10. Experimental Studies in Discourse

BIOGRAPHICAL NOTES
Ted J. M. Sanders, Jet Hoek & Merel C.J. Scholman

Ted J.M. Sanders is a Full Professor of Discourse Studies at the Department of Languages, Literature and Communication and the Institute of Language Sciences, Faculty of Humanities at Utrecht University. He published widely on discourse processing and representation, and on the (cross-) linguistics of coherence, striving for converging evidence by combining various empirical methods: from corpus annotation to eye-tracking experiments. He also has a strong interest in improving communication through comprehensible texts. He has (co-)supervised 25 Ph.D.-candidates. He is currently Vice-Dean of the Faculty, responsible for research and innovation of the Graduate School of Humanities.


Jet Hoek is an assistant professor at the Department of Language and Communication and the Centre for Language Studies at Radboud University Nijmegen. Her research focuses on discourse coherence, studying both relational and referential coherence from a cognitive perspective using both offline and online methodologies. She is especially interested in how discourse expectations influence processing and shape the linguistic realization of a discourse.

Merel Scholman is an assistant professor at the Department of Languages, Literature and Communication at Utrecht University and a researcher at the Department of Language Science and Technology at Saarland University. She is interested in cognitive models of language understanding, focusing on discourse coherence in particular. Using a combination of offline and online methodologies (corpus-based investigations, crowdsourced studies, and reading time experiments), she investigates questions related to the interpretation and processing of discourse.
Abstract
In this chapter, we discuss classic and recent experimental studies of discourse to give an overview of developments in research focus and methodological advancements within this field. The focus is on the fit of methods to research questions. We discuss research related to major questions from the field of discourse coherence, focusing on both relational and referential coherence: how do language users form a coherent representation of a discourse, how does this evolve during online processing, and how is the discourse structure signalled in the linguistic surface code?

Three types of research paradigms can be distinguished, providing information on different aspects of discourse: processing, representation, and analysis. Methods vary from free recall and comprehension for representation, to methods like categorization and insertions tasks for analysis, to methods that focus on online processing like eye-tracking-while-reading and neurological methods such as ERP. We first discuss these methods in relation to questions in the field, after which we provide an easy-to-consult overview of methods in alphabetical order. Finally, we discuss recommendations for practice and future directions of research within the field of discourse.

Acknowledgements
TS is supported by Utrecht University, JH by Nijmegen University, and MS is supported by the European Research Council (ERC) under the European Union’s Horizon 2020 Research and Innovation Programme, Grant 948878 (‘Individualized Interaction in Discourse’).
1. Introduction
People communicate through discourse. We talk to each other, we read the newspaper or interact on social media, and in all these contexts we use discourse; we produce and process messages that consist of more than one clause. The importance of the discourse level for the study of language and linguistics can hardly be overestimated: “Discourse is what makes us human” (Graesser et al., 1997 p.164).
In this chapter, we provide an overview of some classical and recent experimental studies of discourse and their development in terms of research focus and methodological advancements. The focus will be on the fit of methods with research questions.
In the field of discourse studies, the dominant view is that the connectedness of discourse is a characteristic of the cognitive representation of the discourse rather than of the discourse itself (Sanders & Pander Maat, 2006; and many others). The term coherence is used for this type of connectedness. Language users establish coherence by actively relating the different information units in the text.
Generally speaking, there are two respects in which discourse can cohere:

- **Referential coherence**: smaller linguistic units (often nominal groups) may relate to the same mental referent.
- **Relational coherence**: discourse segments (most often conceived of as clauses) are connected by coherence relations like CAUSE-CONSEQUENCE between them.

These two types are illustrated in example (1).

(1) James fired the guy from the accounting department. He was embezzling money.

This example contains an ambiguous pronoun (“he” could refer to James or the guy from accounting) and there is no explicit connective expressing the coherence relation. However, readers likely construct a coherent representation of this simple discourse by interpreting that “he” refers to the guy from the accounting department (referential) and that the reason for being fired is that he was embezzling (relational).
This illustrates how coherence is of a cognitive nature; the processing and representation of coherence phenomena is often based on linguistic signals in the discourse itself. Both coherence phenomena under consideration - referential and relational coherence - have clear
linguistic indicators that can be taken as processing instructions. For referential coherence these are devices such as pronouns (*he*) and lexical Noun Phrases (*the guy*), and for relational coherence these are connectives (*because*) and other lexical markers of relations, such as cue phrases and signalling phrases (*as a result of; the reason for X is*). A major research issue is the relation between the linguistic surface code and aspects of the discourse representation. In this chapter, we discuss research related to this and other major questions from the field of discourse coherence: how do language users form a coherent representation of a discourse, how does this evolve during online processing, and how is the discourse structure signalled in the linguistic surface code?

2. **Historical Perspectives**

Since the 1950s, experimental psychologists have developed techniques that were also put to use in psycholinguistics. One of the first experimental methods that were used to investigate discourse concerns memory: Participants were invited to read a text and then asked what they could remember. Such free recall tasks provide insight in readers’ representation of the textual information. Pioneering studies by Bartlett (1932) showed how people recalled different pieces of information from a text, depending on their cultural background and their previous knowledge of the topic. In the 1970s and early 1980s, **free recall** was often used to investigate whether theories on the relative importance of textual information were correct. Meyer (1975) and colleagues provided hierarchical representations of discourse, predicting that higher ordered information would be recalled better than less important information: the so-called *levels effect*.

Memory measures like free recall have their shortcomings. First, memory performance does not provide insight into the processes of comprehension; the information it provides is limited to the representation of the information. Second, recall is not a very precise method, because recall protocols need to be analysed and evaluated: which information is or is not recalled? Agreement scores between several judges are necessary and scoring is a very laborious task. Another type of experimental method that has been widely used since the 1950s is one in which language users’ discourse knowledge is used to test or corroborate linguistic analyses. In such tasks, participants are invited to reflect consciously on (pieces of) discourse. People can be asked to continue a story, and researchers will study their referential patterns (Gernsbacher, 1989). Intuitive knowledge on coherence can be investigated when people participate in card sorting tasks (going back to Miller, 1969) and put intuitively similar
coherence relations on the same pile (Sanders et al., 1992). Such techniques were refined in recent years, making use of digital platforms and crowdsourcing, as we will see in section 4.

Halfway through the 1970s, more and more on-line measures were used, which could overcome certain shortcomings of offline memory measures. On-line measures focus on the time course of comprehension. A straightforward method is measuring reading times. This can be done, for example, in a self-paced reading paradigm: participants are invited to read a text on the computer screen and to push a button when they have read a segment, which can consist of individual words or clusters of words. The computer registers the reading time per segment. Differences in reading times are interpreted as an indicator of the cognitive effort that readers need to process the information; the longer readers need, the more cognitive effort they need to process the linguistic items. Larger efforts are often correlated with more complexity in the textual information. Illustrative results include the finding that textual information that is important in terms of the hierarchical text structure is processed slower than when this same information is less important (Cirilo & Foss, 1980). For relational coherence, the finding that causally related information is processed faster than non-causally related sentences is key (Myers et al., 1987).

Since the 1980s, such online measures have been developed further. A sophisticated technique for charting the reading process is measurement of people’s eye fixations (Just & Carpenter, 1980). Infrared cameras provide information on the fixations, regressions, and saccades that readers make while reading. This type of eye-tracking research has provided us with many precise insights in the comprehension process. During fixations, information is processed. During saccades – short rapid movements – vision is suppressed so that information cannot be processed. Regressions are a specific type of saccades; those that jump backward through the text. When researchers compare text versions in an eye-tracking experiment, they focus on fixations and regressions. The longer readers fixate on a segment, the more effort they need to process that information. This can be because it is complex, or because it is important. The frequency of regressions as well as their location provides information on processing. Illuminating patterns in discourse processing found with eye-tracking include regressions showing exactly how readers process referential coherence.

While the development of more precise investigations of the online comprehension processes has been crucial to gaining a deeper understanding of discourse comprehension, gaining this understanding still requires the combination of on- and off-line measures. For instance: information is read slower and recalled better when it is important in the discourse compared
to when it is less important, suggesting that readers spend more time encoding important information, which results in better representation of this information as becomes evident from better recall scores.

So far, we have identified three types of research paradigms, providing information on different aspects of discourse: processing, representation, and analysis. In the remainder of this chapter, we will illustrate the use of these paradigms methods in investigating relational and referential coherence.

3. Critical Issues and Topics

We now turn to a discussion of critical topics that have been studied in the field, especially since the 1990s, highlighting which methods are typically used for these investigations. We focus our discussion on relational and referential discourse research separately and discuss the topics in these fields with respect to three paradigms: processing, representation, and analysis.

3.1. Relational coherence

3.1.1. Processing

The representation that people make of a discourse is conceptual by nature and can but need not be linguistically marked by connectives and cue phrases. These linguistic elements are seen as processing instructions that inform readers on how to connect the incoming discourse segments: for example, the usage of because in a sentence makes clear we are dealing with a consequence-cause relation and not with a contrastive or additive relation. Ever since Haberlandt (1982), there is massive evidence for the so-called integration effect: the information that directly follows the connective or cue phrase is processed faster than the same information without the connective, indicating that connectives facilitate the integration of information (see, e.g., Millis & Just, 1994; Murray, 1997; Sanders & Noordman, 2000; Traxler et al., 1997). Self-paced reading was used in these experiments. In the 1990s, there was still some debate on when exactly this integration effect would take place: immediately or toward the end of the sentence? The use of eye-tracking methods (Rayner, 2012) has allowed researchers to be more precise about this type of effect. Virtually all relevant studies with these methods have shown that the integration effect shows up immediately after the connective, while some have found that toward the end of the sentence there is a slowdown
because of the connective, which has been interpreted as the effort readers make to verify the established relation their world knowledge.

Recent eye-tracking studies have also provided more detailed information on the exact nature of the integration process. Using educational texts, van Silfhout et al. (2015) compared texts with and without connectives and found that the connective caused readers to make more regressions (looking back) from the first region after the connective to previous information. When participants went on and reached the information following the connective again, they read this faster than in the implicit version.

This type of result illustrates the benefits of eye-tracking studies: they provide rather precise information on online processing. This also holds for relatively subtle linguistically coded differences. For instance, experiments have shown that lexical markers play a significant role in interpreting two different types of causality, serving as processing instructions. So-called subjective relations (expressing the speaker’s reasoning) are harder to process than objective causal relations (describing relations in the real world). This has been shown for English (Traxler et al., 1997), Dutch (Canestrelli et al., 2013) and French (Zufferey et al., 2018).

Experiments with a Visual World Paradigm (Wei et al., 2019) have shown that the use of specialized subjective connectives, cf. English therefore indeed allow addressees to make predictions about the upcoming discourse.

Neurological measures such as Event-Related Potentials (ERP) provide a clearer view of the neural processes that underly reading. Two ERP-components are particularly relevant: the N400 and the P600. The N400 reflects the predictability of a word in its sentential (e.g., Kutas & Hillyard, 1984) or discourse (e.g., Federmeier & Kutas, 1999; van Berkum et al., 2005) context. The P600 is associated with semantic/pragmatic violations (e.g., Drenhaus et al., 2011; Kuperberg et al., 2003; van Herten et al., 2005). For example, Köhne-Fuetterer et al. (2021) conducted an ERP-experiment to obtain further insight into how and when readers use connectives during discourse processing. They found a P600 effect indicating that, upon encountering a concessive connective (like although), readers immediately update their mental representation from an expected causal relation to an unexpected concessive relation.

3.1.2. Representation

As we have seen so far, there is substantial evidence that (at least certain types of) connectives and cue phrases speed up the processing of subsequent information. Still, one could question the effect of faster processing on the representation of the information. After all, if a given relation is already made clear by overt marking, the language processor must do
less work, which may result in a sloppier representation. Which methods are available to investigate the representation people have made after they have read, listened to, or participated in a discourse?

Pioneering studies on the effect of text structure in the 1970s often used free recall as a dependent variable (Meyer, 1975). Contrary to expectations, connectives and other coherence markers did not affect the amount of information recalled. It looks like this method is not sensitive enough to tap into the representation of a text and pick up effects of coherence marking. In fact, experiments that have applied the free recall method have failed to report effects of relational marking on comprehension (Meyer, 1975; Sanders & Noordman, 2000). On the other hand, other experimental methods have revealed that overt marking of coherence relations might improve the mental representation of a text. It has been shown to lead to more complete summaries (Hyönä & Lorch, 2004; Lorch, 2001), faster responses on verification tasks (Millis & Just, 1994; Sanders & Noordman, 2000) and an overall higher quality of recalled information (Meyer et al., 1980).

Question answering seems a more suitable test to assess text comprehension than memory measures do. Inference questions, often focused on the relation under investigation, do show how linguistic marking leads to more accurate answers (Degand & Sanders, 2002; McNamara, 2001; van Silfhout et al., 2014, 2015). Another method investigating comprehension, which was often criticized in the 1970s and 1980s, was recently revitalized. In a cloze test, participants are invited to fill in gaps in a text. This method was used by Kleijn et al. (2019), who conducted a comprehension study comparing implicit and explicit (connective) versions. They found that contrastive and causal connectives facilitated comprehension, while additive connectives even reduced comprehension. The conclusion is that effects of connectives on text comprehension may be consistent between readers, but not between types of coherence relations, or types of linguistic cues.

Keyword sorting tasks, specifically designed to tap into the situation model representation (Kintsch, 1998), were often used in studying comprehension and learning from text, too: participants are asked to sort crucial concepts from the text in predefined categories that can either be open or that have received a name (McNamara & Kintsch, 1996; McNamara et al., 1996). Results indicate that readers scored higher on sorting tasks when they had read a text with coherence marking than an implicit version (Land, 2009).

Overall, results do not always point into the direction of a positive effect of coherence markers. A crucial question is: can we generalize over a large variety of coherence markers and coherence relations? The answer is: no, we need to differentiate between various types of
relations, such as additive, causal, and contrastive relations. Crucial in doing so is determining which types of relations language users distinguish. In other words: are the categories that we, as linguists, have come up with actually cognitively plausible?

3.1.3. Analysis: Various types of relations and markers

Which types of relations and connectives can be distinguished? This question is seriously debated and investigated in text linguistics, pragmatics, and discourse studies. Based on classical taxonomies of coherence relations (Halliday & Hasan, 1976; Sanders et al., 1992), we could focus on main categories like additive (2), temporal (3), contrastive (4) and causal relations (5); all relations are illustrated here with their prototypical connectives.

(2) Daan worked on his paper and Willem played his guitar.

(3) Jan cooked supper. Afterwards, Jip did the dishes.

(4) Nala likes to play tennis, but Jip prefers soccer.

(5) Jip started crying because Nala stole his car.

Results from theoretical analyses, corpus studies and experiments have shown that adult language users intuitively distinguish between these relational clusters (see Sanders et al., 2018 for an elaborate discussion) and know which connectives to use to express each type of relation. In addition to corpus work, several experimental methods have appeared useful in arriving at this conclusion, such as classification tasks (Sanders et al., 1992) and relation labeling tasks (e.g., Scholman et al., 2016). Analyses of the ‘mistakes’ that analysts make are also informative; for example, people mix up theoretically related labels much more often than less related labels (Sanders et al., 1992, 1993). The disadvantage of such methods is that language users very consciously focus on the relations and must be trained to master the right meta-level terminology: the use of the right relation labels.

Another method to obtain discourse relation interpretations is to make use of relational markers instead of relation labels, thereby exploiting the natural relation lexicon that people use in everyday language (Knott & Dale, 1994). A typical task falls in the insertion paradigm: given two discourse segments, participants are asked to fill in the best fitting connective from a presented list of alternatives (Sanders et al., 1992). This methodology has been applied recently to obtain discourse-annotated data using untrained, crowdsourced participants (Scholman et al., 2022). Coherence relation categorizations can be derived from the connectives that participants have provided to express two relational segments, thereby giving insight into the participants’ mental representation of the relation without requiring meta-level terminology.
3.2. **Referential coherence**

Even though there are clear parallels between relational and referential coherence, there is a crucial difference, too: While each individual coherence relation occurs only once, a referent can be referred to multiple times throughout a discourse. This is why, for referential coherence, the three ‘levels’ we distinguish (representation, processing, analysis), are more intertwined.

3.2.1. Analysis: Different types of referring expressions

Accessibility appears to be a crucial factor in the use of referring expressions (REs; see Arnold, 2010, for an extended discussion): when a referent is not accessible (salient; activated) in the discourse, a longer linguistic form (e.g., a full NP) tends to be used to refer to that referent; when a referent is highly accessible, they are generally referred to using a shorter form (for instance a personal pronoun or a null anaphor). Common experimental methods aimed at investigating the production of REs include **sentence completion** or **story continuation**, where participants are presented with a prompt and asked to write down or utter the first continuation that comes to their mind. A much more restricted variation of this task is **insertion**, where only the RE is left out of a sentence or story, and participants are asked to fill in the blank or, in case of the **forced-choice** version of this task, to select a RE from a list of options. Studies using these paradigms have identified factors that contribute to a referent’s accessibility, such as:

- **Topicality**: topical referents are more often referred to using a reduced RE than non-topical referents (e.g., Cowles & Ferreira, 2012)
- **Grammatical role**: subjects are more often referred to using a reduced RE than non-subjects (e.g., Arnold, 2001; Fukumura & van Gompel, 2010; Stevenson et al., 1994)
- **Thematic role**: e.g., goals are more often referred to using a reduced RE than sources (Arnold, 2001)
- **Animacy**: animate referents are more often referred to using a reduced RE than inanimate referents (Fukumura & van Gompel, 2011; Vogels et al., 2014)
- **The number of other discourse referents**: higher pronominalization rates in contexts with fewer referents (Arnold & Griffin, 2007).

Other factors have been proposed to influence choice of RE as well. Fukumura et al. (2022), for instance, use a spoken production task where participants must describe the
location of an entity in a visual display to an interlocutor to investigate the effect of ambiguity. While this study, like many other studies, did find that people prefer to produce more unambiguous than ambiguous pronouns, the mechanism behind why this is the case is still a topic of discussion (see Hoek et al., 2021 for a discussion). Similarly, an open question is whether the predictability of referents influences pronominalization rates. Fukumura and van Gompel (2010), for instance, find no effect of predictability on choice of RE using a version of the sentence completion task where participants are explicitly asked, in this case by pointing an arrow at one of the referents, to focus their continuation on a specific referent from the preceding discourse. Rosa and Arnold (2017), on the other hand, adapt the standard story continuation method to include a rich narrative and visual context and do find an effect of predictability. This suggests that the selected methodology may impact the obtained results.

3.2.2. Representation: Interpreting referring expressions and determining a referents’ status

When encountering a RE, language users are faced with resolving it to one of the discourse referents. The RE may provide explicit information that helps limit the number of candidate referents, such as person, number, or gender. This can, but does not always, result in an unambiguous RE. In case of ambiguity, resolving the RE to a referent is guided by other factors.

While it may seem intuitive that REs would be interpreted in the same way as they are produced, there are indications that this may not be the case. Stevenson et al. (1994) were the first to use the sentence completion paradigm with both free prompts and pronoun prompts. They found that the number of continuations that focused on the subject of the previous clause was significantly higher in the free prompt than in the pronoun prompt condition, a result that has been replicated numerous times since. Which factors drive the interpretation of REs is therefore a question distinct from the question which factors drive pronoun production, although insights may very well transfer. Rohde and Kehler (2013), using the same experimental method as Stevenson et al. (1994), argue that meaning-related factors (e.g., thematic role, verb semantics, etc.) drive the interpretation of REs more than the production of REs.

Prompts in sentence completion / story continuation tasks may, in addition to just personal pronouns like in the studies mentioned above, also contain other forms of REs, such as demonstratives (e.g., Bader & Portele, 2019), to test the effect of form of mention. However, where these tasks require the researcher to interpret the target referent of the participant’s RE,
there are other tasks that can tap into pronoun interpretations more directly. Participants can for instance be presented with a short discourse containing an ambiguous pronoun, after which they are explicitly asked who the pronoun refers to. Many studies using this **forced choice paradigm** use a **nonce word** in each item to avoid any information that may bias the interpretation of the pronoun beyond the factors being tested (e.g., Hartshorne & Snedeker, 2013). Research testing the interpretation of REs can also make use of visual displays that may or may not correspond to the situation described in the verbal prompts. Participants can then for instance be asked to indicate whether the picture matches the prompt in a **visual verification task** (e.g., Experiment 1 in Kaiser et al., 2009), or to choose which picture best represents what is described in the discourse (e.g., Experiment 2 in Kaiser et al., 2009). In addition, participants can be instructed to **actively engage with the referents**, in either a real-life or digital setting. The study in Brown-Schmidt et al. (2005), for instance, asked participants to “Put the cup on the saucer. Now put it/that over by the lamp.” They then determine whether the pronoun *it* or *that* is interpreted as referring to just the cup or to the composite of the cup plus saucer. Because the instructions are spoken, the experiment can, in addition to varying the specific RE, assess the role of prosody in pronoun resolution. Because referents can be referred to multiple times throughout a discourse, the status of referents in a discourse is dynamic, with referents’ activation levels fluctuating because of, for instance, recency of mention, topic status, and form of mention. Interpreting an RE thus affects the status of referents.

3.2.3. Processing: While the methods discussed in 3.2.2 measure offline how REs are interpreted, online studies are needed to assess how these processes play out during real time processing. Studies using methods such as a **reaction time paradigm**, **self-paced reading** and **eye-tracking** for instance show that it is less costly to resolve unambiguous REs than ambiguous ones (e.g., MacDonald & MacWhinney, 1990; Stewart et al., 2000), less costly to resolve REs with a recent antecedent than REs referring to a more distant referent (e.g., Ehrlich & Rayner, 1983; Patterson, 2013), and less costly to resolve REs referring to topical referents than to non-topical referents (e.g., Clifton & Ferreira, 1987; Gordon & Scearce, 1995). In addition, online experiments have established that languages users do not always passively wait to process whichever RE they encounter but can also actively anticipate upcoming referents. Implicit causality verbs (e.g., Garvey & Caramazza, 1974) have been shown to yield expectations about which referent will be the focus of the subsequent discourse (i.e., bias). In an **eye-tracking-while-reading** study, Koornneef and van Berkum
(2006) show that bias-consistent pronouns are read faster than bias-inconsistent pronouns, which lead to processing delays in various reading measures indicative of early processing delays. In line with these results, the visual world eye-tracking study by Cozijn et al. (2011) shows that upon hearing an implicit causality verb in the auditory input, participants look toward the bias-consistent referent in the visual display before hearing any disambiguating information.

The online studies that are mentioned above all use experimental items containing REs that, expected or not, have a possible antecedent in the discourse. In contrast, other studies investigate REs that, if co-referent with one of the referents from the preceding discourse, constitute a syntactic and/or semantic violation. By comparing event-related potential (ERP) responses, Nieuwland (2014) shows how brain activity differs when encountering felicitous versus infelicitous pronouns. Also using sentences containing violations, Hammer et al. (2007) conduct an fMRI study in German. Since German nouns carry syntactic gender that need not match the semantic / biological gender of the lexical item (e.g., das Mädchen ‘the girl’ has neuter syntactic gender), manipulating what kind of violation a subsequent, supposedly co-referent, pronoun constitutes, allows for disentangling the role of syntax and semantics in the processing of REs.

4. Current Contributions and Research

Linguistic experiments were traditionally mainly conducted in laboratory settings. More recently, however, technological developments have made crowdsourcing platforms increasingly popular for obtaining data (see Chapter 25). The advent of crowdsourcing has also facilitated research into individual differences in language production and comprehension. In the field of discourse, early work has repeatedly shown that language exposure affects language comprehension at the discourse level. In a story continuation study, Scholman et al. (2020) found that comprehenders’ sensitivity to a context-based cue for list relations (e.g., several, a few) depends on their print exposure. Zufferey and Gygax (2020) found in two connective judgments tasks that people’s ability to understand correct and incorrect usages of connectives depends on their print exposure, as well as their grammatical competence. Finally, results from two insertion tasks conducted by Tskhovrebova et al. (2022) indicate that not only participants’ exposure to print and grammatical competence, but also their academic background can predict connective comprehension. Regarding individual differences in referential processing, Arnold et al. (2018) find that pronoun comprehension can be explained by print exposure, but not working memory or theory of mind. Langlois and
Arnold (2020) find that syntactic biases in pronoun comprehension can also be explained by print exposure, but semantic biases cannot. In a non-crowdsourced experiment aimed at investigating the effect of age on pronoun production, Hendriks et al. (2014) show that that children (aged 4-7) and elderly adults (aged 69-87) produce more ambiguous pronouns than young adults (18-35). These recent studies demonstrate the importance of considering individual differences in discourse interpretation and processing in future work.

Related to differences between comprehenders, we note that studies in the field of discourse processing have also traditionally focused mainly on native speakers. Non-native discourse processing is under researched both with regards to relational and referential coherence. The limited work that is available suggests that native and non-native speakers differ greatly in their connective knowledge and usage (e.g., Leedham & Cai, 2013; Wetzel et al., 2020), which in turn affects their discourse processing skills (Crible et al., 2021; Zufferey and Gygax, 2017). Additionally, non-native speakers do not show the same proactive coreference expectation mechanisms as native speakers during discourse processing (Grüter & Rohde, 2021).

Finally, while language processing has been found to be shaped by language-specific properties, most work investigating discourse processing focuses on a single language, rather than on cross-linguistic comparisons between multiple languages. In referential discourse studies, a significant body of work has focused on a variety of languages and on cross-linguistic comparisons; see De la Fuente et al. (2016) for an overview. The field of relational discourse is largely dominated by a focus on a limited set of languages: English, Dutch, French, and German. Accordingly, the field is at risk of missing properties of discourse that are common in a variety of languages but are less common or even absent in (typologically) different languages. Recent years have witnessed studies on Mandarin Chinese (Li et al., 2017; Wei et al., 2019; Xiao et al., 2021), and Spanish (Santana et al., 2018), to broaden the empirical basis. The body of literature on discourse comprehension and interpretation in languages other than English, as well as cross-linguistic comparisons, is now steadily growing. For example, Yi and Koenig (2021) showed that following implicit causality verbs, Korean speakers are less likely to produce explanation continuations in monologues than English speakers, which can be explained by constraints on clause linkage in Korean. The coreference biases associated with implicit causality verbs, on the other hand, were the same as in English. We see the area of cross-linguistic investigations into discourse production and comprehension as a fruitful topic for future research.
5. Main Research Methods

In this section we list the key methodologies used in discourse research, in alphabetical order. For each method, we provide a short description and, if applicable, list important variants of the method. In addition, we point out some considerations in choosing this method. We have included references to studies that have used the method; for more elaborate examples of how a specific method can be used to further our understanding of discourse, we refer to sections 2-4.

5.1. Methodologies by alphabetical order

5.1.1. Classification and labelling tasks

In a classification task, participants are asked to categorize stimuli into groups based on similarity, for instance coherence relations (e.g., Sanders et al., 1992). In a labelling task, participants are presented with a prompt and asked to choose a description that most fittingly describes the prompt (very similar to forced choice interpretation task, but with a set of labels that is fixed throughout the experiment, and often explained in the instructions (e.g., Scholman et al., 2016).

Considerations:
These meta-linguistic tasks ask participants to engage very actively with the prompts, and, as such, may not yield results that reflect what happens when people encounter language in a natural, no-pressure context.

5.1.2. Comprehension questions

Participants are presented with a prompt and asked to answer comprehension questions targeting their representation or interpretation of the prompt (e.g., McNamara et al., 1996). Comprehension questions can consist of, e.g., elaborative or bridging inference questions (requiring participants to make connections not explicitly stated in the text) or keyword sorting task questions (requiring participants to organize a set of key concepts into categories).

Considerations:
Comprehension questions tap into how well readers have understood the situation described by the text. Note that the effect of coherence marking (both connectives and REs) on the mental representation after reading a text is not unequivocal; for example, some studies have
shown a beneficial effect of coherence markers on text comprehension questions (Degand & Sanders, 2002), while others did not find any effect (Spyridakis & Standal, 1987). These varying results might be due to differences in which level of text comprehension the questions target; manipulating whether a relation is or is not explicitly marked by a connective, for instance, may not affect a reader’s comprehension of the entire discourse (see also Kleijn et al., 2019).

5.1.3. Continuation tasks
Participants are presented with a prompt and asked to provide a natural continuation (e.g., Kehler et al., 2008). This allows researchers to study both the content and the form of the continuation. The ending can consist either of a full stop or an ambiguous element.

Considerations:
This paradigm requires researchers to manually annotate the continuations for the phenomenon under investigation, which allows room for subjective interpretations. To ensure that the annotations are reliable, at least a portion of the data should be double coded by two coders. These annotations can then be used to calculate inter-annotator agreement (see Spooren & Degand, 2010).

5.1.4. Directive and act-out tasks
In a directive study, participants are asked to give directives to another entity (e.g., a puppet or another participant). In an act-out task, participants listen to a prompt and are asked to act out the events in the prompt using tools (e.g., puppets).

Considerations:
Given the playful nature of these paradigms and the fact that they typically do not require advanced verbal or reading skills, these methodologies are suitable for child language acquisition research (e.g., Evers-Vermeul & Sanders, 2011; Papakonstantinou, 2015) but can also be applied in studies using an adult sample (e.g., Fukumura et al., 2011; Tourtouri et al., 2019). Responses need to be recorded and manually coded using a predefined coding procedure, which may include determining inter-coder reliability (see also Continuation tasks).

5.1.5. ERP / EEG
Participants are presented with a prompt while their brain waves are being measured through electrodes.

Considerations:
ERP / EEG studies provide insight into activity in the brain as it unfolds in real time, but they are expensive to run and require advanced equipment, which not every researcher might have access to.

5.1.6. Eye-tracking: visual world paradigm
Participants listen to a prompt and simultaneously view a screen with multiple images, while their eye-movements are being recorded through a special camera that tracks the pupils and measures gaze direction (e.g., Kaiser et al., 2009; Köhne-Fuetherer et al., 2021).

Considerations:
An advantage of this method is that it does not require any specific task from the participant; people will just listen to input and unconsciously their eyes will go to the related picture. One disadvantage of this method is that the visual world paradigm primes participants to generate expectations by introducing referents. Other paradigms in the discourse processing field do not have this drawback.

5.1.7. Eye-tracking-while-reading
Participants read prompts while their eye-movements are being recorded (e.g., Koornneef & van Berkum, 2006; van Silfhout et al., 2015).

Considerations:
This method provides very fine-grained insight into the reading process (see for instance Rayner (1983) for a discussion, and Chapter 17, this volume).

5.1.8. fMRI
Participants’ local changes in cerebral blood oxygenation are measured (e.g., Nieuwland & van Berkum, 2006; Xiang & Kuperberg, 2015).

Considerations:
fMRI studies provide insight into activity in the brain as it unfolds in real time, but fMRI studies are expensive to run and require advanced equipment, which not every researcher might have access to.

5.1.9. Forced choice interpretation task
Participants are presented with a prompt and asked to choose from multiple visual or verbal representations (e.g., images or text options) which option presents the intended meaning of the prompt or the intended referent. In addition to choice selection as dependent variable, response times may be considered.

Considerations:
The advantage of a forced choice design is that it forces participants to be explicit about their interpretation of the phenomenon under investigation. However, the choice options should be carefully designed: for instance, one option should not be more salient than the other because of factors unrelated to the research question at hand.

5.1.10. Image description
Participants are presented with one or more images and asked to describe the image in their own words (e.g., Geelhand et al., 2020).

Considerations:
This methodology allows researchers to study natural language production in a controlled context. However, responses need to be manually coded for the phenomena of interest, which requires a detailed coding procedure and an analysis of inter-annotator agreement (see also Continuation tasks).

5.1.11. Insertion
Participants are presented with a prompt and asked to insert a linguistic element, such as a connective or RE (sometimes also referred to as a cloze task). The choice can be free or forced.

Considerations:
Insertion methodologies provide insight into how well participants understand the meaning and usage of specific words (e.g., do they understand which connective should be used in a
specific context?), or it can give insight into their representation of the text (e.g., the connective that participants use reflects the coherence relation they inferred; the RE that they choose reflects how they interpreted the actions and referents in the prompt). The forced choice paradigm is beneficial when studying specific or less frequent phenomena. The free insertion paradigm is suitable for giving insight into relatively natural language usage.

5.1.12. Judgment tasks
Participants are presented with a prompt and asked to judge the prompt regarding a particular quality, for example how acceptable, plausible, logical, coherent, or grammatical the prompt is. Participants can be asked to provide judgments on a binary scale (e.g., *same meaning vs different meaning*, as in Crible & Demberg, 2020), or on a graded scale (e.g., *on a scale of 1-5, rate how much the sentence makes sense*, as in Cain & Nash, 2011).

Considerations:
For judgment tasks, it is particularly important that the instructions are sufficiently clear; participants might differ in their considerations of what constitutes a coherent sentence, for example. Clear examples in the instructions can help, as well as fillers that clearly fall into either of the two binary categories or, in case of a graded scale, on the extreme ends of the scale.

5.1.13. Recall questions
Participants are presented with prompts and asked to recall specific elements of those prompts afterwards (e.g., Jou & Harris, 1990; Sanders & Noordman, 2000).

Considerations:
Recall questions test how well participants remember the actual (surface) content of the original text. For testing a more elaborate understanding of the discourse, comprehension questions might be more suitable.

5.1.14. Self-paced reading
Participants are presented with prompts in masked or hidden “chunks” (words or multi-word regions), which participants reveal by pressing a button (e.g., Koornneef & van Berkum, 2006; Lyu et al., 2020; Wetzel et al., 2022).
Considerations:
This method can be used to study immediate effects of connectives or REs on discourse processing but provides relatively coarse-grained insights. Because this method does not require the usage of special cameras, self-paced reading studies are a common method of crowdsourcing online processing data.

5.1.15. Summarisation and free recall
Participants read a prompt and are afterwards asked to provide a summary, or as many details as they can remember about the text.

Considerations:
All responses must be analysed and scored, which can be a very time-intensive process. In addition, all or part of the data may have to be double coded to ensure reliability (see also Continuation tasks). Summarization is more informative when it comes to participants’ mental representation of the discourse; free recall speaks more to participants’ recollection of surface features of the text.

6. **Recommendations for Practice**
Designing a study does not stop at choosing the appropriate method to answer the research question and creating experimental items. There are various other design issues that deserve careful consideration. One facet of experimental design that can greatly affect the results is participants’ (subconscious) awareness of the phenomenon under investigation. It is crucial to many experimental studies that participants stay unaware to the purpose of the experiment. This holds particularly for processing experiments, as prior studies have shown that participants rapidly adapt to the structure of the linguistic stimuli (e.g., Fine et al., 2013). To minimize the effects of repeated exposure to experimental items, researchers need to ensure that each participant sees every item only once in one condition (i.e., they cannot see every item in every condition), and the proportion of experimental items to *filler items* needs to be chosen carefully. In addition, the data can be checked for any learning effects during statistical analysis.
Since crowdsourcing allows for the recruitment of a larger number of participants, participants can be presented with a limited number of experimental items, to avoid frequent
exposure. However, note that it is important to have enough items in the full set of materials, because repeated measures increase the reliability and generalizability of a study’s results. Another critical facet of experimental research is choosing what type of and how many participants will be tested. The group of participants included in experimental studies should ideally be a representative sample of the population. A related question is how many participants should be sampled, or how much data is enough. (see Chapter 24). Finally, researchers are encouraged to preregister their study online. Preregistration is an open science practice that requires the specification of the study’s hypotheses and planned analyses before the data are inspected. It thereby prevents “p-hacking” or “selective reporting”, a practice whereby researchers try out several statistical analyses and/or data eligibility specifications and then selectively report those procedures that produce significant results. It also reduces the publication bias (i.e., the tendency for statistically significant findings to be published more than non-significant findings), and thereby reduces the importance of significance tests in publication decisions. We refer interested readers to Roettger (2021), who provides more details on preregistration in experimental linguistics.

7. **Future Directions**

Throughout the previous sections, we have touched on fruitful directions for future research. We here reiterate these topics:

- Individual differences in discourse comprehension and production, also related to language learning
- Multi-modal investigations of referential and relational coherence
- Replicating effects found in one paradigm using converging methodologies, experimental but also in corpus research
- Extending discourse research to understudied languages and comparing phenomena between languages

In addition, the following topics deserve more consideration in future research:

- Discourse in the spoken domain; both in terms of discourse production and comprehension
- Ecological validity of results obtained using carefully constructed linguistic items
- Readers’ motivations and the effects on experimentally obtained results of discourse processing and comprehension
8. Further Reading


9. Related topics
Analysing reading with eye-tracking; Analysing spoken language interactions with eye-tracking; Analysing language using brain imaging; Testing in the lab and through the web; eliciting spontaneous linguistic productions; contrasting online and offline measures in experimental linguistics

10. References


