The Problem of Comprehension in Psycholinguistics

Fernanda Ferreira & Zoe Yang

Routledge
Taylor & Francis Group

BRIEF REVIEW

Check for updates

The Problem of Comprehension in Psycholinguistics

Fernanda Ferreira and Zoe Yang

Department of Psychology and Center for Mind and Brain University of California, Davis

ABSTRACT
Most research in psycholinguistics relies on online measures such as reading time to inform and test theories of language comprehension. However, the value of offline measures such as question-answering performance is sometimes overlooked in sentence processing work. Consequently, psycholinguists do not yet understand how the tasks and measures used in online experiments might reflect the content of the representations that are formed by the subject and how our comprehension tasks might influence subjects' reading strategies. We begin this review by briefly discussing evidence that comprehenders often misinterpret language. We then consider some of the reasons why psycholinguists tend not to use offline measures of comprehension to evaluate competing theories of sentence processing and discuss the role of subject engagement and motivation. Finally, we explore what is currently known about the relationships among task, depth of comprehension, and reading measures.

Introduction

A wide range of theoretical questions is studied as part of the enterprise of investigating language processing. One set of questions relates to lexical processing and focuses on questions concerning the ease with which the visual or auditory form of a word is recognized and located in the lexicon so that its meaning can be retrieved. Another set of questions pertains to the issue of how sentences are combined to form a coherent discourse, which includes investigations of topics such as establishing co-reference between anaphors and their antecedents, drawing inferential connections among ideas, and creating a macrostructure for multisentence texts. Between these representational levels we find the study of sentence comprehension, which, as the term implies, tends to emphasize the processing of individual sentences and which has traditionally focused on issues relating to syntax and aspects of meaning closely tied to sentence structure. Unfortunately, these subareas of the psychology of language are not as well linked as they could be. In this article we highlight one of the consequences of the separation between the fields of sentence and discourse processing. Although discourse processing research has made clear the value of comprehension measures such as question-answering accuracy and text recall, those insights have not informed studies of sentence comprehension. The result is that our theories of sentence comprehension are not as informative about the products of processing as they could be, and, in addition, our methodological practices lack a firm empirical foundation that would allow psycholinguists to rely on more than intuition when selecting comprehension measures for assessing not just the online processing of some sentence-level phenomenon but also its interpretation.

To illustrate this point, imagine you are a psycholinguist and you decide to set up an experiment to investigate how people interpret a sentence containing a type of linguistic expression, say a quantified noun phrase such as *every student*. If you have been trained in the standard approach,

you would begin by creating a number of unique sentences with and without the quantified expression, together with a set of unimportant filler sentences that will probably go unanalyzed and that serve mainly to mask the properties of the stimuli created to test the critical hypothesis. Because it is important that subjects in your experiment pay attention to the stimuli, you would likely also create some comprehension items, which might be yes or no questions about some superficial aspect of the sentences. Often these questions are presented on only some proportion of the trials, and in many cases comprehension items are administered only for the filler items, to avoid contaminating or influencing the processing of the experimental items. The subjects typically would not receive any feedback on their question-answering accuracy. The stimuli would then be presented to subjects on a computer screen and reading times measured. The responses to the comprehension items might be analyzed and reported, but these data are usually provided simply to reassure readers that comprehension levels were acceptable and that any differences across conditions are consistent with the hypotheses.

What is somewhat odd about this scenario is that it suggests the field of sentence comprehension does not take its comprehension measures particularly seriously. Instead, competing theories are tested mainly based on results from online measures, even though we lack a clear understanding of what those reading times or reaction times reveal about the content of the representations that presumably are built during comprehension. Moreover, our decisions about the number and types of comprehension items to include are largely intuitive and based on rules of thumb that have been passed on from lab to lab, with little solid empirical justification. As a result, there are at least two major gaps in our knowledge: First, we know little about how widely used online measures such as reading time reflect the content of the representations generated during processing, and, second, we do not know how various comprehension tasks affect comprehension strategy. The problem, then, is this: Imagine that in our study of sentences containing quantified noun phrases we observe that people read the quantified phrase more slowly than some control expression. Although we might draw some inferences about when information contained in the quantified phrase is accessed or integrated into the rest of the sentence, can we draw firm conclusions about the reader's interpretation of the quantified phrase? And, more fundamentally, in what sense are we studying comprehension if we don't assess people's interpretation of the critical phrase?

Our goal in this article is to challenge this research strategy and urge the study of sentence comprehension to proceed along the lines of what has taken place in the field of discourse processing, which transitioned from focusing mainly on offline, outcome measures to the use of online processing measures without downplaying the importance of either type (van Den Broek & Gustafson, 1999). Our article is organized into five sections. In the first we summarize evidence pointing to systematic comprehension failures on the part of typical college students. The second section describes some of the explanations that have been given for these comprehension failures. In the third section we address the question why psycholinguists do not make better use of comprehension measures such as question-answering and memory tasks. From there we discuss the issue of engagement and how it might affect language processing. Finally, we address the extent to which we understand the connections among three key aspects of psycholinguistic research: comprehension tasks, levels of comprehension, and effects on measures such as reading time. Our conclusion is that we have a poor understanding of how these concepts are interrelated, highlighting the need for further research assessing people's interpretations.

People often fail to comprehend or comprehend superficially

In the 1998 film *There's Something About Mary*, an exchange takes place between two lead characters —Ted, played by Ben Stiller, and Dom, played by Chris Elliott—that seems almost deliberately designed to showcase misinterpretations. In the sequence Ted looks around at his married friend's home and says wistfully, "It must be wonderful having all this." His friend, who is less happy than appearances suggest, replies, "Each day is better than the next." At first the response seems perfectly

reasonable and intended to communicate that Dom grows increasingly happy with each passing day, but closer examination makes clear that the sentence means just the opposite (replace *next* with *last* and the misinterpretation will immediately pop out). In my own experience using this example in talks as a real-life demonstration of "good-enough" language processing effects (Ferreira, Bailey, & Ferraro, 2002; Ferreira & Patson, 2007; Sturt, Sanford, Stewart, & Dawydiak, 2004), few people get the actual meaning.

These sorts of misunderstandings have also been elicited in the lab in a number of classic experiments. For instance, readers tend to overlook the anomaly in the sentence *No head injury is too trivial to be ignored* (Wason & Reich, 1979) as well as *This book fills a much-needed gap in the literature* (Johnson-Laird, 1981). Another illustration is the famous "Moses illusion" (Barton & Sanford, 1993; Erickson & Mattson, 1981; Reder & Kusbit, 1991), in which readers who are asked questions such as *How many animals of each sort did Moses put on the ark*? tend to reply by saying "two" instead of objecting to the presupposition behind the question. Yet another example is that although most people know that survivors are people fortunate enough to live through a calamity, a surprising number of subjects will respond to a question asking where survivors should be buried by offering useful suggestions instead of, again, challenging the questioner (Barton & Sanford, 1993). Pullum refers to some of these sentences as Plausible Angloid Gibberish (Language Log, 2004), pointing to examples such as *More people have been to Russia than I have*, which at first seems fine but on reflection makes no sense (see Wellwood, Pancheva, Hacquard, & Phillips, 2018 for a recent investigation of these intriguing forms).

In a series of experiments summarized elsewhere, my colleagues and I have provided evidence that misinterpretations are common and systematic. Based on these results, we argued for a framework we call Good-Enough Language Processing (Ferreira et al., 2002). One key finding (Christianson, Hollingworth, Halliwell, & Ferreira, 2001) is that subjects presented with a garden-path sentence such as *While Mary bathed the baby played in the crib* tend to mis-interpret it to mean that Mary bathed the baby, even though the global form of the sentence specifies that Mary is bathing herself and not the baby. Readers obtain this incorrect reading even though subtle syntactic tests reveal that these subjects have correctly analyzed and revised the syntactic structure so as to make the first clause intransitive, and they have parsed *the baby* as the subject of the main clause. The problem seems to be that the initial misinterpretation built incrementally as the sentence is processed lingers in memory after the sentence is over and even influences the processing of a subsequent sentence (Slattery, Sturt, Christianson, Yoshida, & Ferreira, 2013). A surprising and underappreciated aspect of these results is that subjects who make these errors are confident their interpretations are right; indeed, they are as confident in their incorrect as in their correct answers (Christianson et al., 2001). Thus, this is not a matter of people simply getting confused and guessing in response to comprehension questions; instead, people seem unaware that they might not have properly understood the sentence. This pattern of results is reminiscent of findings from the text processing literature, where it has been shown that readers fail to update their representations when initial information is later contradicted (Albrecht & O'Brien, 1993; O'Brien, Rizzella, Albrecht, & Halleran, 1998). It seems that once a proposition or a set of propositions is built in memory, it is tenacious and difficult to erase. Infrequent forms such as the passive are often also frequently misinterpreted (Ferreira, 2003). For example, *The dog was bitten by the man* is often misinterpreted to mean that the dog bit the man, indicating that people's priors (i.e., their belief that dogs bite people, not the other way around) can swamp out the meaning associated with the compositionally derived reading (see also Bader & Meng, 2018).

In summary, many studies—far more than we have space to review here—have provided evidence for misinterpretations. These results should be taken as evidence that researchers cannot simply assume the meaning that subjects in experiments will obtain based solely on the words and form of the sentence but instead should directly assess interpretations using measures that tap into the content of the representation.

What causes misinterpretations?

The tendency for interpretations to deviate from the meaning actually specified by the sentence can be attributed to two general categories of causes. The first category assumes that misinterpretations arise because the language system sometimes fails to work properly due to resource limitations or noise, and the second draws on the notion that the language system might have been designed to handle certain forms well but at the expense of other linguistic forms. Beginning with the first category, certainly no one would be surprised to find that a 3-year-old child does not fully understand some passive constructions or garden-path sentences, since children are unlikely to have the cognitive skills or linguistic experience to handle these more challenging forms. Theories of language understanding have long emphasized the importance of frequency in ease of processing, including lexical frequency (Balota & Chumbley, 1985) and, more recently, form frequency as well (Arnon & Snider, 2010; MacDonald, Pearlmutter, & Seidenberg, 1994). Evidence also suggests that learning takes place during the psycholinguistic experiment itself: For example, garden-path effects often shrink over successive trials (Farmer, Fine, & Jaeger, 2011; Wells, Christiansen, Race, Acheson, & MacDonald, 2009; but see Harrington Stack, James, & Watson, 2018), and a recent report demonstrates that the costs associated with understanding sarcastic statements is evident in the first block of an experiment but not in the second (Olkoniemi, Johander, & Kaakinen, 2019).

Related to this idea is the observation that language processing is noisy: Comprehension often takes place in environments unlike our quiet testing rooms, such as in restaurants in which multiple conversations are happening simultaneously, possibly accompanied by music and other ambient noise. Under these conditions listeners might not be confident about what they have heard, which incentivizes them to default to their priors (Gibson, Bergen, & Piantadosi, 2013; Kuperberg & Jaeger, 2016). Of course, it is also true that reading tends to happen in far less noisy environments than listening, and the input is usually more intelligible. At the same time we often find ourselves having to decipher text that is illegible or difficult to read, either because of messy handwriting or poor typing skills. With the recent widespread adoption of written forms of informal communication such as text messaging, some genres of written language encountered in everyday life have increasingly come to resemble casually generated speech, with all its errors and deviations from standards of well-formedness. We also know that producers are fallible and occasionally make errors when they speak and even when they write. Recent Noisy Channel models of comprehension assume that comprehension includes some built-in "auto-correct" functions that repair deviant input (Gibson et al., 2013). The connection between Noisy Channel normalization and misinterpretations is that the comprehender can never be certain that an atypical form arose from speaker error; instead, the problem may lie with the comprehender, who has normalized a speaker's intended form because it falls outside the comprehender's knowledge or experience or because it is inconsistent with the comprehender's priors and expectations.

The second general category of explanation for misinterpretations does not appeal to the idea of individual difference characteristics such as language experience or to the idea that listeners might mishear, misread, or normalize input due to noise or errors; instead, this other category assumes that certain linguistic forms are like visual illusions: They reveal that the language comprehension system generally operates efficiently but fails on forms for which it essentially was not designed. One example is the missing-verb phrase effect (Gibson & Thomas, 1999), which arises during the comprehension of the notoriously difficult multiple center-embedded structure (King & Just, 1991; Miller & Chomsky, 1963). Gibson and Thomas found that subjects actually prefer the ungrammatical *The ancient manuscript that the graduate student who the new card catalog had confused a great deal was missing a page* to the grammatical version containing the obligatory verb phrase (see also Gimenes, Rigalleau, & Gaonac'h, 2009). The explanation for these findings is not yet clear (for one proposal, see Frank, Trompenaars, & Vasishth, 2016), but one idea is that the comprehension system finds it easier to process a sentence if can skip over words whose integration makes nearly impossible demands on working memory, and this

preference overrides the need for grammatical well-formedness. A similar explanation has been proposed to explain linguistic illusions such as *More people have been to Russia than I have* (Wellwood et al., 2018).

But shouldn't comprehenders realize that their interpretations can't be right because even though bits of it seem fine, the sentence overall is ungrammatical? Here we come to another fundamental bias of the language processing system, which is that it seems to weight local coherence over global well-formedness. For example, evidence from a word-monitoring task indicates that listeners are bothered more by disruptions to the integrity of local prosodic and syntactic phrases than to distortions of global sentence structure (Tyler & Warren, 1987). This bias toward local coherence has also been shown using sentences such as *The coach smiled at the player tossed the frisbee*: Readers analyze the sequence *the player tossed the frisbee* as a coherent constituent, in defiance of the global structure which is inconsistent with that analysis (Tabor, Galantucci, & Richardson, 2004). The comprehension system seems eager to identify and group a lexical head and its dependents (e.g., a verb and its arguments) into a coherent package, but when it comes to combining the packages into a global form, a sort of "good enough" association or construal (Frazier & Clifton, 1996) seems sufficient.

At the same time these findings concerning the dominance of local over global coherence are based primarily on what has been observed from reaction time and reading time measures, so we know little about the meanings those subjects construct from the locally coherent but globally problematic structures. For example, are separate local packages simply conjoined semantically, or are they hierarchically arranged? To answer this question, it is probably necessary to give readers or listeners comprehension questions or a memory task that can reveal the content of the representations they have constructed. Moreover, it is possible that a post-sentence task that assessed comprehension would entice subjects to process the sentences more carefully, eliminating some of these local-over-global effects, or at least providing information about the conditions under which they occur. Here we see another example of how a narrow focus on what can be learned from online measures might make us less aware of the advantages of combining these measures with comprehension assessments. A helpful model of how we might proceed comes from the field of discourse comprehension, in which a similar debate regarding the priority of local versus global processing took place, except the units in question were not phrases within a sentence but rather sentences within a discourse, and the central question was whether priority during processing is given to local coherence, connecting nearby sentences to each other, or global coherence, forming distant connections across a text. The emerging consensus is that both local and global processing take place and that global coherence is in part supported by passive, resonance-based memory processes (Albrecht & O'Brien, 1993; O'Brien & Albrecht, 1992; O'Brien et al., 1998). These conclusions from the discourse processing literature are based on studies that carefully combined the results of outcome and online processing measures.

Why the bias against offline measures?

Given the number of demonstrations of misinterpretations as well as the theoretical proposals meant to explain them, it is somewhat surprising that the content of people's interpretations are not assessed more directly in standard psycholinguistic research. One reason for avoiding the use of measures such as question-answering accuracy or recall is psycholinguists' concern that the results might be dismissed as "merely" reflecting the products of comprehension, the worry that they are not providing what is theoretically critical, and that is information about moment-by-moment, online processing. Reviewers may even recommend that an article not be published until the investigators conduct additional empirical work showing the reported effect is detectable in an online measure, arguing that otherwise the study is not sufficiently informative to justify publication. How did this bias arise, given that surely the domain of psycholinguistics includes understanding the contents of people's mental representations of language as well as the time it takes to create them? A complete answer to this question would require far more space than we have here, but this shift

toward online measures probably arose during the battles of the 1980s between proponents of modular models of language processing and advocates of more interactive approaches. As the theoretical debate moved away from questions about the content of representations to those concerning the incremental retrieval and integration of information sources during comprehension, offline measures came to be seen as essentially irrelevant, except for the role they might play in ensuring that subjects paid at least some attention to the stimuli shown in the experiment. The online measures were emphasized because predictions concerning timing of information use distinguished competing models, and those could only be assessed using online processing measures.

Of course, the ideal research strategy would be to combine the two types of measures: Offline measures such as question-answering accuracy or recall would establish the content of the representation the comprehender generated, and online measures such as reading times or probability of regressions would tell us how those representations are built in real time. The field of discourse comprehension again provides a helpful model. As a reviewer of this article pointed out, two approaches have been used for putting together the two types of findings: one is simply to compare the patterns intuitively and the other relies on analytical tools such as regression or multilevel modeling to evaluate statistically how the measures relate to one another (Magliano & Graesser, 1991; Magliano, Trabasso, & Graesser, 1999; Richter, 2006). This statistical method could be a powerful tool for the investigation of sentence comprehension as well and could help shed light on exactly how processing times and comprehension outcomes relate to one another. For example, one pattern observed in sentence comprehension studies is that readers will sometimes spend little time reading a sentence that is especially difficult (e.g., some garden-path sentences) because it is argued that comprehenders engage in "triage," that is, they quickly give up and move on. A technique that combined data suggesting this tendency with actual outcome measures would be a valuable tool for understanding this triage process and on what level of comprehension it is based.

Another reason for the unpopularity of comprehension tasks as prime dependent measures is the worry that they are "metalinguistic": that is, that they do not reflect the operation of the language processing system itself but instead reveal what readers or listeners know about their own comprehension processes, or that they reflect people's ability to perform abstract, essentially meaningless tasks. For example, grammaticality or acceptability judgment decisions have been part of the language community's toolkit for decades, but it requires some level of abstract thinking to intuit what is behind an instruction such as "ignore the meaning of this sentence and just indicate whether the form seems okay." Although this concern is worth keeping in mind, it is important to appreciate that answering questions about a sentence or text is not outside the range of most people's everyday experience, and certainly college students can execute these instructions naturally, without learning any specific skill for the experiment. Additionally, although debates regarding how linguistic and general cognitive systems interact are not yet resolved (and perhaps never will be), most would agree that there is no clear-cut boundary between the language and the reasoning systems, for example, and therefore much if not all knowledge used for higher level thinking and reasoning is probably also recruited during language comprehension, at least at later stages of processing. The distinction, then, between a linguistic and a metalinguistic task or judgment is likely not categorical but more of a continuum, and the utility of any specific task will depend on the theoretical question that is at issue.

Finally, researchers may suspect that comprehension tasks will affect and change the way the subject approaches the task of understanding sentences, as is already known to occur during discourse comprehension (Britt, Rouet, & Durik, 2018; McCrudden & Schraw, 2007). For example, a common objection to the Christianson et al. (2001) results concerning misinterpretations of garden-path sentences has been to suggest that the question asked after the sentence might have reinstated the original garden-path interpretation. To address this concern we conducted a follow-up study in which people were asked simply to recall the garden-path sentences and found results consistent with the original report (Patson, Darowski, Moon, & Ferreira, 2009). However, we could view this potentially biasing aspect of comprehension tasks as a feature, not a bug; that is, if we assume that language is always understood or processed for some purpose and to some criterion

level of depth or detail, then comprehension tasks are an essential tool for determining how readers flexibly adjust their reading strategies depending on their goals. In one of our studies on comprehension of globally ambiguous sentences such as *The man saw the servant of the actor who was on the balcony* (Swets, Desmet, Clifton, & Ferreira, 2008), we varied the frequency and difficulty of questions to determine whether people always definitively attach an ambiguous relative clause (i.e., do they know whether the servant or actor was the one on the balcony). We found that when questions were simple and did not query the interpretation of the relative clause, reading times for fully ambiguous sentences were faster than for disambiguated controls. In contrast, when questions could not be answered without assigning some attachment site to the relative clause, reading times for the sentences slowed down dramatically, suggesting the need to know what entity the relative clause was modifying altered the subjects' reading strategies (see also Potts, Keenan, & Golding, 1988).

It appears, then, that the goals of the reader or listener influence how they process language. Of course, this should come as no surprise to psycholinguists and cognitive scientists. Although many aspects of language comprehension are automatic, others are difficult and demand attentional resources. This point brings us to the topic of engagement, which we consider next.

Engagement

A theme throughout this piece has been the importance of measuring comprehension directly, not only to assess the content of the representations built during online processing but also to address the critical issue of reader or listener motivation in language processing. Unfortunately, the design of our experiments often requires us to ask subjects to read material that is rather boring: The sentences are about entities and events they don't care about, sentences across trials are unconnected, and the testing room is more conducive to napping than to deep comprehension. If subjects are not asked to answer questions about the texts or to recall the material, they potentially will check out completely. Research on mind-wandering during reading (McVay & Kane, 2012) has revealed that subjects' attention frequently drifts away from what they are supposed to be reading, and those mind-wandering episodes are likely associated with more superficial comprehension.

But aren't the effects of some text characteristics impervious to these variations in listener attention? For example, let's consider lexical frequency effects: Isn't lexical processing essentially automatic, so that frequency differences can be assumed to affect variables such as lexical decision or reading time regardless of comprehender engagement? The answer to this question is apparently not. For example, frequency effects disappear when people mind-wander (Reichle, Reineberg, & Schooler, 2010). Frequency effects also are larger when people proofread compared with when they read for comprehension (Kaakinen & Hyönä, 2010; Schotter, Bicknell, Howard, Levy, & Rayner, 2014), in part because the demand to scrutinize text and detect errors leads readers to pay close attention to lexicality (a word with a zero frequency is likely to be a mistake). If a variable such as word frequency is vulnerable to task demands, then higher level linguistic features including syntactic properties and phrase-level meaning are likely even more susceptible. We cannot put the point better than Schotter et al. (2014): "Our findings imply that, in the future, researchers should anticipate the way in which the instructions they give to subjects and the types of questions they ask of them might change the way they approach the task of reading and subsequently the way in which they process words and sentences" (p. 19).

If subjects are made to read a series of disconnected sentences in a darkened, quiet room for nearly an hour with no comprehension task or with just occasional questions, it is likely they will be disengaged and susceptible to mind-wandering and might resort to simply skimming the presented materials enough to answer those occasional questions, making it difficult to detect the influence of our linguistic manipulations. What is necessary to ensure that subjects process the stimuli from our experiments in a way that resembles how they normally process everyday language is to make sure to

include comprehension questions or a memory recall task that subjects cannot perform by guessing, using their world knowledge, or identifying a few key content words.

Another critical issue regarding the use of comprehension items is the role of feedback. Surprisingly, it is rare for psycholinguistic studies that include comprehension items to give subjects information about whether their answers are right or wrong (notable exceptions include Chen, Gibson, & Wolf, 2005; Garnsey, Pearlmutter, Myers, & Lotocky, 1997; Pearlmutter, Garnsey, & Bock, 1999), which means subjects are not incentivized to perform well even when questions are included. Of course, again, researchers probably worry that difficult comprehension items combined with feedback on performance will cause comprehenders to adopt "special strategies," but our response to this is twofold: First, as stated before, goals and motivations are an important aspect of language processing that we ought to understand better, and, second, surely we want our theories to be about the interpretation of materials our subjects are engaged with rather than stimuli they are merely skimming (unless, of course, our goal is to understand skimming strategies). Thus, we would argue that this aspect of our experimental protocols probably needs a rethink as well, as we discuss in the next section.

A positive consequence of emphasizing engagement in comprehension is that it encourages the use of more naturalistic materials in psycholinguistic research, as has been true in the field of discourse processing for decades (e.g., Magliano & Graesser, 1991). Recently, Hasson, Egidi, Marelli, and Willems (2018) made a compelling case for adopting this approach if we want to develop theories of language processing that take into account the context in which comprehension takes place. This is now a trend in some subareas of psycholinguistics, especially among researchers interested in how variations in predictability affect language processing. The traditional method has been to write sentences containing a word that is either predictable or unpredictable, but the exciting alternative is to use computational tools to assign predictability values to all words in some naturally occurring text and then relate those continuously varying levels of predictability to online measures of processing such as reading time, electrophysiological responses, and the hemodynamic response (Brennan, 2016; Brennan, Stabler, Van Wagenen, Luh, & Hale, 2016; Henderson, Choi, Luke, & Desai, 2015; Luke & Christianson, 2016; Willems, Frank, Nijhof, Hagoort, & van Den Bosch, 2016).

Interconnections among task, comprehension level, and online measures

Ultimately, our goal should be a theory of language processing that links the three concepts we have highlighted here: (1) depth of processing or comprehension, which can be assessed using measures such as question-answering accuracy or recall/recognition; (2) engagement or motivation, which is influenced by task (among other factors); and (3) online processing, as revealed in measures such as fixation times, regression probability, and reaction times. These concepts have arguably been investigated pairwise in past research (e.g., we know a bit about how engagement affects depth of processing, particularly from the education literature [Wigfield et al., 2008]), but what we currently lack is a complete understanding of how engagement affects depth of processing and in turn influences what we should expect to find in our online measures.

Fortunately, researchers are beginning to evaluate this sequence that is hypothesized to proceed from engagement to comprehension depth to effects on dependent measures. One eye-tracking study conducted to assess the importance of rereading to comprehension levels reported that subjects who could not make regressive eye movements during reading made more errors on comprehension questions than did those who could regress (Schotter, Tran, & Rayner, 2014). More recently, difficult comprehension questions have been linked to more rereading and regressions but not to longer initial reading times (Weiss, Kretzschmar, Schlesewsky, Bornkessel-Schlesewsky, & Staub, 2017; Wotschack & Kliegl, 2013).

A theoretical understanding of these interconnections is obviously critical for a basic science of language processing, and it is essential in domains such as education, assessment, and remediation.

In addition, if these links were understood, we could design our studies more intelligently and be guided by data and theoretically motivated hypotheses when we decide how we will assess comprehension. At the moment these critical scientific decisions are made based on rules of thumb passed on from investigator to investigator, with no strong empirical basis for those choices. Our hope is that future research will be directed toward solving some of these important challenges in psycholinguistic research and that better connections can be formed among subfields of the psychology of language, including investigations of sentence and discourse processing.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Albrecht, J. E., & O'Brien, E. J. (1993). Updating a mental model: Maintaining both local and global coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(5), 1061–1071. doi:10.1037//0278-7393.19.5.1061

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67–82. doi:10.1016/j.jml.2009.09.005

Bader, M., & Meng, M. (2018). The misinterpretation of noncanonical sentences revisited. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *44*(8), 1286–1311. doi:10.1037/xlm0000519

Balota, D. A., & Chumbley, J. I. (1985). The locus of word-frequency effects in the pronunciation task: Lexical access and/or production. *Journal of Memory and Language*, *24*(1), 89–106. doi:10.1016/0749-596X(85)90017-8

Barton, S. B., & Sanford, A. J. (1993). A case study of anomaly detection: Shallow semantic processing and cohesion establishment. *Memory & Cognition*, *21*(4), 477–487. doi:10.3758/BF03197179

Brennan, J. (2016). Naturalistic sentence comprehension in the brain. *Language and Linguistics Compass*, *10*(7), 299–313. doi:10.1111/lnc3.v10.7

Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W. M., & Hale, J. T. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, *157*, 81–94. doi:10.1016/j.bandl.2016.04.008

Britt, M., Rouet, J., & Durik, A. (2018). *Literacy beyond text comprehension: A theory of purposeful reading*. New York, NY: Routledge.

Chen, E., Gibson, E., & Wolf, F. (2005). Online syntactic storage costs in sentence comprehension. *Journal of Memory and Language*, *52*(1), 144–169. doi:10.1016/j.jml.2004.10.001

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, *42*(4), 368–407. doi:10.1006/cogp.2001.0752

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Memory and Language*, *20*(5), 540.

Farmer, T., Fine, A., & Jaeger, T. (2011). Implicit context-specific learning leads to rapid shifts in syntactic expectations. *33rd Annual Meeting Conference Cognitive Science Society*, *8*, 2055–2060.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, *47*(2), 164–203. doi:10.1016/S0010-0285(03)00005-7

Ferreira, F., Bailey, K., & Ferraro, V. (2002). Good enough representations in language comprehension. *Current Directions in Psychological Science*, *11*(1), 1–15.

Ferreira, F., & Patson, N. D. (2007). The "good enough" approach to language comprehension. *Language and Linguistics Compass*, *1*(1–2), 71–83. doi:10.1111/j.1749-818X.2007.00007.x

Frank, S. L., Trompenaars, T., & Vasishth, S. (2016). Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? *Cognitive Science*, *40*(3), 554–578. doi:10.1111/cogs.2016.40.issue-3

Frazier, L., & Clifton, C. (1996). *Construal*. Cambridge, MA: MIT Press.

Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, *37*(1), 58–93. doi:10.1006/jmla.1997.2512

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(20), 8051–8056. doi:10.1073/pnas.1216438110

Gibson, E., & Thomas, J. (1999). Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes*, *14*(3), 225–248. doi:10.1080/016909699386293

Gimenes, M., Rigalleau, F., & Gaonac'h, D. (2009). When a missing verb makes a French sentence more acceptable. *Language and Cognitive Processes*, 24(3), 440–449. doi:10.1080/01690960802193670

Harrington Stack, C. M., James, A. N., & Watson, D. G. (2018). A failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition*, 46(6), 864–877. doi:10.3758/s13421-018-0808-6

Hasson, U., Egidi, G., Marelli, M., & Willems, R. M. (2018). Grounding the neurobiology of language in first principles: The necessity of non-language-centric explanations for language comprehension. *Cognition*, 180, 135–157. doi:10.1016/j.cognition.2018.06.018

Henderson, J. M., Choi, W., Luke, S. G., & Desai, R. H. (2015). Neural correlates of fixation duration in natural reading: Evidence from fixation-related fMRI. *NeuroImage*, 119, 390–397. doi:10.1016/j.neuroimage.2015.06.072

Johnson-Laird, P. N. (1981). Comprehension as the construction of mental models. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 295(1077), 353–374. doi:10.1098/rstb.1981.0145

Kaakinen, J. K., & Hyönä, J. (2010). Task effects on eye movements during reading. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 36(6), 1561–1566. doi:10.1037/a0020693

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, 30(5), 580. doi:10.1016/0749-596X(91)90027-H

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, 31(1), 32–59. doi:10.1080/23273798.2015.1102299

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, 88, 22–60. doi:10.1016/j.cogpsych.2016.06.002

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution [corrected]. *Psychological Review*, 101(4), 676–703. doi:10.1037/0033-295X.101.4.676

Magliano, J.P., & Graesser, A.C. (1991). A 3-Pronged method for studying inference generation in literary text. *Poetics*, 20(3), 193–232. doi:10.1016/0304-422X(91)90007-C

Magliano, J.P., Trabasso, T., & Graesser, A.C. (1999). Strategic processing during comprehension. *Journal of Educational Psychology*, 91(4), 615–629. doi:10.1037/0022-0663.91.4.615

McCrudden, M., & Schraw, T. (2007). Relevance and goal-focusing in text processing. *Educational Psychology Review*, 19(2), 113–139. doi:10.1007/s10648-006-9010-7

McVay, J. C., & Kane, M. J. (2012). Why does working memory capacity predict variation in reading comprehension? On the influence of mind wandering and executive attention. *Journal of Experimental Psychology. General*, 141(2), 302–320. doi:10.1037/a0025250

Miller, G. A., & Chomsky, N. (1963). Finitary models of language users. In D. Luce (Ed.), *Handbook of mathematical psychology* (pp. 2–419). London, UK: John Wiley & Sons.

O'Brien, E. J., & Albrecht, J. E. (1992). Comprehension strategies in the development of a mental model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(4), 777–784.

O'Brien, E. J., Rizzella, M. L., Albrecht, J. E., & Halleran, J. G. (1998). Updating a situation model: A memory-based text processing view. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 24(5), 1200–1210. doi:10.1037/0278-7393.24.5.1200

Olkoniemi, H., Johander, E., & Kaakinen, J. K. (2019). The role of look-backs in the processing of written sarcasm. *Memory & Cognition*, 47(1), 87–105.

Patson, N. D., Darowski, E. S., Moon, N., & Ferreira, F. (2009). Lingering misinterpretations in garden-path sentences: Evidence from a paraphrasing task. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 35(1), 280–285. doi:10.1037/a0014276

Pearlmutter, N. J., Garnsey, S. M., & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and Language*, 41(3), 427–456. doi:10.1006/jmla.1999.2653

Potts, G. R., Keenan, J. M., & Golding, J. M. (1988). Assessing the occurrence of elaborative inferences: Lexical decision versus naming. *Journal of Memory and Language*, 27(4), 399–415. doi:10.1016/0749-596X(88)90064-2

Pullum, G. K. (2004, May 6). *Plausible Angloid gibberish* [Blog post]. Retrieved from http://itre.cis.upenn.edu/~myl/languagelog/archives/000860.html

Reder, L. M., & Kusbit, G. W. (1991). Locus of the Moses illusion: Imperfect encoding, retrieval, or match. *Journal of Memory and Language*, 30(4), 385–406. doi:10.1016/0749-596X(91)90013-A

Reichle, E. D., Reineberg, A. E., & Schooler, J. W. (2010). Eye movements during mindless reading. *Psychological Science: A Journal of the American Psychological Society/APS*, 21(9), 1300–1310. doi:10.1177/0956797610378686

Richter, T. (2006). What is wrong with ANOVA and multiple regression? Analyzing sentence reading times with hierarchical linear models. *Discourse Processes: A Multidisciplinary Journal*, 41(3), 221–250. doi:10.1207/s15326950dp4103_1

Schotter, E. R., Bicknell, K., Howard, I., Levy, R., & Rayner, K. (2014). Task effects reveal cognitive flexibility responding to frequency and predictability: Evidence from eye movements in reading and proofreading. *Cognition*, 131(1), 1–27. doi:10.1016/j.cognition.2013.11.018

Schotter, E. R., Tran, R., & Rayner, K. (2014). Don't believe what you read (only once): Comprehension is supported by regressions during reading. *Psychological Science*, 25(6), 1218–1226. doi:10.1177/0956797613518350

Slattery, T. J., Sturt, P., Christianson, K., Yoshida, M., & Ferreira, F. (2013). Lingering misinterpretations of garden path sentences arise from competing syntactic representations. *Journal of Memory and Language*, *69*(2), 104–120. doi:10.1016/j.jml.2013.04.001

Sturt, P., Sanford, A. J., Stewart, A., & Dawydiak, E. (2004). Linguistic focus and good-enough representations: An application of the change-detection paradigm. *Psychonomic Bulletin & Review*, *11*(5), 882–888. doi:10.3758/BF03196716

Swets, B., Desmet, T., Clifton, C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory & Cognition*, *36*(1), 201–216. doi:10.3758/MC.36.1.201

Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, *50*(4), 355–370. doi:10.1016/j.jml.2004.01.001

Tyler, L. K., & Warren, P. (1987). Local and global structure in spoken language comprehension. *Journal of Memory and Language*, *26*(6), 638–657. doi:10.1016/0749-596X(87)90107-0

van Den Broek, P., & Gustafson., M. (1999). Comprehension and memory for texts: Three generations of reading research. In R. Goldman, A. C. Graesser, & P. van Den Broek (Eds.), *Narrative comprehension, causality, and coherence: Essays in honor of Tom Trabasso* (pp. 15–34). Mahwah, NJ: Erlbaum.

Wason, P. C., & Reich, S. S. (1979). A verbal illusion. *The Quarterly Journal of Experimental Psychology*, *31*(Pt 4), 591–597. doi:10.1080/14640747908400750

Weiss, A. F., Kretzschmar, F., Schlesewsky, M., Bornkessel-Schlesewsky, I., & Staub, A. (2017). Comprehension demands modulate re-reading, but not first pass reading behavior. *Quarterly Journal of Experimental Psychology*, *71*(1), 198–210.

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence comprehension: Statistical learning, working memory and individual differences. *Cognitive Psychology*, *58*(2), 250–271. doi:10.1016/j.cogpsych.2008.08.002

Wellwood, A., Pancheva, R., Hacquard, V., & Phillips, C. (2018). The anatomy of a comparative illusion. *Journal of Semantics*, *35*(3), 543–583. doi:10.1093/jos/ffy014

Wigfield, A., Guthrie, J. T., Perencevich, K. C., Taboada, A., Klauda, S. L., McRae, A., & Barbosa, P. (2008). Role of reading engagement in mediating effects of reading comprehension instruction on reading outcomes. *Psychology in the Schools*, *45*(5), 432–445. doi:10.1002/(ISSN)1520-6807

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & van Den Bosch, A. (2016). Prediction during natural language comprehension. *Cerebral Cortex*, *26*(6), 2506–2516.

Wotschack, C., & Kliegl, R. (2013). Reading strategy modulates parafoveal-on-foveal effects in sentence reading. *Quarterly Journal of Experimental Psychology*, *66*(3), 548–562. doi:10.1080/17470218.2011.625094