley and Fink 2003), or modality-independent emotions (Peelen et al. 2010). Taken together, it appears that an enhanced "inner voice" sensory experience during silent reading of direct speech remains the best explanation of Yao et al.'s (2011) data.

But how does this sensory experience relate to prosody? In fact, the *prosodic* nature of such "inner voices" was illuminated in a follow-up fMRI study by Yao et al. (2012). In Yao et al.'s (2011) study, there was no acoustic stimulation as a reference alongside the silent reading task (except for the functional localizer procedure). It was hence difficult to specify what types of acoustic representations may constitute the "inner voice" experiences during silent reading of direct speech. Interestingly, however, the acoustic processing literature indicates that the right auditory cortex areas appear to be specialised in processing slow-pitch modulations, including speech melody (Scott et al. 2000), musical melody (Patterson et al. 2002; Zatorre et al. 1994, 2002), and emotional prosody (Mitchell et al. 2003; Wildgruber et al. 2005). Thus, the specifically right-lateralized activation pattern observed in Yao et al. (2011) might be taken to suggest a suprasegmental prosodic nature of the “inner voices” experiences in silent reading of direct speech.

To verify this conjecture, Yao et al. (2012) sought to examine the neural correlates of “top-down” suprasegmental prosodic processing during *auditory* comprehension of reported speech. If these neural correlates show substantial overlap with the differential brain activation regions found in silent reading (Brück et al. 2014; Yao et al. 2011), this would lend support to the hypothesis that the latter may be of a suprasegmental prosodic nature. To this end, Yao and colleagues prepared audio recordings of the same short stories as in Yao et al. (2011). Crucially, both the direct and indirect speech utterances in these recordings were deliberately spoken in a *monotone* which is usually more felicitous for indirect rather than direct speech. The following is an example story:

(7) *Luke and his friends were watching a movie at the cinema. Luke wasn't particularly keen on romantic comedies, and he was complaining a lot after the film.*
[DIRECT SPEECH] He said: “God, that movie was terrible! I've never been so bored in my life.”
[INDIRECT SPEECH] *He said that the movie was terrible and that he had never been so bored in his life.*

This example story describes *Luke*'s terrible experience with a boring film. Normally, one would expect Luke to sound rather impatient and moany (e.g., “GOD, that movie was t-EEE-rible!”[1]), depicting how much *Luke* regretted watching the film. In stark contrast, the direct speech quotation was actually spoken in a steady tone which sounded emotionally detached (perhaps even sarcastic), and did not fit into the overall context (recordings can be found at: http://www.psy.gla.ac.uk/~boy/fMRI/samplerecordings/). Acoustically, this *monotone* manipulation preserved (sub)-segmental acoustic information (e.g., the phonological representations of words) but severely curtailed rich suprasegmental prosodic information (e.g., varied intonation patterns) that is typically expected of direct speech quotations. Yao et al. (2012) hypothesized that the brain may actively compensate for monotonously spoken direct speech by “filling in” suprasegmental prosodic information (i.e., expressive prosody) that is missing from the actual input. Such “filling in” processes should be reflected in increased brain activity within the TVAs. Comprehension of monotonous indirect speech utterances, however, is unlikely to involve such processes. Unlike its direct speech counterpart, indirect speech is typically spoken in a more neutral, less varied prosody (e.g., Yao 2011, described earlier). Thus, the brain does not need to compensate for monotonously spoken indirect speech utterances.

Using fMRI, Yao et al. (2012) measured participants' brain activity when they were listening to the monotonously spoken stories illustrated above. The

[1] The capitalization and repetition of letters represent emphases in intensity and length.

participants' individual TVAs were determined using the same voice localizer task as before (Belin et al. 2000). Neural activity within the TVAs was determined while listening to the critical direct speech or indirect speech utterances (underscored sentences in the above example). As expected, it was found that monotonously spoken direct speech elicited significantly higher brain activations within the right TVAs than monotonously spoken indirect speech. Most intriguingly, the increased activations for direct speech were located in virtually the same brain areas (i.e., the posterior, middle, and anterior parts of the right STS) as those previously observed in silent reading of direct versus indirect speech (see Fig. 2).

However, it remained unclear whether these differential brain activations indeed reflected enhanced "top-down" prosodic processing when listening to monotonous direct speech, or whether they were merely evoked "bottom-up" by differential acoustic characteristics of direct versus indirect speech utterances. To address this question, three variables (or "parametric modulators") were specified to potentially account for these increased rSTS activations. These were (a) the *acoustics* of the recordings (i.e., parameters such as pitch, intensity, duration, etc.), (b) the subjectively perceived *vividness* of the speech utterances without

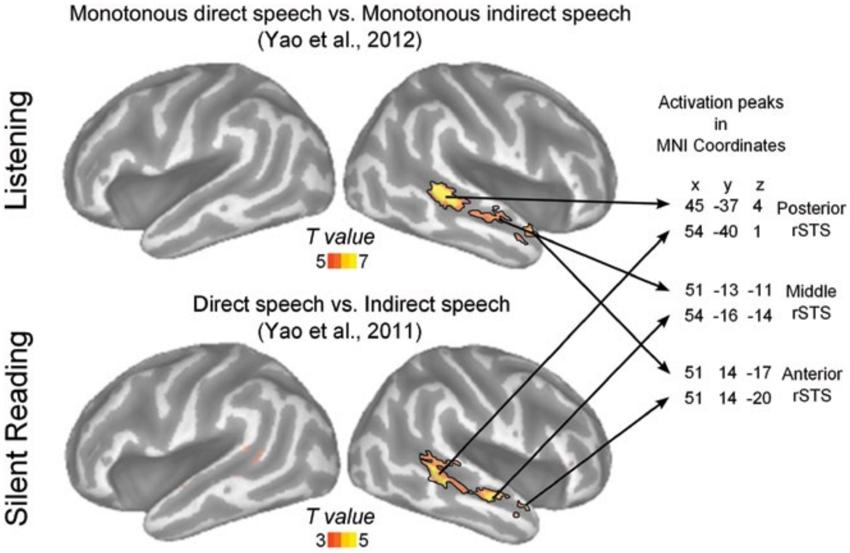

Fig. 2 Consistent findings between the two fMRI studies (only the effects within the TVAs are shown). The *top* panel shows the contrast between the monotonous direct speech and the monotonous indirect speech conditions during listening (Yao et al. 2012). The *bottom* panel shows the contrast between direct speech and indirect speech during silent reading (Yao et al. 2011). The *arrows* point to the peak voxel coordinates (in MNI space) in the activation clusters. The peak voxels were paired with their anatomical counterparts between the two studies. The thresholds for the two contrasts were adjusted to better illustrate the activation clusters

context (established via ratings), and (c) the *contextual congruency* of the speech utterances within the given story contexts (i.e., to what extent the speech utterances were perceived as congruent with a given context or not—again, this variable was established via ratings). *Acoustics* (a) and *vividness* (b) were taken as *objectives,* respectively *subjective* measures of the acoustic differences between the direct and indirect speech conditions without considering the context that the quotations were embedded in. Both modulators were expected to account for differences in "bottom-up" acoustic processing. In contrast, *contextual congruency* (c) was taken to index the degree of mismatch between the actual (monotonous) speech input and perceivers' "top-down" *expectation* for expressive speech prosody in the given story context. *Contextual congruency* was expected to explain differential "top-down" prosodic processing between conditions. It was found that the increased rSTS activations in monotonous direct speech (relative to monotonous indirect speech) was in fact most reliably explained by *contextual congruency* but not by the *acoustics* or *vividness* of the speech utterances. The analyses confirmed that the rSTS activation pattern indeed reflected "top-down" prosodic processing when listening to monotonously spoken direct speech utterances, as if the brain was actively trying to "fill in" prosodic information that was missing from the actual speech input. By reconciling the findings of the two fMRI studies, we conjecture that the "inner voices" observed during silent reading of direct speech may also involve suprasegmental prosodic information similar to that in auditory processing.

The prosodic nature of such an "inner voice" during silent reading of direct speech was further demonstrated behaviourally by Yao and Scheepers (2011). They examined whether the speech-related representations activated during silent reading of direct speech could be characterized in time (or speed). Time is an important dimension of prosody. It determines the rhythm, stresses (e.g., length of articulation), and global dynamics of speech. If prosodic representations were activated during silent reading of direct speech, they should reflect the implied speaking rate of the quoted speech. A potential behavioural consequence of this is that readers may adjust their reading rates in accordance with how fast a quoted speaker would speak in a given context.

Previous research by Alexander and Nygaard (2008) has already suggested that reading speed may be influenced by auditory imagery. In their study, they first familiarized participants with audio recordings of voices from either fast or slow speakers. In subsequent reading sessions, they told participants to imagine those speakers as authors of the materials given for reading. They observed that during both oral and silent reading, participants were faster to read the presented text materials when they were told that the author of the text was a previously introduced "fast speaker". Alexander and Nygaard's findings demonstrated that explicitly *encouraged* auditory imagery of an author's speaking rate had an influence on how fast one would read written text that was supposedly produced by that author. This, in turn, suggests that "inner voices" during silent reading of direct speech—a more *spontaneous* form of auditory imagery—could equally interact with reading behavior.

To test this idea, Yao and Scheepers (2011) prepared short stories containing direct and indirect speech utterances (see below for an example). Each story started with a narrative vignette which set up either a fast-speaking (i.e., where the speaker was likely to speak very quickly) or a slow-speaking scenario. The scenario led to a reported speech utterance that employed either direct speech or indirect speech. The story was then concluded by an additional sentence. Crucially, the critical speech sentences (e.g., the underscored sentences in the example) were identical between the fast-speaking and slow-speaking stories and were largely equivalent between direct speech and indirect speech conditions. Thus, differences in reading rate could not plausibly be attributed to differential wording across conditions.

(8) [FAST-SPEAKING] *It was a typical British day, rainy and gloomy. Sixteen-year-old pianist Bobby was going to play in the quarter-finals of a local talent competition. He was extremely nervous before his performance.*

[DIRECT SPEECH] *His mother encouraged him but he was all shaking and said: "No! I can't do it! This is the end of the journey because it is unlikely that I will make it this time."*

[INDIRECT SPEECH] *His mother encouraged him but he was all shaking and said that he couldn't do it and that it was the end of the journey because it was unlikely that he would make it this time.*

His mother tried to calm him down, saying that it's not the winning that counts, but the taking part.

(9) [SLOW-SPEAKING] *It was a typical British day, rainy and gloomy. At Glasgow Royal Infirmary, an old man was dying, and too weak to sit up. His family members were sitting around the bed, feeling sad. He wanted to say something, so his daughter placed a cushion under his head.*

[DIRECT SPEECH] *Slowly, he looked around and said: "I'm grateful you're all here. This is the end of the journey because it is unlikely that I will make it this time."*

[INDIRECT SPEECH] *Slowly, he looked around and said that he was grateful for their coming and that it was the end of the journey because it was unlikely that he would make it this time.*

Then he closed his eyes and everyone burst into tears.

Yao and Scheepers (2011) tested these materials in both oral and silent reading. In the oral reading task, participants were instructed to read aloud the stories in one go and as naturally and fluently as possible. Importantly, participants were not explicitly told to act out the reported speaker's voice during reading. Oral reading rates during the critical quotation passages were measured in syllables per second. A different group of participants were given the stories for silent reading while their eye movements were continuously monitored. Participants in the silent reading task were told to read the stories carefully for comprehension, and their reading rates were indexed by go-pass reading times (in milliseconds) on the critical direct or indirect speech sentences. In line with the predictions, it was found that in both oral and silent reading, participants spontaneously adjusted their reading rates to the contextually implied speech rate of the quoted speaker, but only when reading direct speech quotations and not when reading indirect speech passages. Most interestingly, Yao and Scheepers (2011) observed a high by-item correlation ($r = 0.56$, after accounting for effects of stimulus length) of reading rates across the two reading tasks. This suggests a strong temporal relation between "explicit prosody" (oral reading) and "implicit prosody" (silent reading) for the processing of both direct and indirect speech utterances.

In a more recent eye-tracking study, Stites et al. (2013) showed very similar effects during silent reading of direct speech, but again, not during silent reading of indirect speech. Interestingly, they found that these effects can be triggered by a single adverb (e.g., *John walked into the room and said* "energetically" *vs.* "nonchalantly"…) before the critical quotation passages. That is, direct quotations that were described as being said "quickly" were read faster than those described as being said "slowly".

In summary, the research on direct versus indirect speech has provided neuroimaging and behavioural evidence of "top-down" prosodic processes during language comprehension, in particular during silent reading of direct speech quotations. For the prosodic representations that are mentally simulated during silent reading of direct speech, we will use the term SIP to distinguish it from DIP that we shall discuss later. SIP appears to be primarily processed along the rSTS areas of the auditory cortex which are part of the TVAs (Belin et al. 2000). One important aspect of SIP is reflected in the close relationship between modulations of speaking rate (oral reading) and modulations of reading rate (silent reading) on the same language materials. In a broader context, these findings support the demonstration theory of direct speech (Clark and Gerrig 1990) from the perspective of language comprehension, highlighting the fact that direct speech is intrinsically more expressive than its indirect speech counterpart. The findings also extend embodied theories of language comprehension in several respects. First, the reviewed evidence for implicit prosody during silent reading (presumably in the form of mental simulations of actual speech, or at least involving speech-related mental representations) extends embodied theories to the auditory perceptual domain at the sentence/discourse level, which so far has received limited attention in the literature (previous research has mostly focused on sound-related words, see Kiefer et al. 2008 for example). Second, while most empirical research on embodied language comprehension focuses on the grounding of the linguistic meaning in perception and action, the research reviewed here involves differences in language pragmatics (direct speech as demonstration; indirect speech as description) and the consequences of such differences for processing semantically comparable reporting styles. In verbal communication, direct speech usually coincides with vivid demonstrations of the reported speech act whereas indirect speech is reported in a less vivid fashion. The present research shows that this vividness distinction is also reflected in how language is processed, and that direct speech is more likely to evoke mental simulations of voices or voice-related representations than indirect speech. Third, the reviewed fMRI research revealed that the posterior, middle, and anterior parts of the rSTS are potentially involved in mental simulations of suprasegmental prosodic representations. These data would motivate more sophisticated research on the neural mechanisms of implicit prosody and the neural configurations of the TVAs of the auditory cortex in general.

4 Open Questions and the Relation Between "Simulated" and "Default" Implicit Prosody

Many questions remain as to the detailed nature, mechanisms and functions of SIP. One interesting avenue for future research might be to probe its characteristics in other dimensions such as pitch, accent, and speaker identity.

Other questions relate to the durability of SIP representations. The studies above have mostly employed *online* methods (such as eye-tracking and fMRI) that probed into the ongoing processing of reported speech. In contrast, studies using *offline* methods such as probe-reaction *after* reading of quotations, appeared to be less sensitive in detecting differences between direct and indirect speech processing (Eerland et al. 2013; Yao 2011, Chap. 4). Given the temporal correlation between SIP and explicit prosody (Yao and Scheepers 2011), one might infer that effects related to SIP are relatively short-lived. More sophisticated testing is therefore needed to characterise the temporal properties of SIP in greater precision, specifying its onset, saturation, and offset.

Further questions for future research concern the function of SIP during silent reading of quotations. For example, is SIP beneficial to reading and memory? Given that aspects of SIP were shown to influence reading speed (Stites et al. 2013; Yao and Scheepers 2011), it appears worthwhile to further explore its role in eye movement control during silent reading.

While the research on direct and indirect speech is interesting in its own right, one interesting question arises as to how SIP during silent reading (particularly of direct speech) would inform the *implicit prosody hypothesis* (IPH) for silent reading (e.g., Fodor 1998, 2002; Quinn et al. 2000). The IPH assumes that a DIP is projected during silent reading of text, with potential consequences for syntactic processing. Such a default prosodic contour is very similar to the usual "explicit" prosodic contour for actual speech: It implements pauses, emphases, etc., thereby suggesting a prosodic grouping of a sentence during silent reading. These "implicit" prosodic groups appear to influence the syntactic parsing of a sentence, and may even determine its preferred interpretation in the face of syntactic ambiguity. Evidence for DIP processing during silent reading is provided, for example, by a rich body of research on relative clause (RC) attachment. Consider the English sentence "*Someone shot the servant of the actress who was on the balcony*", which is ambiguous as to whether the RC "*who was on the balcony*" should be attached *high* to the complex noun phrase "*the servant of the actress*" or low to the simpler and more recent noun phrase "*the actress*". It has been shown that native speakers of English tend to prefer a low attachment interpretation when silently reading a sentences such as the one quoted above (e.g., Carreiras and Clifton 1993, 1999). By contrast, speakers of other languages such as Spanish (Carreiras and Clifton 1993, 1999), French (Zagar et al. 1997), and German (e.g., Hemforth et al. 1998) prefer a high attachment interpretation for equivalent structures. The IPH provides a promising explanation for such RC attachment biases in different languages. When no other disambiguation cues (e.g., gender agreement, case marking, or semantic constraints) are available, DIP contours (which may differ across languages) provide structural information that aids syntactic ambiguity resolution. This claim has been *indirectly* supported by research on the effect of explicit prosody on RC attachment disambiguation. For example, Quinn et al. (2000) asked participants to read and interpret ambiguous RC sentences silently and then read the sentences again aloud. They analyzed the F_0 (fundamental frequency) values of N1 and N2 in sentences disambiguated for high/low attachment. They found that pitch accents (i.e., peaks in F_0) on the critical noun phrases (NPs) were related to preferred RC attachment. That is, in an NP1–NP2–RC structure, pitch accents on NP1 were more strongly associated with high attachment, whereas pitch accents on NP2 were more strongly associated with low attachment of the RC. They suggested that in silent reading, RC attachment may be disambiguated by the prominence relations of the NPs and RC that are marked by the purported implicit default prosody. Other prosodic factors such as prosodic breaks or pauses have also been found to influence RC attachment interpretations in speech. It has been established that a prosodic break before an RC generally prompts high attachment of the RC (e.g., Clifton et al. 2002; Lovrić et al. 2000, 2001; Maynell 1999). In a silent reading study, Lovrić et al. (2001) manipulated the duration of NP1 and NP2 in order to trigger implicit prosodic breaks at different locations of an NP1–NP2–RC structure. They found that the lengthening of NP1 (prompting a prosodic break before NP2) resulted in a low attachment preference; the lengthening of NP2 before a long RC (prompting a prosodic break between NP2 and RC) increased probability of high attachment interpretations. Such correlations between DIP breaks and RC attachment preferences also lend support to the IPH.

Although DIP has been established behaviourally in different languages (e.g., Koizumi 2009; Shafran 2011; Shaked 2009), the cognitive and neural mechanisms underlying the projection of DIP remain largely unknown. By its very nature, DIP is not easy to manipulate or to measure, and it has yet to offer a comprehensive explanation for crosslinguistic variation in RC attachment. We believe that theories such as the IPH could potentially benefit from systematic analyses of what we called SIP during silent reading of direct (vs. indirect) speech.

In the following, we will discuss potential relations between DIP (as primarily revealed in research on ambiguity resolution) and SIP (as discussed in the context of reported speech processing). One possibility is that DIP and SIP are two instantiations of the same cognitive process, involving largely the same mental representations. This seems plausible because both refer to prosodic representations that are generated “internally”, i.e., without external auditory stimulation. Research has shown that (at least aspects of) DIP and SIP are correlated with explicit prosody during actual speech (e.g., Lovrić et al. 2000, 2001; Yao and Scheepers 2011). This might indicate that DIP and SIP share the same sensory grounding. Moreover, it is evident that the SIP activated (particularly) during direct speech processing may be an enhanced form of DIP which is activated during indirect speech processing and/or the processing of materials that do not involve reported speech. In fact, Yao et al.’s (2011) fMRI study on silent reading of direct versus indirect speech indicated that *both* direct *and* indirect speech processing lead to increased rSTS activation compared to a baseline condition where only a fixation cross was presented (no reading). This additional observation suggests that even silent reading of indirect speech may not be completely "silent" in that it also involves some form of implicit prosodic processing, although to a much lesser extent when compared to silent reading of direct speech. It therefore appears plausible to speculate that SIP during silent reading of direct speech may be a special, enriched form of the more generic prosody (DIP) assumed by the IPH. One way to test the relations between DIP and SIP might be to embed ambiguous RC structures in direct speech quotations, and examine whether RC attachment preferences during silent reading are in some way "*enhanced*" compared to RC attachment in isolated sentences or sentences introduced as indirect quotes. For example, one could test whether the NP1-NP2-RC structure in *When asked by the police, she said, "Someone shot the servant of the actress who was on the balcony"* would result in a stronger low attachment preference in English than when it is not in direct quotes. If we observed such interaction between RC disambiguation and reporting style, it would add weight to the hypothesis that DIP and SIP share aspects of the same mental representation.

In addition, DIP and SIP both interact with language processing and it seems that a common function of them is to facilitate comprehension. It is well established that DIP can help resolve syntactic ambiguity during silent reading by providing prosodic cues to the configurational interpretation of linguistic structure when other cues (e.g., syntactic or semantic) are not available (Fodor 2002). However, RC attachment is by no means the only processing domain where implicit prosody becomes relevant. For example, a recent eye-tracking study by Ashby and Clifton (2005) examined the effects of lexical stress on eye movements during silent reading. Participants read sentences containing words with a single stressed syllable or words with two stressed syllables. With other factors controlled, it was found that two-syllable words took longer to read compared to one-syllable words. The findings are in line with the IPH, suggesting that a prosodic contour is routinely constructed during silent reading, affecting not only sentence-level processing but also lexical access.

In a similar vein, SIP during silent reading of direct speech also appears to be beneficial to language processing. The notion of SIP essentially refers to the addition, or enhancement, of another sensory (i.e., auditory) layer during silent reading, which is particularly noticeable in direct speech processing. This layer enriches the mental representations of direct speech in many respects, including the emotional states of the quoted speakers, speech pragmatics, speech styles, and so on. For example, consider the following two sentences:

(10) *Mary said with excitement, "This dress is absolutely beautiful!"*
(*This dress is ABSOLUTELY BEAUUU-tiful*)
(11) *Mary said with excitement that the dress was absolutely beautiful.*
(*The dress was absolutely beautiful*)

The sentences in parentheses illustrate how the speech utterances in (10) and (11) may be interpreted prosodically during silent reading. The capital letters in (10) represent a hypothetical increase in pitch and volume (accents), and the repetition of the letter *U* represents the lengthening of the vowel/ju:/ in *beautiful*. Semantically, both sentences describe that *Mary* found a dress very beautiful. In (10), however, the more "dramatic" prosodic contour adds a sensory layer that allows the brain to *perceptually experience* the excitement in speech. This additional sensory information creates an enriched representation of the emotional state of the quoted speaker, causing (10) to be more accessible and engaging. In contrast, although (11) characterizes the emotionality of the speaker semantically, the more generic, default prosodic contour in (11) does not reinforce this representation. As a result, (11) is likely to be perceived as being more distant and emotionally disconnected.

The perceptually enriched representation of direct speech might explain why direct speech appears to be associated with deeper processing than indirect speech (Bohan et al. 2008; Eerland et al. 2013). Bohan et al. (2008), for example, visually presented participants with a direct or an indirect speech sentence like the following:

(12) *John said, "I needed some nine-inch nails so I went to B&Q".*
(13) *John said he needed some nine-inch nails so he went to B&Q.*

Immediately after the initial presentation, they showed the same sentence again, and asked participants whether or not this sentence was different from the one that had just been shown. In half of the trials, the second sentence was indeed exactly the same as the first sentence. In the other half of the trials, however, the second sentence presentation involved a very subtle text change within the critical quotation passage (e.g., replacing the verb "*went*" with a close semantic relative such as "*walked*"). Bohan et al. (2008) found that such subtle verb exchanges were reliably more detectable when they occurred within a direct speech rather than an indirect speech text passage, suggesting deeper processing (or enhanced verbatim memory) of direct speech. Eerland et al. (2013) later extended these findings to cases where the text changes were not restricted to verbs. Both studies consistently showed a memory advantage for direct speech as compared to indirect speech. These findings support the idea that covert prosody enhances the representations of direct speech. However, the link between such a memory advantage and SIP is yet to be established.

While DIP and SIP appear to be highly comparable from a phenomenological and functional perspective, it is equally conceivable that they actually entail two distinctive cognitive processes. In fact, a rather complex picture emerges as to the potential mechanisms underlying DIP and SIP. By definition, DIP is routinely generated and projected during silent reading. It can be viewed as a regular prosodic channel which informs the configurational interpretation of language when disambiguating cues from other channels (e.g., syntax, semantics) are not available. DIP has been shown to be informed by a default prosodic contour (i.e., phonology) of a given language, as well as surface visual features such as punctuation (e.g., Steinhauer and Friederici 2001; Steinhauer 2003), phrase length (e.g., Lovrić et al. 2001), or line breaks (e.g., Koizumi 2009). In contrast, SIP appears to be highly dependent on linguistic context and pragmatics (Stites et al. 2013; Yao et al. 2012; Yao and Scheepers 2011), and operates at a deeper, semantic level in a "predictive" manner. In line with embodied theories (Barsalou 1999, 2008), SIP is the speech experience that is *mentally simulated* during comprehension of (particularly) direct speech, as part of a more vivid mental representation of the latter. Mental simulations not only re-enact sensory, motor, and introspective experiences for representing language that is currently being processed; more importantly, they also place the perceiver in the simulated situations, thereby producing continual predictions about events likely to be described, actions likely to take place and introspections likely to result in the incoming language stimuli (Barsalou 2009). As evidence for the predictive aspect of SIP, the findings by Yao et al. (2012) showed that when direct speech quotations are spoken in a context-inappropriate monotone, the perceiver's brain automatically "talks over" such boring quotes by actively projecting context-appropriate prosodic structure that is missing from the input. It appears that during listening, SIP can serve as a top-down predictor of actual speech.

The similarities and differences between DIP and SIP may be reconciled in partially overlapping processing models for the two phenomena. Considering their comparable correlations with explicit prosody, it seems plausible to conjecture that DIP and SIP share a common neural network for representing prosodic contours. However, their potentially distinctive cognitive origins (projection of default prosodic contours on the one hand vs. perceptual simulation of voice and speech on the other) may be reflected in differential engagement of brain regions within this common network and/or engagement of additional brain regions that modulate this network. Only future research can tell the exact differences and commonalities between DIP (as reflected in research on ambiguity resolution) and SIP (as revealed by differences in processing direct versus indirect speech).

5 Conclusions

In this chapter, we have examined the mental representations of direct speech (e.g., *Mary said, "This dress is absolutely beautiful!"*) versus indirect speech (e.g., *Mary said that the dress was absolutely beautiful*). We showed that the brain is more likely to generate enriched suprasegmental prosodic representations of the reported speaker during comprehension of direct speech as opposed to meaning-equivalent indirect speech. We dubbed this specific "inner voice" phenomenon SIP. We have presented consistent neuroimaging evidence showing that SIP is primarily processed at the posterior, middle, and anterior areas of the rSTS of the auditory cortex—also parts of the TVAs (Belin et al. 2000). One aspect of SIP becomes evident in processing rates for direct speech quotations, as reflected in modulations of explicit speaking rates during oral reading as well as in eye movements during silent reading. The findings provide empirical support for the theory of direct speech as demonstration (Clark and Gerrig 1990) and embodied theories of language comprehension (e.g., Barsalou 1999, 2008).

What are the implications of these findings for the IPH? The IPH proposes that a default prosodic contour is generated internally and projected onto visual texts during silent reading. We have termed this kind of projected information DIP. DIP provides prosodic cues (e.g., emphases, prosodic breaks) that benefit configurational interpretations of ambiguous language structures (e.g., relative clause attachment) when other types of cues (e.g., syntactic, semantic) are not available. By their nature, DIP and SIP are both internally generated prosodic representations without external auditory stimulation, and are correlated with prosody in actual speech. Moreover, DIP and SIP both appear to be beneficial to language processing, although in their own ways. While DIP aids in structural interpretation, SIP perceptually enriches the mental representation of language, resulting in deeper processing of (or enhanced verbatim memory for) direct speech compared to indirect speech. With respect to the mechanisms of DIP and SIP, we recognize that they may be derived from distinctive cognitive processes. Based on the existing evidence, we conjecture that DIP operates relatively independently at a surface level of linguistic representation, routinely informing structural interpretations of language. In comparison, SIP appears to be a mentally simulated sensation of voice that is highly dependent on semantic and pragmatic context. We attempt to reconcile the similarities and discrepancies between DIP and SIP by conjecturing partially overlapping processing networks for these two phenomena.

Although research on SIP in silent reading of direct speech is still in its infancy, it complements the research on DIP by providing a potential platform to address how implicit prosody may operate at the neural, cognitive, and behavioural level. By investigating the similarities and discrepancies between DIP and SIP, future research has the potential to venture beyond simple demonstrations of these phenomena by seeking the evidence necessary to develop explicit mechanistic models of the two processes. An interdisciplinary approach would be very useful in pursuing this ambition. For example, a combination of eye tracking with fMRI and electroencephalography (EEG) or with magnetoencephalography (MEG) would allow us to delineate the neural circuitry underlying DIP and SIP processing in high spatiotemporal precision during real-time silent reading. This could illuminate where DIP and SIP originate from and whether they indeed converge into overlapping neural circuits, resulting in comparable prosodic sensations. The precise neural dynamics and parameters provided would lay the biological and empirical foundation for cognitive modelling of DIP and SIP, leading to more sophisticated theories in both domains.

References

Alexander, J. D., & Nygaard, L. C. (2008). Reading voices and hearing text: Talker-specific auditory imagery in reading. *Journal of Experimental Psychology-Human Perception and Performance, 34*(2), 446–459. doi:10.1037/0096-1523.34.2.446.

Ashby, J., & Clifton, C. (2005). The prosodic property of lexical stress affects eye movements during silent reading. *Cognition, 96*(3), B89–B100. doi:10.1016/j.cognition.2004.12.006.

Banfield, A. (1973). Narrative style and grammar of direct and indirect speech. *Foundations of Language, 81*(4), 1–39.

Banfield, A. (1982). *Unspeakable sentences: Narration and representation in the language of fiction*. Boston: Routledge.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70*(3), 614–636. doi:10.1037/0022-3514.70.3.614.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*(4), 577–660.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*, 617–645. doi:10.1146/annurev.psych.59.103006.093639.

Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society B-Biological Sciences, 364*(1521), 1281–1289. doi:10.1098/rstb.2008.0319.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature, 403*(6767), 309–312. doi:10.1038/35002078.

Bohan, J., Sanford, A. J., Cochrane, S., & Sanford, A. J. S. (2008). *Direct and indirect speech modulates depth of processing*. Poster presented at the 14th Annual conference on architectures and mechanisms for language processing (AMLaP), Cambridge, UK.

Brück, C., Kreifelts, B., Gößling-Arnold, C., Wertheimer, J., & Wildgruber, D. (2014). Inner voices: The cerebral representation of emotional voice cues described in literary texts. *Social Cognitive and Affective Neuroscience*. doi:10.1093/scan/nst180.

Carreiras, M., & Clifton, C. (1993). Relative clause interpretation preferences in Spanish and English. *Language and Speech, 36*(4), 353–372. doi:10.1177/002383099303600401.

Carreiras, M., & Clifton, C. (1999). Another word on parsing relative clauses: Eyetracking evidence from Spanish and English. *Memory & Cognition, 27*(5), 826–833. doi:10.3758/BF03198535.

Clark, H. H., & Gerrig, R. J. (1990). Quotations as demonstrations. *Language, 66*(4), 764–805.

Clifton, C., Carlson, K., & Frazier, L. (2002). Informative prosodic boundaries. *Language and Speech, 45*(2), 87–114. doi:10.1177/00238309020450020101.

D'Argembeau, A., Feyers, D., Majerus, S., Collette, F., Van der Linden, M., Maquet, P., & Salmon, E. (2008). Self-reflection across time: Cortical midline structures differentiate between present and past selves. *Social Cognitive and Affective Neuroscience, 3*(3), 244–252. doi:10.1093/scan/nsn020.

Eerland, A., Engelen, J. A. A., & Zwaan, R. A. (2013). The influence of direct and indirect speech on mental representations. *PLoS ONE, 8*(6), e65480. doi:10.1371/journal.pone.0065480.

Fodor, J. D. (1998). Learning to parse? *Journal of Psycholinguistic Research, 27*(2), 285–319. doi:10.1023/A:1023258301588.

Fodor, J. D. (2002). *Prosodic disambiguation in silent reading*. In *PROCEEDINGS-NELS* (*Vol. 1*, pp. 113–132).

Friederici, A. D., Meyer, M., & von Cramon, D. Y. (2000a). Auditory language comprehension: An event-related fMRI study on the processing of syntactic and lexical information. *Brain and Language, 74*(2), 289–300. doi:10.1006/brln.2000.2313.

Friederici, A. D., Wang, Y. H., Herrmann, C. S., Maess, B., & Oertel, U. (2000b). Localization of early syntactic processes in frontal and temporal cortical areas: A magnetoencephalographic study. *Human Brain Mapping, 11*(1), 1–11.

Hemforth, B., Konieczny, L., Scheepers, C., & Strube, G. (1998). Syntactic ambiguity resolution in German. In D. Hillert (Ed.), *Sentence processing: A crosslinguistic perspective—syntax and semantics* (*Vol. 31*, pp. 293–312). San Diego: Academic.

Kiefer, M., Sim, E. J., Herrnberger, B., Grothe, J., & Hoenig, K. (2008). The sound of concepts: Four markers for a link between auditory and conceptual brain systems. *Journal of Neuroscience, 28*(47), 12224–12230. doi:10.1523/jneurosci.3579-08.2008.

Koizumi, Y. (2009). *Processing the not-because ambiguity in English: The role of pragmatics and prosody*. Dissertation, City University of New York.

Li, C. N. (1986). Direct speech and indirect speech: A functional study. In F. Coulmas (Ed.), *Direct and indirect speech* (pp. 29–45). Berlin: Mouton de Gruyter.

Lovrić, N., Bradley, D., & Fodor, J. D. (2000). *RC attachment in Croatian with and without preposition*. Poster presented at the 6th Annual Conference on architectures and mechanisms for language processing (AMLaP), Leiden.

Lovrić, N., Bradley, D., & Fodor, J. D. (2001). *Silent prosody resolves syntactic ambiguities: Evidence from Croatian*. Paper presented at the 2nd SUNY/CUNY/NYU Conference, Stonybrook, NY.

Maynell, L. A. (1999). *Effect of pitch accent placement on resolving relative clause ambiguity in English*. Poster presented at the 12th Annual CUNY Conference on human sentence processing, New York.

Mitchell, R. L. C., Elliott, R., Barry, M., Cruttenden, A., & Woodruff, P. W. R. (2003). The neural response to emotional prosody, as revealed by functional magnetic resonance imaging. *Neuropsychologia, 41*(10), 1410–1421. doi:10.1016/s0028-3932(03)00017-4.

Ogawa, S., & Lee, T. M. (1990). Magnetic-resonance-imaging of blood-vessels at high fields: In vivo and in vitro measurements and image stimulation. *Magnetic Resonance in Medicine, 16*(1), 9–18. doi:10.1002/mrm.1910160103.

Ogawa, S., Lee, T. M., Kay, A. R., & Tank, D. W. (1990a). Brain magnetic-resonance-imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences of the United States of America, 87*(24), 9868–9872. doi:10.1073/pnas.87.24.9868.

Ogawa, S., Lee, T. M., Nayak, A. S., & Glynn, P. (1990b). Oxygenation-sensitive contrast in magnetic-resonance image of rodent brain at high magnetic-fields. *Magnetic Resonance in Medicine, 14*(1), 68–78. doi:10.1002/mrm.1910140108.

Partee, B. (1973). The syntax and semantics of quotation. In S. R. Anderson & P. Kiparsky (Eds.), *A festschrift for Morris Halle* (pp. 410–418). New York: Holt, Reinhart and Winston.

Patterson, R. D., Uppenkamp, S., Johnsrude, I. S., & Griffiths, T. D. (2002). The processing of temporal pitch and melody information in auditory cortex. *Neuron, 36*(4), 767–776. doi:10.1016/s0896-6273(02)01060-7.

Peelen, M. V., Atkinson, A. P., & Vuilleumier, P. (2010). Supramodal representations of perceived emotions in the human brain. *Journal of Neuroscience, 30*(30), 10127–10134. doi:10.1523/jneurosci.2161-10.2010.

Quinn, D., Abdelghany, H., & Fodor, J. D. (2000). *More evidence of implicit prosody in reading: French and Arabic relative clauses*. Poster presented at the 13th Annual CUNY Conference on human sentence processing, La Jolla, CA.

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. S. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain, 123*(12), 2400–2406. doi:10.1093/brain/123.12.2400.

Shafran, R. W. (2011). *Prosody and parsing in a double PP construction in Hebrew*. Dissertation, City University of New York.

Shaked, A. (2009). *Attachment ambiguities in Hebrew complex nominals: Prosody and parsing*. Dissertation, City University of New York.

Steinhauer, K. (2003). Electrophysiological correlates of prosody and punctuation. *Brain and Language, 86*(1), 142–164. doi:10.1016/S0093-934x(02)00542-4

Steinhauer, K., & Friederici, A. D. (2001). Prosodic boundaries, comma rules, and brain responses: the closure positive shift in ERPs as a universal marker for prosodic phrasing in listeners and readers. *Journal of Psycholinguistic Research, 30*(3), 267–295.

Stites, M. C., Luke, S. G., & Christianson, K. (2013). The psychologist said quickly, “Dialogue descriptions modulate reading speed!” *Memory & Cognition, 41*(1), 137–151. doi:10.3758/s13421-012-0248-7.

Tannen, D. (1986). Introducing constructed dialogue in Greek and American conversational and literary narrative. In F. Coulmas (Ed.), *Direct and indirect speech* (pp. 311–332). Berlin: Mouton de Gruyter.

Tannen, D. (1989). “Oh talking voice that is so sweet”: Constructing dialogue in conversation. In D. Tannen (Ed.), *Talking voices: Repetition, dialogue, and imagery in conversational discourse*. Cambridge: Cambridge University Press.

Vogeley, K., & Fink, G. R. (2003). Neural correlates of the first-person-perspective. *Trends in Cognitive Sciences, 7*(1), 38–42. doi:10.1016/S1364-6613(02)00003-7.

Wierzbicka, A. (1974). The semantics of direct and indirect discourse. *Research on Language & Social Interaction, 7*(3–4), 267–307.

Wildgruber, D., Riecker, A., Hertrich, I., Erb, M., Grodd, W., Ethofer, T., & Ackermann, H. (2005). Identification of emotional intonation evaluated by fMRI. *Neuroimage, 24*(4), 1233–1241. doi:10.1016/j.neuroimage.2004.10.034.

Yao, B. (2011). *Mental simulations in comprehension of direct versus indirect quotations*. PhD thesis, University of Glasgow.

Yao, B., & Scheepers, C. (2011). Contextual modulation of reading rate for direct versus indirect speech quotations. *Cognition, 121*(3), 447–453. doi:10.1016/j.cognition.2011.08.007.

Yao, B., Belin, P., & Scheepers, C. (2011). Silent reading of direct versus indirect speech activates voice-selective areas in the auditory cortex. *Journal of Cognitive Neuroscience, 23*(10), 3146–3152. doi:10.1162/jocn_a_00022.

Yao, B., Belin, P., & Scheepers, C. (2012). Brain "talks over" boring quotes: Top-down activation of voice-selective areas while listening to monotonous direct speech quotations. *NeuroImage, 60*(3), 1832–1842. doi:10.1016/j.neuroimage.2012.01.111.

Zagar, D., Pynte, J., & Rativeau IV, S. (1997). Evidence for early closure attachment on first pass reading times in French. *The Quarterly Journal of Experimental Psychology, 50*(2), 421–438. doi:10.1080/713755715.

Zatorre, R. J., Evans, A. C., & Meyer, E. (1994). Neural mechanisms underlying melodic perception and memory for pitch. *Journal of Neuroscience, 14*(4), 1908–1919.

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: music and speech. *Trends in Cognitive Sciences, 6*(1), 37–46. doi:10.1016/s1364-6613(00)01816-7.