J Cult Cogn Sci (2019) 3:113–139
https://doi.org/10.1007/s41809-019-00035-3

REVIEW

Fixations in the visual world paradigm: where, when, why?

James S. Magnuson

Received: 8 May 2019 / Revised: 12 August 2019 / Accepted: 31 August 2019 / Published online: 23 September 2019

Abstract Over the last 25 years, the visual world paradigm has enabled discoveries and theoretical advances in spoken language processing. However, the intuitive interpretation of fixations in the visual world paradigm—that fixations directly reflect over-time processes of activation and competition governing cognitive and language processing—deserves scrutiny. This paper provides a selective review of studies that suggest that the relations between fixations and ongoing processing are more complex than suggested by the intuitive interpretation. A particular challenge is explaining why context sometimes appears to have deep effects on language processing, while other times fixations appear to violate strong contextual constraints. I discuss implications of these seemingly contradictory patterns for theories of real-world language processing, and practical implications for using the visual world paradigm. Along the way, I review four possible linking hypotheses for connecting measures in the paradigm to theories of language and cognition. This review leads to the conclusion that implemented computational models will be needed to assess to what degree different linking hypotheses generate distinguishable predictions.

Keywords Visual world paradigm · Eye tracking · Psycholinguistics

J. S. Magnuson (✉)
Psychological Sciences, University of Connecticut, Storrs, CT 06269-1020, USA
e-mail: james.magnuson@uconn.edu

Introduction

In the visual world paradigm (VWP), participants' eye movements are tracked as they hear spoken language while they view scenes of real objects (e.g., Tanenhaus et al. 1995), or computer screens displaying images (Allopenna et al. 1998), or blank screens (Richardson and Spivey 2000). The accompanying speech may be instructions to interact with items in the display (the original Tanenhaus method), it may be about the display, but with no explicit task (Altmann and Kamide 1999), it may be about a scene or image that has been removed (as in the blank screen paradigm), or it may intentionally mismatch the display (e.g., Huettig and McQueen 2007). The modern VWP[1]

[1] A variant of the VWP was first reported by Cooper (1974), but this work went largely unnoticed. As of January 13, 2019, Cooper (1974) had been cited 389 times (not counting one citation that mysteriously predated Cooper's paper by 6 years; verified on scopus.com), but only 5 of those citations predate the independent development of the VWP by Tanenhaus et al. (1995) (which had 1245 citations as of January 13, 2019). The remainder of Cooper's citations follow a citation in Tanenhaus and Spivey-Knowlton (1996). Cooper presented the technique as one having great potential, but did not apply it to any theory-driven questions, and not even Cooper himself applied the technique in any later work. In their comprehensive review of the VWP, Huettig et al. (2011b) suggest that the impact of Cooper (1974) was minimal while that of Tanenhaus et al. (1995) was transformative partly because of the cumbersome and expensive nature of eye tracking in the 1970s, and partly due to theoretical debates that emerged in the 1980s to which Tanenhaus et al. applied the technique. The latter seems key

was introduced by Tanenhaus et al. (1995). The original article was followed rapidly by several more from Tanenhaus and colleagues and was soon being used by other labs as well (see Huettig et al. 2011b for a comprehensive review). There is no question that the VWP has had a transformative impact on psycholinguistics. Timecourse data has provided traction or resolution in longstanding debates in sentence processing (Tanenhaus et al. 1995), spoken word recognition (e.g., Allopenna et al. 1998; Dahan et al. 2001a, b), and speech perception (McMurray et al. 2002).

But what drives fixations in the VWP? How can we interpret them? What alternative interpretations are plausible? In addressing these questions, this paper follows in the footsteps of two previous reviews: the early articulation of a clear linking hypothesis for the VWP (Tanenhaus et al. 2000) and a comprehensive critical review of the paradigm by Huettig et al. (2011b). My focus is reexamining some issues these earlier papers covered, along with some complementary ones. As such, this paper incudes only a selective review of the VWP. I will discuss four possible linking hypotheses for the VWP, as well as challenges and limitations of each one. As we shall see, there is a strong case to be made for pervasive, bidirectional interaction between language, vision, and integrative decision processes. In the final section, I will summarize some remaining challenges and possible ways to address them, with an emphasis on how constraints that arise from linguistic, task, or even cultural context may complicate the interpretation of fixations in the VWP.

Linking theories to dependent variables in the VWP

No index of cognitive processes, whether behavioral (button press, eye movement), sensory (blink, pupil dilation), physiological (heart rate, galvanic skin response), or neural (scalp potential, hemodynamic response), has a transparent meaning. The experimenter must devise a *linking hypothesis* that operationally ties the response to the hypothesized process (Tanenhaus et al. 2000), although in many cases we may rely on implicit or vague linking hypotheses. For example, some studies of priming operationally define priming as speeding or slowing of responses contingent upon the presentation of an item (the prime) that is somehow related to the target (the probe). On a simple linking hypothesis, if contingent speeding or slowing is observed, priming has occurred. A detailed linking hypothesis would articulate details of the process, such as: (a) the prime is activated by bottom-up match to the input; (b) classes of items with specific relations to the target are hypothesized to subsequently receive positive activation (facilitation, or positive priming) or inhibitory activation (inhibition, or negative priming).

A step beyond this would be model-based simulations predicting the relative activations of primes and primed items. Figure 1 displays a schematic of a hypothetical simulation-based prediction. Such simulations would include concrete details about timing (presentation of prime, bottom-up activation of the prime, and impact on subsequently presented related vs. unrelated items (possibly sustained, and/or possibly decaying, positive or negative). The simulation model would need to include a concrete mechanism for activation flow from primes to primed items (direct connections, semantic mediation, etc.). While such a model may seem transparently related to a variety of tasks, the linking hypothesis is incomplete without a task-specific response model. The linking hypothesis must relate a task-specific response and dependent measure (timing or accuracy of a button press, or eye movement, or timing or magnitude of an event-related potential) in studies with human subjects to measurable details of the model (e.g., lexical activations). Note that one might only simulate the prime presentation, and infer degree of priming *potential* from the states of items following that presentation, in which case the linking hypothesis would include an assumption that speeded or slowed activation of subsequent presentation of the prime (repetition), a possibly related probe, or an unrelated baseline item, would be proportional to state after the initial prime presentation. Or one could explicitly simulate that step as well, implementing a more direct linking hypothesis.

Visual world studies (including some in my own publications) often leave the linking hypothesis

Footnote 1 continued
(which raises interesting questions about the convergence of tools and theories), as does the rapid extension of the VWP by the Tanenhaus lab following their 1995 paper to theoretically-motivated questions in multiple domains of psycholinguistics.

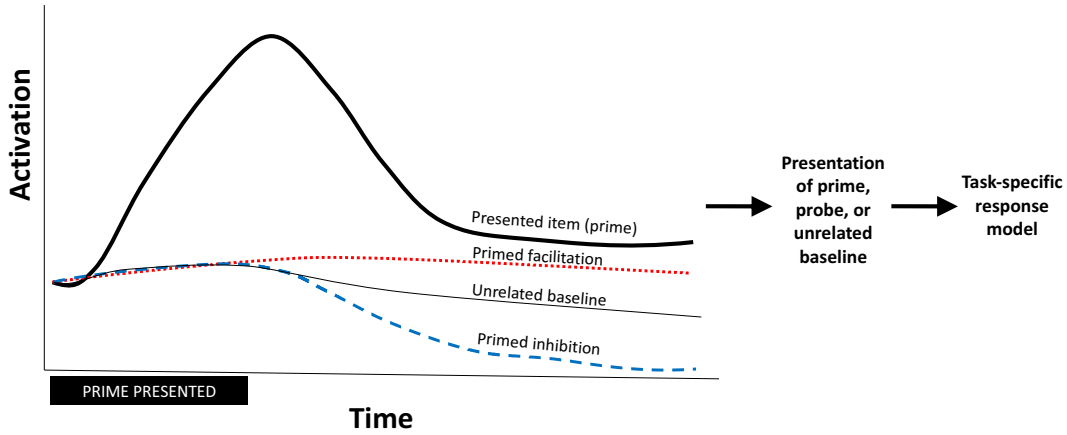

Fig. 1 Schematic of possible priming simulations, coupled with a decision/response model (Magnuson 2019a)

implicit. The default interpretation seems to be that greater fixation proportions indicate greater activation in the underlying language processing system, whether manipulations are phonological, semantic, syntactic, etc. If a linking hypothesis is not specified, we might infer that the authors assume the informal linking hypothesis proposed by Tanenhaus et al. (2000):

> Informally, we have automated behavioral routines that link a name to its referent; when the referent is visually present and task relevant, then recognizing its name accesses these routines, triggering a saccadic eye movement to fixate the relevant information. (p. 565)

This informal definition leaves out at least four key aspects of the VWP. (1) Saccades are typically made based on partial information, prior to word recognition (e.g., Allopenna et al. 1998). (2) Fixation likelihood increases for items that share phonological or semantic features of lexical items activated by the spoken input (which is a larger set than only the target word; e.g., Huettig and Altmann 2005; Yee and Sedivy 2006). (3) Anticipatory fixations may precede bottom-up specification (e.g., Altmann and Kamide 1999; Kukona et al. 2011). (4) Some manipulations cause a *reduction* or complete suppression of fixations to particular item types (e.g., Magnuson et al. 2008).

In the following sections, I review four possible linking hypotheses and their ability to address these complications. (1) Linguistic and visual channels are processed independently with cascaded integration at a decision level, allowing for parallel-contingent processing (quite similar to the linking hypothesis of Tanenhaus et al. 2000); (2) displays in the VWP trigger the loading of working memory with visual, phonological, and semantic features of displayed objects prior to linguistic input; (3) behavior in the VWP and other tasks suggests internal modeling of context and language (mental world hypothesis); and (4) "just-in-time" deep interaction, with bidirectional constraints between language, vision, and decision, but minimal internal representation. In the final sections, I compare these views and also discuss challenges that remain for all of them.

Linking Hypothesis 1: parallel-contingent independence

Independence

An assumption of independence is that visual and language processing go on in parallel, with continuous, cascading integration (McClelland 1979). It is contrasted most notably with suggestions that fixations in the visual world provide a contaminated index of language processing, on the assumption that, for example, names of visible objects are automatically activated.

A simple example of an independence assumption comes from Allopenna et al. (1998), who modeled spoken word recognition with an encapsulated computational model (TRACE; McClelland and Elman 1986). Activations in TRACE were then linked to the four-alternative forced choice (4AFC) VWP task at a decision level: activations for the four displayed items

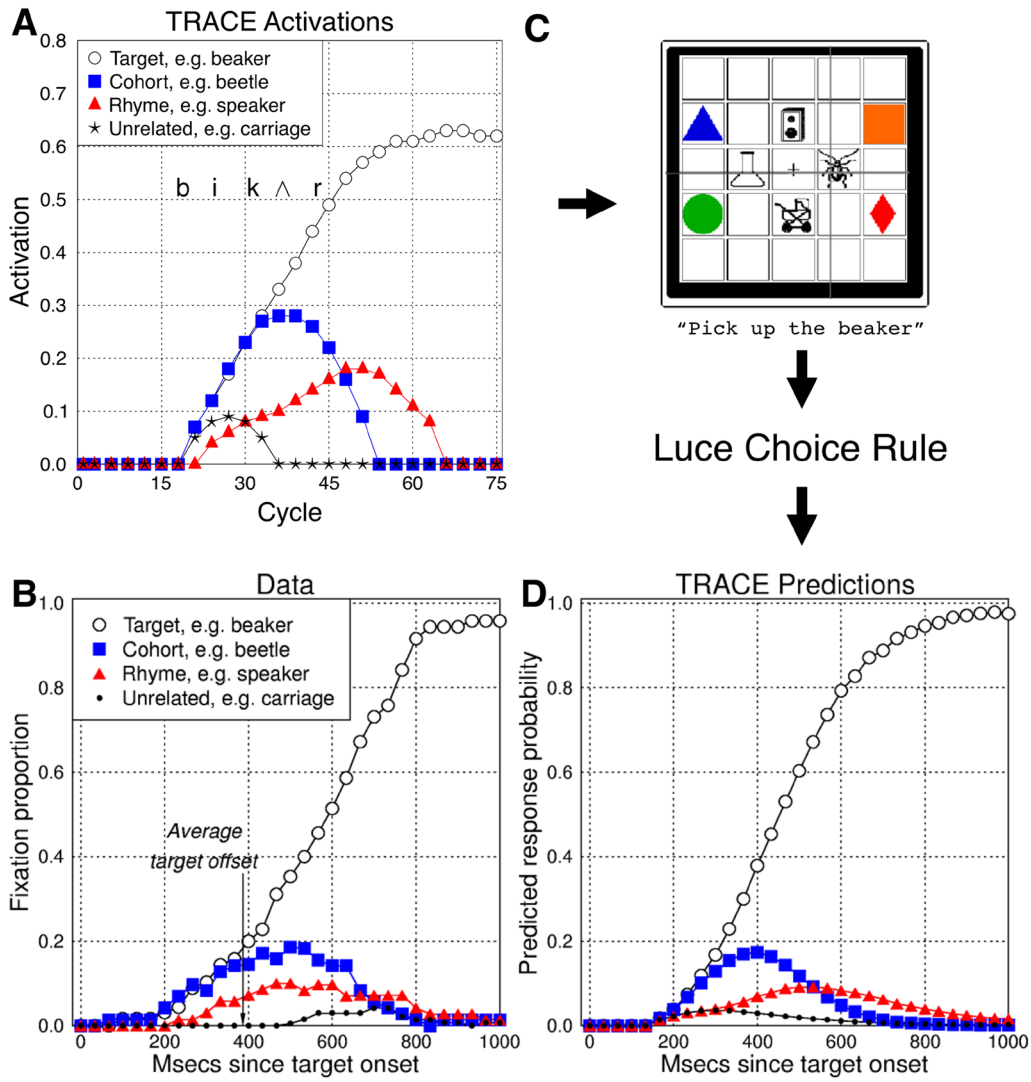

Fig. 2 The VWP linking hypothesis introduced by Allopenna et al. (1998). Raw TRACE activations (**a**) are clearly similar to human performance data (**b**). A formal linking hypothesis that takes into account the four-alternative forced choice task (**c**) performed by human subjects using the Luce Choice Rule results in tight fits between the model and human performance

were plugged into the R.D. Luce (1959) Choice Rule (LCR; see Fig. 2).

First, a simulation was conducted with the target item as input to TRACE. Activations reflected the dynamics of the full (212-word) lexicon without any consideration of visual context. To link those activations to the task faced by participants, the activations of the four displayed items at each $\sim$ 10 ms time step were extracted from TRACE. Then, response probabilities were calculated at each time step using the LCR. On the basic approach, each word's activation at each time step is converted to a response strength, calculated as e^{ak}, where e is the base of the natural logarithm, a is the activation, and k is a constant that determines scaling of values. At low values of k, larger values are only slightly amplified; as k increases, the response strength function becomes increasingly exponential. Then response probabilities are calculated by normalizing response strengths at each time step (that is, the response probability for an item is the ratio of its strength to the summed strengths of all four items). Thus, as k increases, it has an increasingly stronger inhibitory impact; as soon as one item has a slightly higher value than the others, the difference

between it and the others will be amplified with higher values of k.

Allopenna et al. used a slightly more complex version of the LCR, where k increased over time. In their Table 3, Allopenna et al. show that this sigmoidal k provides substantially stronger fits than a constant value (of 7) when model response probabilities are compared to human data. This provided excellent fits to human performance data (see Fig. 2), but in subsequent papers (e.g., Dahan et al. 2001a, b; Li et al. 2019; Simmons and Magnuson, accepted), we have typically used the simpler approach, with a constant value of $k = 7$ [note that a value of 7 works well for 4AFC, but larger k values are appropriate for choices among dozens or hundreds of words; Frauenfelder and Peeters (1998) found that a value of 20 worked well for 200–900 word lexica]. Another consideration is how to relate processing time in the model to real time. A simple heuristic that works quite well is to calculate the mean duration of target items used with human subjects and divide that value by the mean number of phonemes per target. Then one does the same thing with TRACE analogs: determine the mean target duration in TRACE time steps, and divide by the mean number of phonemes per word. One can then relate number of milliseconds per phoneme in the human items to number of cycles in the TRACE items.

Some of these details are far removed from theoretical commitments, such as relating model time to real time or the specific value of k to use in the LCR, or even *how* to model the decision level. Also, the choice to model central tendencies—probabilities over time—rather than saccades must be acknowledged (though in unpublished simulations, I have found it quite easy to quickly recover central tendencies like those in panel D of Fig. 2 from a simple saccade model; thus, I do not believe this is a *crucial* consideration). Here, however, our concern is with the theoretical commitments within this linking hypothesis. There appear to be four. (1) Language processing proceeds independently of vision. (2) Names of visual objects are not automatically activated. (3) Even without being fixated, visual objects (those not too peripheral from the point of gaze) are recognized to at least a coarse level that allows gradual matching; as speech is processed, and lexical items are activated in memory from phonological form, fixations are drawn to objects as a function of their relatedness to visual features associated with activated lexical representations. (4) Thus, on this linking hypothesis, fixation proportions over time provide an essentially direct index of lexical activation: the probability of fixating an object increases as the likelihood that it has been referred to increases.

Several findings support independence. First, in Allopenna et al. (1998), fixation proportions over time map closely onto phonetic similarity over time. While this does not preclude the possibility that pictures activate or prime their names, potentially *boosting* phonological competition effects, if the competition effects were completely *driven* by visual-based activations, it is not clear that fixation proportions should map so tightly to over-time predictions of purely phono-lexical models like TRACE or TISK (Hannagan et al. 2013). Later work also showed that the VWP is exquisitely sensitive to phonetic detail, such as misleading coarticulation (Dahan et al. 2001b) or subtle differences in vowel duration and other prosodic cues to word length (Salverda et al. 2003).

Second, the VWP is sensitive to lexical dimensions, not just the contents of the display. In their Experiment 2, Dahan et al. (2001a) displayed high- and low-frequency targets among three unrelated distractors and found robustly steeper target fixation proportion increases for high-frequency items. Magnuson et al. (2007) used 4AFC displays where items had negligible phonological, semantic, or visual overlap (e.g., *fox, chisel, sofa, blender*). They crossed target frequency, neighborhood density, and cohort density of target words.[2] Targets were displayed among three unrelated items, and participants followed a similar procedure as in Allopenna et al. (1998): they followed a spoken instruction (e.g., *click on the fox*) to select the named target item. Magnuson et al. found significant effects of all three factors (target fixation time series were steeper for high frequency or low cohort-density items, and there was an early advantage for high neighborhood density but a late advantage for low neighborhood density targets that appeared to result from differences in the proportion of neighbors that were also cohorts). Thus, whatever the influence of

[2] Neighborhood density was defined as the ratio of a target word's log frequency to the summed log frequencies of all words differing by no more than one phoneme; cohort density was defined as the ratio of a target word's log frequency to the summed log frequencies of all words overlapping in the first two phonemes with the target.

items in the visual display, the VWP is sensitive to lexical factors that are independent of the display; the speed with which participants are likely to fixate a pictured target depends on its frequency *and* the number and nature of its phonological competitors, even when none of them are displayed.

Third, Dahan et al. (2001b) demonstrated how the contents of the VWP can impact the magnitude of effects. Their study used subcategorical mismatches—target items were cross-spliced such that a medial vowel either had consistent or mismatching coarticulation for the final segment, and mismatches could either map to a word or nonword (e.g., for the target *net,* the final segment was always /t/, but the vowel was altered such that the /ɛ/ had coarticulation consistent with /t/, /k/ [*neck*] or /p/ [nonword **nep*]). In their Experiment 1, target objects were displayed with three unrelated distractors, and the TRACE-predicted pattern of consistent > nonword mismatch > word mismatch was observed. In their Experiment 2, a picture corresponding to the mismatching word (e.g., *neck* when the target was *net*) was included in the display, and the effects were substantially stronger. This is easily accounted for by the assumption that fixation decisions (saccadic targets) are based on applying something like the LCR to encapsulated lexical activations; when there is another object in the display with high relatedness to the emerging phonological form, it will attract fixations proportionate to its relatedness to the input. Note that these first three findings were the extent of evidence available when Tanenhaus et al. (2000) reviewed the VWP and articulated their linking hypothesis, which was essentially that language and vision could be considered more or less independent, with integration at a post-perceptual decision (saccadic targeting) stage. Additional relevant findings have been observed since, as I review next.

Fourth, looking to an item based on phonological information entails mapping phonology to the visual features of a displayed item. It would follow that competition should also be observed for items with similar visual features. Dahan and Tanenhaus (2005) found just such effects, such as an elevated probability of fixating *snake* when the instruction is to find the *rope*. The timing is similar to that for competition based on initial phonology (e.g., *beaker, beetle*), which is not surprising, since looking at a phonological competitor (or a target) entails sufficient visual processing to extract fundamental visual features. Dahan and Tanenhaus also note that this result casts doubt on the possibility that objects must be visually recognized and their names retrieved before language can guide fixations to display locations. If that were the case, one would expect competition to be strongly (possibly exclusively) driven by phonology. Having retrieved the names for both *snake* and *rope*, for example, shape features would have to be given extremely heavy weight in saccade targeting for shape features to overwhelm phonology. Instead, it appears that: phonetic information is mapped incrementally to lexical items; as lexical items are activated by phonetic input, corresponding visual features are activated and guide saccades.

Fifth, phonology also activates semantic features, and fixations can also be drawn to objects based on semantic overlap. For example, Huettig and Altmann (2005) reported that when participants heard a sentence that mentioned a *piano*, and there was also a *trumpet* in the display, participants were significantly more likely to fixate that semantic competitor than unrelated items. We know from cross-modal semantic priming (e.g., Marslen-Wilson and Zwitserlood 1989) that performance on orthographic lexical decision can be modulated by spoken primes via phonological links (e.g., hearing *beaker* can prime *insect*, a semantic associate of *beaker*'s phonological relative, *beetle,* suggesting that hearing *beaker* sufficiently activates *beetle* that activation spreads to its semantic relatives). Yee and Sedivy (2006) asked whether the VWP would be sensitive to such complex interactions. In a task where participants simply touched the screen location corresponding to a spoken word, semantic relatives had an elevated fixation proportion, as in Huettig and Altmann's (2005) study (e.g., on hearing *lock*, participants were more likely to fixate a picture of *key* than an unrelated item like *apple*). In addition, they found (somewhat less) elevated fixation proportions to an item like *key* when the target was instead *logs*, a phonological relative of *key*'s semantic relative, *lock*. This would seem to be another case where language leads vision; participants did not anticipatorily fixate semantic or phono-semantic relatives. Those fixations lagged behind speech, suggesting that phonological features activated semantic features and both were subsequently mapped to visual features associated with items with phonetic or semantic relations to the spoken input.

Notably, the VWP appears to be more sensitive to subtle semantic overlap than other methods. Mirman and Magnuson (2009b) followed up on earlier work (Mirman and Magnuson 2008, 2009a) investigating how graded semantic distance affects semantic judgements [using features elicited by humans (McRae et al. 2005; McRae et al. 1998) and modeled in attractor networks (Cree et al. 1999)]. In their VWP study, there was a strongly elevated probability of fixating near neighbors (*celery* given *broccoli* as target), and a more modest but still significantly elevated fixation probability for distant neighbors (*banana* given *broccoli*). This was remarkable given prior failures to detect reliable priming for distant neighbors in more conventional button-press paradigms (e.g., Cree et al. 1999).

Flexible parallel contingency

A likely and sometimes implicit corollary of the independence linking hypothesis is that linguistic input and visual scenes have a parallel-contingent (Turvey 1973) relationship, where processing occurs in parallel for two dimensions, but one is contingent upon the other. If the participant is following spoken instructions to interact with the display, her fixations must be guided by the spoken input. If we assume coarse object recognition proceeds in parallel but without automatic activation of object names, fixations will clearly be *driven* by linguistic input. Unlike some instances of parallel contingency [e.g., accommodation of talker differences in speech perception (Magnuson and Nusbaum 2007) or aspects of vision (Turvey 1973)], the contingency could reverse in the VWP. Hypothetically, if the experimenter instructed the participant to click on each *red* object as the experimenter named it, vision could lead language in that attention could be directed to scene locations where red was present. Indeed, Spivey et al. (2004) added a linguistic twist to a classic visual search paradigm, where time to respond varied with set size and target presence (there either was or was not a target item with a unique conjunction of color and orientation in the display). A spoken instruction (e.g., *Is there a green vertical?*) either preceded the visual display (auditory first) or was offset such that the first dimension (e.g., *green*) was aligned just after visual display onset (A/V concurrent). Participants were much less affected by set size (number of distractors) in the A/V concurrent condition. Spivey et al. proposed that in the concurrent condition, the timing of the spoken input allowed participants to serially filter the display, first identifying items matching in color and then in orientation; essentially, it allowed participants to conduct two single-feature searches. In the auditory first condition, having all of the details (both features) in advance relatively impaired performance because it promoted a conjunction (dual-feature) search (see also Chiu and Spivey 2014; Reali et al. 2006). Thus, the nature of the parallel contingency is flexible and can flux with simple parameters like relative timing.

Critical evaluation of Linking Hypothesis 1

The first linking hypothesis does not take a strong stand on *how* language and vision interact. In the case of spoken word recognition, excellent fits are obtained when a model like TRACE is treated as an independent system with no input from vision, and instead integration occurs at a decision (saccade targeting) process with cascading visual and linguistic inputs [modeled with the Luce (1959) Choice Rule]. In more complex cases (sentence processing, interpretation of scalar adjectives, etc.), so long as a behavioral 'outlet' is available to distinguish between interpretations (that is, objects that should be fixated at precise, different times according to competing hypotheses), the same approximate linking hypothesis holds. Such early-separation and late (if cascading) integration is interestingly at odds with theoretical positions embraced by many researchers who adopt Linking Hypothesis 1. For example, Tanenhaus et al. (1995) explicitly reject encapsulated linguistic processing based on their findings. All the same, a reader of Tanenhaus et al. (2000) might infer that the spoken word recognition linking hypothesis developed by Allopenna et al. (1998) and refined by Dahan et al. (2001a, b) is also proposed for more complex linguistic processing. However, the emphasis in the 2000 paper is on evidence from studies of spoken word recognition that visual displays do not distort language processing (as fixation proportions closely follow phonetic similarity over time, and effects of non-displayed linguistic knowledge are robustly observed); no strong claim is made for independence of visual and linguistic information for more complex processing (indeed, the paper ends with a strong case

that language processing is not encapsulated from other aspects of perception and cognition).

Nonetheless, there are at least two strong challenges to this view. First, the independence assumption for spoken word recognition remains at odds with theoretical frameworks such as interactive or constraint-based processing, as well as the rejection of encapsulation by Tanenhaus et al. (2000) themselves. Indeed, Allopenna et al. (1998) described independence with later integration as a provisional simplifying assumption. They noted (p. 438): "It will be important in future work to evaluate the degree to which this assumption is viable, and if it needs to be modified, how the modifications could limit the generality of conclusions from the visual world paradigm." Second, it is very difficult to *prove* that visual stimuli in the VWP do not activate linguistic representations prior to linguistic input. If they did, this would falsify the independence assumption. As we shall see next, several researchers, most notably Huettig and his colleagues, have challenged the independence assumption with emphasis on the issue of whether linguistic representations are automatically activated by visual displays (given adequate preview time).

Linking Hypothesis 2: displays load working memory and activate representations prior to speech

Huettig et al. (2011a, b) have proposed a starkly different linking hypothesis in which visual displays in the VWP alter language processing in important ways (labelled the "cascaded activation model of visual–linguistic interactions" by de Groot et al. 2016). On this account, when conditions allow sufficient time, comprehensive scene analysis is carried out, with a "restricted number" of displayed items loaded into visual working memory (approximately four, as Huettig et al. (2011a), review evidence that four is the average "maximum number of objects that can be efficiently prioritized, processed, cued, tracked, counted, and actively remembered"). On this account, the 'various representations' (phonological, semantic, and visual features; de Groot et al. 2016) of displayed items are activated when displayed objects are recognized. This presumably fundamentally alters language processing, in that these pre-activations will give an advantage to a displayed item if its name is heard (with associated complications for items related to the spoken word). Prior to spoken input, these pre-activations would presumably also drive spreading activation based on phonological, semantic and visual features. Thus, on this view, although processing is not *limited* to the displayed items, pre-activation makes language processing in the VWP contingent on the display contents. When speech occurs, fixations could be directed by partial matches to displayed items, though processing would be parallel-contingent, since matching to any feature would have to happen via phonological links. Furthermore, on this view, fixating specific objects based on phonological, semantic, or visual properties requires an internal representation that binds those dimensions to specific locations in the visual world. Figure 3 presents an elaborated schematic based on this proposal [cf. Figure 2 of Huettig et al. (2011a)]. On its face, this proposal is compatible with independent flexible parallel contingency: either visual or spoken input could drive processing, and both routes can operate in parallel; significantly, however, when vision leads language, language becomes (at least partially) contingent on vision.

Huettig et al. (2011a) propose that visual objects must be recognized and bound to both lexical representations and locations in visuospatial working memory before the VWP can reveal phonological form-based linguistic competition. The proposal is motivated in part by results reported by Huettig and McQueen (2007), who varied the visual preview time prior to the onset of target words in a "target-absent," "look-and-listen" paradigm. Targets were not displayed, and there was no task. Four items were displayed. One item shared visual features with the target word, one shared semantic features, one overlapped phonologically (in the first one or two phonemes), and one was unrelated. Targets were presented in neutral sentences with no connection to any of the displayed items. The experiment was carried out in Dutch, but an English analog to one item set would be a sentence like *Eventually, she looked at the belt in front of her*, while the pictures were displayed were *snake* (shape overlap), *sandal* (semantic associate), *bell* (phonological competitor), and *ashtray* (unrelated).

When pictures were displayed at sentence onset (presumably allowing a preview of 700–1000 ms), fixation proportions were as one might predict from previous studies, with fixations to phonological

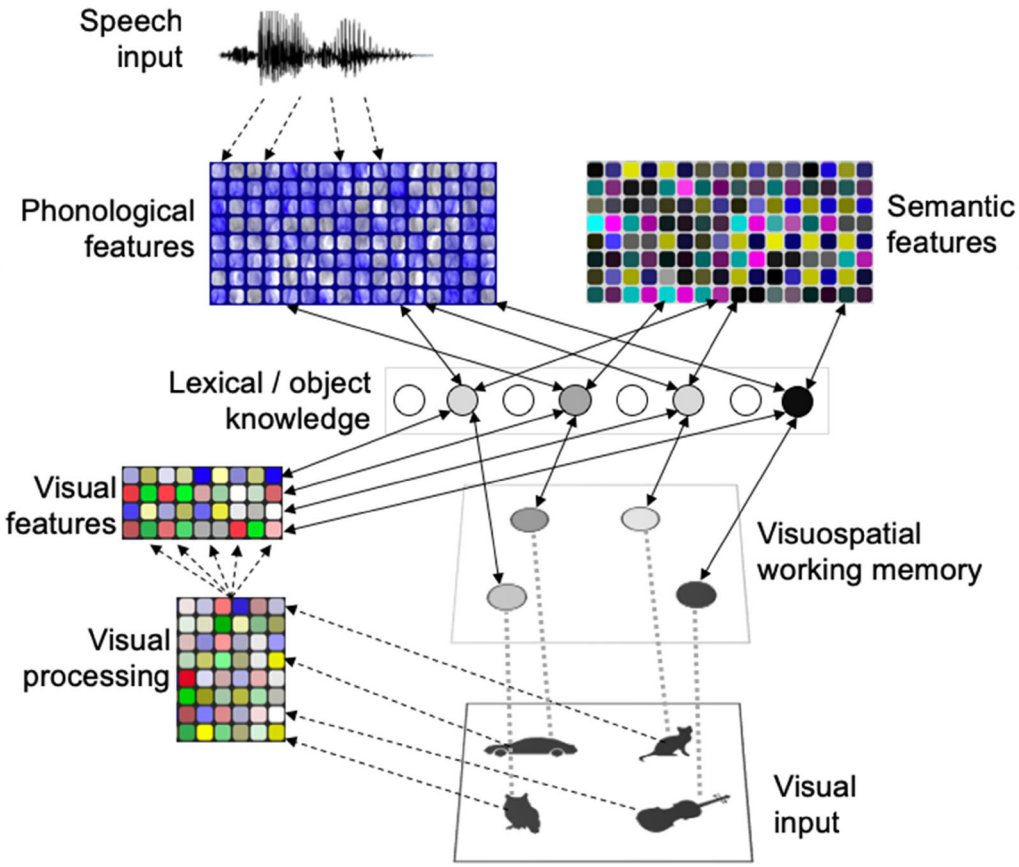

Fig. 3 An elaborated schematic of the working memory linking hypothesis from Huettig et al. (2011a) from Magnuson (2019b). Object positions in the visual input "seed" visuospatial working memory locations. Object recognition requires visual processing, which extracts visual features, which activate lexical-object representations. Speech input activates phonological forms, which also feed lexical-object representations. Lexical-object representations link to positions in visuospatial working memory (additional connections might be called for from visual processing and/or to visuospatial positions to allow binding, but are left out for clarity). Dashed arrows indicate unidirectional pathways. Dotted lines from objects to positions indicate connections based on coarse coding. Bidirectional pathways to lexical-object connections allow activation to spread even in the absence of input (e.g., phonological forms can be activated by visual objects, given sufficient time). Thus, for example, lexical-object representations can activate visual features, which can, through resonance, activate other lexical-object representations associated with those features

competitors elevated in comparison with unrelated items approximately 250–650 ms after target onset, and elevated shape- and semantic-relative fixation proportions from approximately 350 ms onward. However, when the preview time was reduced to 200 ms, fixation proportions to phonological competitors did not differ statistically from unrelated items, though fixation proportions were still elevated to shape- and semantic-relative items (from approximately 500 ms and 600 ms, respectively). Huettig et al. (2011a) interpret this as indicating that phonologically-guided fixations will only be observed once the phonological forms of the names of objects in the display have been retrieved.

Critical evaluation of Linking Hypothesis 2

On this account, names (phonological forms) cannot be bound to locations in visuospatial working memory until enough time goes by for visual object recognition to sufficiently activate them (see Fig. 3). However, this interpretation suggests a paradox. The basis for fixations to items overlapping with a target like *belt* in shape or semantic features is *its phonological form.*

This means that in the 200 ms preview case, 200 ms is sufficient for visual and semantic features to be associated with locations in visuospatial working memory. Paradoxically, this entails that the phonology of the target word is activated sufficiently to activate visual and semantic features associated with it, *and visual features of semantically-related items*, but *not* visual features of an item overlapping phonetically at word onset. However, for semantic features to drive fixations to competitors, those features must be linked to appropriate visual features in order to target a saccade. The critical puzzle is that there is no way for the spoken word recognition system to "know" the target is *belt* and not *bell* until the /t/ is heard (leaving aside potential coarticulatory information for this case). If the target were displayed, one would expect that fixations to the target would be equal to or greater than the fixations to any other item, given the bottom-up match to the target's name. Since a token of *bell* should be virtually indistinguishable from a token of *belt* in the first three segments, it is puzzling that the probability of fixating phonological (cohort/onset) competitors was no greater than the probability of fixating an unrelated distractor in the early time course.

Second, note that the VWP variant used by Huettig and McQueen (2007) differs in two important respects from the original VWP studies. First, it is a "look-and-listen" paradigm where participants are not asked to perform any task. As discussed by Vivianni (1990), a task-free paradigm is open to myriad strategies, and is possibly more susceptible to "good subject" effects (where participants aim to produce the behavior they infer is desired by the experimenter) than other paradigms (such as instructions to perform visually-guided motor tasks, as in the original version of the Tanenhaus VWP: Tanenhaus et al. 1995; Allopenna et al. 1998; etc.). Of course, given the large number of studies using look-and-listen paradigms, it seems the proof is in the pudding; given that fixations over time still map onto emergent speech input and its relation to the visual scene when the participant has no explicit task to perform, it appears to be safe to assume that look-and-listen paradigms do not necessarily engender more strategic processing (and see Mishra et al. 2013, who make the case that look-and-listen paradigms are not more susceptible to strategy because language-mediated fixations are an overlearned, semi-automatic behavior).

However, the second divergence from original VWP studies may be problematic; the use of the "target-absent" paradigm, where the target is not included in the display. What is the linking hypothesis in this case? Intuitively, fixations should be drawn to items in the display to the degree that their visual, semantic, or phonological features have been activated. However, what strategies might emerge given this unfamiliar task, where scenes are accompanied by spoken language that (at least occasionally, depending on fillers) does not refer to items in the display? Although it seems unlikely that participants would become explicitly conscious of shape, semantic, or phonological relations between sentences and displays, it is difficult to know what strategies might arise when targets are not present.

The impact of the relative timing of pictures and words also must be considered with respect to the needs of visual and linguistic processing, as well as saccadic control. A recent review estimates the time needed to recognize visual objects at 80–240 ms (Contini et al. 2017). Estimates that one needs approximately 200 ms to plan and launch an eye movement do not include prerequisites such as time needed to recognize individual objects or at least coarsely code elements of a visual scene. Speculatively, a preview of 200 ms may simply be too fast to allow first fixations to reflect integration of visual and phonological information (similar to how it is difficult to detect word frequency effects in the early time course of spoken word recognition; Dahan et al. 2001a). If participants are likely to launch at least one saccade (to a random location) during the 200 ms preview, signal-driven fixations would be likely to be second or third fixations,[3] and would likely be triggered past the point of disambiguation between the target and its phonological competitor [Huettig and McQueen (2007) reported that those items overlapped by an average of 192 ms]. If the opportunity to launch a saccade is after the point of disambiguation, the probability of fixating the phonological competitor might well be reduced.

Of course, a fundamental premise of Linking Hypothesis 2 is that visual displays activate linguistic representations prior to speech input. McQueen and Huettig (2014) review several studies that suggest that

[3] I thank Michael Spivey for pointing out the implications of this preview time for saccade timing.

visual objects activate corresponding linguistic representations. For example, Noizet and Pynte (1976) reported a task where participants were directed to sequentially fixate three items and silently identify them, and found that fixation durations were proportional to word length. Dell'Acqua and Grainger (1999) found priming from isolated pictures to isolated printed words when pictures were presented prior to words, and participants made semantic judgements about the words. McQueen and Huettig (2014) presented a series of studies where participants saw a visual prime (image or printed word) and then heard a spoken token and made a lexical decision to it. Robust inhibitory priming was observed when visual prime names were phonologically similar to the following spoken word, while robust facilitation was observed for semantic relatedness; McQueen and Huettig concluded that participants must have activated phonological forms in response to visual primes. Mani and Plunkett (2010) found faciliatory priming for phonologically-related picture primes with 18-month olds, and drew the same conclusion. I will return to a discussion of these findings after discussing one other.

A study by Zelinsky and Murphy (2000) delved deeper into the question of whether names are obligatorily accessed and stored in working memory when pictures are presented, using a paradigm closer to the VWP than in the studies just reviewed. Zelinsky and Murphy used variations of the VWP to create tasks that varied in the potential utility of using object or person names. Given displays of four objects to examine (where half had one-syllable names and half had three-syllable names) and a subsequent single-item display with the task of indicating whether or not the single item had been in the previous display, number and duration of fixations were proportional to the phonological length of an object's name. The same result pertained when faces were used rather than objects, after participants were trained to map names to eight unfamiliar faces, where half the names were one-syllable long and half were three-syllables long. However, there was no effect of name length when the task was reversed (with a single target item displayed first, and the task was verifying whether it was included in a subsequent four-item display, although this variation was only carried out with newly-learned faces). This suggests that when the task does not *require* phonological encoding of display items, phonological encoding does not occur, although evidence from a study employing a task closer to the standard psycholinguistic VWP might be more compelling.

The Zelinsky and Murphy finding also challenges the findings reviewed above that suggest pictures activate names since pictures can phonologically prime spoken words. The common concern in each of those studies is that isolated or sequential images were followed by speech, over and over. The tasks do not require pictures to be named (in contrast to the condition in Zelinsky and Murphy where verbal recoding promotes performance), but they differ substantially from the VWP and may induce internal responses (strategic or not) that would not be induced by the VWP.

Intuitively, the assumption that seeing objects automatically activates their names does not seem plausible in the real world. Consider the experience of walking through a room, driving down a road, sitting down at a desk strewn with objects, or glancing at a row of photographs in a hallway (maybe even including people you know). While intuition is a notoriously poor foundation for theorizing, it seems implausible that the names of objects are implicitly activated as you pass trees, trestles, tracks, and trailers on the highway, or view staplers, staples, pens, papers, cups and computers on a desktop. The proposal that seeing all objects in a scene, or even attending to a subset of objects, automatically triggers retrieval of phonological word forms is reminiscent of a heads-up display, like that available to "the Terminator" in the 1984 film, where objects in the visual world are tagged with orthographic labels. In my own experience, I am certainly unaware of activating the names of objects or people I see. (Indeed, quite the opposite; many years ago, I had the mortifying experience of happily chatting with my great aunt, when the occasion arose that I should introduce her to a guest; I drew a complete blank after "Aunt".) Similarly, evidence for cross-language competition also raises plausibility concerns. Spivey and Marian (1999) found that Russian-English bilinguals showed competition between markers and stamps in English-only and Russian-only contexts due to phonological overlap between the English word *marker* and the Russian word for *stamp*, *marka*. A challenge is that this result, on the logic of Linking Hypothesis 2, implies that names of all displayed items in all languages a participant is fluent in should be loaded into working

memory. On the one hand, this may seem implausible, especially for individuals who speak more than two languages, but on the other, it suggests a testable hypothesis: on Linking Hypothesis 2, performance should degrade more quickly for multilingual vs. monolingual participants as the number of elements in the display increases (since the amount of phonological material in working memory will be larger for individuals with more than one language).

Finally, note that at least one prior study suggests that many (but not all) aspects of the dynamics in the VWP are unaffected by display timing when targets are present in the display. Sedivy et al. (1999) found similar patterns of on-line competition in a VWP study of context-dependent scalar adjective interpretation (e.g., *tall glass*) with previews of 20 s or less than a second (though fixations were initiated sooner given a 20-s preview).

This discussion suggests that the inferences drawn by Huettig and McQueen (2007) and Huettig et al. (2011a) should be considered provisional. However, it may be that the VWP is less like the real world than one might hope, and repeated presentations of four objects and then an instruction to click on one may induce unnatural modes of processing (as I suggested might happen for repeated single picture-spoken word sequences). A study that used a target-present, visually-guided motor task (e.g., *click on the belt*) with an incremental manipulation of preview time would provide a stronger test of the proposal that phonological competition effects in the VWP can only be detected once visual stimuli have driven the binding of visual and phonological forms via visuospatial working memory. Another avenue for refining this account would be to link it to an implemented computational model, which might reveal predictions that could robustly distinguish this account from others.

Linking Hypothesis 3: the mental world hypothesis

Altmann and Kamide (2007) proposed an additional linking hypothesis that we might call the "mental world" hypothesis, which follows from their perspective on what is implied by anticipation in the VWP and other tasks. The classic example of anticipation comes from a study by Altmann and Kamide (1999). In their example display, a boy was pictured in a room, with several objects nearby (cake, ball, toy car, toy train). If participants heard a sentence that began *the boy will move*, they were equally likely to anticipatorily fixate any of the objects, but if the sentence began *the boy will eat*, they were much more likely to anticipatorily fixate the cake. There have been several extensions of this seminal finding (e.g., Knoeferle and Crocker 2006, 2007, who documented how interactions between scene and speech allow anticipatory activation of thematic roles and other functions). I will focus on one example from Kamide, Altmann and Haywood (2003). They used displays with two agents (man, girl), two rideable items (motorbike, carousel), and two tasteable items (beer, sweets). Participants heard sentences like, *the man/girl will ride/taste the…* In these conditions, when the verb is heard, the subject (girl or man) rapidly shifts fixations to the more age-appropriate object (e.g., to *motorbike* given *the man will ride*). Kamide et al. proposed that this suggests "an incremental processor that establishes the fullest possible interpretation at each moment in time" (p. 153).

Altmann and Kamide (2007) extended the "fullest possible interpretation" idea to account for the inferences participants appear to make in VWP studies when they integrate various kinds of context. Altmann and Kamide and other authors have documented anticipatory eye movements that reveal predictions based on complex contingencies among displayed objects and linguistic details and discourse-constrained affordances (e.g., looking to an empty or full glass depending on tense [*will drink*], and compatibility of agents and actions, etc.). Altmann and Kamide (2007) propose that anticipatory eye movements in the VWP do not simply "reflect the unfolding language processing; they reflect an unfolding (mental) world" (p. 515). They develop a linking hypothesis that is similar to Linking Hypothesis 1, but extend it to a focus on 'conceptual representations' where the role of phonological form has no special status—it is just one of many aspects of human knowledge about objects and the world. They add to the list of conceptual dimensions the context- (visual and discourse) specific *affordances* of an object: roughly, what it can be used for in service of an agent's goals. They argue that internal conceptual representations that take into account visual and discourse context are required to explain, for example, sensitivity to tense (*will drink* promotes looks to full vessels, while *drank* promotes looks to empty ones) implies internal conceptual representations modulated by discourse

implications, and saccade targeting that follows from automatic guidance of visual attention by conceptual representations. For the listener-viewer to be sensitive to how a scene relates to unfolding discourse, a "mental world" must be represented, where conceptual representations for displayed or discourse-referenced items or actions not present in the visual display can be updated.

Critical evaluation of Linking Hypothesis 3

Notably, the Altmann and Kamide (2007) mental world hypothesis appears to take a step back from the Kamide et al. (2003) proposal that anticipatory and discourse-guided fixations imply "an incremental processor that establishes the fullest possible interpretation at each moment in time". Altmann and Kamide (2007) emphasize that activation of conceptual representations and subsequent "language-mediated eye movements are little different theoretically (and perhaps no less automatic behaviorally) than priming effects which have elsewhere been explained in terms of spreading activation and/or conceptual overlap" (p. 514). However, they do not explicitly discuss their previous "fullest possible interpretation at each moment" proposal. This leaves a small but important theoretical gap: an account that appeals to a process like priming may be starkly different from an active search mechanism that maximizes prediction by forecasting upcoming language. Huettig (2015) discusses a possibility that may offer a resolution: the possibility that there is a "smart" but slow, active and controlled process [similar to Kahneman's (2011) *System 2*] and a fast but "dumb", mostly automatic process (or set of processes) that operate in the priming-like fashion suggested by Altmann and Kamide (2007), akin to Kahneman's *System 1*. (See Huettig 2015, for a review of prediction and trenchant assessments of theoretical implications that are beyond the scope of the current review.)

A study by Kukona et al. (2011) may support this proposal. The study originated when a group of students at the University of Connecticut were skeptical of the "fullest possible interpretation" claim, and the active forecasting it implies. For example, in Kamide et al. (2003), agent-verb fit did not extinguish fixations to agent-inappropriate objects (participants fixated the motorbike more after *the girl will ride* than after *the girl will taste*, even if they fixated it less than after *the man will ride*). They hypothesized that semantic and/or statistical relations between verbs, agents, and patients could contribute to such findings via a more passive process—possibly priming (and thus possibly compatible with the Altmann and Kamide 2007, proposal). To address this possibility, Kukona et al. (2011) assembled sets of verbs, agents, and patients that were strongly thematically related (e.g., *arrest-policeman-crook*). They presented participants with a scenario in which they were going to hear about the adventures of "Toby", an adventurous graduate student. Every sentence would be about something Toby would do. Toby was pictured at the center of the display on every trial. Every sentence had the same pattern: *Toby will* VERB *the* PATIENT (e.g., *Toby will arrest the crook*). Critical displays included images of the agent and patient associated with the verb (e.g., policeman and crook) and two unrelated distractors. Since Toby *always* filled the agent role, fixations guided by an incremental processor establishing the fullest possible interpretation at each moment should not be directed at the verb's associated agent. However, we found virtually identically elevated fixation proportions to the verb's associated patient (as predicted) *and* agent in an anticipatory region preceding the point where saccades were likely to be influenced by the bottom-up specification of the patient. In a second experiment with different participants, the scenario changed to one where all kinds of things happened *to* Toby; every sentence was about something that happened to him, with a form like, *Toby will be arrested by the policeman*. Here, a predictive processor should rule out the possibility that the verb's associated patient would be heard, since Toby always filled the patient role. In this case, with more syntactic cues to the upcoming role and more time between the verb and second noun, there were again virtually identically elevated fixation proportions to the verb's associated agent and patient initially (in the *by the* region) but a significant advantage emerged for the agent during the first 200 ms of the agent in the speech (too early for the changes in fixation proportions to have been driven by that word).

Kukona et al. (2014) pursued a different line in considering evidence for anticipation. They used items that echoed those of Altmann and Kamide (1999). For example, one critical display had a slice of white cake, a slice of brown cake, a white car, and a brown car. This necessitated a complex referring expression for

unambiguous reference, e.g., *the boy will eat the white cake*. Given a sentence like this, Kukona et al. (2014) replicated the main finding of Altmann and Kamide (1999): there were anticipatory fixations to both pieces of cake following *eat*. However, they also observed a small but significant elevation in fixations to the verb-incompatible white *car* following *white*, suggesting a continuing interplay of bottom-up signal and top-down expectations (in a second experiment, they replicated this result while eliminating the possibility of unusual activation of implicit contrast from the pair of cars of different colors, replacing the brown car with a toy train). Kukona et al. (2014) interpreted these results as consistent with previous evidence for *local coherence effects* (e.g., Tabor et al. 2004; Tabor and Hutchins 2004), where competition is observed between global and local context scopes in sentences. For example, in *the coach smiled at the player tossed a frisbee*, the surface form of *the player tossed a frisbee* is itself a grammatical sentence in English. Of course, in the longer context, it is a reduced relative clause (*the coach smiled at the player [who was] tossed a frisbee*). But comprehenders experience difficulties consistent with competition between the globally- vs. locally-consistent parses. If the processing system were making maximal use of context, it should not consider parses or (direct objects) rendered impossible by context. Kukona et al. also note that rhyme competition reflects an analogous local coherence at the level of mapping speech sounds to phonological word forms (the global context includes the word's onset, which is incompatible with the rhyme; nonetheless, rhymes are still substantially activated due to their *subsequent* high degree of similarity with a spoken target).

Another interesting theoretical question with respect to the mental world hypothesis is what aspects of processing in the visual world require internal models. I agree with Altmann and Kamide (2007) that cases where listener-viewers show sensitivity to a change of state implied in the future or past are difficult to explain without appeal to updated internal representations. However, they may attribute more work to internal representations than is necessary with respect to affordances. Altmann and Kamide reject the Gibsonian view (Gibson 1979):

> Similarly, the affordances of an object are not 'out there', but are experientially-based encodings requiring a representational substrate that encodes information that goes beyond that conveyed by the scene itself. Thus, anticipatory eye movements do not reflect the mapping of language onto the scene itself, but rather, onto some dynamically interpretable representation of that scene and the affordances it contains. (Altmann and Kamide 2007, p. 511)

It is notable that ecological theories do not require affordances to be "out there" independent of the perceiving-acting organism (indeed, they are only definable with respect to potential environment-organism interactions). They may fall on a continuum from those that are as apparent as any visual feature (e.g., a doorway tall enough that I can pass through without bending, or a handle on a small mug large enough for me to grasp) to others that are more like "hidden" (non-visual) semantic features that indeed depend on experience and may need to be learned (Gibson and Pick 2000) and "recognized" (see Greeno 1994, for an extended discussion)—possibly learning what plants and berries are edible, or learning what kind of trees have branches sufficiently non-brittle to pry a rock or what kinds of stone are slippery when wet, or, possibly, learning about thematic relations (who/what is likely to interact in certain ways with other actors or objects). The fact that language can be used to specify goals or contexts that modulate the relevance of various affordances, potentially including ad-hoc ones that one would not expect to be "present" otherwise (*the boy needed [something to stand on] to reach the cookie*), indeed poses interesting challenges for ecological *and* cognitive theories, and Altmann and Kamide (2007) have taken an important step in this direction.

Should theorists strive to avoid or minimize proposing internal elements? There are examples from other fields that suggest this is the case. Major advances in computer and human vision followed from two revolutionary insights. In computer vision, "animate" and "active" approaches abandoned the goal of constructing and maintaining comprehensive models of the visual world; instead, seeming limitations of biological visual systems (e.g., limited high-resolution foveae, limited field of view) were embraced in robotic vision systems, allowing massive reductions in computational complexity (e.g., Ballard 1991). In studies of human vision, it became clear that we often maintain only very coarse internal

representations of the world around us (e.g., Rensink et al. 1997).[4] It appears that a highly successful biological strategy that also happens to be extremely computationally efficient is to use the world as memory, sampling from it when information is needed. Thus, I would suggest that radically internalist and externalist views provide a useful tension when it comes to the representations required by the VWP, vs. how much comes for free by being present in the world (even when cognition and perception have become attuned to external affordances through experience and learning); the truth likely lies somewhere in the middle, but having diametrically opposed theories can motivate strong, falsifiable hypotheses.

I will make two final points on different aspects of Linking Hypothesis 3. First, consider a connection between Linking Hypotheses 2 and 3. If we were to take the Huettig et al. (2011a) hypothesis to a possible extreme, and merge it with the mental world hypothesis, would propositions relevant to possible events afforded by a scene need to be pre-activated linguistically before fixations could be guided by event descriptions? This seems unlikely; perhaps this should make us skeptical about pre-activation claims even for object names, but perhaps the time scales and complexities of these two situations are incommensurate.

Finally, a gap with the mental world hypothesis (as with Linking Hypothesis 2) is the lack of an implemented model. Implementing models can reveal complications or simplifications that are not apparent before one grapples with operationalizing theoretical proposals (see Magnuson et al. 2012, for a discussion). Without such a model, it is possible that updating and reactivating conceptual representations may be more (or less) challenging in practice than we might anticipate.

Linking Hypothesis 4: just-in-time deep interaction

A fourth hypothesis is compatible with the general framework suggested in parallel distributed processing approaches (especially interactive activation, e.g., McClelland and Elman 1986; McClelland and Rumelhart 1981), Spivey's (2007) *continuity of mind* perspective, and constraint-based theories of linguistic processing (e.g., MacDonald et al. 1994; McRae et al. 1998; Trueswell and Tanenhaus 1994). Let me note in advance that the proposal developed as Linking Hypothesis 4 will have much in common with all three previous linking hypotheses, although it will also have important differences.

On an interactive activation view, bidirectional interactions are pervasive and continuous throughout biological systems (see Fig. 4). Spivey's (2007) continuity of mind perspective merges this view with radically ecological and embodied approaches to perception, action, and cognition. Internal encoding is avoided unless the task requires it (e.g., if one can anticipate that displays will be brief, or that a memory task is upcoming). Instead, the world can be used as memory, since the visual world is normally persistent. Also, on such views, there is no discrete store of lexemes or concepts. There is, as Elman (2009, 2011) famously put it, lexical knowledge without a lexicon. Instead of such discrete units as the locus where features are bound, the *experience* of hearing, seeing, or multi-modally perceiving an object arises from coordinated, over-time patterns of activation of lexical (phonological, semantic, thematic…) and sensory features. A gap in the schematic in Fig. 4 is how the over-time flux of multimodal featural activations are related to spatial locations. However, this gap is also a concern for Fig. 3 (or Fig. 2 in Huettig et al. 2011a). There, it is stipulated that locations are encoded in working memory. Here, we would need to stipulate that visual features must mediate the *effective* (not literal) binding of non-visual features (e.g., phonological and semantic features) to scene locations; learned associations or causal relations between an object's visual feature and its name, affordances, and "hidden" semantic features link the latter features with locations where relevant visual features are present *to the degree that they associate* with those features.

On this view, language processing and visual processing have dual parallel-contingent relations (as allowed under the three preceding linking hypotheses). To know which object is being referred to, speech input must activate phonological features, which will activate corresponding visual and semantic features; thus, phonological activations will clearly lead visual

[4] However, *attended* scene details are less subject to change-blindness and more likely to be available in long-term memory after processing (e.g., Hollingworth and Henderson 2002), although susceptibility to change-blindness is also modulated by in-the-moment needs for perception and action (Ballard et al. 1997).

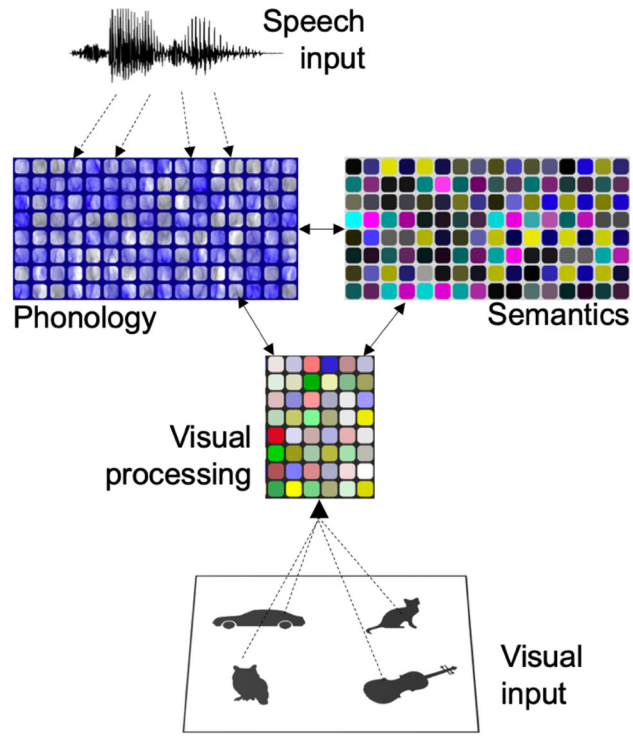

Fig. 4 An interactive activation/continuity of mind perspective on the VWP (Magnuson 2019c). External inputs activate complexes of features, but by default, there is no internal reconstruction of objects or scenes. The world serves as memory, and features are related to scene elements via visual routines and processes

feature activations given speech input. Conversely, visual features will lead given visual input. The relative timing of visual and speech inputs will result in a "tug of war" among phonological, semantic, and visual information, as Huettig and McQueen (2007) put it. Degree of linguistic vs. visual dominance will depend on timing and task. Crucially, we must keep in mind that neither word nor object processing is simply categorical; as Balota (1990) put it, our theories appear often to assume that there is a "magical moment" when a word is recognized. Balota and others (e.g., Elman) have made a compelling case that word recognition is gradual. In fact, in models like TRACE or a Simple Recurrent Network (SRN; Elman 1990), there is no model-internal recognition state. Rather, activations wax and wane over time. Similarly, it is increasingly well-established that visual object recognition is gradual (Contini et al. 2017), with neural responses to visual objects sharpening over time (e.g., after 50–100 ms, neural signals distinguish faces from bodies, and artifacts from natural kinds, and finer distinctions emerge over as much as another 100 + ms).

However, as I discussed regarding Linking Hypothesis 2, there may still be reason to expect language to dominate in the VWP, and to be skeptical about visual objects automatically activating object names. Furthermore, there would be no need to do so if one can rapidly map speech inputs onto relevant items through associative and other relations between linguistic and visual knowledge, or actively retrieve names 'just in time' when they are needed. On this view, in the VWP, if objects are close enough to the point of gaze and each other (or are sufficiently large if they are further in the periphery), coarse visual information is enough to begin mapping phonology to visual features. When correspondences to one location vs. another are strong enough, a saccade can be launched.

A key aspect of this view is the possibility of *deep interaction*. Sufficient visual context could drive anticipatory linguistic activations and vice versa. The triangle-model-like connectivity in Fig. 4 (cf. Harm and Seidenberg 1999, 2004; Plaut et al. 1996; Seidenberg and McClelland 1989) allows for potentially strong interactions between feature types (though note that it is easy to parameterize interactive activation models to preserve bottom-up priority and avoid feedback-driven hallucination; McClelland and Elman 1986). Note also that the triangle model also implements *lexical knowledge without a lexicon*.

There are no lexical units; just pattern nodes for phonology, orthography and semantics, with banks of hidden nodes between pattern nodes. When input is applied to one or more pattern layers, if the pattern corresponds to a word the model has learned, it will gradually settle into appropriate attractor states through waves of interaction between pattern layers. Whether such an approach can be generalized to a model based on Fig. 4, where persistent visual inputs provide the "glue" to bind across dimensions, is a challenge for future research.

Next, let us consider evidence for deep interaction. Deep interaction would be cases where, for example, the contents of the visual display would alter linguistic processing (or vice versa). One might reasonably argue that deep interaction was observed in the very first VWP study by Tanenhaus et al. (1995). In that study, the premise was that if language processing proceeds in modular stages, with syntax evaluated prior to integration of meaning, so-called garden path effects should still be observed even if a visual context could potentially constrain interpretation. On critical trials, participants heard an instruction like *put the salt shaker on the envelope in the bowl*. A sentence like this should require reanalysis, since on the simplest incremental parse *on the envelope* should be the target location for the salt shaker. On hearing *in*, participants who have strongly activated that incremental parse would have to reanalyze the utterance, most likely as a reduced relative (*put the salt shaker [that is] on the envelope in the bowl*). Tanenhaus et al. paired such instructions with "helpful" and "unhelpful" displays. In both, there would be a salt shaker on top of an envelope, but also an isolated envelope—a plausible target location for *on the envelope* (until *in the bowl* is encountered). In the helpful display, there was an additional salt shaker (on a napkin), along with two unrelated items. In the unhelpful display, an unrelated item replaced the second salt shaker. In the unhelpful condition, participants appeared truly garden pathed. At *salt shaker*, they fixated the sole salt shaker. At *on the envelope*, they fixated the empty envelope. At *in the bowl*, they appeared to be quite confused, looking back and forth between the empty envelope and the salt shaker, before usually picking up the salt shaker and placing it in the bowl. In the helpful condition, participants showed no sign of being garden pathed. At *salt shaker*, they were equally likely to fixate either salt shaker, but at *on the envelope*, they settled on the salt shaker that was on an envelope, and almost never fixated the empty envelope before fixating the bowl at *in the bowl*. This appears to be an instance of deep interaction; the syntactic structure incrementally assumed by the listener as she hears the sentence is modulated by the pragmatic implications of the display. Given a single salt shaker, it is somewhat infelicitous to specify any more detail than "salt shaker", whereas, given two salt shakers, a more complex referring expression is required to indicate *which* is the intended target.

One might debate whether this should really suggest deep interaction, rather than parallel, cascading scene analysis and sentence processing, as in either Linking Hypotheses 1 or 2. Has the visual context truly altered linguistic processing, or "merely" biased it in favor of one parse once linguistic input has been encountered? The challenge in relating this to parallel-contingent processing is that the helpful condition visual display appears to completely *preempt* (Magnuson 2017) the destination parse (i.e., *on the envelope* is where one should put the salt shaker) for the reduced relative clause parse as early as can be measured. Listener-viewers appear to process the sentence wholly within the pragmatic context set up by the visual display, with virtually no evidence of competition with the garden path parse *prior to bottom-up linguistic support for the reduced relative* (i.e., *in the bowl* in this example). Without being able to anticipate the specific form of the sentence (due to the use of unambiguous sentences with unreduced relative clauses; Tanenhaus et al. 1995), listener-viewers appear to be sensitive directly to the probability of a reduced relative clause conditioned on the contents of the display. Thus, like overt anticipation (e.g., Altmann and Kamide 1999), preemption entails prediction; competition is able to be suppressed because of expectations that preactivate expected representations and/or suppress context-incongruent ones.

Preemption may be more compelling in a case where the visual display appears to alter a lower-level process than sentence processing, such as spoken word recognition. One example comes from Chambers et al. (2004), who found that action-based affordances of objects and instruments alter fixation patterns in the VWP. For example, given a display with an egg in its shell and a liquid egg visible in a glass bowl, participants would be equally likely to fixate either of the eggs given an instruction beginning *put the egg*.

However, given an instruction to *pour the egg*, fixations were much more likely to the liquid egg. In another experiment, Chambers et al. had participants wield a hook. Given a display with two whistles, one with a "hookable" string and one without, participants were much more likely to fixate the hookable whistle given an instruction beginning *put the whistle*. Thus, there can be cases where a 100% match between a displayed item and its spoken name can be substantially reduced based on action-based implications.

Another example comes from Magnuson et al. (2008), who found that syntactic and pragmatic constraints implied by a visual display can impact phonological competition. Magnuson et al. used an artificial lexicon paradigm (based on Magnuson et al. 2003) to instantiate names for novel nouns (referring to shapes) and adjectives (textures applied to the shapes). The lexicon included phonological cohort pairs (that is, items that overlapped at onset) like /pibo/ and /pibu/. Pairs could both be nouns or adjectives, or one of each. Displays either had four unique shapes, in which case a simple referring expression was possible (*click on the pibo*), or two instances of two different shapes with different textures, in which case both the adjective and noun were needed (*click on the tadu pibo*). When displays had four unique shapes (noun required condition), there was clear competition if a noun target's noun cohort was displayed. Similarly, in the adjective-required condition, there was clear cohort competition if an adjective cohort of the target's adjective was in the display. However, there was no competition with cross-class potential competitors. For example, if /tadu/ was a texture and /tadi/ was a shape, in the adjective-required condition, there were no fixations to the /tadi/ shape when the /tadu/ adjective was heard. Conversely, in the noun-required condition, if the target was /tadi/, there were no fixations to an item with the /tadu/ texture. We expected reduced competition across form classes, but the contingency between display contents and the expected noun phrase (bare noun or adjective-noun) was sufficient to wipe out cross-class competition. This may say more about language processing in general than the VWP (i.e., the implication that low-level phonological competition can be constrained by form class expectations), but it is also an instance of deep interaction; details of the visual display quickly reveal syntactic constraints that appear to extinguish phonological competition across form classes. Pirog Revill et al. (2008) reported very similar findings with a novel lexicon where competition was modulated by action-based contingencies. A possibly even more compelling example comes from Brown-Schmidt and Tanenhaus (2008), who studied unscripted conversation-based coordination in pairs of naïve participants. Cohort competition effects came and went as a dyad's referential context emerged, suggesting that interpretation and possibly lexical access can be modulated by the demands and constraints of complex communicative interaction.

Again, the three sets of studies I have just reviewed illustrate *preemption effects* (Magnuson 2017). In all three cases, we might have expected graded effects, with context diminishing but not wiping out competition. Strand et al. (2017) have recently observed such graded effects by extending this line of inquiry to real English words. They used a classic 4AFC VWP with grammatically unconstrained sentences (*the word is rug*, vs. *the word is run*) vs. grammatically constrained sentences (*they thought about the rug*, vs. *they began to run*). In contrast to the complete absence of cross-class competition found by Magnuson et al. (2008), Strand et al. found graded effects, with strong competition between cross-form class cohorts given grammatically unconstrained sentences, and greatly reduced (but still robust) cross-class competition given the grammatically constrained sentences.

Such graded effects are predicted on a constraint-based approach, and suggest that we should replace the term *preemption* with a more general one, such as *damping*. Strand et al.'s (2017) demonstration of graded effects is particularly key in making a case for deep interaction. If we only had evidence for preemption, this might suggest a form of *subsumption* (Brooks 1991) rather than interaction, where one hierarchically-superior modular subsystem completely dominates another (e.g., in the Tanenhaus et al. 1995 case, vision or executive function might dominate and control language processing; see Linking Hypothesis 4). Graded damping is consistent with constraint-based (interactive, parallel distributed processing, etc.) theories with pervasive, bidirectional information flow between levels of organization.

Finally, I hypothesize that another key aspect of perception and cognition in the VWP is the implication that internal representations are activated *in the moment* only when needed or directly stimulated in a *just-in-time* fashion. Consider the fact that a

wastebasket could be turned upside-down to be used as a step-stool to reach an object beyond arm's length, or that it could be used as a stool for sitting, or that it could be used to ferry water to a fire. In appropriate contexts, these affordances could become apparent. But we would not expect them (and all other possible unconventional uses) to be automatically activated when we see a wastebasket, hear the word *wastebasket*, or even interact with one physically. The just-in-time idea applies the same logic to core aspects of object knowledge. In a real-world situation, until I want to throw something away (or reach something, or sit when there are no chairs, or ferry water to a fire, or answer when someone asks where I found such a fine wastebasket), a wastebasket in my environment has little meaning or relevance, except to avoid colliding with it as I move around. Thus, what I *need* to activate about the wastebasket when it is not relevant to action or language is very minimal visual features *when it is in the visual field.* Minimal, coarse features will provide that gateway if the wastebasket becomes relevant for action; there is no need to pre-activate full object representations (phonological form, conceptual knowledge, etc.), and doing so might have negative consequences for perception and action *in the moment*. Note that I have emphasized *in the moment* again. This is because it is possible that in more unusual situations where I am presented with a scene for a substantial period of time prior to any other instruction or incidental language, it may be that representations of scene elements may be activated.

Critical evaluation of Linking Hypothesis 4

One could justifiably object that *deep interaction* is a slight variation on any of the preceding linking hypotheses: Linking Hypothesis 1 shed of the independent component assumption; Linking Hypothesis 2 without the premise that representations of features of scene elements, including names, are automatically activated and loaded into working memory; or Linking Hypothesis 3 without explicit accommodation for modulation of conceptual representations. The point of Linking Hypothesis 4 is to make a strong commitment to pervasive interaction rather than merely allowing the possibility. That said, Linking Hypothesis 4 suffers from the same gap as Linking Hypotheses 2 and 3: there is no implemented computational model that could be used to explore parameter spaces and determine whether, e.g., removing the independence assumption of Linking Hypothesis 1 actually generates predictions capable of distinguishing Linking Hypotheses 1 and 4. In particular, a challenging question is whether preemption and damping are truly language-internal modulations. It is possible that competition still goes on within the entire lexicon, but visual attention can be constrained by visual and discourse constraints. Again, an implemented model will be key in determining whether such results can be modeled as parallel-contingent independence, with integration of constraints at a decision level, or whether they strongly support deep interaction.

The same point holds for the *just-in-time* proposal. Another concern with that proposal is the hedge that sustained displays without language or action demands might activate representations. Does this imply a categorical distinction (activated or not)? Under what conditions (both in terms of task demands and timing) would the system shift from just-in-time to full activation? Would the shift be gradual or more like a phase transition? Again, an implemented model may be required to specify predictions that would distinguish Linking Hypothesis 4 from others.

Comparing the linking hypotheses

Deep interaction is clearly different from the basic form of Linking Hypothesis 1 (parallel-contingent activation), where language and vision are independent component systems with outputs that are *integrated* at a decision stage. On that view (which was intended as a provisional simplifying assumption), the two systems do not interact. Activation and competition occur in spoken word recognition in the context of the entire lexicon, and any constraints provided by the display (e.g., the possible behavioral 'outlets' provided by displayed objects as fixation targets) impact a post-perceptual decision stage.

Deep interaction also differs subtly from Linking Hypothesis 2 (*displays load working memory and activate representations prior to speech*). On that view, the interaction of scene and language is a fundamental assumption, as the claim is that lexical and visual features of displayed items are activated prior to any speech input. One could argue this is a claim of even deeper interaction than what I have discussed for Linking Hypothesis 4. This premise is the crucial distinction between Linking Hypotheses 2

and 4. In my opinion, it must be the case either that displayed items do not automatically activate lexical representations (and names in particular, given facts reviewed above, such as that fixations map onto phonetic similarity over time [Allopenna et al. 1998], the VWP is sensitive to non-displayed details of target items [e.g., word frequency (Dahan et al. 2001a) and non-displayed competitors (Magnuson et al. 2007), and even fairly high-level processing is largely unaffected by long display previews (Sedivy et al. 1999)], or that the standard 4AFC VWP provides a grossly distorted view of language processing. I say this because it seems implausible that as one scans a scene or simply moves through the world that the names of all objects one encounters are activated. Thus, if it were the case that names are automatically activated in the VWP, this might mean that language processing in the VWP is qualitatively different from real-world language processing. In addition, as discussed under Linking Hypothesis 2, the strongest evidence for name activation comes from highly constrained tasks that may promote covert naming. Aside from the issue of whether elements of visual scenes are automatically named, there is little difference between Linking Hypotheses 2 and 4, although the *just-in-time* proposal represents a significant and explicit difference in theoretical commitment.

This takes us to Linking Hypothesis 3, the mental world hypothesis. In fact, there is little apparent difference between Linking Hypotheses 2, 3 and 4, except for the assumption under Linking Hypothesis 2 that names of scene elements are automatically activated, and the commitment in Linking Hypothesis 4 to minimize internal representation. Indeed, both Linking Hypotheses 2 and 3 appear compatible with deep interaction, although this idea may not be a core concern of either. There is also the potentially unresolved difference between the proposal of an incremental processor that achieves the fullest possible interpretation at each moment, including predictions of upcoming words or structures (Kamide et al. 2003), and the more passive notion of a priming-like basis for language to guide visual attention (Altmann and Kamide 2007).

Challenges

In this section, I will set aside the details of linking hypotheses to consider challenges for interpreting visual world paradigm data, which will reinforce the need for implemented computational models to guide understanding of the task itself, as well as the perception and cognition it is meant to measure. Clearly, fixations in the VWP can be directed based on phonetic details in spoken input. This entails mapping linguistic features to visual features. "Features" is a carefully chosen term. Selecting a saccadic target does not require *recognizing* a phonological form *or* a visual form; fixations can be directed based on partial phonetic information in an incremental fashion to objects that are associated with visual features of a spoken word, its phonological competitors (Allopenna et al. 1998), or its semantic relatives (Huettig and Altmann 2005; Yee and Sedivy 2001, 2006). In many cases, linguistic processing—and therefore fixations—can be constrained by visual context, as in the classic original demonstration by Tanenhaus et al. (1995). It is perhaps underappreciated that Tanenhaus et al. (1995) already established that, under certain conditions, context overwhelms bottom-up input. Given a display with two salt shakers (one on an envelope, one on a napkin, and with an "empty" envelope and bowl also present), and an instruction to *put the salt shaker on the envelope in the bowl*, bottom-up control would predict a substantial proportion of fixations to the unoccupied envelope. Instead, fixations to the second envelope were virtually extinguished by syntactic expectations driven by discourse constraints embedded in the visual context.

As Tanenhaus and his colleagues have argued (e.g., Tanenhaus et al. 2000), the complex relationships of linguistic and visual information at different time scales entail explicit linking hypotheses between fixations in the VWP and theories of language processing. Since fixations are not simply linked directly to the bottom-up processing of each word form, a theory must specify predictions appropriate for linguistic and visual context. As I reviewed in discussions of the four linking hypotheses, a theory must predict fixation behavior under at least four contingencies: fixations that follow bottom-up specification, fixations that follow from semantic or visual features associated with phonetic or grammatical features of words in the input, fixations that precede

bottom-up specification (anticipation), and fixations that are absent (or diminished) despite bottom-up (or feature-associated) match to the signal (i.e., preemption or damping). But other challenges remain.

Do fixations indicate linguistic and/or visual competition?

It is a conundrum that fixations may be drawn to objects for multiple reasons. Items might be fixated because they correspond to linguistic entities that are in competition, are simply co-activated, or that actually facilitate one another. By default, all of these cases, including the last one, would be interpreted as indicating competition in the VWP. In spoken word recognition, we assume that fixations to phonological relatives (cohorts, rhymes) reflect both co-activation and competition, since our theories (especially TRACE; McClelland and Elman 1986) include mechanisms for direct competition between co-activated items (e.g., lateral inhibition in TRACE). We appear to find evidence for competition from both the close fit between TRACE's predictions and human fixation proportions over time as a function of phonological similarity (since TRACE predicts competition), but also from changes in fixations as a function of the distractors, with reduced target fixation proportions when objects corresponding to a cohort or rhyme are included in the display. However, recent simulations (Magnuson, in preparation) suggest that rhymes, at least for longer words, may have an overall faciliatory impact on each other. That is, having more rhymes is associated with faster reaction times in TRACE, TISK (Hannagan et al. 2013; You and Magnuson 2018), and a simple recurrent network (SRN; Elman 1990) trained on inputs like those used with TRACE. This appears to follow from feedforward-feedback resonance between lexical and sublexical representations in a model like TRACE, shifting from a competitive dynamic to a cooperative one over the course of processing longer words. Thus, even in such a simple case, fixations may be directed to items that are either actively inhibiting or facilitating the activation of the target item, but both cases will appear to indicate competition in the VWP because fixations can be directed to only one item at a time. Fixations to any item aside from the target, for whatever reason, necessarily reduce total fixation proportions to the target.

The same holds, of course, for semantically-mediated fixations. When looks are directed to items related to a target or one of its phonological relatives in features, function, or association, this likely reflects complex dynamics that involve faciliatory, inhibitory, lateral, forward and backward information flow. Specific proposals about the direction and degree of activation and resulting predictions (akin to the cartoon schematic in Fig. 1) will be required to link fixation patterns to theories of language processing. Doing so will require experimental innovations. Perhaps the locus question could be addressed, for example, by specifying a concrete theory of priming (as in Fig. 1) and checking for residual effects on items assumed to have been boosted or suppressed following an instance of preemption. In any case, we must be mindful that the nature of eye movements means there is always a competitive element to visual selection, although underlying language processing may reflect dynamic combinations of competitive and cooperative interactions.

Memory for and impact of visual displays

Another challenge is the role and impact of the visual display, and its relation to memory. As I discussed above, there is room for skepticism about the proposal that items in the visual world (even a simple, 4AFc display) are or even must be recognized and encoded in working memory before they could provide a basis for integrating visual and linguistic information (Huettig et al. 2011a). Figure 4 is a schematic of a hypothetical organization where the visual items themselves are the basis for integration. So long as they are close enough to one another that coarse visual features can be detected without direct fixation, there may be no need to encode them into memory—at least in a simple task (more on this shortly). Crucially, this implies that it is not *necessary* for visual objects to be fully recognized or for phonological forms for each object to be fully retrieved before linguistic input can be incrementally mapped onto the visual world. The kinds of experiments that could resolve this issue have not been conducted. For example, careful manipulation of visual preview times has only been done in a "target absent," "look-and-listen" variant of the VWP (Huettig and McQueen 2007); a study is needed where such a manipulation is done with the "classic" variant

of the task so as to isolate the effects of timing manipulations.

Similarly, we must be mindful of task complexity as we develop linking hypotheses for the VWP. I agree with Huettig and colleagues that the likelihood of participants becoming explicitly aware of visual, semantic, and linguistic details of depicted objects in the VWP will increase over time, though the degree to which this happens will also depend on the nature of the task used. Given a simple task that can be completed in a under a second (as in a typical spoken word recognition variant of the VWP), the likelihood (and need) to recognize objects and retrieve their details could remain low. But with a task that involves a longer-duration display or (likely more importantly) a *series* of linguistically-guided actions, the likelihood of activating details of displayed objects likely increases substantially, along with the potential need to encode details of the scene in memory. For example, Chambers and San Juan (2008) labeled specific scene regions and gave participants series of instructions such as: *Move the chair to area 2; now move the chair to area 5; now return the chair to area 2*. On hearing *now return*, participants made anticipatory saccades to areas that had previously been referred to. Such behavior entails memory for scene locations and discourse history, although it does not necessarily entail a full or complex mental model. Ballard et al. (1997) provided compelling evidence that memory for scene locations in complex visual scenes could be in a form analogous to fixation coordinates rather than the detailed contents of a location or a detailed model of the entire display. Spivey et al. (2004) argue that such a system could provide the basis for understanding fixation behavior in the VWP without a complex mental model, including task variants where displays are removed but participants continue to make fixations to the previous locations of displayed items (e.g., Richardson and Spivey 2000).

Context and culture

One way to summarize VWP studies of language processing is to say that listeners are sensitive to every potentially useful (i.e., predictive) constraint that has been tested as early as we can measure. As Spivey and Spevack (2017) discuss, humans are embedded in multiple, nested levels of environment, including not just the physical environment, but other organisms and culture as well. Participants in the VWP are embedded not just in the experimental context of our studies, but in a much more complex environment that includes other minds and culture. In the VWP, participants must coordinate in the moment with the experimenter, as well as with the task previously devised by the experimenter, which embodies implicit assumptions about the meaning of stimuli, and is premised on implicit assumptions that the task constraints will allow the task to be performed, and that the participant will comply with what is asked. In addition, experiences shared by individuals within a cultural group impact interpersonal interaction dynamics, and can have surprising impact on presumably low-level aspects of cognition, such as visual attention (Nisbett et al. 2001). As visual world studies become more complex—including two or more active participants engaged in communication and/or other cooperative tasks (e.g., Hanna et al. 2003; Keysar et al. 1998), or more complex, realistic displays and unscripted interactions (e.g., Brown-Schmidt and Tanenhaus 2008), ideally studied cross-culturally—such concerns will need to be integrated into theories and models.

Damping, anticipation and constraint satisfaction

Preemption (Magnuson 2017) or damping of fixations is a particular challenge. Does it reflect deep penetration of language processing by visual and/or linguistic context? For example, when linguistic and/or visual context predicts a noun vs. an adjective (Magnuson et al. 2008) or a noun vs. a verb (Strand et al. 2017) and fixations to items with matching phonology but mismatching form class decline or vanish, has a filter been applied to lexical processing (if we assume a TRACE-like architecture, perhaps by boosting resting levels of the expected form class, or lowering resting levels of other form classes)? Or does damping happen at the decision stage, where saccade targeting occurs? Might different instances of preemption or damping have different loci? For example, perhaps a constraint like form class can be applied to lower-level processing, but ruling out an egg in solid, unpourable form given an instruction to *pour the egg*… cannot have the same locus, as the word form *egg* that must be activated matches both an egg in its shell and a liquid egg (Chambers et al. 2004).

Consider again the results of Kukona et al. (2011, 2014), which present us with a paradox. Why

can listeners flexibly launch saccades in anticipation (Altmann and Kamide 1999; Kamide et al. 2003) or damp (Strand et al. 2017) or fully preempt them in some cases (Magnuson et al. 2008; Pirog Revill et al. 2008), consistent with processing in the mode of an efficient prediction machine, but in other cases, cannot suppress clearly irrelevant fixations? The complement to preemption is what we might call *irrational* or *leaky* fixations. These are fixations that, from a perspective like rational analysis (e.g., Anderson 1991), must be considered degenerate or at least suboptimal. As Kukona et al. (2014) pointed out, these include fixations to rhymes (looking to *speaker* on hearing *beaker* implies rejecting /b/ as evidence for *beaker* and against *speaker*; Allopenna et al. 1998), as well as to nouns whose likely role has already been filled (e.g., looking to *policeman* on hearing *Toby will arrest the*… in a study where every sentence has *Toby* as the agent requires that an active prediction mechanism would reject this incredibly robust basis for anticipation; Kukona et al. 2011), or to an inedible object (*white car*) with featural overlap with a clearly-specified target (*white cake*; Kukona et al. 2014). On reflection, these fixations are not mysterious. They are diagnostic of a constraint-based system, rather than an optimal forecasting system. When links between verbs and nouns are relatively weaker [as in *the boy will eat*… (Altmann and Kamide 1999) or the *the girl will ride*… (Kamide et al. 2003)], strong (but not complete, overwhelming) anticipation is observed. As Kukona et al. (2014) showed, we can easily see the interplay between top-down and bottom-up influences that lead to anticipatory and 'irrational' fixations (see also Strand et al. 2017).

The overall picture that emerges is one consistent with a constraint-based view of language processing (MacDonald et al. 1994; McRae et al. 1998; Trueswell and Tanenhaus 1994), possibly reliant on at least two sets of processes, as suggested by Huettig (2015): a fast, semi-automatic but "dumb" set (akin to priming), and a slow but "smart" set under active control. However, the proposal of two systems does not resolve the challenge here. It remains unclear why in some cases preemption is possible (e.g., Magnuson et al. 2008; Tanenhaus et al. 1995) and in other cases potentially extremely strong constraints are ignored (e.g., Kukona et al. 2011). It is unsatisfactory to attribute results to the smart and dumb systems post hoc based on outcomes; if this account can be developed to the point where it can predict when and why each system dominates, it has the potential to provide a substantial theoretical advance.

Another consideration is how individual differences in linguistic and cognitive abilities relate to effects in the VWP. Among the *putatively* positive characteristics of the VWP are its naturalistic structure (participants perform a linguistically-guided task rather than make metalinguistic judgements) and its relatively low demands. If we consider leaky fixations reported by Kukona et al. (2014) (i.e., looks to a *white car* after anticipatory fixations have already been made to *white cake* given the sentence *the boy will eat the white cake*), what might we expect from an individual differences approach? If suppressing such possibly irrelevant saccades is governed at the decision (saccade targeting) level, one might expect individual differences in visual attention or other aspects of executive function to predict rates of irrelevant saccades. However, Kukona et al. (2016) found that individual differences in language ability (reading- and comprehension-related measures) were the most important predictors (see also Li et al. 2019, who found that the relative weight of phonetic vs. lexical details shifted with language ability). This is consistent with an intuitive account that puts the locus of preemption/damping within the language processing system, but such a conclusion must be viewed with caution without a complete linking hypothesis relating individual differences to the VWP task.

Progress

Why do linking hypotheses matter? The preceding discussions are intended to make clear that we risk attributing experimental outcomes to the wrong level (linguistic, visual, or decision) if we do not articulate and formalize linking hypotheses. The schematic approach taken by Huettig et al. (2011a), which I have attempted to extend here (Figs. 3 and 4), is a promising start. Deep understanding may require the implementation of integrated models of the full task performed by participants in visual world studies: a language processing model, a visual processing model, and a mechanism for generating eye movements. Our intuitions and even schematics may otherwise mislead us. For example, extremely plausible, logical intuitions about what a model of language

processing *should* do have often turned out to be wrong once the model is actually tested (Magnuson 2008; Magnuson et al. 2012). Without taking our models to this level of specificity, we risk repeating the kinds of errors we (as a field) have made with respect to the component processes underlying performance in the visual world paradigm. Altmann and colleagues (e.g., Altmann and Kamide 1999, 2007; Kamide et al. 2003) as well as Knoeferle and Crocker (2006, 2007) have identified great challenges that face the field: we need theories and linking hypotheses that extend to highly complex aspects of discourse and event *understanding*. A crucial prerequisite may be the development of formal computational models capable of generating falsifiable predictions about the interaction of language processing, visual processing, and event understanding. A recent model by Venhuizen et al. (2019) that grapples with the interaction of world knowledge and language processing has the potential to be the foundation of such a model.

Acknowledgements This paper is based on a talk presented at the *Attentive Listener in the Visual World* meeting in Trondheim, Norway, in August, 2018. I am grateful to Falk Huettig, Mila Vulchanova, Valentin Vulchanov, Inge-Marie Eigsti, and Kenny Coventry for stimulating discussions that reshaped this paper. Preparation of this paper was supported in part by U.S. National Science Foundation Grants 1754284, *Computational approaches to human spoken word recognition*, and 1735225, *Science of learning, from neurobiology to real-world application*.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38,* 419–439.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73,* 247–264.

Altmann, G. T. M., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language, 57,* 502–518.

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences, 14*(3), 471–517.

Ballard, D. H. (1991). Animate vision. *Artificial Intelligence, 48,* 57–86.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *Behavioural and Brain Sciences, 20*(4), 723–742.

Balota, D. A. (1990). The role of meaning in word recognition. In D. A. Balota, G. Flores D'Arcais, & K. Rayner (Eds.), *Comprehension processes in reading* (pp. 9–32). Hillsdale, NJ: Lawrence Erlbaum Associates.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence, 47,* 139–159.

Brown-Schmidt, S., & Tanenhaus, M. K. (2008). Real-time interpretation of referential domains in unscripted conversation: A targeted language game approach. *Cognitive Science, 32,* 643–684. https://doi.org/10.1080/03640210802066816.

Chambers, C., & San Juan, V. (2008). Perception and presupposition in real-time language comprehension: Insights from anticipatory processing. *Cognition, 108,* 26–50.

Chambers, C. G., Tanenhaus, M. K., & Magnuson, J. S. (2004). Actions and affordances in syntactic ambiguity resolution. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 30,* 687–696.

Chiu, E. M., & Spivey, M. J. (2014). Timing of speech and display affects the linguistic mediation of visual search. *Perception, 43,* 527–548.

Contini, E. W., Wardle, S. G., & Carlson, T. A. (2017). Decoding the time-course of object recognition in the human brain: From visual features to categorical decisions. *Neuropsychologia, 105,* 165–176.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology, 6,* 84–107.

Cree, G. S., McRae, K., & McNorgan, C. (1999). An attractor model of lexical conceptual processing: Simulating semantic priming. *Cognitive Science, 23,* 371–414.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001a). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology, 42,* 317–367.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001b). Tracking the time course of subcategorical mismatches: Evidence for lexical competition. *Language and Cognitive Processes, 16*(5/6), 507–534.

Dahan, D., & Tanenhaus, M. K. (2005). Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. *Psychonomic Bulletin and Review, 12,* 453–459.

De Groot, F., Huettig, F., & Olivers, C. N. L. (2016). When meaning matters: The temporal dynamics of semantic influences on visual attention. *Journal of Experimental Psychology: Human Perception and Performance, 42*(2), 180–196. https://doi.org/10.1037/xhp0000102.

Dell'Acqua, R., & Grainger, J. (1999). Unconscious semantic priming from pictures. *Cognition, 73*(1), B1–B15.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14,* 179–211.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science, 33,* 1–36.

Elman, J. L. (2011). Lexical knowledge without a lexicon? *The Mental Lexicon, 6*(1), 1–33.

Frauenfelder, U. H., & Peeters, G. (1998). Simulating the time course of spoken word recognition: An analysis of lexical competition in TRACE. In J. Grainger & A. M. Jacobs (Eds.), *Localist connectionist approaches to human cognition* (pp. 101–146). Mahwah, NJ: Erlbaum.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin Company.

Gibson, E. J., & Pick, A. D. (2000). *An ecological approach to perceptual learning and development*. New York: Oxford University Press.

Greeno, J. G. (1994). Gibson's affordances. *Psychological Review, 101*(2), 336–342.

Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). *Journal of Memory and Language, 49,* 43–61.

Hannagan, T., Magnuson, J. S., & Grainger, J. (2013). Spoken word recognition without a TRACE. *Frontiers in Psychology, 4,* 563.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review, 106*(3), 491–528.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychological Review, 111*(3), 662–720.

Hollingworth, A., & Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance, 28*(1), 113–136. https://doi.org/10.1037/0096-1523.28.1.113.

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research, 1626,* 118–135. https://doi.org/10.1016/j.brainres.2015.02.014.

Huettig, F., & Altmann, G. T. M. (2005). Word meaning and the control of eye fixation: semantic competitor effects and the visual world paradigm. *Cognition, 96*(1), 23–32.

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language, 57*(4), 460–482. https://doi.org/10.1016/j.jml.2007.02.001.

Huettig, F., Olivers, C. N. L., & Hartsuiker, R. J. (2011a). Looking, language, and memory: Bridging research from the visual world and visual search paradigms. *Acta Psychologica, 137,* 138–150. https://doi.org/10.1016/j.actpsy.2010.07.013.

Huettig, F., Rommers, J., & Meyer, A. S. (2011b). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica, 137,* 151–171. https://doi.org/10.1016/j.actpsy.2010.11.003.

Kahneman, D. (2011). *Thinking*. Fast and Slow: Macmillan.

Kamide, Y., Altmann, G. T. M., & Haywood, S. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49,* 133–156.

Keysar, B., Barr, D. J., & Horton, W. S. (1998). The egocentric basis of language use: Insights from a processing approach. *Current Directions in Psychological Science, 7,* 46–50.

Knoeferle, P., & Crocker, M. W. (2006). The coordinated interplay of scene, utterance, and world knowledge: Evidence from eye tracking. *Cognitive Science, 30,* 481–529.

Knoeferle, P., & Crocker, M. W. (2007). The influence of recent scene events on spoken comprehension: Evidence from eye-movements. *Journal of Memory and Language, 57,* 519–543.

Kukona, A., Braze, D., Johns, C. L., Mencl, W. E., Van Dyke, J. A., Magnuson, J. S., et al. (2016). The real-time prediction and inhibition of linguistic outcomes: Effects of language and literacy skill. *Acta Psychologica, 171,* 72–84.

Kukona, A., Cho, P. W., Magnuson, J. S., & Tabor, W. (2014). Lexical interference effects in sentence processing: Evidence from the visual world paradigm and self-organizing models. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 40*(2), 326–347.

Kukona, A., Fang, S., Aicher, K. A., Chen, H., & Magnuson, J. S. (2011). The time course of anticipatory constraint integration. *Cognition, 119,* 23–42.

Li, M. Y. C., Braze, D., Kukona, A., Johns, C. L., Tabor, W., Van Dyke, J. A., et al. (2019). Individual differences in subphonemic sensitivity and phonological skills. *Journal of Memory and Language, 105,* 195–215.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*(4), 676–703.

Magnuson, J. S. (2008). Nondeterminism, pleiotropy, and single word reading: Theoretical and practical concerns. In E. Grigorenko & A. Naples (Eds.), *single word reading* (pp. 377–404). Mahweh, NJ: Erlbaum.

Magnuson, J. S. (2017). Mapping spoken words to meaning. In G. Gaskell & J. Mirkovic (Eds.), *Speech Perception and spoken word recognition* (pp. 76–96). New York: Routledge.

Magnuson, J. S. (2019a). Schematic of the time course of priming. figshare. Figure. https://doi.org/10.6084/m9.figshare.9465416.v1

Magnuson, J. (2019). Working memory visual world linking hypothesis (Version 2). figshare. Figure. https://doi.org/10.6084/m9.figshare.8019518.v2

Magnuson, J. (2019c). Deep interaction visual world paradigm linking hypothesis (Version 1). figshare. Figure. https://doi.org/10.6084/m9.figshare.8020184.v1

Magnuson, J. S. (in preparation). Comparative modeling of spoken word recognition.

Magnuson, J. S. (in preparation). Similar microstructure of spoken word recognition across computational architectures.

Magnuson, J. S., Dixon, J., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science, 31,* 133–156.

Magnuson, J. S., Mirman, D., & Harris, H. D. (2012). Computational models of spoken word recognition. In M. Spivey, K. McRae, & M. Joanisse (Eds.), *The Cambridge handbook of psycholinguistics* (pp. 76–103). Cambridge: Cambridge University Press.

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance, 33,* 391–409.

Magnuson, J. S., Tanenhaus, M. K., & Aslin, R. N. (2008). Immediate effects of form-class constraints on spoken word recognition. *Cognition, 108*(3), 866–873.

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., & Dahan, D. (2003). The time course of spoken word recognition and learning: Studies with artificial lexicons. *Journal of Experimental Psychology: General, 132*(2), 202–227.

Mani, N., & Plunkett, K. (2010). In the infant's mind's ear: Evidence for implicit naming in 18-month-olds. *Psychological Science, 21,* 908–913.

Marslen-Wilson, W., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 576–585.

McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review, 86,* 287–330.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1–86.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88,* 375–407.

McMurray, B., Tanenhaus, M., & Aslin, R. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition, 86*(2), B33–B42.

McQueen, J. M., & Huettig, F. (2014). Interference of spoken word recognition through phonological priming from visual objects and printed words. *Attention, Perception and Psychophysics, 76,* 190–200. https://doi.org/10.3758/s13414-013-0560-8.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, 37,* 547–559.

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence processing. *Journal of Memory and Language, 38,* 283–312.

Mirman, D., & Magnuson, J. S. (2008). Attractor dynamics and semantic neighborhood density: Processing is slowed by near neighbors and speeded by distant neighbors. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 34,* 65–79.

Mirman, D., & Magnuson, J. S. (2009a). The effect of frequency of shared features on judgments of semantic similarity. *Psychonomic Bulletin & Review, 16*(4), 671–677.

Mirman, D., & Magnuson, J. S. (2009b). Dynamics of activation of semantically similar concepts during spoken word recognition. *Memory and Cognition, 37,* 1026–1039.

Mishra, R. K., Olivers, C. N. L., & Huettig, F. (2013). Spoken language and the decision to move the eyes: To what extent are language-mediated eye movements automatic? In V. S. C. Pammi & N. Srinivasan (Eds.), *Progress in brain research: Decision making: Neural and behavioural approaches* (pp. 135–149). New York: Elsevier.

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic vs. analytic cognition. *Psychological Review, 108,* 291–310.

Noizet, G., & Pynte, J. (1976). Implict labeling and rediness for pronunciation during the perceptual process. *Perception, 5,* 217–223.

Pirog Revill, K., Tanenhaus, M. K., & Aslin, R. N. (2008). Context and spoken word recognition in a novel lexicon. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 34*(5), 1207–1223.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103,* 56–115.

Reali, F., Spivey, M. J., Tyler, M. J., & Terranova, J. (2006). Inefficient conjunction search made efficient by concurrent spoken delivery of target identity. *Perception and Psychophysics, 68,* 959–974.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science, 8*(5), 368–373.

Richardson, D. C., & Spivey, M. J. (2000). Representation, space, and Hollywood Squares: Looking at things that aren't there anymore. *Cognition, 76*(3), 269–295.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition, 90,* 51–89.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition, 71,* 109–147.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96,* 523–568.

Spivey, M. J. (2007). *The continuity of mind*. New York: Oxford University Press.

Spivey, J. J., & Marian, V. (1999). Cross talk between native and second languages: Partial activation of an irrelevant lexicon. *Psychological Science, 10*(3), 281–284.

Spivey, M. J., Richardson, D. C., & Fitneva, S. A. (2004). Thinking outside the brain: Spatial indices to visual and linguistic information. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. 161–189). New York, NY, US:Psychology Press.

Spivey, M. J., & Spevack, S. C. (2017). An inclusive account of mind across spatiotemporal scales of cognition. *Journal of Cultural Cognition, 1,* 25–38. https://doi.org/10.1007/s41809-017-0002-6.

Strand, J. F., Brown, V. A., Brown, H. E., & Berg, J. J. (2017). Keep listening: Grammatical context reduces but does not eliminate activation of unexpected words. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 44*(6), 962–973.

Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language, 50,* 355–370. https://doi.org/10.1016/j.jml.2004.01.001.

Tabor, W., & Hutchins, S. (2004). Evidence for self-organized sentence processing: Digging in effects. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 30,* 431–450. https://doi.org/10.1037/0278-7393.30.2.431.

Tanenhaus, M. K., Magnuson, J. S., Dahan, D., & Chambers, C. (2000). Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Journal of Psycholinguistic Research, 29,* 557–580.

Tanenhaus, M. K., & Spivey-Knowlton, M. J. (1996). Eye-tracking. *Language and Cognitive Processes, 11,* 583–588.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268,* 1632–1634.

Trueswell, J. C., & Tanenhaus, M. K. (1994). Toward a lexicalist framework for constraint-based syntactic ambiguity resolution. In C. Clifton, L. Frazier, & K. Rayner (Eds.), *Perspectives in sentence processing* (pp. 155–179). Hillsdale, NJ: Lawrence Erlbaum Associates.

Turvey, M. (1973). On peripheral and central processes in vision: Inferences from an information-processing analysis of masking with patterned stimuli. *Psychological Review, 80,* 1–52.

Venhuizen, N. J., Crocker, M. W., & Brouwer, H. (2019). Expectation-based comprehension: Modeling the interaction of world knowledge and linguistic experience. *Discourse Processes, 56*(3), 229–255.

Vivianni, P. (1990). Eye movements in visual search: Cognitive, perceptual, and motor control aspects. In E. Kowler (Ed.), *Eye movements and their role in visual and cognitive processes. Reviews of oculomotor research V4* (pp. 353–383). Amsterdam: Elsevier.

Simmons, E. S., & Magnuson, J. S. (accepted with minor revisions). Word length, proportion of overlap, and the time course of phonological competition in spoken word recognition: An empirical and computational investigation. *Cognitive Science.*

Yee, E., & Sedivy, J. (2001). Using eye movements to track the spread of semantic activation during spoken word recognition. *Paper presented to the 13th annual CUNY sentence processing conference*, Philadelphia.

Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 32,* 1–14.

You, H., & Magnuson, J. S. (2018). TISK 1.0: An easy-to-use Python implementation of the time-invariant string kernel model of spoken word recognition. *Behavior Research Methods*. https://doi.org/10.3758/s13428-017-1012-5.

Zelinsky, G. J., & Murphy, G. L. (2000). Synchronizing visual and language processing: An effect of object name length on oculomotor behavior. *Psychological Science, 11,* 125–131.