ISSN 0163-853X Volume 57, Number 10 December 2020

Discourse Processes

A Multidisciplinary Journal

The Official Journal of The Society for Text and Discourse

DP

Editor
David N. Rapp

Associate Editors
Rick Dale
Sidney D'Mello
Johanna Kaakinen
Debra L. Long

Routledge Taylor & Francis Group

Routledge Taylor & Francis Group

Discourse Processes

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hdsp20

Individual differences in expecting coherence relations: Exploring the variability in sensitivity to contextual signals in discourse

Merel C. J. Scholman , Vera Demberg & Ted J. M. Sanders

To cite this article: Merel C. J. Scholman , Vera Demberg & Ted J. M. Sanders (2020) Individual differences in expecting coherence relations: Exploring the variability in sensitivity to contextual signals in discourse, Discourse Processes, 57:10, 844-861, DOI: 10.1080/0163853X.2020.1813492

To link to this article: https://doi.org/10.1080/0163853X.2020.1813492

Published online: 02 Oct 2020.

DISCOURSE PROCESSES
2020, VOL. 57, NO. 10, 844–861
https://doi.org/10.1080/0163853X.2020.1813492

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Individual differences in expecting coherence relations: Exploring the variability in sensitivity to contextual signals in discourse

Merel C. J. Scholman[a], Vera Demberg[b], and Ted J. M. Sanders[c]

[a]Department of Language Science and Technology, Saarland University; [b]Department of Language Science and Technology, Department of Computer Science, Saarland University; [c]Utrecht Institute of Linguistics OTS, Utrecht University

ABSTRACT

The current study investigated how a contextual list signal influences comprehenders' inference generation of upcoming discourse relations and whether individual differences in working memory capacity and linguistic experience influence the generation of these inferences. Participants were asked to complete two-sentence stories, the first sentence of which contained an expression of quantity (*a few, multiple*). Several individual-difference measures were calculated to explore whether individual characteristics can explain the sensitivity to the contextual list signal. The results revealed that participants were sensitive to a contextual list signal (i.e., they provided list continuations), and this sensitivity was modulated by the participants' linguistic experience, as measured by an author recognition test. The results showed no evidence that working memory affected participants' responses. These results extend prior research by showing that contextual signals influence participants' coherence-relation-inference generation. Further, the results of the current study emphasize the importance of individual reader characteristics when it comes to coherence-relation inferences.

When comprehenders understand discourse, they understand more than what is explicitly stated: They create connections between clauses and sentences that are left implicit in the discourse. These connections, known as *coherence relations*, are needed to create a coherent mental representation of the text (Sanders & Noordman, 2000; Sanders et al., 1992). Comprehenders can make use of different sources of information to infer coherence relations. The most-well-studied type of cues that direct the interpretation process is the prototypical relational marker: connectives and cue phrases such as *because* and *for example*. Relations that are signaled by such markers are referred to as *explicit relations*.

Many relations are in fact left *implicit*; that is, they are not marked by a connective or cue phrase (over 50% of relations in the Penn Discourse Treebank, see Prasad et al., 2007). For such relations, comprehenders can exploit regularities of signals that tend to co-occur with specific types of relations (these regularities will be referred to as *relational signals* from now on). Relational signals include lexico-semantic word pairs such as *good–bad* and *many–one of these* (Park & Cardie, 2012; Pitler et al., 2009); implicit causality verbs such as *praise* or *hate* (Ehrlich, 1980; Kehler et al., 2008; Koornneef & Van Berkum, 2006); and transfer-of-possession verbs such as *give* or *bring* (Elman et al., 2006; Stevenson et al., 2000).

The existing research into relational signals has focused on local, segment-internal markers (markers that occur in the first argument of the relation, such as implicit causality verbs) or intersegmental markers (markers that occur between two segments of a relation, such as the connective *but*). Much less is known about whether global, cross-sentence markers influence the generation of a specific type of coherence-relation inference. The current study therefore aims to extend previous

CONTACT Ted J. M. Sanders t.j.m.sanders@uu.nl Utrecht University

*All authors contributed equally.

research by investigating whether a more global cue, located before the start of the first segment of the relation, leads comprehenders to generate inferences of a specific type of upcoming coherence relation.

The type of cue that is investigated in the current study is an expression of quantity, such as *a few* or *several*.[1] A wealth of studies provides evidence that expressions of quantity not only convey quantity but also serve broader discourse functions. In particular, quantifiers provide subtle information that influences the kind of inferences that a comprehender makes (see Moxey & Sanford, 2000; Paterson et al., 2009, for a review of studies). In the current study, it is assumed that such expressions can signal upcoming list relations. Consider Example (1).

(1) The woman experienced several unfortunate events last night.

A logical continuation of this passage would be a specification of multiple unfortunate events, for example, *She got wine thrown at her by her dining companion and the heel of her Jimmy Choo broke.* Providing only one instance of an event could render the story incomplete because several instances were evoked in (1). The current study presents a first investigation into whether comprehenders are sensitive to expressions of quantity located in the context. Several individual difference measures were included to investigate whether sensitivity to the list signal is modulated by individual comprehender characteristics.

Coherence relational signals

The linguistic marking of coherence relations is known to have a significant effect on the processing of such relations. A variety of studies has shown that comprehenders use discourse markers such as *because* and *however* to anticipate what type of information they will encounter in the second argument of the relation (see, e.g., Canestrelli et al., 2013; Drenhaus et al., 2014; Köhne & Demberg, 2013; Sanders & Noordman, 2000; Van Silfhout et al., 2014; Xiang & Kuperberg, 2015). A second stream of research has shown that cues found in the first argument of the relation can influence what type of relation comprehenders expect to encounter in the discourse (see, e.g., Rohde & Horton, 2014; Scholman et al., 2017). However, most research has focused on the effect of the prototypical markers connectives and cue phrases. Less is known about which other types of relational cues exist and how they influence processing.

To address this gap in the literature, recent studies have made an effort to inventorize other relational signals. Das and Taboada (2018), for example, identified five categories of linguistic cues that can signal coherence relations: reference, semantic, lexical, syntactic, and graphical features. Hoek et al. (2018) identified three distinct ways in which segment-internal elements, such as negation words and verb tense, can systematically interact with connectives to express a relation. Although these recent efforts have provided more insight into the types and functioning of segment-internal signals that can be identified in texts, more work is needed to investigate their effects on comprehenders.

Moreover, as little as we know about segment-internal signals, we know even less about contextual signals. In the field of discourse coherence, context is assumed to be important for interpretation and comprehension. Context grounds the discourse that comprehenders construct (Cornish, 2009), which means that the interpretation of any sentence (other than the first) in a discourse is constrained by the preceding context (Song, 2010). This preceding context has significant effects on different facets of discourse comprehension, such as determining the rhetorical role each sentence plays in the discourse, and the temporal relations between the events described (Lascarides et al., 1992; Sanders et al., 1992; Spooren & Degand, 2010). Context is therefore considered crucial in discourse processing.

The evidence so far regarding the effects of context on interpretation is mixed. Upadhyay et al. (2019) showed that quantifiers located in the preceding story context (such as *a lot of weak students*) influenced readers' comprehension of subsequent negative quantifiers (such as *few students*). Sanders (1997) found that readers took into account the context when determining the subjectivity of coherence relations but only for those relations that were ambiguous between an objective and subjective reading. By contrast, Canestrelli et al. (2016) did not find evidence that contextual subjective markers, such as *fantastic* and *horribly*, facilitated the processing of subjective causal relations. Moreover, Scholman and Demberg (2017) found no significant difference in interpretations of coherence relations presented with or without their linguistic context. Note that this might be caused by the fact that the context and items were not manipulated: it could be that there were no strong relational cue words in the context of enough experimental items to reveal an effect.

In sum, there is ample evidence that comprehenders are able to use connectives to generate coherence-relation inferences, and a growing body of work shows that intersegmental signals frequently co-occur with specific relation types. However, the effect of contextual signals on coherence-relation interpretation is limited. Context is considered to be important for discourse comprehension, but the few studies that have investigated the effects of context have provided mixed results. The current study targets a contextual list signal to test whether comprehenders are sensitive to such a global relational signal. Quantity expressions located in the context are not restrictive cues, and the degree of sensitivity to such a cue can therefore vary between individuals. The next question that arises is whether all comprehenders are sensitive to cues.

Individual differences in language comprehension

Traditionally, individual variation is often overlooked in the field of coherence relations (but see Kamalski et al., 2008; McNamara et al., 1996; Van Silfhout et al., 2014, for notable examples). However, a large body of research has shown that individual differences affect language processing and comprehension in various areas of linguistics, such as syntax, phonology, and second language acquisition (e.g., Afflerbach, 2015; Fuchs et al., 2015; Kidd et al., 2018).

A coherent representation of an unfolding discourse is made possible through the coordination of multiple cognitive and linguistic processes. These processes vary across individuals and throughout the lifespan and are related to developmental and environmental variables, such as linguistic experience and the quality of linguistic input (Kidd et al., 2018). In the current study, we explore individual differences in both cognitive and linguistic processes.

Working memory

Regarding cognitive factors, most of the research on individual differences has focused on the influence of (verbal) working memory capacity. Working memory (WM) is a limited-capacity, short-duration system that is responsible for the temporary storage of words, phrases, and ideas and for the processing of new and already stored information (Baddeley & Hitch, 1974; Just & Carpenter, 1992). The WM system is integral to maintaining activated discourse representations of words and sentences and computing relations among these representations. For example, compared to high-WM comprehenders, low-WM comprehenders experience more difficulty with maintaining both an overall passage representation and more-local–sentence-to-sentence connections (Whitney et al., 1991).

The most commonly used task in the field of linguistics is the verbal working memory task (VWM), often referred to as the reading span task (or RSpan) (see, e.g., Caplan & Waters, 1999; Just & Carpenter, 1992; Waters & Caplan, 1996). For this task, participants judge the acceptability of sentences, and after judging several sentences, the readers are instructed to recall the last word of each sentence in the series in the order that they read the sentences. The task thus includes both language processing and storage components.

However, the verbal working memory capacity has been critiqued for its dependence on linguistic experience (see, e.g., Traxler et al., 2012). To tap into a more-language-independent component of working memory, researchers make use of nonverbal tasks. A commonly used nonverbal task is the operation span test (or OSpan), for which participants are asked to remember lists of letters while they solve simple arithmetic problems (Conway et al., 2005; Turner & Engle, 1989). The absence of linguistic stimuli in this test means that there is less overlap between the operation span test and typical linguistic experiments.

In the current study, both the reading span test and the operation span test were included to explore the possible individual effects of both verbal and nonverbal components of working memory. Given that comprehenders with low WM capacity have a reduced ability to temporarily store and manipulate information and have difficulty maintaining more-local–sentence-to-sentence connections, they might be less able to adequately generate inferences of upcoming coherence relations using a cross-sentence cue (that is, a cue that is located outside of the relational segments). For the current study, we therefore predict that participants with a lower WM capacity would be less capable of applying the contextual list signal to generate and maintain a list relation inference. High-WM capacity participants would thus be more sensitive to the contextual signal than low-WM capacity participants. No specific predictions are made regarding a possible difference between the effects of verbal and nonverbal working memory capacity, given that both measures have been found to impact language processing.

Linguistic experience

Historically, individual variability in language processing and comprehension has been attributed mainly to variability in working-memory capacity. However, some have argued that correlations between working-memory capacity and language comprehension might have been spurious, in so far as other measures that capture relevant aspects better have not been collected (see, e.g., Van Dyke et al., 2014). It has also been argued that the variability in capacity can partly be attributed to differences in linguistic experience (see, e.g., Farmer et al., 2017).

To be able to tease apart the relationships between working memory and linguistic experience, a print-exposure test and reading-habits survey were included in this study. Previous research has shown that variability in print exposure (i.e., the amount of text people had read) can account for differences in reading comprehension (Cipielewski & Stanovich, 1992). Individuals with more print exposure are more likely to learn about low-frequency structures compared to individuals with less print exposure (Freed et al., 2017). Moreover, increased print exposure leads readers to learn more about semantic relations and concepts and to acquire skills such as logical reasoning (Scribner, 1981; Stanovich & Cunningham, 1993).

The most commonly used measure for print exposure is the Author Recognition Test (ART). For this test, participants are presented with a list of author and nonauthor names and asked to indicate which names they recognize to be authors. The test has been found to be a strong predictor of reading skill, likely because author knowledge is often acquired through reading or other forms of print exposure (Freed et al., 2017). Nevertheless, because the ART does not directly measure reading habits, it remains possible that high print-exposure scores could be achieved by individuals who rarely read. To account for this, we also included a reading-habits survey (Scales & Rhee, 2001). This survey provides a measure of how often and what comprehenders like to read and whether they enjoy reading in general.

Given that comprehenders with more linguistic experience also have more experience with low-frequency discourse structures and various semantic relations, it is hypothesized that these participants will be more sensitive to the contextual list signal in the current study. This would mean that participants with more print exposure and a higher reading frequency would be more likely to generate inferences of list relations based on the list signal than those with low print exposure and reading frequency.

Method

The current study investigates whether comprehenders are sensitive to a contextual list signal and whether this sensitivity is modulated by individual reader characteristics. To measure comprehenders' sensitivity to the list signal, a method was needed to show how they interpret the discourse with the list signal. More traditional reading tasks, such as offline comprehension questions or online reading-time measures, do not provide insight into exactly which type of coherence relation comprehenders construct as part of their mental representation of the discourse. We therefore decided to run a story-continuation study to answer the first question. Story-continuation tasks require participants to read and comprehend a prompt and then provide a written continuation in accordance with their interpretation of that prompt. Participants' written responses are thus used to provide insight into their mental representation of the discourse (cf. Hoek, 2018; Kehler et al., 2008; Scholman et al., 2017). To address the individual variability question, we conducted two working-memory tasks and two linguistic-experience tasks.

Participants within the age range of 25–35 years were invited to participate. This age range was chosen to ensure that the version that we used of the Author Recognition Task would be relevant (because it is a generation-sensitive task) and that the participants were comparable in terms of working memory, which varies over the lifespan. Participants were recruited online via the crowdsourcing platform Prolific.[2] They were asked to participate in several studies and were paid a bonus upon completion of the final study. Crowdsourcing was chosen because it allowed us to recruit participants with the demographic background of interest for this study (i.e., native English speakers from the United States, between the ages of 25 and 35, from various educational backgrounds). Moreover, crowdsourcing provided access to a sample of naïve participants that is likely to vary more in terms of individual differences than typical college-population samples. Finally, it facilitated the repeated-measures design of the study by allowing for easy collection of data for each study on different days.

Although untraditional, the use of crowdsourcing to collect linguistic data is becoming more common in the field. An important concern that arises when using crowdsourced participants is the reliability of the data. However, various studies, starting with Snow et al. (2008), have shown that crowdsourcing can be used to obtain reliable data for linguistic studies. Crowdsourced participants have been used successfully in a variety of discourse tasks, including story continuation studies (Andersson & Spenader, 2014; Hoek, 2018; Kehler & Rohde, 2017). Individual-differences tests are most commonly done in a laboratory, but even for these studies we are experiencing rapid technological developments. Most notably, Hicks et al. (2016) developed working-memory tests that can be conducted via crowdsourcing. Unsworth et al. (2005) and Von der Malsburg and Vasishth (2013) have also developed automated versions of the operation-span test that can be used online (but were still applied in the laboratory). The current study represents one of the first efforts to collect these data in a crowdsourced setting. When presenting the results, we will also present comparisons with in-lab studies to verify the validity of the data.

Participants

One hundred sixty-three native-English-speaking participants (age range, 25–35 years; mean age, 30 years; 83 female), registered as "participants" on the Prolific website, took part in all five separate tasks. All participants indicated that they were born in the United States and were currently living there. Sixty-three participants had completed postsecondary-level education (had earned an undergraduate degree, a graduate degree, or a doctorate); 100 participants had completed high school or had no formal qualifications.

Story-continuation study

Participants were asked to provide one- or two-sentence completions to 20 experimental passages.[3] Each passage consisted of two sentences. The first sentence was the context sentence, which introduced a referent and described the situation the referent was in. This sentence was manipulated to create two versions. In the control (nonlist signal) condition, the context sentence described the place or situation the referent was in, without referring to multiple instances. In the list-signal condition, the context sentence always contained an expression of quantity. Included expressions were *a couple, a few, multiple*, and *several*. The second sentence of the items started with a pronoun that referred to the referent and described a situation or event with a second referent. This sentence was identical over the two conditions. Passage 2 shows an example of an experimental item.

(2) *List* The woman experienced several unfortunate events last night. She got wine thrown at her by her dining companion.

Control The woman went out for dinner last night. She got wine thrown at her by her dining companion.

The items were constructed in such a way that a causal continuation was most likely in the control condition (i.e., participants were most likely to provide the reason or result of the wine-throwing as a continuation of the control condition in Passage 2). This was done to ensure that the items were comparable to each other in terms of the type of expectations that they would elicit. The causal continuation likelihood was manipulated by including implicit causality verbs or events that elicit an expectation of a causal relation (such as in Passage 2)(cf. Mak & Sanders, 2013). This likelihood was pretested in a story-continuation task with a different group of participants. The pretest included 32 items, of which the 20 items that received the most list continuations in the list condition were selected. All items received a majority of causal continuations in the control condition.

Ten filler items were included to decrease the proportion of list signals in the experiment. The filler items consisted of the same structure as items in the control condition. The experimental and filler items were counterbalanced across two lists, with each experimental passage appearing in a different condition in each list. All participants saw all experimental passages and filler passages. Participants were randomly assigned to one of the lists, and the items in a list were presented in a unique order to each participant.

Participants were instructed to complete the stories with one or two sentences that were complete, grammatical, and longer than two or three words. This method allowed them to provide a causal as well as a list continuation, if they created an inference for both a causal relation and a list relation based on the prompts. The instructions also emphasized that participants should write the continuation that came to mind first and should treat every story separately.

Coding procedure

For the analysis, the completions were labeled as list, causal (cause or result), or other (temporal, elaboration, contrast, concession) in relation to the prompt. Passages (3) through (8) present examples of continuations that participants provided in response to the list condition in Passage (2), above.

(3) Afterward, she tried to go home but the taxis all sped by, spraying her with dirty street water as they flew past. When she finally got home she found out that her house had burned down while she was away.
(4) She shouldn't have stolen his french fries. No one eats Joey's food.
(5) The wine ruined her suede sweater.
(6) Nothing she could do about it now, though. She was hours away from home, and was determined to not let a stain ruin her evening.
(7) Her companion stormed out angrily.
(8) It was straight out of a scene of *Real Housewives*.

Relations within the continuation were not annotated (e.g., the contrastive relation marked by *but* in Example (3)); these are outside of the scope of investigation. Rather, the analysis focused on whether the continuation referred to a second item of the list initiated by the list signal, a cause for the event in the prompt, a result of the event in the prompt, or none of these (in which case it was labeled as "other").

A list continuation was defined as a continuation whereby a second instance was provided in relation to the list signal in the prompt. For example, the continuation in Example (3) made reference to two more unfortunate events: a taxi spraying her with dirty street water and her house burning down. Note that the connective *Afterward* also marked a temporal relation with the previous sentence, but we here focus on whether a list item, cause, or result was provided in relation to the events in the prompt.

A causal continuation was defined as a continuation whereby either a cause or a result of either sentence in the prompt was given. Example (4) presents a cause for the companion throwing wine at our protagonist: she stole his french fries. Note that the continuation was not framed as a typical causal relation (e.g., *she got wine thrown at her because she stole his french fries*); rather, the continuation stated the error she made, which was in fact the reason for the wine-throwing event. Example (5) presents the result of the wine-throwing event: her sweater is ruined.

Continuations that did not provide a list item, cause, or result were marked as "other." Such continuations were often contrastive (Example (6)), temporal (Example (7)), or another type of elaboration or evaluation (Example (8)).

Of the 3,260 provided continuations, 87% consisted of one sentence; 0.3% (10 continuations in total) consisted of more than two sentences. The full continuations were taken into consideration, even though they had more than two sentences. On average, continuations consisted of 12 words. In less than 5% of the data, the continuations contained multiple relations with the prompt; for example, one segment provided a list item in response to the contextual signal and another segment provided a cause for the event (see Example (9)). In such cases, if one of the two relation types was list, then the continuation was labeled as a list continuation (this occurred in 26 instances). If one of the two relation types was causal and the other one was "other," then the continuation was annotated as causal.

(9) *Prompt*: The teenager had disagreements with a couple of people yesterday. He got into an argument with his mother. *Continuation*: He also got into an argument with his girlfriend. Both of those arguments were about his drug use.

The continuations were coded by one of the authors. Of the 3,260 continuations, a random subset of 10% was double-coded by a second, independent, experienced coder who was unrelated to the current study. The agreement between the coders was high: 89%, κ= 0.80. All disagreements were instances wherein one annotator assigned the label "causal" whereas the other assigned "other" to a continuation. The agreement on the label "list" was 100%.

Working memory tests

Reading-span test

Verbal working memory was measured by an automated version of the reading-span task (RSpan), with items similar to those used by Waters et al. (1987) and Waters and Caplan (1996). Participants read several sets of sentences and judged the acceptability of each sentence. After the participants judged all sentences in a set, they were instructed to recall the last word of each sentence in the order that they read the sentences. The task thus included both processing and storage components.

Fifty-six sentences of eight to 11 words were created. Half of the sentences contained a verb that required an animate subject (e.g., *donate, forget*) and half contained a verb that required an animate object (e.g., *fascinate, disappoint*). Examples (10) and (11) illustrate these sentences.

(10) It was the yellow notebook that the girl brought.
(11) The boy envied the friend that bought a game.

Half of the sentences was acceptable and half was unacceptable. Unacceptable sentences were formed by inverting the animacy of the subject and object noun phrases (e.g., *It was the girl that the yellow notebook brought*).

The automated version of the reading span was developed following recommendations by Conway et al. (2005), Von der Malsburg and Vasishth (2013), and Unsworth et al. (2005). The test consisted of two practice phases and the main phase. In the first practice phase, participants judged the acceptability of sentences without remembering the final words. The reaction time was measured for each judgment. The average reaction time plus two standard deviations was used as a cutoff (i.e., a time-out) for answering each sentence in later stages, to prevent participants from taking extra time to rehearse the words and from writing down the word elsewhere. This individual time-out period allowed participants to work at their own pace.

In the main test, set sizes from 2 to 5 sentences were presented; there were four sets of each size (cf. Von der Malsburg & Vasishth, 2013). The sets were presented in random order so that participants could not anticipate how many words they had to remember before they were instructed to recall them. Participants were instructed to perform the acceptability task very accurately and then perform as best as they could on the recall task (cf. Waters & Caplan, 1996).

Working-memory capacity was calculated using the partial-credit unit (PCU) scoring procedure (Conway et al., 2005; Friedman & Miyake, 2005). PCU expresses the mean proportion of words within a set that were recalled correctly and ranges from 0 to 1. Typos were accounted for by allowing for a one-character difference between the target word and the response.

Operation-span test

The design of the operation-span task (OSpan) resembled that of the reading span task, but instead of checking the acceptability of sentences, participants were asked to check the validity of mathematical equations (e.g., (1+3) * 3 = 9). After each equation, a letter was shown for participants to memorize. Participants judged the validity of 56 equations. Set sizes ranged from two to five equations, and there were four sets of each size. The sets were presented in random order. Working-memory capacity on the OSpan was calculated using the partial-credit unit (PCU) scoring procedure: the mean proportion of letters within a set that were recalled correctly.

Linguistic-experience tests

Author Recognition Test

An automated version of the Author Recognition Test was used as a measure of print exposure. Participants were presented with a list of 130 potential authors names; 65 were real author names from the Acheson et al. (2008) version of the test, and 65 were fake (nonauthor) names taken from the Martin-Chang and Gould (2008) adaptation of the ART. Acheson et al. (2008) developed a list that reflected a mix of classic and popular authors at the time of the study (2008). Their participants' average age was 20.3 years. In the current study (data collected in 2018), the participants' average age was 30 years (range, 25–35 years) to ensure the relevance of the material.

The names were presented one at a time in alphabetical order by last name. To ensure that participants would not be able to use search engines to determine whether the name corresponds to an existing author, participants were given 10 seconds to decide whether a name was an author name. Participants were instructed to answer whether they recognized the name as that of a real author. They were instructed not to guess and only select names that they were absolutely certain to be author names since their score would be penalized for falsely identifying nonauthor names as authors.

Print-exposure scores were calculated by subtracting the total number of nonauthor names that were falsely identified from the total number of authors that were correctly identified.

Reading-habits survey

The reading habits survey (adapted from Scales & Rhee, 2001) consisted of seven items that assessed how much participants enjoy reading, the amount of time participants typically spend reading different types of materials, and which types of reading they typically enjoy. Participants were also asked to indicate how many books they had read in the past 12 months. This question was not included in the Scales and Rhee (2001) survey.

We included participants' responses to the items SurveyQ1: "How often do you read?" (5-point scale from *never* to *very often*) and SurveyQ4: "How many books did you read in the past 12 months?" (6-point scale: *1 = 0 books*, 2 = 1–9 books, 3 = 10–19 books, 4 = 20–29 books, 5 = 30–49 books, 6 = *more than 50 books*) as a measure of reading frequency and linguistic experience.

Procedure

The order of the tasks was held constant across participants: they first completed the Author Recognition Test, followed by the operation-span task, the reading-span task, the story-completion task, and finally the reading-habits survey. The tasks were administered on separate days, with the exception of the story-completion task and the reading-habits survey, which were administered on the same day.

The main concern with using crowdsourcing paradigms concerns the reliability of the results—in particular, whether participants are gaming the system (providing random answers or cheating in another way). This is especially likely to occur with tasks that make it easy to provide random answers, such as with the ART task (in comparison to tasks for which participants have to provide written responses). To control for this, we implemented several measures. First, we excluded several participants from participating in further studies based on their results on the ART. A review of previous studies using in-lab ART have shown that typical ART scores are not lower than 0. We therefore excluded participants with a negative ART score (i.e., those who selected more nonauthors than real authors) and participants who showed a random distribution in their selection of fake names versus real author names. Such participants, for example, correctly identified 45 real authors and falsely identified 40 nonauthors. Their score could end up a positive number (i.e., 5), but the underlying distribution is not typical of that found in ART studies. Originally, data for 180 people were collected with the ART task. After exclusion, a second round of data was collected to replace participants. In total, 316 participants took part in the ART. Of these, 102 were excluded from continuing with the other tests based on their ART scores.[4] After exclusion, the scores mirror those of other studies, as reported in the next section.

We also included timers in the ART and the working-memory test to ensure that participants would not be able to copy, write down, or look up any information. Finally, we disabled the text to ensure that participants would not be able to copy and paste the words or letters into a search engine or another document.

The Author Recognition Test took 3 minutes on average and participants were paid 0.45 GBP upon completion. Average completion time for the operation-span task and the reading-span task was approximately 13 minutes, and participants were paid 1.5 GBP per task. Average completion time for the story-completion task was 20 minutes, and participants were paid 2.50 GBP. Finally, average completion time for the reading-span task was 1 minute, and participants were paid 0.20 GBP. After having completed all tasks, participants were paid a bonus of 2.50 GBP each. The total reward for taking part in the studies therefore added up to 8.65 GBP.

Results

First, we provide descriptive data regarding the individual characteristics of the participants. Next, we present the results regarding readers' sensitivity to the contextual list signal, followed by the effect of individual reader characteristics on this sensitivity.

Scores on individual difference measures

The means and standard deviations for each individual difference measure appear in Table 1.

The RSpan capacity of our group of participants ranged between .12 and .94, with a mean of .75. This range compares relatively well to the range of working memory reported in Friedman and Miyake (2005), although it is more extreme: the working-memory capacity of students participating in their study ranged from .46 to .9, with a mean of .68. The increased variability in the range of PCU can be explained by the difference in educational level (no formal qualification up to a doctorate) and possibly also age (mean age 30 years) in our population compared to that of Friedman and Miyake (undergraduate students, mean age not reported but likely to fall around 20 years for college-age students). The OSpan score ranged from .31 to 1 in the current study, with a mean of .94. This mean is higher than that of the RSpan, but results from James et al. (2018) indicate that this is not unusual. We therefore assume that the distribution of working memory found in this study is relatively typical of participants with our demographic characteristics.

Regarding the ART, the scores for participants in the current study ranged from 2 to 56, with a mean of 19.64 and a proportion correct of .30. These scores compare well to that of other studies (e.g., Acheson et al., 2008; Farmer et al., 2017; Moore & Gordon, 2015). The scores for the SurveyQ1 (reading frequency) of our participants also are in line with previous studies (Freed et al., 2017; Scales & Rhee, 2001). The mean score in the current study is 3.92, which falls in the category of reading “sometimes (2–3 times per week).” Finally, regarding SurveyQ4 (number of books read in the past 12 months), the results show that participants on average had read from one to nine books in the past year. The Pew Research Center reports that in 2017, Americans read an average of 12 books, with the typical (median) American reading four books per year (Perrin, 2018). These numbers compare well to the self-reported numbers of participants in the current study.

Table 2 presents the correlations among these measures, showing that all measures correlate significantly with each other.

The reading-span task correlated significantly with the operation-span task, demonstrating that participants with a greater verbal working memory also have a greater nonverbal working memory. This is consistent with previously reported relationships between these two types of working memory (e.g., James et al., 2018; Kane et al., 2004). The ART correlates significantly with the other two measures of linguistic exposure (reading frequency and the number of books read in the past 12 months), indicating that participants who knew more author names also had more reading experience. This is also consistent with previous research (e.g., Freed et al., 2017). Finally, the working-memory measures and linguistic-exposure measures all correlated significantly with each other. This

Table 1. Descriptive Statistics for Individual Difference Measures

Variable	Possible range	Observed range	Mean	*SD*
RSpan	0 to 1	.12 to .94	.75	.15
OSpan	0 to 1	.31 to 1	.94	.08
ART	0 to 65	2 to 56	19.64	10.88
SurveyQ1: reading frequency	1 to 5	1 to 5	3.92	1.10
SurveyQ4: no. of books read in 12 months	1 to 6	1 to 6	2.90	1.31

Note. RSpan = reading span, OSpan = operation span, ART = Author Recognition Test.

Table 2. Correlations Among Scores of Each Individual Difference Measure

Variable	RSpan	OSpan	ART	SurveyQ1	SurveyQ4
RSpan	–				
OSpan	.45*	–			
ART	.22*	.10*	–		
SurveyQ1	.23*	.14*	.23*	–	
SurveyQ4	.23*	.07*	.30*	.39*	–

*Correlation significant at the .01 level.

indicates that, in our sample, participants with a greater verbal working memory also had more linguistic experience. Previous studies have revealed similar correlations between a combination of these studies—for example, between the ART and verbal working memory (e.g., Farmer et al., 2017; Payne et al., 2012) and between the reading survey and the verbal working memory (Freed et al., 2017).

Sensitivity to contextual list signal

Next, we turn to the results of the story-completion task to determine whether the participants were sensitive to the contextual list signal that was present in the list condition. Table 3 presents the percentage of continuation type per condition.

Items in the control condition received mainly causal continuations (69%). In 1% of continuations in the control condition, participants provided a list continuation without being prompted to do so. Example (12) illustrates such an occurrence: The participant listed another high grade, without there being a list signal present that prompted several high grades.

Items in the list condition received a higher proportion of list continuations (35%), but the majority of continuations that participants provided were causal (46%). In such instances, the continuation focused on the cause or result of the event, without further reference to the list that was evoked, as in Example (13). Note that participants were allowed to write two sentences and could therefore have added another sentence with a list item.

(12) *Prompt*: The student came home happy from school. She got an A for a geography exam. *Continuation*: She also got an A for a math exam.

(13) *Prompt*: The student received multiple good grades at school. She got an A for a geography exam. *Continuation*: Her parents took her out to a fancy dinner to celebrate.

Of all list continuations, 73% were marked with a connective or cue phrase. The most common connective used was *also* (53% of list connectives), followed by *then* (24%). Other connectives or signals used were *and, another, later, after/afterward, as well, next*, and *too*. Many of these reflect the temporal nature of the stories (e.g., *He apologized to his mother. Next/Later/Afterward, he apologized to his brother.*).

Items in both the list condition and the control condition received a relatively high proportion of "other" continuations (19% and 30%, respectively). Most of these other continuations provided background to the event in the prompt (Example (14)), a continuation to the story (Example (15)), or a contrast to the event in the prompt (Example (16)). Given the variety of responses that fell under "other" and the free nature of the task, the proportion of "other" responses is not considered unexpected.

(14) *Prompt*: The child went to swim practice this morning. He got kicked underwater by one of the older kids. *Continuation*: He had been getting bullied a lot recently.

(15) *Prompt*: The congresswoman worked in Washington, DC, last week. She made a fool of herself during a debate. *Continuation*: The news played the sound bite over and over of her gaff.

(16) *Prompt*: The receptionist got told off by several people. She got scolded by a customer. *Continuation*: Her boss saw she looked upset, and gave her a bonus.

To statistically measure the difference in distributions between the continuations in the list and the control conditions, we conducted a logistic mixed-model analysis using the statistical software R (R

Table 3. Average Percentage of Continuation Type per Condition

Condition	% List	% Causal	% Other	Raw total number
List	35	46	19	1,630
Control	1	69	30	1,630

Development Core Team, 2008; *lme*4 package, Bates & Sarkar, 2007). A GLMER model with a binary variable for continuation (list versus nonlist continuations) confirmed the significant effect of condition on the continuation (β = 6.31; *SE* = .87; *z* = 7.24; *p* < .001; random slope of participant under condition was removed for convergence): More list continuations were provided in the list condition than in the control condition.

These results indicate that the contextual list signal led participants to provide continuations that function as elements of a list relation. However, and as expected, the list signal was not strong enough to elicit 100% list continuations. This could mean that some items invoked a stronger list response (i.e., there was variability in the strength of the items) or that not all participants were sensitive to the list signal (i.e., there was variability in the sensitivity to the signal) or both.

Looking at the spread of the items in the list condition, the results show that the items received between 9% and 65% list continuations, with a mean of 33% (mean raw number of list continuations = 26.65; *SD* = 13.69). Three items received less than 20% list continuations. This indicates that, although there is a difference in the effect of items on participants' continuations, some variability likely occurs from individual differences between participants. Looking at the effect of specific signal types (*a couple, a few, multiple*, and *several*) displayed in Table 4, there seems to be some variation between the proportion of list continuations per signal: *A few* elicited most list continuations, whereas *multiple* elicited fewest list continuations. *A few* therefore seems to be the strongest expression of quantity. However, these results need to be interpreted with caution, as the study was not designed for such an analysis. This will be addressed in the discussion.

Looking at the proportion of list continuations provided per participant, the results indicate that there was variability in participants' sensitivity to the list signal. The frequency of list continuations ranged between 0% and 100%. The mean percentage of list continuations provided was 35% (mean raw number of list continuations = 3.38; *SD* = 2.73). In total, 29 out of 163 participants did not provide any list continuations. The results presented here thus suggest that participants differed in their sensitivity to the list signal.

Individual variability in sensitivity to contextual list signal

Next, we explored the relationship between the sensitivity to the contextual list signal effect and each of the individual difference variables, using a GLMER model and backward selection. The binary continuation relation types (list or nonlist) for the list condition continuations were regressed onto the main effects of trial ID (order in which the items appeared), RSpan, OSpan, ART, reading frequency, and number of books read in the past 12 months. The participants' educational background (higher education versus high school or no formal qualification) was also included in the full model. Each individual difference variable was centered and the educational background was contrast coded.

The full model contains all individual difference variables; a final model was selected using backward selection. After selecting a final model, random slopes for trial ID under participant and item were added and for the significant individual difference measure under item. The results of the full model and the final model[5] are summarized in Table 5. In the full model, there is an effect of trial ID, which suggests that participants became more sensitive to the list signal during the course of the experiment (most likely due to repeated exposure). This effect is no longer significant in the final model, after random slopes for trial ID under participant and item are added.

Table 4. Average of List Continuations per Signal Type

Signal	% List
A couple	34
A few	47
Multiple	21
Several	35

Table 5. Regression Coefficients and Test Statistics From the Generalized Linear Mixed-Effects Model Including All Individual Difference Variables

Variable	Full model			Final model		
	β	*SE*	*z*	β	*SE*	*z*
(Intercept)	−1.16	0.33	−3.**52**	−1.19	0.35	−3.**41**
Trial ID	0.27	0.07	**3.68**	0.14	0.09	1.50
RSpan	0.05	0.18	0.28			
OSpan	−0.05	0.18	−0.31			
ART	0.73	0.17	**4.35**	0.83	0.15	**5.49**
Reading frequency	−0.03	0.17	−0.17			
Number of books read	0.13	0.17	0.76			
Educational background	0.04	0.32	0.12			

Note. The outcome variable is the continuation type; β represents the odds. for numbers printed in bold *p* <.01.

Regarding the individual difference measures, both the full model and the final model reveal a main effect of the Author Recognition Test: participants with a higher ART score also provide more list continuations. This significant effect of ART is visualized in Figure 1. None of the other individual difference measures nor the educational background were found to be significant predictors of continuation type.

Discussion and conclusion

In building a coherent discourse representation, comprehenders can make use of a variety of linguistic cues. A wealth of processing research over the last 30 years has highlighted the key role of coherence relational signals such as cue phrases and connectives. However, much less is known about the impact of other discourse signals such contextual cues, even though researchers in discourse studies and text linguistics have stressed the importance of such cues on discourse interpretation. The current study presented a first exploration of whether comprehenders are sensitive to global cues in the context. A story-continuation study revealed that participants were indeed sensitive to a contextual list signal:

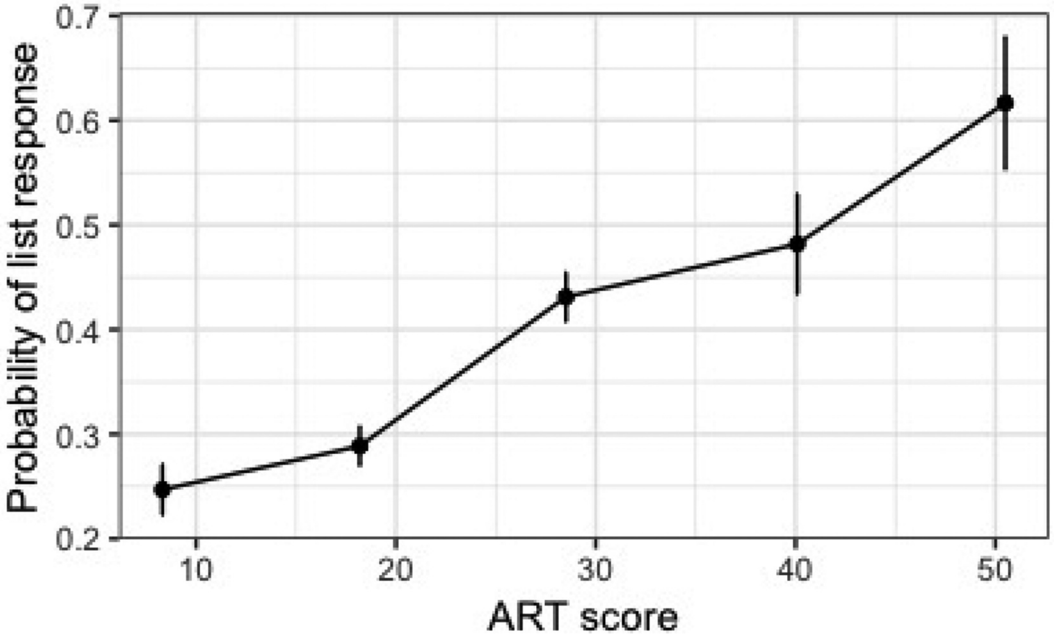

Figure 1. Probability (and standard error) of a list response as a function of the ART scores sorted into five bins.

Quantity expressions such as *a few* and *several* influenced coherence relation expectations by leading participants to provide continuations that function as elements of a list relation.

In cases in which the continuation did not contain a list item, it was assumed that the participant did not generate the list inference. However, it is possible that participants did generate a list inference, but they also generated another, stronger inference for a different type of relation. To control for this, participants were allowed to provide one or two sentences as their continuation. This meant they would be able to provide both a causal continuation and a list continuation, if they felt prompted to do so. The results showed that participants provided only one sentence in most continuations, and they rarely provided both a list and a different type of continuation in one. Future studies could further investigate the multitude of inferences by forcing participants to write more.

The results also suggested that there was variability in the effect of different signals: Items containing the signal *a few* received 47% list continuations; whereas, items containing *multiple* received only 21% list continuations. It should be noted here that the current study was not designed to look into the different effects of such signals, but these results suggest it might be interesting to look more closely at whether participants respond differently to various quantifying expressions. Such a study would include type of list signal as a condition and examine participants' responses to prompts varying only in type of list signal (rather than absence or presence of any signal).

Finally, we note that the context sentences of the items between conditions were not identical. It would not have been possible to simply remove the quantity expression from the list condition to create the control condition, since this would render a list continuation incoherent. Modifying, for example, *The woman experienced several unfortunate events last night* to become *The woman experienced an unfortunate event last night* would mean that we would not be able to measure the proportion of list continuations without a list signal prompt. Instead, the context sentence described the place or situation the referent was in. However, in the majority of the items, the context sentence ended up being slightly longer in the list condition than in the control condition (an average of 9.7 words versus 9.3 words). We believe it is unlikely that this small difference affected the outcome of the study, but especially if a similar study were conducted as an online experiment, care should be taken to minimize the difference in sentence length and word use.

We addressed a second, largely neglected factor in psycholinguistic studies of discourse coherence: individual differences. Research in the field of discourse coherence often has not considered the effect of individual variability in processing capabilities. However, discourse can be difficult to process, especially when looking at global, higher-level structures, and it can therefore be expected that large differences can occur between comprehenders with higher and lower levels of literacy. The current study investigated whether sensitivity to contextual list signals was modulated by individual reader characteristics. The results showed that participants indeed differed in how sensitive they were to the list signals presented in the context.

Several individual difference measures that could explain this sensitivity were collected. The results revealed an effect of the Author Recognition Test (ART) scores, which represent a measure of print exposure and, thereby, linguistic experience. Participants with a higher score on the ART (and thus more linguistic experience) also provided more list continuations.

Reading frequency and number of books read in the previous 12 months, although correlated with ART and representing linguistic experience, did not have a significant effect on the type of continuation that was provided for items. A possible explanation for this discrepancy is the nature of the test: The survey requires participants to make self-judgments; whereas, the ART is more objective in nature. The lack of effect of reading frequency and number of books read might therefore be due to those results being less precise.

An alternative interpretation is that the ART scores might be a better indicator of literacy than the reading questionnaire: Participants who score high on ART might be the ones who have sustained reading habits for a longer period during their lifespan. Additionally, the type of books that participants read might also be a factor—reading complex literary language may train discourse predictions better than reading easy entertainment literature. In future work, a closer assessment of the complexity of texts that participants typically read—and their long-term reading habits—might therefore be an interesting avenue for a more detailed exploration of factors affecting discourse expectations.

The results furthermore did not show evidence that verbal and nonverbal working memory affected the continuation type that participants provided. However, both working-memory measures correlated with the ART. The lack of effect of working memory may, in fact, be related to the ART: The added speed component in the online ART version may have accounted for some of the correlation between WM and ART. Consequently, it may have contributed to WM not being statistically significant in the final model.

Alternatively, it is possible that working-memory capacity would have a greater effect on online interpretations, compared to the offline interpretations that were collected in the current study—the reason for this being that participants were not constrained in time during the story-continuation task and therefore had the opportunity to process the input for longer than they would for online tasks.

The current study extends prior signaling research by showing that contextual signals influence participants' coherence-relation generation. The influence of context on the anticipation and interpretation of coherence relations deserves more consideration. Future research could further investigate what types of relational cues occur outside of the relation and whether these contextual signals influence comprehension in both offline and online processing. One interesting avenue to pursue would be whether comprehenders need less time to process a list relation in the list condition than in the control condition.

Further, the results of the current study emphasize the importance of individual reader characteristics to coherence-relation inferences. The possibility of systematic individual differences has been largely overlooked in the field of coherence. Traditionally, discourse studies use college students as participants, who can be considered to be experienced readers. Individual variation is often considered to be "error variance" in experiments (see also Kidd et al., 2018). However, there is an increasing body of evidence indicating that comprehenders display significant variation in other areas of language comprehension. The results from the current study extend these findings to the coherence-relation field: Comprehenders display individual variability in expectations they generate regarding upcoming coherence relations, and this variability is related to their linguistic experience.

Given the results of the current study, prior studies may have overgeneralized the beneficial effect of signals. In other words, comprehenders from the general population might be less sensitive to discourse signals than they have been assumed to be. Some variance in the effect of relational signals is already well known: There is an interaction effect of world knowledge and relational signals, such that readers who have little domain knowledge of a text topic under discussion are known to benefit more from connectives and cue phrases than experts in the domain (Kamalski et al., 2008; McNamara et al., 1996). Such findings indeed seem to substantiate the idea of variance with readers' characteristics, rather than the generalization of a beneficial effect of connectives and cue phrases. Still, the conclusion that experienced readers are more sensitive to relational cues than inexperienced readers is challenged by studies among young and relatively inexperienced readers in primary and secondary school, who benefit from the presence of connectives during processing and in comprehension (Cain & Nash, 2011; Van Silfhout et al., 2014). Such results, together with the findings of the current study, underline that individual variability should be taken into account systematically to develop cognitively plausible discourse theories. The possibility of individual differences in coherence-relation expectations and in interpretation and processing we hope will receive more attention in future research efforts.

Notes

1. This excludes *no, none*, and *all*, which can be considered to be expressions of quantity but do not raise the same list expectations.
2. See www.prolific.ac.
3. All materials and instructions used in this study are available in an online repository at https://tinyurl.com/y83panj5.

4. An additional 51 participants "dropped out" at some point during the study because they did not participate in a specific study within the time frame that was provided (they were given a week to participate, before being invited to the next study).
5. The final model: Y˜ trialid + ART + (1+ trialid|participant) + (1+ trialid|item) + (1+ ART|item)). *Bobyqa* was used as optimizer.

Acknowledgments

This research was funded by the German Research Foundation (DFG) as part of SFB 1102 "Information Density and Linguistic Encoding."

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Acheson, D. J., Wells, J. B., & MacDonald, M. C. (2008). New and updated tests of print exposure and reading abilities in college students. *Behavior Research Methods*, *40*(1), 278–289. https://doi.org/10.3758/BRM.40.1.278

Afflerbach, P. (2015). *Handbook of individual differences in reading: Reader, text, and context*. Routledge.

Andersson, M., & Spenader, J. (2014). Result and Purpose relations with and without 'so'. *Lingua*, *148*, 1–27. https://doi.org/10.1016/j.lingua.2014.05.001

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation*, *8*, 47–89. http://dx.doi.org/10.1016/S0079-7421(08)60452-1

Bates, D., & Sarkar, D. (2007). The lme4 package. *R Package Version 2*. https://cran.r-project.org/web/packages/lme4/lme4.pdf

Cain, K., & Nash, H. M. (2011). The influence of connectives on young readers' processing and comprehension of text. *Journal of Educational Psychology*, *103*(2), 429–441. https://doi.org/10.1037/a0022824

Canestrelli, A. R., Mak, W. M., & Sanders, T. J. M. (2013). Causal connectives in discourse processing: How differences in subjectivity are reflected in eye movements. *Language and Cognitive Processes*, *28*(9), 1394–1413. https://doi.org/10.1080/01690965.2012.685885

Canestrelli, A. R., Mak, W. M., & Sanders, T. J. M. (2016). The influence of genre on the processing of objective and subjective causal relations: Evidence from eye-tracking. In N. Stukker, W. Spooren & G. Steen (Eds.), *Genre in language, discourse and cognition* (pp. 51–73). Walter de Gruyter.

Caplan, D., & Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences*, *22*(1), 77–94. https://doi.org/10.1017/S0140525X99001788

Cipielewski, J., & Stanovich, K. E. (1992). Predicting growth in reading ability from children's exposure to print. *Journal of Experimental Child Psychology*, *54*(1), 74–89. https://doi.org/10.1016/0022-0965(92)90018-2

Conway, A. R. A., Kane, M. J., Bunting, M., Hambrick, D., Wilhelm, O., & Engle, R. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786. https://doi.org/10.3758/BF03196772

Cornish, F. (2009). "Text" and "discourse" as context. *Working Papers in Functional Discourse Grammar (WP-FDG-82): The London Papers I, 2009* (pp. 97–115).

Das, D., & Taboada, M. (2018). RST signalling corpus: A corpus of signals of coherence relations. *Language Resources and Evaluation*, *52*(1), 149–184. https://doi.org/10.1007/s10579-017-9383-x

Drenhaus, H., Demberg, V., Köhne, J., & Delogu, F. (2014). Incremental and predictive discourse markers: ERP studies on German and English. In *Proceedings of the 36th annual conference of the cognitive science society (CogSci)* (pp. 403–408). Québec City, Canada.

Ehrlich, K. (1980). Comprehension of pronouns. *The Quarterly Journal of Experimental Psychology*, *32*(2), 247–255. https://doi.org/10.1080/14640748008401161

Elman, J. L., Kehler, A., & Rohde, H. (2006). Event structure and discourse coherence biases in pronoun interpretation. In *Proceedings of the annual meeting of the cognitive science society (CogSci)* (pp. 697–702). Vancouver, Canada.

Farmer, T. A., Fine, A. B., Misyak, J. B., & Christiansen, M. H. (2017). Reading span task performance, linguistic experience, and the processing of unexpected syntactic events. *The Quarterly Journal of Experimental Psychology*, *70*(3), 413–433. https://doi.org/10.1080/17470218.2015.1131310

Freed, E. M., Hamilton, S. T., & Long, D. L. (2017). Comprehension in proficient readers: The nature of individual variation. *Journal of Memory and Language*, *97*, 135–153. https://doi.org/10.1016/j.jml.2017.07.008

Friedman, N. P., & Miyake, A. (2005). Comparison of four scoring methods for the reading span test. *Behavior Research Methods*, *37*(4), 581–590. https://doi.org/10.3758/BF03192728

Fuchs, S., Pape, D., Petrone, C., & Perrier, P. (2015). *Individual differences in speech production and perception* (Vol. 3). Peter Lang Publishing Group.

Hicks, K. L., Foster, J. L., & Engle, R. W. (2016). Measuring working memory capacity on the web with the online working memory lab (the OWL). *Journal of Applied Research in Memory and Cognition*, *5*(4), 478–489. https://doi.org/10.1016/j.jarmac.2016.07.010

Hoek, J. (2018). *Making sense of discourse: On discourse segmentation and the linguistic marking of coherence relations* (Volume 509) [Doctoral dissertation, Utrecht University]. LOT Dissertation Series.

Hoek, J., Zufferey, S., Evers-Vermeul, J., & Sanders, T. J. (2018). The linguistic marking of coherence relations: Interactions between connectives and segment-internal elements. *Pragmatics & Cognition*, *25*(2), 276–309. https://doi.org/10.1075/pc.18016.hoe

James, A. N., Fraundorf, S. H., Lee, E.-K., & Watson, D. G. (2018). Individual differences in syntactic processing: Is there evidence for reader-text interactions? *Journal of Memory and Language*, *102*, 155–181. https://doi.org/10.1016/j.jml.2018.05.006

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*(1), 122–149. https://doi.org/10.1037/0033-295X.99.1.122

Kamalski, J., Sanders, T. J. M., & Lentz, L. (2008). Coherence marking, prior knowledge, and comprehension of informative and persuasive texts: Sorting things out. *Discourse Processes*, *45*(4–5), 323–345. https://doi.org/10.1080/01638530802145486

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*(2), 189–217. https://doi.org/10.1037/0096-3445.133.2.189

Kehler, A., Kertz, L., Rohde, H., & Elman, J. L. (2008). Coherence and coreference revisited. *Journal of Semantics*, *25*(1), 1–44. https://doi.org/10.1093/jos/ffm018

Kehler, A., & Rohde, H. (2017). Evaluating an expectation-driven question-under-discussion model of discourse interpretation. *Discourse Processes*, *54*(3), 219–238. https://doi.org/10.1080/0163853X.2016.1169069

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in Cognitive Sciences*, *22*(2), 154–169. https://doi.org/10.1016/j.tics.2017.11.006

Köhne, J., & Demberg, V. (2013). The time-course of processing discourse connectives. In *Proceedings of the 35th annual meeting of the cognitive science society (CogSci)* (pp. 2760–2765). Berlin, Germany.

Koornneef, A. W., & Van Berkum, J. J. (2006). On the use of verb-based implicit causality in sentence comprehension: Evidence from self-paced reading and eye tracking. *Journal of Memory and Language*, *54*(4), 445–465. https://doi.org/10.1016/j.jml.2005.12.003

Lascarides, A., Asher, N., & Oberlander, J. (1992). Inferring discourse relations in context. In *Proceedings of the 30th annual meeting on association for computational linguistics (ACL)* (pp. 1–8). Delaware, US.

Mak, W. M., & Sanders, T. J. M. (2013). The role of causality in discourse processing: Effects of expectation and coherence relations. *Language and Cognitive Processes*, *28*(9), 1414–1437. https://doi.org/10.1080/01690965.2012.708423

Martin-Chang, S. L., & Gould, O. N. (2008). Revisiting print exposure: Exploring differential links to vocabulary, comprehension and reading rate. *Journal of Research in Reading*, *31*(3), 273–284. https://doi.org/10.1111/j.1467-9817.2008.00371.x

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14* (1), 1–43. https://doi.org/10.1207/s1532690xci1401_1

Moore, M., & Gordon, P. C. (2015). Reading ability and print exposure: Item response theory analysis of the author recognition test. *Behavior Research Methods*, *47*(4), 1095–1109. https://doi.org/10.3758/s13428-014-0534-3

Moxey, L. M., & Sanford, A. J. (2000). Communicating quantities: A review of psycholinguistic evidence of how expressions determine perspectives. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *14*(3), 237–255. https://doi.org/10.1002/(SICI)1099-0720(200005/06)14:3<237::AID-ACP641>3.0.CO;2-R

Park, J., & Cardie, C. (2012). Improving implicit discourse relation recognition through feature set optimization. In *Proceedings of the 13th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)* (pp. 108–112). Seoul, South Korea.

Paterson, K. B., Filik, R., & Moxey, L. M. (2009). Quantifiers and discourse processing. *Language and Linguistics Compass*, *3*(6), 1390–1402. https://doi.org/10.1111/j.1749-818X.2009.00166.x

Payne, B. R., Gao, X., Noh, S. R., Anderson, C. J., & Stine-Morrow, E. A. (2012). The effects of print exposure on sentence processing and memory in older adults: Evidence for efficiency and reserve. *Aging, Neuropsychology, and Cognition*, *19*(1–2), 122–149. https://doi.org/10.1080/13825585.2011.628376

Perrin, A. (2018, September 25). *Nearly one-in-five Americans now listen to audiobooks*. Retrieved ofrm https://www.pewresearch.org/fact-tank/2018/03/08/nearly-one-in-five-americans-now-listen-to-audiobooks/

Pitler, E., Louis, A., & Nenkova, A. (2009). Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (pp. 683–691). Suntec, Singapore.

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A. K., Robaldo, L., & Webber, B. (2007). The Penn discourse Treebank 2.0 annotation manual [Computer software manual].

R Development Core Team. (2008). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. http://www.R-project.org

Rohde, H., & Horton, W. S. (2014). Anticipatory looks reveal expectations about discourse relations. *Cognition*, *133*(3), 667–691. https://doi.org/10.1016/j.cognition.2014.08.012

Sanders, T. J. M. (1997). Semantic and pragmatic sources of coherence: On the categorization of coherence relations in context. *Discourse Processes*, *24*(1), 119–147. https://doi.org/10.1080/01638539709545009

Sanders, T. J. M., & Noordman, L. G. M. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse Processes*, *29*(1), 37–60. https://doi.org/10.1207/S15326950dp2901_3

Sanders, T. J. M., Spooren, W. P. M. S., & Noordman, L. G. M. (1992). Toward a taxonomy of coherence relations. *Discourse Processes*, *15*(1), 1–35. https://doi.org/10.1080/01638539209544800

Scales, A. M., & Rhee, O. (2001). Adult reading habits and patterns. *Reading Psychology*, *22*(3), 175–203. https://doi.org/10.1080/027027101753170610

Scholman, M. C. J., & Demberg, V. (2017). Crowdsourcing discourse interpretations: On the influence of context and the reliability of a connective insertion task. In *Proceedings of the 11th linguistic annotation workshop (LAW)* (pp. 24–33). Valencia, Spain.

Scholman, M. C. J., Rohde, H., & Demberg, V. (2017). "On the one hand" as a cue to anticipate upcoming discourse structure. *Journal of Memory and Language*, *97*, 47–60. https://doi.org/10.1016/j.jml.2017.07.010

Scribner, S. (1981). *The psychology of literacy*. Harvard University Press.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on empirical methods in natural language processing (EMNLP)* (pp. 254–263). Honolulu, Hawaii.

Song, L. (2010). The role of context in discourse analysis. *Journal of Language Teaching and Research*, *1*(6), 876–879. https://doi.org/10.4304/jltr.1.6.876-879

Spooren, W. P. M. S., & Degand, L. (2010). Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory*, *6*(2), 241–266. https://doi.org/10.1515/cllt.2010.009

Stanovich, K. E., & Cunningham, A. E. (1993). Where does knowledge come from? Specific associations between print exposure and information acquisition. *Journal of Educational Psychology*, *85*(2), 211. https://doi.org/10.1037/0022-0663.85.2.211

Stevenson, R., Knott, A., Oberlander, J., & McDonald, S. (2000). Interpreting pronouns and connectives: Interactions among focusing, thematic roles and coherence relations. *Language and Cognitive Processes*, *15*(3), 225–262. https://doi.org/10.1080/016909600386048

Traxler, M. J., Long, D. L., Tooley, K. M., Johns, C. L., Zirnstein, M., & Jonathan, E. (2012). Individual differences in eye-movements during reading: Working memory and speed-of-processing effects. *Journal of Eye Movement Research*, *5*(1), 1–16. https://doi.org/10.16910/jemr.5.1.5

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, *28*(2), 127–154. https://doi.org/10.1016/0749-596X(89)90040-5

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*(3), 498–505. https://doi.org/10.3758/BF03192720

Upadhyay, S. S. N., Houghton, K. J., & Klin, C. M. (2019). Is "few" always less than expected?: The influence of story context on readers' interpretation of natural language quantifiers. *Discourse Processes*, *56*(8), 708–727. https://doi.org/10.1080/0163853X.2018.1557006

Van Dyke, J. A., Johns, C. L., & Kukona, A. (2014). Low working memory capacity is only spuriously related to poor reading comprehension. *Cognition*, *131*(3), 373–403. https://doi.org/10.1016/j.cognition.2014.01.007

van Silfhout, G., Evers-Vermeul, J., Mak, W. M., & Sanders, T. J. (2014). Connectives and layout as processing signals: How textual features affect students' processing and text representation. *Journal of Educational Psychology*, *106*(4), 1036–1048. https://doi.org/10.1037/a0036293

von der Malsburg, T., & Vasishth, S. (2013). Scanpaths reveal syntactic underspecification and reanalysis strategies. *Language and Cognitive Processes*, *28*(10), 1545–1578. https://doi.org/10.1080/01690965.2012.728232

Waters, G. S., & Caplan, D. (1996). The measurement of verbal working memory capacity and its relation to reading comprehension. *The Quarterly Journal of Experimental Psychology Section A*, *49*(1), 51–79. https://doi.org/10.1080/713755607

Waters, G. S., Caplan, D., & Hildebrandt, N. (1987). Working memory and written sentence comprehension. In M. Coltheart (Ed.), *Attention and performance xii: The psychology of reading* (pp. 531–555). Lawrence Erlbaum Associates, Inc.

Whitney, P., Ritchie, B. G., & Clark, M. B. (1991). Working-memory capacity and the use of elaborative inferences in text comprehension. *Discourse Processes*, *14*(2), 133–145. https://doi.org/10.1080/01638539109544779

Xiang, M., & Kuperberg, G. (2015). Reversing expectations during discourse comprehension. *Language, Cognition and Neuroscience*, *30*(6), 648–672. https://doi.org/10.1080/23273798.2014.995679