Inferences in Discourse, Psychology of

Leo GM Noordman, Tilburg University, Tilburg, The Netherlands
Wietske Vonk, Max Planck Institute for Psycholinguistics and Radboud University Nijmegen, Nijmegen, The Netherlands

Abstract

An inference is defined as the information that is not expressed explicitly by the text but is derived on the basis of the understander's knowledge and is encoded in the mental representation of the text. Inferencing is considered as a central component in discourse understanding. Experimental methods to detect inferences, established findings, and some developments are reviewed. Attention is paid to the relation between inference processes and the brain.

When listeners or readers understand discourse, they understand much more than what is stated explicitly in the sentences. On the basis of their knowledge of the language and the world, they are able to understand what has been left implicit in the text but is intended to be communicated.

Consider the following text (a):

(a) There were municipal elections yesterday. Because the majority of the lower town voted for the local party, there was a shift toward the left in the city council. But the right-wing party was not completely disappointed. It had expected to lose much more.

In text (a), readers can infer that the local party is a left-wing party; that there is a causal relation between many people voting for a particular party and that party getting a stronger representation; that there are height differences in the town and that the town is probably located on a river bank; that the city council used to be more right oriented; that there is a contrast between the shift toward the left and not being completely disappointed, and consequently that the shift is a reason for being disappointed and that there are other reasons (worse expectations) why the right-wing party was not so disappointed; and that 'it' refers to the right-wing party and not, for example, to the city council. All these pieces of information are not stated explicitly in the text, but readers do understand them. They are called inferences.

An inference is defined as information that is not expressed explicitly by the text but that can be derived from the text on the basis of the comprehender's knowledge and that is encoded into the representation the comprehender constructs of the text. In this article, the notion of inference is restricted to the derivation of implicit information that occurs in spontaneous language processing. These inferences are distinguished from logical deductions in reasoning and problem solving (Johnson-Laird, 1983; see also *Reasoning with Mental Models*), such as the logical deduction 'Some artists are chemists' from the premises 'Some beekeepers are artists. All beekeepers are chemists.' Logic is concerned with the assessment of the validity of arguments in reasoning. Although listeners and readers certainly make logical deductions, and some inferences can be considered as logical inferences, everyday language behavior is not characterized by the evaluation of the validity of the arguments in reasoning.

Two kinds of inferences might be distinguished. The first kind is an inference that is the derivation of new information. This is what in ordinary language is called an inference. For example, from the sentence 'John selected his stranston shoes because there was much mud,' the reader can infer that apparently stranston is a material or brand that has advantages if there is much mud. Given that the reader is not familiar with stranston, the reader cannot know that the conjunction 'because' is correctly used. However, by assuming that the sentence makes sense, the reader can derive the inference as new information. The second kind of inference is an activation of available knowledge. Examples from text (a) are that 'it' refers to the right-wing party, and that, given the contrastive relation indicated by 'but,' a shift toward an opponent party is in general a reason for disappointment. This kind of inference in general is not called an inference in everyday language, but most of the psycholinguistic research focuses on this kind of inference.

Issues in Inference Research

There are several issues that make inferences an important topic of research in language understanding. In understanding a discourse, the number and variety of inferences that may be derived from the discourse seem to be almost unlimited. This may lead to a computational explosion. Yet, the human processing system has limited capacity, and comprehension is accomplished too quickly for many time-consuming inferences to be made. Therefore, the first issue is how to account for the control of inferences: Which inferences are made and which ones are not made? This issue will be discussed in Section Some Established Findings in Inference Research. The second issue concerns the inference as a process. What constitutes the process of making an inference; how does the inferred information get activated and deactivated; how does the information in the text interact with the reader's knowledge? Some aspects of the process of inferring will be discussed in Section Toward a More Differentiated View on Inferences. Two additional issues will be discussed that reflect the development in inference research in the last decade. First, until recently, inferences were considered as a rather isolated phenomenon, while in current research inferences are considered as one of the

components of the comprehension process (Section Inferencing as One Component of Discourse Comprehension). Second, the advent of methods to measure brain activity leads to a better understanding of the relation between inferences and the brain (Section Inferences and the Brain). This overview focuses on inferences in reading, but most of the results apply to listening as well.

Methods in Inference Research

Whether an inference is made and when it is made can be investigated by measuring specific aspects of the reading behavior. The measurements are made during the reading of the words or sentences that trigger the inference (*online methods*) or after reading those words or sentences (*off-line methods*).

Off-line Methods

The assumption underlying off-line methods is that the inferences are encoded into the mental text representation, similar to information that is expressed explicitly by the text. Therefore, they can be investigated in tasks such as reproduction, recognition, and verification. In a reproduction task, the reader is required to reproduce the text after having read it. Information that was not stated explicitly in the text but that is reproduced by the reader is supposed to be inferred by the reader. In a recognition task, the reader has to judge whether particular words or sentences occurred in the text. The rationale of the recognition task is similar to that of the reproduction task. Information that is incorrectly judged as having appeared in the text is assumed to be inferred. Many people who had studied the sentences 'John was trying to fix the birdhouse. He was pounding the nail when his father came out to watch him and to help him do the work' incorrectly recognized the sentence 'John was using the hammer to fix the birdhouse when his father came out to watch him and to help him do the work.' This indicates that an instrumental inference (hammer) was made (Bransford and Johnson, 1973). Reproduction and recognition tasks can indicate whether the inferences are made but in general cannot distinguish whether the inferences are made during reading (at encoding) or during the measurement (at retrieval). In a verification task, readers judge whether the content of a sentence is true or false with respect to the content of the text. Not only the accuracy, but also the reaction time for the verification or the recognition is important. If inferred information is recognized or verified as quickly as explicit information, this is an indication that the inference is made during reading, and not at the moment of the off-line task. Off-line methods are used frequently in combination with online methods to answer the question of when inferences are made.

Online Methods

Online methods are employed during reading and aim to detect the ongoing inference process immediately. The assumption is that inference processes require time. Therefore, they are detected by a long reading time at the moment the inference is made, relative to a control condition in which no inference is made. Reading times are measured in a self-paced reading task in which the reader by pressing a button exposes successive units of text (words or clauses) in a window on a computer screen. The interval between button presses is defined as the reading time for the unit of text. For example, for understanding the second sentence in 'Mary got the picnic supplies out of the car. The beer was warm,' the inference is required that there was beer among the picnic supplies. Therefore, the reading time for the second sentence in this sequence should be longer than in the sequence 'Mary got the beer out of the car. The beer was warm' (Haviland and Clark, 1974). Reading times are also measured in eye-tracking recordings that reveal what the reader looks at and how long. This makes it possible to measure fairly exactly when the inferences are made. Just and Carpenter (1978) found that the gaze duration on 'killer' in 'The killer left no clues for the police to trace' is longer if this sentence is preceded by the sentence 'The millionaire died on a dark and stormy night' than if it is preceded by the sentence 'The millionaire was murdered on a dark and stormy night.' Reading the word 'killer' requires the information that the person was murdered; this information has to be inferred after 'die' and is explicit in 'murdered.'

Another online method is measuring event-related potentials (ERPs), changes in the electric encephalogram due to activities in the brain. This technique is very time sensitive and therefore can reveal ongoing reading processes. Van Berkum et al. (2005) showed that hearers of a discourse make predictions of an upcoming noun. The noun was preceded by a gender-marked adjective whose suffix matched or mismatched the predicted noun. For prediction-inconsistent adjectives, a positive inflection in the ERP wave was observed 50–250 ms after onset of the inflection (before the noun), indicating that listeners anticipate an upcoming word. In addition, a negative inflection was observed at about 400 ms from noun onset, which is generally observed when a word does not fit very well in the context.

Other methods are administered immediately after the reading of the word or sentence that triggers the inference and can also be considered as online methods. In these tasks, a word is presented as a probe during reading, at or shortly after the moment that the inference is supposedly made. The probe is related to the information that is presumably inferred. An example is the probe word 'break' after the sentence 'No longer able to control his anger, he threw the delicate porcelain vase against the wall.' The task may be to decide whether or not the string of letters of the probe forms a word (*lexical decision*), to decide whether or not the probe word had been presented in the previous sentence(s) (*probe recognition*), or to pronounce the probe word (*naming*). The assumption is that if the inference is encoded into the text representation, the lexical decision time for the probe and its naming time are shorter and the time needed to indicate that 'break' did not occur in the text is longer than in a control condition (see Section The Process Character of an Inference) in which the inference is not made. The status of the different lexical techniques is not quite undisputed. In some investigations, it is claimed that these lexical tasks are sensitive to transient activations (i.e., inferences that are activated only for a short time) rather than to inferences. In particular, naming is sometimes considered a task that does not

reflect the nature of the text representation and, accordingly, does not reflect inferences that are encoded in the text representation. In other studies, it is suggested that lexical decision and probe recognition, in contrast to naming, register the fit of the probe with the preceding context rather than the encoding of inferences.

Some Established Findings in Inference Research

The issue that has stimulated much psycholinguistic research on inferences since the early 1970s is the following question: Which inferences are made and which ones are not made? This issue was investigated by looking for classifications of inferences, on the assumption that some kinds of inferences are made during reading and some other kinds are not.

A common distinction is between *necessary* and *elaborative inferences*. Inferences are called necessary if without them the text representation is not coherent. Inferences that are not necessary for coherence are called elaborative, optional, or embellishing. Two aspects of coherence can be distinguished. *Referential coherence* is achieved by the fact that a sentence in a text deals with entities that are expressed earlier in the text. Linguistic devices to indicate referential coherence are referential (or anaphoric) expressions, such as pronouns and definite noun phrases that refer to an antecedent in the text (*see Anaphora Resolution*). *Relational coherence* is achieved by the fact that the content of a sentence has a conceptual relation with other sentences, such as a causal, contrastive, or concessive relation. Relational coherence can be expressed by conjunctions. If, in text (a) above, the inference about the coreferentiality between 'it' and 'party' or the inference about causality between 'voting' and 'shift' is not made, there is no coherence between the sentences. The pronoun 'it' is considered as an instruction to find a particular referent; the conjunction 'because' is considered as an instruction to find a causal relation in the context. If such instructions are not executed, the representation is not coherent. In this way, inferences triggered by referential expressions and conjunctions contribute to the coherence of the representation.

The notion of *necessary inference* is sometimes defined in a different way and contrasted with *pragmatic inference*. An inference is necessary if it follows logically from the propositions in a text. From 'John is taller than Pete and Pete is taller than Bill,' it follows that John is taller than Bill. The causal inference between the propositions with 'voting' and 'shift' in text (a) is not only necessary for the coherence, but also necessarily follows from the text. It can be deduced from the conjunction 'because' and is called a conventional implicature.

Inferences that do not follow logically from the text are called pragmatic inferences. An example is the inference that the vase broke from the sentence 'John slipped on the wet floor and dropped the vase.'

Another distinction that is made frequently is that between *backward* and *forward inferences*. Backward inferences relate the current part of the text to an earlier part. Forward inferences anticipate information that might be expressed in the subsequent text. Backward inferences contribute in general to the coherence of the text and are in that sense necessary inferences, while forward inferences are not. Backward inferences are much more restricted than forward inferences; the reader can anticipate a great number of things.

Inferences can also be classified with respect to their content. This yields a list of inferences that can be extended *ad libitum*. It includes inferences about instruments, causes, consequences, goals, time, place, and protagonists.

Established findings in inference research (Singer, 1994, 2007; Van den Broek, 1994; Garrod and Sanford, 1994) are that necessary inferences – that is, inferences that achieve coherence between sentences – are made during reading. These inferences are in general backward inferences. Elaborative, embellishing inferences are in general not made during reading, except when they are highly constrained (O'Brien et al., 1988; Garrod et al., 1990); they are not required for comprehension. They are in general forward inferences. There are many possible forward inferences that can be made, and since they do not achieve coherence, there is no motivation for making them. Inferences that can necessarily be derived from the text are far from always made; in general, it depends on whether they contribute to the coherence or not.

Toward a More Differentiated View on Inferences

The research discussed in the previous section can be characterized by two views on inferences. First, inferences were considered as dichotomous entities: They are made or not made. Second, inferences were described with respect to their function in the text: Whether they are made or not depends on whether they contribute to the coherence of the text. Since the late 1980s and early 1990s, there has been a gradual change in inference research. These changes will be discussed in two subsections. First, the dichotomous view on inferences is being replaced by a process view on inferences: What happens if an inference is made? Second, attention has shifted from the function of inferences in the text to the availability of information and knowledge as determinants of inferences, and in this way to the role of the reader.

The Process Character of an Inference

The shift in attention toward a process view on inferences was stimulated by research on forward inferences. Presumably, the reason was that backward inferences were well established and that the idea that forward inferences are not made was not quite convincing. Several studies (McKoon and Ratcliff, 1986) demonstrated that forward inferences about very predictable events can be made, but are only partially or minimally encoded. After presentation of several sentences, among them, 'The director and the cameraman were ready to start shooting when suddenly the actress fell from the 14th floor,' the word 'dead' was presented as a probe for recognition. If an inference about dying were encoded in memory, it would interfere with the correct answer 'no.' When the target word was preceded by another word from the same sentence, an interfering effect did indeed occur, but when the target word was preceded by a neutral word (e.g., 'ready'), no interference occurred. This result was interpreted as evidence for a minimal encoding of the inference, such as 'something bad happened,' because a prime word from the same sentence

was necessary to strengthen the match between the probe word and the memory representation.

Other research demonstrated that inferences are built up and decay over time. First, forward inferences are not generated until some time after the presentation of the words the inference is based on. Experiments in which the time between the context that primes the inference and the presentation of a target word to detect the inference is varied indicate that a delay of about 1 s is necessary for the inference to be built up (Calvo et al., 1999). This observation may explain why no evidence for forward inferences was obtained in most of the earlier studies. If, in a reading experiment, subsequent information that does not support the inference is presented very quickly, the inference is not built up. However, if readers can read at their own pace or if subsequent information supports the inference, forward inferences can be built up. Second, an inference can be made for a brief period of time. It need not remain active. Keefe and McDaniel (1993) found that the naming latency for the word 'break' was shorter after 'One day, no longer able to control his anger, he threw a delicate porcelain vase against the wall' than after the control sentence 'One day, unable to control his impulses, he went out and purchased a delicate porcelain vase' when the probe word was presented with a short delay after the target sentence, but there was no facilitation when there was an intervening sentence between target sentence and probe. They argued that predictive inferences are drawn but are quickly deactivated. The likelihood that predictive inferences are encoded seems to increase if the information to be inferred is foregrounded and salient, and if the inference concerns a causal relation (Klin et al., 1999). In addition, the specificity of predictive inferences increases as the contextual support increases (Lassonde and O'Brien, 2009). With respect to the predictive inferences found by Van Berkum et al. (2005), one might hypothesize that since listening is slower than reading, more inferences are made during listening than during reading.

Availability of Information and Reader's Knowledge as Determinants of Inferences

Whether inferences are made or not depends not only on the function of the inference in the *text*, but also on whether the *reader* has information and knowledge available. This has been acknowledged in two well-known theories of inferences: the *minimalist theory* and the *constructionist theories*. In the minimalist theory, language processing and inferencing are described largely in terms of bottom-up processing of the information in the text. In the constructionist theories, the reader's search for meaning and the reader's knowledge play an important role, so that understanding entails a considerable amount of top-down processing.

In their minimalist theory on inferences, McKoon and Ratcliff (1992) argue that only two kinds of inference are drawn spontaneously during reading: inferences that serve to achieve *local coherence* and inferences that are based on *easily available information*. Local coherence refers to coherence between propositions that are not farther apart in the text than one or two sentences. Information is easily available if it is expressed in the current two or three sentences or if it is well-known general knowledge, such as instances of categories (collie – dog). In this way, the availability of information in the text is defined in terms of the linear structure of the text.

In constructionist theories, the knowledge of the reader plays a much more important role than in minimalist theory. According to constructionist theories, readers construct a *mental model* or a *situational model*. Such a model contains not only information in the propositions of the text but also information that is constructed by the reader, including elaborative inferences and global inferences (Johnson-Laird, 1983; Garnham and Oakhill, 1996; Zwaan and Radvansky, 1998). Many participants (Bransford and Johnson, 1973) who heard a sentence such as 'Three turtles rested on a floating log and a fish swam beneath them' recognized, incorrectly, the sentence 'Three turtles rested on a floating log and a fish swam beneath it,' but participants who had heard the sentence 'Three turtles rested beside a floating log and a fish swam beneath them' did not often recognize incorrectly the sentence with 'it.' Apparently, comprehension requires the activation of knowledge about spatial relations.

The central idea in the constructionist theory of Graesser et al. (1994) is that reading is considered as a search for meaning. The readers' goal and knowledge guide the reading process. Readers try to construct a meaning representation that is coherent both at the local and at the global level. They try to explain the actions, events, and states mentioned in the text; that is, readers try to answer 'why' questions with respect to the text. According to this theory, inferences concerning causal antecedents and superordinate goals are made because they explain actions in the text, whereas inferences about causal consequences and subordinate goals and states are not made because they do not contribute to the explanation. In evaluating minimalism and constructionism, one should keep in mind that even local inferences are not made if they do not address information that is part of the reader's knowledge (Noordman and Vonk, 1992).

Inferencing as One Component of Discourse Comprehension

In the studies reviewed up till now, inferences were considered as a rather isolated phenomenon. The main question was whether inferences are made or not, and the answer was considered to depend on the kind of inferences: bridging, forward, elaborating, instrumental, etc. In an earlier overview on inferences in discourse, Vonk and Noordman (2001) concluded that progress in inference research will be made only if inferences are no longer considered as a rather isolated phenomenon, but a component in discourse comprehension. Inferences should then be considered in relation to models of discourse comprehension and attention should be paid to the crucial role of world knowledge. This is indeed what happened to certain extent in the last decade. Once one considers inferences as a component in the comprehension process, it becomes clear that minimalism and constructionism are compatible. Discourse comprehension is, like all cognitive activities, *memory-based*: The processes take place in working memory and make

contact with information in long-term memory, which comprises information both from the previous discourse and from world knowledge. A crucial concept in the minimalist theory of inferences is 'easily available information.' But since easily available information includes previous discourse, world knowledge, knowledge about the topic of the text, about the goals of the protagonist (who can be referred to by a pronoun, testifying for its accessibility), and about the text structure, it does not follow from this notion that only a few inferences are made. For example, given that immediately available information includes world knowledge related to a specific word, it is not clear why minimalism would predict that the inference 'spoon' is not made when reading a sentence 'John stirred the coffee.' Interesting in this respect are data from *embodied cognition* showing that when reading sentences that describe a particular action similar activity is observed in the brain as when performing that action. Why would minimalism claim that the instrumental inference for 'stirring coffee' is not made, while inferences about the waiter, the menu, and the bill are made in a restaurant script story? Whether the inferences are made or not is, of course, an empirical question, but the prediction does not follow from the minimalist position. Other examples are studies by Albrecht and O'Brien (1993) and Myers et al. (1994). Given that the protagonist and his or her attributes are active during the whole story, it is not surprising – and it does not refute minimalism – that there is an increase in reading time for 'she ate a cheeseburger' if the protagonist was introduced earlier in the discourse as a vegetarian. To further strengthen the conciliation between minimalism and constructionism: it is not clear why some findings are considered to support constructionism and to argue against minimalism. Consider the beginning of a story 'Valerie left early for the birthday party' vs 'Valerie left the birthday party early' followed by a number of sentences including 'She spent an hour shopping at the mall' (Graesser et al., 1994). World knowledge about 'left for' and 'left' makes clear that the birthday party is an upcoming episode or a terminated episode, respectively. Consistent with the view that readers keep track of the temporal development in a story (Bestgen and Vonk, 1995; Zwaan and Radvansky, 1998), the birthday script is relevant and available only in the 'left for' version, which was evidenced by a test sentence. This explanation of the results is not in conflict with minimalism at all. It shows that the constructionist theory is not incompatible with memory-based processing. And this claim can be made stronger. In the constructionist view, reading is considered as a search after meaning and explanations. The goal dictates what kind of knowledge the reader is after and activates specific knowledge and makes it available. It determines what information is in focus. The reader's goal can affect the processes in working memory and interact with long-term memory. The search after meaning of this explanation-based approach can be accommodated in the memory-based approach. The position of Gerrig and O'Brien (2005) is in agreement with our proposal. "There is no need to define categories of inferences that readers will typically encode and those that readers are less likely to encode. Inferences are encoded to the extent that information in active memory makes contact with relevant or necessary information from inactive portions of the discourse model and general world knowledge" (p. 236). They also suggest that the search-after-meaning of the explanation-based view can be accommodated in the memory-based view: "within the memory-based view, the concept of search-after-meaning is not actually a search of memory; instead it is an attentional, resources-consuming … process" (p. 237). What is less clear, however, is how this memory-based view on inferences can be invalidated, since all cognitive processes are memory processes.

Less progress has been made in relating inferences to models of discourse comprehension and to the crucial role of world knowledge. Most of the models of discourse incorporate world knowledge, but not in a computational way (Frank et al., 2008). In a memory-based model, world knowledge of the previous discourse and world knowledge are supposed to resonate. The Resonance model (Myers and O'Brien, 1998) is the example *par excellence* of memory-based processing. But the model does not incorporate world knowledge and does not claim to do so. The Construction–Integration model (Kintsch, 1988, 1998) has been a fruitful framework for the interpretation of experimental data and the generation of further research. The model accommodates inferences, but the implementation of world knowledge is rather ad hoc and restricted to text-relevant knowledge. This also is true for the Landscape model (Van den Broek et al., 1999) and the model by Langston and Trabasso (1999). On the other hand, the Distributed Situation Space model (Frank et al., 2003) encodes world knowledge based on events in a microworld. Therefore, world knowledge in this model is not introduced ad hoc and the model predicts inferences in the microworld in a computational way. The question, of course, is whether the model can be scaled up to more realistic amounts of knowledge. This problem applies to the other models as well.

Inferences and the Brain

The advent of techniques that measure brain activity is an important step in the study of discourse processing. One such technique is measuring ERPs. The ERP signal is a pattern of positive and negative peaks in the EEG. Hagoort et al. (2004) showed that when a word in a sentence violates world knowledge, this is reflected by the N400 in the brain wave. In the study by Van Berkum et al. (2005) cited earlier, differential ERP effects were found when an upcoming word matched vs mismatched a prediction.

Kuperberg et al. (2011) investigated inference processes in understanding sentences that have different degrees of causal relatedness. Lexicosemantic co-occurrence was held constant across the conditions. Critical words in the causally unrelated sentences evoked a larger N400 than critical words in both the highly related and intermediately related sentences. This was the case both when the critical words occurred before and at the end of the sentences. These data demonstrate that readers are immediately sensitive to coherence breaks. Causal coherence at the situation level immediately influences inferencing during semantic processing of incoming words. At many of the electrode sites, the amplitude of the N400 to critical words in intermediately related sentences fell in

between the amplitude of the N400 in the unrelated and highly related conditions. Interestingly, the linear relation between causal relatedness and the amplitude of the N400 differs from curvilinear relations obtained in a functional magnetic resonance imaging (fMRI) study by Mason and Just (2004) in brain activation and Myers et al. (1987) in recall. One possible interpretation, according to Kuperberg et al. (2011), might be the difference in temporal resolution between ERP and fMRI.

An interesting topic is the mapping of cognitive processes to regions in the brain. In general, it is assumed that processes in language understanding, such as word coding and syntactic analysis, activate the left hemisphere. There is increasing evidence, however, that the right hemisphere is also involved in language processes and in particular in inferencing. Beeman et al. (2000) showed in a visual hemifield study that predictive inferences activate the right hemisphere. Reading the sentence 'The space shuttle sat on the ground, waiting for the signal' might lead the reader to the predictive inference that the shuttle is taking off. This would lead to facilitation of the target word 'launch.' When this word was presented to the left visual hemifield (corresponding to the right hemisphere), the naming latency for this word was shorter than for an unrelated control word. But presenting the target word to the right visual hemifield, there was no difference between target and control word. According to the coarse semantic coding theory (Beeman et al., 1994), the right hemisphere is involved in the representation of wider connections of word meanings and less central meanings of words. This representation includes distant associations between words. Since these associations play a role in the processing of inconsistencies and in inferencing, in particular when there is no direct overlap between adjacent pieces of text, it is supposed that the right hemisphere is involved in inferences. Interestingly, for causal inferences that require bridging inferences, Beeman et al. (2000) found that the left hemisphere is responsible for the inferences after the shuttle sat on the ground. When the shuttle disappeared into space,' the word 'launch' is processed faster than a control word when it is presented to the right visual hemifield, but not when it is presented to the left visual hemifield.

Virtue et al. (2006b) investigated in a visual hemifield study bridging and predictive inferences that were either strongly or weakly constrained by the context. A target word related to the inference was presented in one of the visual fields for lexical decision. For strongly constrained bridging and predictive inferences, facilitation (shorter lexical decision time for target words in the inference condition than in a control condition) was obtained in both hemispheres. For weakly constrained bridging and predictive inferences, there was greater facilitation in the right hemisphere than in the left hemisphere, but for predictive inferences there was some facilitation in the left hemisphere as well.

Mason and Just (2004) used fMRI to study the neuroanatomical basis for inference processes in reading. This is a technique that measures activity in specific brain areas by detecting changes in blood flow. It is less sensitive to temporal resolution than ERPs. Mason and Just (2004) used sentence pairs of Keenan et al. (1984) and Myers et al. (1987) but only in three versions: the distantly related, moderately related, and highly related versions. In the right hemisphere, there was greater activity for the moderately related sentence pairs than for the highly related and the distantly related pairs. In the left hemisphere, there was no difference between the three versions and the activation for the highly related and distantly related versions was greater than in the right hemisphere. These results were interpreted to mean that the left hemisphere is concerned with processing the sentences, while the right hemisphere is responsible for establishing the coherence and presumably for making the inferences.

Kuperberg et al. (2006) used in an fMRI study similar items as Mason and Just, but they had items of three sentences long instead of two. In addition, they changed the experiment in some methodological ways (e.g., increasing the number of items and testing only on the last sentence instead of on two sentences). In the intermediately related condition, processing required more time (participants had to judge the coherence of the sentences). For the intermediately related sentences (compared to the other two conditions) they found increased activity in both hemispheres (left lateral temporal/inferior parietal/prefrontal cortices, the right inferior prefrontal gyrus, and bilateral superior medial prefrontal cortices). Apparently, understanding information that is not explicitly stated in the discourse is accompanied by the activation of a large cortical network in both hemispheres.

Interesting as it is to identify the areas that are activated, more interesting is what localization data can contribute to a theory of inferencing. A distinction is frequently made between the generation of inferences and the integration of the inferred information in long-term memory (Mason and Just, 2004). Is there experimental evidence from brain studies for this distinction? Mason and Just suggest that these processes correspond to two large-scale cortical networks: a reasoning system in dorsolateral prefrontal cortex for generating inferences and specific parts of the right-hemisphere for integrating the possible inferences that have been generated. Kuperberg et al. (2006) are even more specific in relating cognitive processes to brain regions. They state that inferencing is accompanied by the activation of many regions in the brain. These regions act in consort, but building upon previous studies, Kuperberg et al. (2006) suggest that parts of this network may have distinct roles in causal inferencing. The temporal regions are involved in the activation of stored semantic information; the inferior prefrontal regions may mediate the retrieval and selection of semantic information; the posterior inferior prefrontal regions and posterior dorsolateral prefrontal regions may play a role in the maintenance of semantic information in working memory as it is integrated in long-term memory; activity in the superior medial prefrontal regions may reflect directed search for meaning and the examination of temporal relationships between events to generate inferences. Furthermore, they suggest that the activation in the right temporal and inferior prefrontal regions play a role in detecting incoherence in the discourse. These suggestions are speculative, but they indicate how imaging studies may advance theories about inference processes.

The neural correlates of discourse comprehension processes have only recently been investigated. This is a promising line of research. But we need more relevant empirical data, as well as

replications of earlier data, because consistency and convergence are important issues. This is particularly relevant since studies differ in many respects: the method, the task for the reader/listener, the materials that are used in the experiment. To illustrate this point: A topic of quite a few studies is coherence, including lack of coherence, coherence breaks, and inferences to establish coherence. Different studies find quite different locations of activations; they differ even with respect to whether there is activation in the left or right hemisphere. But that is not so surprising if one considers that under the guise of (in)coherence processing quite different operationalizations of coherence and quite different tasks are hid, that give rise to quite different cognitive processes.

In Mason and Just (2004), for example, coherence is operationalized as different degrees of causal relatedness between two consecutive sentences. Moderate coherence resulted in greater activation in specific areas of the right hemisphere (middle and superior temporal gyrus, inferior temporal gyrus, inferior frontal gyrus, and inferior parietal area). The task was to read the texts and to answer probe questions for filler texts.

Robertson et al. (2000), on the other hand, operationalized coherence as presence of a definite or an indefinite article. The indefinite article produced more activity than the definite article in the right hemisphere, in particular in the prefrontal region, but not in the left hemisphere. The task was to read sets of sentences.

Ferstl and Von Cramon (2001), for their part, operationalized coherence in terms of (un)related sentences. The task was to judge whether the two sentences have something to do with each other. They found that the coherence manipulation resulted in activation in the left frontal gyrus. They did not find any additional activation in the right hemisphere (see also Ferstl et al., 2008).

In Virtue et al. (2006a), coherence break is operationalized as a place in the text where an inference is required: A sentence had to be connected to the preceding sentence; in one condition an unspecific verb ('got to work') occurred in the preceding sentence, in the other condition a specific verb (started ironing). The task was to listen to the text and answer a question, not related to the coherence break. They found increased activation for the unspecific over the specific event in the right and left superior temporal gyrus and inferior frontal gyrus.

Kim et al. (2012) used strong coherence and weak coherence stories of three sentences in a positron emission tomography study. This technique, as fMRI, measures brain activity on the basis of local blood flow. The last sentence in the weak condition required an inference, but the last sentence in the strong condition did not. The task was to read the sentences and to judge the plausibility of the last sentence. In the strong coherence condition, the dorsomedial prefrontal cortex was activated relative to a control condition. This was interpreted as reflecting coherence processing (but in other studies activation at this location is obtained in low coherence conditions). In the weak coherence condition, the left middle temporal gyrus was activated. This was interpreted as reflecting a bridging inference.

The operationalization of coherence in these studies might very well lead to quite different cognitive processes. The use of indefinite articles in the Robertson et al. (2000) study violates (Gricean) maxims of conversation, in particular of referring. Moreover, some of the references are ambiguous. The reader has to deal with these problems. It is not very likely that establishing reference in this task is done on the basis of world knowledge activation. Quite different are the Mason and Just (2004) and the Virtue et al. (2006a) studies. In these studies, coherence can be achieved by activating world knowledge, making bridging inferences, and integrating them into the discourse representation. And indeed, Mason and Just argue that these processes take place and that they can be distinguished as located in different brain networks. In the Ferstl and Von Cramon (2001) study, participants perform a metalinguistic decision task. To reach a criterion that two sentences have nothing to do with each other, in comparison with the sets of sentences that are connected, it is not necessary that participants engage in world knowledge activation and inference generation, as in the Mason and Just study. Finally, in the Kim et al. (2012) study, the metalinguistic task of judging the plausibility of a sentence introduces an additional task and probably introduces additional processes.

The upshot is that different materials and different tasks may lead to different processes in coping with 'incoherence.' These examples illustrate the need for carefully conducted empirical studies on inference processes. Only then will brain studies increase the understanding of the different subprocesses in inferencing and how they are related to each other. Brain research will then not only reveal which areas in the brain are activated in specific tasks, but may also yield data that contribute to a theory of inferences and discourse comprehension.

See also: Psychology of Inferences; Sentence Comprehension, Psychology of.

Bibliography

Albrecht, J.E., O'Brien, E.J., 1993. Updating a mental model: maintaining both local and global coherence. Journal of Experimental Psychology: Learning, Memory, and Cognition 19, 1061–1070.

Beeman, M.J., Bowden, E.M., Gernsbacher, M.A., 2000. Right and left hemisphere cooperation for drawing predictive and coherence inferences during normal story comprehension. Brain & Language 71, 310–336.

Beeman, M.J., Friedman, R.B., Grafman, J., Perez, E., Diamond, S., Lindsay, M.B., 1994. Summation priming and coarse coding in the right hemisphere. Journal of Cognitive Neuroscience 6, 26–45.

Bestgen, Y., Vonk, W., 1995. The role of temporal segmentation markers in discourse processing. Discourse Processes 19, 385–406.

Bransford, J.D., Johnson, M.K., 1973. Considerations of some problems of comprehension. In: Chase, W.G. (Ed.), Visual Information Processing. Academic Press, New York, pp. 383–438.

Calvo, M.G., Castillo, M.D., Estevez, A., 1999. On-line predictive inferences in reading: processing time during versus after the priming context. Memory and Cognition 27, 834–843.

Ferstl, E.C., Neumann, J., Bogler, C., Von Cramon, D.Y., 2008. The extended language network: a meta-analysis of neuroimaging studies on text comprehension. Human Brain Mapping 29, 581–593.

Ferstl, E.C., Von Cramon, D.Y., 2001. The role of coherence and cohesion in text comprehension: an event-related fMRI study. Cognitive Brain Research 11, 325–340.

Frank, S.L., Koppen, M., Noordman, L.G.M., Vonk, W., 2003. Modeling knowledge-based inferences in story comprehension. Cognitive Science 27, 875–910.

Frank, S.L., Koppen, M., Noordman, L.G.M., Vonk, W., 2008. World knowledge in computational models of discourse comprehension. Discourse Processes 45, 429–463.

Garnham, A., Oakhill, J., 1996. The mental models theory of language comprehension. In: Britton, B.K., Graesser, A.C. (Eds.), Models of Understanding Text. Erlbaum, Mahwah, NJ, pp. 313–339.

Garrod, S.C., O'Brien, E.J., Morris, R.K., Rayner, K., 1990. Elaborative inferences as an active or passive process. Journal of Experimental Psychology: Learning, Memory, and Cognition 16, 250–257.

Garrod, S.C., Sanford, A.J., 1994. Resolving sentences in a discourse context: how discourse representation affects language understanding. In: Gernsbacher, M.A. (Ed.), Handbook of Psycholinguistics. Academic Press, San Diego, CA, pp. 675–698.

Gerrig, R.J., O'Brien, E.J., 2005. The scope of memory-based processing. Discourse Processes 39, 225–242.

Graesser, A.C., Singer, M., Trabasso, T., 1994. Constructing inferences during narrative text comprehension. Psychological Review 101, 371–395.

Hagoort, P., Hald, L., Bastiaansen, M., Petersson, K.M., 2004. Integration of word meaning and world knowledge in language comprehension. Science 304, 438–441.

Haviland, S.E., Clark, H.H., 1974. What's new? Acquiring new information as a process in comprehension. Journal of Verbal Learning and Verbal Behavior 13, 512–521.

Johnson-Laird, P.N., 1983. Mental Models. Cambridge University Press, Cambridge, UK.

Just, M.A., Carpenter, P.A., 1978. Inference processes during reading: Reflections from eye fixations. In: Senders, J.W., Fisher, D.F., Monty, R.A. (Eds.), Eye Movements and the Higher Psychological Functions. Erlbaum, Hillsdale, NJ, pp. 157–174.

Keefe, D.E., McDaniel, M.A., 1993. The time course and durability of predictive inferences. Journal of Memory and Language 32, 446–463.

Keenan, J.M., Baillet, S.D., Brown, P., 1984. The effects of causal cohesion on comprehension and memory. Journal of Verbal Learning and Verbal Behavior 23, 115–126.

Kim, S., Yoon, M., Kim, W., Lee, S., Kang, E., 2012. Neural correlates of bridging inferences and coherence processing. Journal of Psycholinguistic Research 41, 311–321.

Kintsch, W., 1988. The role of knowledge in discourse comprehension: a construction-integration model. Psychological Review 95, 163–182.

Kintsch, W., 1998. Comprehension: A Paradigm for Cognition. Cambridge University Press, Cambridge, UK.

Klin, C.M., Guzmán, A.E., Levine, W.H., 1999. Prevalence and persistence of predictive inferences. Journal of Memory and Language 40, 593–604.

Kuperberg, G.R., Lakshmanan, B.M., Caplan, D.N., Holcomb, P.J., 2006. Making sense of discourse: an fMRI study of causal inferencing across sentences. NeuroImage 33, 343–361.

Kuperberg, G.R., Paczynski, M., Ditman, T., 2011. Establishing causal coherence across sentences: an ERP study. Journal of Cognitive Neuroscience 23, 1230–1246.

Langston, M.C., Trabasso, T., 1999. Modeling causal integration and availability of information during comprehension of narrative texts. In: Van Oostendorp, H., Goldman, S.R. (Eds.), The Construction of Mental Representations during Reading. Erlbaum, Mahwah, NJ, pp. 29–69.

Lassonde, K.A., O'Brien, E.J., 2009. Contexual specificity in the activation of predictive inferences. Discourse Processes 46, 426–438.

Mason, R.A., Just, M.A., 2004. How the brain processes causal inferences in text. Psychological Science 15, 1–7.

McKoon, G., Ratcliff, R., 1986. Inferences about predictable events. Journal of Experimental Psychology: Learning, Memory, and Cognition 12, 82–91.

McKoon, G., Ratcliff, R., 1992. Inference during reading. Psychological Review 99, 440–466.

Myers, J.L., O'Brien, E.J., 1998. Accessing the discourse representation during reading. Discourse Processes 26, 131–157.

Myers, J.L., O'Brien, E.J., Albrecht, J.E., Mason, R.A., 1994. Maintaining global coherence during reading. Journal of Experimental Psychology: Learning, Memory, and Cognition 20, 876–886.

Myers, J.L., Shinjo, M., Duffy, S.A., 1987. Degree of causal relatedness and memory. Journal of Memory and Language 26, 453–465.

Noordman, L.G.M., Vonk, W., 1992. Reader's knowledge and the control of inferences in reading. Language and Cognitive Processes 7, 373–391.

O'Brien, E.J., Shank, D.M., Myers, J.L., Rayner, K., 1988. Elaborative inferences during reading: do they occur on-line? Journal of Experimental Psychology: Learning, Memory, and Cognition 14, 410–420.

Robertson, D.A., Gernsbacher, M.A., Guidotti, S.J., Robertson, R.W.W., Irwin, W., Mock, B.J., Campana, M.E., 2000. Functional neuroanatomy of the cognitive process of mapping during discourse comprehension. Psychological Science 11, 255–260.

Singer, M., 1994. Discourse inference processes. In: Gernsbacher, M.A. (Ed.), Handbook of Psycholinguistics. Academic Press, San Diego, CA.

Singer, M., 2007. Inference processes in discourse comprehension. In: Gaskell, M.G. (Ed.), The Oxford Handbook of Psycholinguistics. Oxford University Press, Oxford, pp. 343–359.

Van Berkum, J.J.A., Brown, C.M., Zwitserlood, P., Kooijman, V., Hagoort, P., 2005. Anticipating upcoming words in discourse: evidence from ERPs and reading times. Journal of Experimental Psychology: Learning, Memory, and Cognition 31, 443–467.

Van den Broek, P.W., 1994. Comprehension and memory of narrative texts: inferences and coherence. In: Gernsbacher, M.A. (Ed.), Handbook of Psycholinguistics. Academic Press, San Diego, CA, pp. 539–588.

Van den Broek, P., Young, M., Tzeng, Y., Linderholm, T., 1999. The landscape model of reading: inferences and the online construction of a memory representation. In: Van Oostendorp, H., Goldman, S.R. (Eds.), The Construction of Mental Representations during Reading. Erlbaum, Mahwah, NJ, pp. 71–98.

Virtue, S., Haberman, J., Clancy, Z., Parrish, T., Beeman, M.J., 2006a. Neural activity of inferences during story comprehension. Brain Research 1084, 104–114.

Virtue, S., Van den Broek, P., Linderholm, T., 2006b. Hemispheric processing of inferences: the effects of textual constraint and working memory capacity. Memory and Cognition 34, 1341–1354.

Vonk, W., Noordman, L.G.M., 2001. Inferences in discourse, psychology of. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social and Behavioral Sciences. Elsevier, Amsterdam, pp. 7427–7432.

Zwaan, R.A., Radvansky, G.A., 1998. Situation models in language comprehension and memory. Psychological Bulletin 123, 162–185.