CHAPTER 17

WHAT CAN NEUROSCIENCE TELL US ABOUT REFERENCE?

BERIT BROGAARD

17.1 INTRODUCTION

WHEN seen against the background of traditional formal semantics, the question of whether neuroscience can tell us anything about reference is easily answered. Cognitive neuroscience is concerned with how the brain processes information. Within formal semantics, answers to semantic questions have no bearing on brain processing or mental representations. Although traditional semantic frameworks, such as Montague Grammar, use a representational language, the level of representation is not supposed to reflect mental representation and is, in fact, dispensable (Montague 1970b). So, if traditional formal semantics provides the best account of natural language, then neuroscience can only tell us about the psychology of language processing.

Linguists and philosophers, however, are gradually starting to come to terms with the fact that formal semantics cannot provide an adequate semantics for natural language. Many linguistic phenomena depend on larger non-linguistic and linguistic contexts, including quantifier scope, anaphora, VP-ellipsis, presuppositions, and resolution of lexical ambiguity. Such discourse-dependent phenomena cannot be adequately accounted for by traditional formal semantics but require a dynamic semantic framework that can model discourse updates. Two of the most influential dynamic semantic theories and also the first on the linguistic scene are Discourse Representation Theory (DRT), developed by Hans Kamp (1981) and file change semantics, developed independently by Irene Heim (1982). Here I shall focus primarily on DRT but most of what I say will apply more broadly.

Traditional formal semantics took the primary focus of semantic theory to be reference, truth, and satisfaction. The meaning of a sentence is its truth-conditions.

Within DRT and related dynamic approaches, the most central concepts are not those of reference, truth, and satisfaction, but that of information, or representation. The meaning of a sentence consists of both its potential to change a given discourse representation structure (DRS) into a new one and its truth-conditional import in the representation that results from a potential update (Kamp et al. 2011). In DRT a sentence thus has meaning only in a derivative sense. As an illustration of the difference, consider Chomsky's grammatical but nonsensical "Colorless green ideas sleep furiously". Whereas this sentence would be associated with a truth-condition in formal semantics, viz. *that green ideas sleep furiously*, the sentence is unlikely to have any potential to change a given discourse representation and hence will have no truth-conditional import in the representation that results. It is therefore meaningless in the context in which it occurs.

While reference and truth are not the most central notions in DRT, they are nonetheless important notions. Proper names, indefinite and definite descriptions, certain other quantifiers, demonstratives, and pronouns can all function as referential expressions that contribute a discourse referent and a condition to the discourse representation. Discourse referents are a special kind of variable that are bound by discourse-wide existential quantifiers. So, referential expressions refer to an individual when there is an individual that can serve as a value of the variable introduced by the expression. Reference requires satisfying not only the conditions introduced by the referring expression but also those introduced by co-referring expressions. For example, 'he' as it occurs in 'A delegate arrived. He was dressed in black' contributes a discourse referent x and the condition that the referent must be male. But assuming that 'he' and 'a delegate' both occur in the updated discourse representation and that 'he' refers back to 'a delegate', 'a delegate', and 'he' refer to a particular individual S only if there is a mapping from 'male(x) and delegate(x)' to S. Since discourse representations are continually updated, an expression that doesn't refer may come to refer after a discourse update. For example, if the hearer learns that the delegate who arrived is a woman, then 'a delegate' doesn't refer to anything, as there is no mapping from 'male(x) and delegate(x)' to an individual. But the new information may result in the revised discourse representation *A delegate arrived. She was dressed in black*, in which case 'a delegate' acquires an actual referent.

Since traditional formal semantics is not concerned with how the brain processes information, neuroscience cannot provide evidence for or against traditional formal semantics. Evidence from neuroscience only becomes semantically significant once it has been determined on independent grounds that dynamic semantics is required to provide an adequate account of linguistic phenomena in natural language. Neuroscience can then shed light on how the human brain actually interprets natural language. In what follows I will first review some of the arguments for thinking that DTR or a related dynamic framework is required to account for various linguistic phenomena in natural language. I will then review whether the neuroscientific findings lend support to reference-related aspects of the theory. Finally, I will consider some methodological concerns that have been raised about the existing neuroscientific approaches to natural language interpretation.

17.2 DRT VERSUS TRADITIONAL SEMANTICS

Kamp's Discourse Representation Theory (DRT) and Irene Heim's (1982) related file change semantics have been enormously successful as models of linguistic phenomena that depend on larger linguistic contexts, including quantifier scope, anaphora, VP-ellipsis, presuppositions, and resolution of lexical ambiguity. Anaphora has proven particularly difficult to account for within traditional formal semantics. For example, traditional semantics seems unable to provide an account of anaphoric discourse of what has become known as 'donkey anaphora':

(1) If Pedro owns a donkey, he beats it

One might attempt the following analysis of (1):

(2) $\exists x$: (donkey(x) & owns(pedro, x)) $\rightarrow$ beats(pedro, x)

But this is not a well-formed formula, as the variable in the consequent 'beats(pedro, x)' is outside the scope of the existential quantifier. The only option is an analysis, using a universal quantifier:

(3) $\forall x$: (donkey(x) & owns(pedro, x)) $\rightarrow$ beats(pedro, x)

But the structure of this analysis is different from that of (1) and interprets the existential quantifier as a universal, which is not the natural interpretation of (1). In DRT, an interpretation is a mental representation that consists of discourse referents and conditions that put constraints on the values of the discourse referents. Unlike the indefinite description 'someone named "Pedro"', the name 'Pedro' carries a presupposition to the effect that Pedro(x) is already part of the existing DRS. If Pedro has not already been introduced, the presupposition is accommodated by adding the discourse referent and condition introduced by the name. In Kamp's (1981) model, this is represented by placing discourse referents for proper names outside of the scope of the conditional to reflect that the discourse referent is supposed to be part of the already processed parts of the text. So, (1) can be represented as:

(4) x: Pedro x & [donkey y & owns (x, y) $\rightarrow$ beats(x, y)]

Because the discourse referent for 'donkey' occurs in the antecedent of a conditional, any donkey can serve as a value, so the truth-condition comes out as 'For every donkey Pedro owns, Pedro beats it'.

Donkey anaphora is not a decisive factor in choosing a dynamic semantic theory over traditional semantics. Although traditional semantics cannot account for donkey anaphora using standard first-order logic, alternative accounts have been developed that can account for the basic cases. An alternative analysis is Stephen Neale's (1990: 121) D-type analysis. On Neale's view, donkey pronouns are numberless

definite descriptions constructed from the content of the antecedent clause. On this view, (1) comes out as:

(5) If Pedro owns a donkey, he beats the donkey(s) he owns.

While this analysis does not require dynamic resources, it is unclear that it can be extended to all cases of donkey anaphora. Adverbs of quantification introduce new levels of difficulties (Brogaard 2007). For example, 'if Pedro owns a donkey, he usually beats it' can be read as saying that Pedro beats most of the donkeys he owns. However, the sentence 'if Pedro owns a donkey, he usually beats the donkey(s) he owns' can only be given the temporal interpretation that if Pedro owns any donkey, then he beats all of the donkeys he owns most of the time.

DRT also has an advantage in terms of accounting for the difference in the meaning of sentences that are rendered truth-conditionally equivalent by standard semantics. For example, even though 'one of the ten marbles is not in the bag' and 'nine of the ten marbles are in the bag' are truth-conditionally equivalent, according to standard semantics, only the former licenses anaphora:

(6) a. One of the ten marbles is not in the bag. It is under the sofa.
b. *Nine of the ten marbles are in the bag. It is under the sofa.

Because DRT predicts that only 'one of the ten marbles is not in the bag' introduces a discourse reference, it can account for why the two sentences differ in meaning despite being truth-conditionally equivalent.

A further problem for traditional semantics is that it cannot straightforwardly account for the fact that the interpretation of one sentence can depend on what is uttered at a later time. Consider the following example, taken from Asher et al. (2001):

(7) A nurse saw every patient. Dr. Smith did too.

Although there is a scope-ambiguity in the first sentence, the preferred reading after the second sentence is processed is one where 'a nurse' takes wide scope, viz. $\exists x(\forall y$ (patient $y \rightarrow x$ saw $y))$. Asher et al. (2001) argue that the more specific common themes between sentences is preferred over the wider one. The common theme between the second sentence and the wide-scope reading of the first sentence is more specific than the common theme that would result from a reading where 'every patient' takes wide scope. Traditional semantics stipulates that the first sentence in (7) should have the meaning it does independently of discourse that is added later.[1] DRT, on the other hand, predicts that discourse that is uttered later can affect the meaning of the existing DRS.

DRT is thus able to provide a more comprehensive account of a multiplicity of linguistic phenomena than standard semantics. Moreover, because DRT equates meaning with mental representation, it is suitable as an object of investigation from a neuroscientific perspective. The aspects of reference that have been most extensively

[1] This is not to say that traditional semantics does not recognize that natural language often is highly ambiguous.

investigated by neuroscience are anaphoric reference and presupposition accommodation. While philosophers of language have traditionally drawn a sharp distinction between anaphoric reference and reference to extra-mental entities, the two are not clearly separable from a language-processing perspective. Empirical studies of reference to extra-mental entities might examine which entity in a visually presented narrative a speaker takes an expression to refer to. Analogously, studies of anaphora might examine which entities in a verbally presented narrative a speaker takes an expression to refer to. While there might be crucial differences between the two approaches, it seems clear that studies of anaphora and related phenomena should be able to shed light on reference to extra-mental entities. After reviewing some general neurophysiological studies of discourse representation I will review the evidence pertaining to anaphoric reference and presupposition accommodation and then turn to the reference of proper names, a topic that has been of particular interest to philosophers of language.

17.3 Electrophysiological studies of discourse updates

Electrophysiological measurements of event-related brain potentials (ERPs) have been used to investigate how the brain processes discourse. ERPs are average brain responses to sensory, cognitive, or motor events measured with electroencephalography (EEG), which record continuous electrical activity in the brain via electrodes placed on the scalp. ERP signals take the form of a temporal sequence of negative and positive voltage deflections compared to a pre-stimulus baseline. These deflections vary in polarity (negativity or positivity), amplitude, latency, duration, and distribution over the scalp, and these variations can be used to distinguish different cognitive processes on the basis of their ERP signature. Because EEG can provide fine-grained temporal information, EEG (and the newer, but related, MEG) provides a more accurate picture of how the brain processes discourse in real time compared to most other forms of neuroimaging and behavioral studies of language. For example, studies that compare response times to different stimuli are end-state measures and hence are unable to track language processing in real time (Kutas and Federmeier 2011).

The most extensively studied linguistic phenomena are syntactically (e.g., 'the man prepared herself for the operation') and semantically (e.g., 'he spread the warm bread with socks') anomalous sentences and discourses. Syntactic violations (e.g., 'the man prepared herself for the operation'), syntactic ambiguities (e.g., 'a nurse visited every patient'), syntactic complexity (e.g., 'the boat sailed down the river sank') and new lexical information elicit a P600 and sometimes an early left anterior negative (ELAN) effect or a left anterior negative (LAN) effect (Osterhout and Mobley 1995; Callahan 2008). A P600 effect is a positive-going deflection in frontal or parietal brain regions,

starting around 500 milliseconds after stimulus onset, peaking at 600 milliseconds, and lasting for approximately another 100–200 milliseconds. The ELAN effect is a negative-going wave that peaks around 200 milliseconds or less after stimulus onset and is most frequently elicited by stimuli that violate phrase structure (e.g., 'he is the in office'), whereas the LAN effect is a negative-going wave that peaks around 300–500 ms post word onset and is most frequently observed in cases of word-category violations (e.g. 'she likes to food').

The first ERP study that successfully manipulated semantic variables was conducted by Kutas and Hillyard (1980). They found a correlation between large-amplitude negative ERP components with the semantic plausibility of a word given the preceding sentence context. Consider, for example:

(8) a. He spread the warm bread with butter.
b. He spread the warm bread with socks.

Relative to baseline (8a), sentences such as (8b) trigger a large-amplitude negative ERP component in the centroparietal region starting around 250 ms after word onset, peaking at 400 ms, and lasting for approximately another 150 ms. This is also known as the N400 effect. Even sentences with congruent sentence final words elicit some degree of N400 activity but the amplitude is significantly larger with incongruent sentence final words than with congruent sentence final words. This relative negative shift elicited by inappropriate words compared to appropriate ones was not observed with other unexpected occurrences, such as variation of the physical attributes of the stimulus (e.g., larger font or different font type). For example, 'I shaved off my mustache and BEARD', with unexpected capital letters at the end, did not elicit an N400 effect but only the more general P300 that is observed in response to unexpected stimuli more generally. The N400 effect has subsequently been found in a wide range of variations on the original experimental setup. It was observed for spoken words, American Sign Language, and non-linguistic meaningful stimuli, such as pictures and drawings, when primed by linguistic contexts. But it was not elicited by other structured stimuli, such as music (Kutas and Federmeier 2011).

Kutas and Hillyard (1980) proposed on the basis of their original results that the N400 may be an electrophysiological marker of the interpretation of semantically anomalous information. However, it was later found that semantic anomaly is not required for the N400 effect to be elicited. For example, it was found when one word is unexpected relative to another plausible word (Kutas and Hillyard 1984). Although (9a) and (9b) are both plausible, (9b) elicits a larger N400 amplitude than (9a):

(9) a. John turned on the faucet and poured water in his glass.
b. John turned on the faucet and poured beer in his glass.

Rather than reflecting semantic anomaly, the N400 amplitude is now commonly taken to reflect the level of difficulty of integrating new semantic information into an existing semantic representation (Hagoort et al. 2009) or the level of difficulty of identifying a discourse referent for a definite expression (Burkhardt 2006). When understood in this way, the N400 effect can also be treated as an indicator of the extent to which a DRS

needs to be revised. When a hearer receives the information 'John turned on the faucet and poured...', a model is generated in which John is pouring water in his glass. The unexpected noun requires a revision of the model, which yields a larger N400 amplitude compared to baseline.

In (9b) 'beer' is unexpected relative to background information rather than previous linguistic context. Various other empirical studies have found that a word or sentence that is unexpected given world knowledge yields the same effect as a word or sentence that is unexpected given the linguistic context. For example, Hagoort et al. (2004) found that 'The Dutch trains are white and very crowded' and 'The Dutch trains are sour and very crowded' elicited the same N400 effect in Dutch speakers who know that trains are yellow. The finding that integration of new information into both existing lexico-semantic knowledge and world knowledge occurs within 500 ms, suggests that the brain does not integrate lexico-semantic knowledge prior to world knowledge. Hagoort et al. (2004) argue that this suggests that the existing DRS consists of both lexico-semantic information, expected information, and background information. New semantic information elicits an N400 response when it clashes with the existing information and an update is required. The update can consist in a rational reconstruction of the DRS or a rejection of new information if the information is considered incoherent in the context in which it occurs (Kamp et al. 2011).

In line with this suggestion, Baggio et al. (2008) found that discourse models represent the outcome of inferences anticipating the goal state of actions and events prior to receiving information about whether they take place. They compared sentences of the following kind:

(10) a. The girl was writing a letter when her friend spilled coffee on the table cloth.
b. The girl was writing a letter when her friend spilled coffee on the paper.

Spilling coffee on the tablecloth is likely neutral with respect to the writing activity. So, from (10a) the reader would be expected to conclude that the girl will accomplish her goal. The reader should thus be expected to assent to 'The girl has written a letter'. However, the inference to the goal state is *defeasible* or *non-monotonic*. As spilling coffee on the paper is likely to lead to a termination of the writing activity, the inference from that to the accomplishment of writing the letter will likely be suppressed in (10b). The reader should thus be expected to assent to 'The girl has not written a letter'. Because 'tablecloth' is less semantically expected in this context, the authors predicted that (10a) would evoke a greater N400 response. This prediction was borne out. Nouns elicited a larger N400 effect in neutral than in disabling clauses. They further found that the amplitude of the ERP effect evoked by disabling clauses is correlated with the frequency with which readers inferred that the goal state was not attained. The results suggest that the default expectation is that the actual world is an inertial world, that is, a world that is identical to the actual world up until the present moment and that continues in the way that is most compatible with the history of the world up until the present moment (Dowty 1979). The disabling clause thwarts that expectation.

Inference-dependent N400 attenuations have also been observed with multi-sentence text (Van Berkum 2012). Consider:

(11) a. Mark and John were having an argument. Mark began to hit John hard.
b. Mark and John were having an argument. Mark got more and more upset.
c. Mark and John were gambling at the casino. They won every game of blackjack.
d. The next morning John had many bruises.

A smaller N400 effect was observed when (11d) follows the explicitly supportive text in (11a) compared to the unsupportive text in (11c). However, the implicitly supportive text in (11b) also elicited a smaller N400 effect than the unsupportive text in (11c), indicating that the hearer fills in anticipatory information. The ERP studies thus appear to confirm the tenet underlying DRT to the effect that language interpretation proceeds by incremental updating of an ever-growing discourse model.

17.4 Anaphoric Reference

The majority of neuroscientific research related to reference has focused on the question of how speakers determine anaphoric reference. In cases of movement, the speaker encounters a phrase in a non-canonical position and initiates a search in upcoming context for a corresponding anaphor, that is, any pronoun or pro-form with little descriptive content (e.g., 'there is a girl in our class who is so endearing that every boy adores her'). In co-referential anaphora, the speaker encounters an anaphor and initiates a search in previous context for a linguistic antecedent (Callahan 2008). Several constraining conditions of the anaphor guide the search for an antecedent, including gender, number, and descriptive information. The difficulty of selecting an antecedent and integrating an anaphor into the existing DRS is determined by these features as well as the distance between the anaphor and the antecedent, the syntactic roles of anaphor and antecedent, the salience of the antecedent, and the number of competitors. A LAN effect or an N400 effect is elicited when it is more difficult to select and retrieve an antecedent from working or long-term memory, whereas a P300 is elicited when it is easier to select an antecedent. For example, a pronoun referring to a less frequent antecedent elicits a larger P300 effect, reflecting that the low frequency of the antecedent makes it easier to retrieve from working memory (Heine et al. 2006). Difficulties integrating an anaphor owing to increased distance between the anaphor and the antecedent are associated with a larger N400 amplitude. For example, Streb et al. (2004) found that a more distant antecedent (12b) elicited a larger N400 amplitude compared to the closer antecedent (12a), reflecting the difficulty of integrating the anaphor into the previous discourse context.

(12) a. Beate besitzt eine kleine Tierpension. Überall im Haus sind Tiere. Tom$_i$ ist ein alter Kater. Heute hat Tom/er$_i$ der Frau die Tür zerkratzt.
'Beate has a small boarding-home for animals. Everywhere in the house are animals. Tom is an old cat. Today Tom/it scratched the door of the woman.'

b. Lisa$_i$ schlender über einen Basar. Peter verkauft Edelsteine an Touristen Die Steine sind hervorragend geschliffen. Nun wird Lisa/sie$_i$ dem Händler einen Diamenten abkaufen.
'Lisa strolls across a bazaar. Peter sells gems to tourists. The gems are cut excellently. Then Lisa/she will buy a diamond from the trader.'

A similar N400 effect has also been found when the antecedent plays a difficult syntactic role compared to the anaphor. For example, the dative '*ihm*' in (13a) yields a more prominent N400 effect than the nominative '*er*', indicating greater difficulty of integrating an anaphor that plays a different syntactic role compared to the antecedent.

(13) a. Peter$_i$ besucht Julia in der Klinik. Dort hat Peter/er$_i$ dem Arzt eine Frage gestellt.
'Peter visits Julia in the hospital. There Peter asked the doctor a question.'

b. Peter$_i$ besucht Julia in der Klinik. Dort hat die Schwester Peter/ihm$_i$ das Zimmer gezeigt.
'Peter visits Julia in the hospital. There the nurse showed Peter/him the room.'

The difficulty of integrating an anaphor into an existing DRS has also been found to vary with whether the anaphor is a surface or a deep anaphor. The distinction between surface and deep anaphora was introduced by Hankamer and Sag (1976). Surface anaphors require a linguistic antecedent, whereas deep anaphors do not. Consider:

(14) a. Scenario: A man sees a woman about to jump off a bridge.
Man: 'Don't do it!' ('do it' - deep anaphor)
*Man: 'Don't do so!' ('do so' - surface anaphor)

b. Man: 'A woman was about to jump off a bridge . . .
. . . and I told her not to do it.'
. . . and I told her not to do so.'

Unlike 'do it', 'do so' is acceptable only with a linguistic antecedent, as shown in (14a)–(14b). Because it requires a linguistically represented antecedent, it is a surface anaphor. Sag and Hankamer (1984) proposed that unlike deep anaphors, surface anaphors are sensitive to syntactic parallelism, between the antecedent and the anaphor. They offered the following example:

(15) The children asked to be squirted with the hose, so
a. they were []. (VPE, surface)
b. *we did []. (VPE, surface)
c. we did it. [sentential it, deep]

Because the antecedent phrase is passive, the surface anaphor is acceptable only when it is also passive (15a). Hankamer and Sag furthermore predicted that surface anaphora, unlike deep anaphors, are sensitive to intervening discourse. Consider:

(16) a. John raked the leaves in the back yard.
(i) Later, Bill did too. (surface)
(ii) Later, Bill did it too. (deep)

b. John raked the leaves in the back yard. This was much more fun than studying for exams.
(i) ?Later, Bill did too. (surface)
(ii) Later, Bill did it too. (deep)

In (16b), in which there is intervening discourse between the anaphor and the antecedent, the surface anaphor is unacceptable.

Hankamer and Sag (1976) speculated that these and other differences between surface and deep anaphors are due to a difference in how surface and deep anaphors are processed. In the case of deep anaphors hearers access the antecedents using a non-linguistic discourse-level interpretation of the antecedent. In the case of surface anaphors, hearers access the antecedents at a linguistic level determined by surface syntactic structure. They do this by assigning the antecedent's VP structure to the site of the null anaphoric VP. For example, 'John told her not to do so' is interpreted at a linguistic level as 'John told her not to jump off the bridge'. Subsequent empirical studies, however, do not support this mechanism for surface anaphora. Woodbury (2011) carried out two fMRI studies combined with naturalness ratings and reading time measurements to test Sag and Hankamer's claims that distance affects surface anaphora but not deep anaphora and that syntactic parallelism, particularly surface word order, is required for surface anaphora but not for deep anaphora. She found that increased distance between the antecedent and the anaphor affected surface anaphora more than deep anaphora. Increased distance resulted in increased activity in several language-related areas of the brain for surface anaphors but had no added effect on those areas for deep anaphora. One explanation of this may be that the representation of surface syntax decays over intervening discourse while semantic or pragmatic information does not, which explains why only surface anaphors are unacceptable in the case of intervening discourse. The second study of semantic parallelism (i.e., word order) did not confirm the prediction that surface word order has a greater adverse effect on surface anaphora compared to deep anaphora. The most likely explanation for this is that neither surface anaphora nor deep anaphora depend on exact word order for correct processing. As Woodbury (2011) points out, this is consistent with a model of surface anaphora, according to which the syntactic form of the antecedent is not copied into the location of the anaphor, but is processed via the antecedent's deep-level syntactic representation in semantic memory. This would explain why surface anaphora is sensitive to certain factors that affect syntactic information but not semantic or pragmatic information, yet is not sensitive to surface syntax.[2]

[2] It should be noted that since the copying analysis—like syntactic analyses in general—was not intended to a have a direct analogue in processing, the above finding is not a direct contradiction of Hankamer and Sag's proposal analysis. Their primary contribution was in noting the distinction between surface and deep anaphora.

Together the neuroimaging data provide some indication that the treatment of anaphora provided by traditional formal semantics needs to be modified or augmented. Anaphoric reference cannot always be resolved in the same sentence. Moreover, syntactic models of anaphora that predict that the antecedent is copied into the location of the anaphor does not seem to be empirically supported by neuroscience. The empirical data suggests that anaphora is resolved by eliciting a search for an antecedent in the previous discourse context and that integration is affected by factors, such as the prominence and frequency of the antecedent and the distance between the anaphor and the antecedent.

17.5 Definite descriptions and presupposition accommodation

Definite descriptions can be considered a special case of anaphora in that they often seem to require an existing discourse referent to refer back to. Consider:

(17) a. John talked to a bouncer
b. John talked to the bouncer

Unlike (17a), which introduces a new discourse referent, (17b) seems to require that the existing DRS provides a discourse referent for 'the bouncer' to refer back to (Heim 1982).[3] The definite description thus triggers the presupposition that an antecedent for it can be found in the previous discourse context (existence/familiarity). Furthermore, it triggers a uniqueness presupposition to the effect that there is only one salient bouncer in the previous discourse context (uniqueness) (Roberts 2003; Abbott 2008a). By triggering these presuppositions the definite description acts like a pointer to information provided earlier.

When the existing DRS does not provide a discourse referent for a definite description to refer back to, the standard view among discourse representation theorists is that a discourse referent is added via an inferential process, known as 'presupposition accommodation'. This process updates the DRS by adding the required discourse referent (Lewis 1979a; Heim 1982). Presupposition accommodation makes definite descriptions different from anaphoric pronouns. While the latter also trigger presuppositions, these presuppositions are not as easily accommodated. Compare the following discourse fragments, taken from Kamp et al. (2011):

(18) a. Bill is a donkey owner. The donkey is not happy.
b. Bill is a donkey owner. ?It is not happy.

The first sentence implies that Bill owns one or more donkeys. In (18a) 'the donkey' triggers a uniqueness presupposition to the effect that Bill owns only one donkey. Since

[3] The referent could, of course, also be provided by non-linguistic context. But I shall set that aside here.

there is no single discourse referent for 'the donkey' to refer back to, this presupposition is accommodated by adding a discourse referent to the discourse context. In (18b) the anaphoric pronoun 'it' also triggers a uniqueness and existence presupposition. However, in spite of the fact that 'it' introduces the condition 'non-person(x)', which rules out that the pronoun refers back to Bill, the presupposition triggered by it is not accommodated. So, (18b) is infelicitous.

The presuppositions triggered by definite descriptions can be accommodated when there is a plausible inference from the previous discourse context to the presupposition. If, however, the presupposition is not a plausible inference from the DRS, then it cannot easily be accommodated by the DRS. For example, if the context of (17b) is that John went to the local elementary school, then (17b) is much harder to interpret, as in (19b):

(19) a. John went to a local club. He talked to the bouncer.
b. John went to the local elementary school. ?He talked to the bouncer.
c. John went to the local elementary school. He talked to one of the other parents, a bouncer from the local club. The bouncer said that he was there to complain about his kid's new teacher.

There is evidence from neuroscience supporting the theory of presupposition accommodation. A larger N400 amplitude has been found for definite descriptions that require accommodation (19a) compared to definite descriptions for which there is a salient discourse referent to refer back to (19c).[4] The N400 effect is most prominent when there is no plausible inference from the immediate discourse context that would allow construction of a new discourse referent for the definite description (19b) (Burkhardt 2006). A P600 effect has been observed both when an indefinite description is used to introduce a new discourse referent (19c) and when presupposition accommodation is required to process a definite description (19a) but not when a discourse referent was already available (Burkhardt 2006; Schumacher 2009).

Van Berkum et al. (1999) examined the uniqueness presupposition of definite descriptions. Participants were presented with sentences that contained a singular definite description embedded in stories that varied in the number of suitable referents they introduced for a singular definite, as in:

(20) a. De aardige reus werd onderweg vergezeld door een elfje (een fee) en een kabouter. Het elfje (de fee) had zich vastgeklampt aan zijn bovenarm, de kabouter had zich genesteld in een comfortabele broekzak. De reus waarschuwde het elfje dat ze niet moest vallen.

[4] This shows that discourse referents that have been linguistically introduced (and are still in short-term memory) require less effort to associate with a definite form than those that require an extra inferential process which makes an indirect association. But it would follow equally well from a theory that simply requires association with a unique referent. Constructing a new unique reference by inference would also require more effort than identification with a recently introduced referent that is still in short-term memory; so while these facts do support the theory of presupposition accommodation, they don't support it over alternative theories of definiteness that require unique identifiability, not previous familiarity.

'On the road, the gentle giant was accompanied by an elf (fairy) and a goblin. The elf (fairy) had clung [itself] to his upper arm, the goblin had ensconced itself in a comfortable trouser-pocket. The giant warned the elf that she shouldn't fall off.'

b. De aardige reus werd onderweg vergezeld door twee elfjes (feeën). Het ene elfje (de ene fee) had zich vastgeklampt aan zijn bovenarm, het andere (de andere) had zich genesteld in een comfortabele broekzak. De reus waarschuwde het elfje dat ze niet moest vallen.

'On the road, the gentle giant was accompanied by two elves (fairies). One of the elves (fairies) had clung [itself] to his upper arm, the other had ensconced itself in a comfortable trouser-pocket. The giant warned the elf that she shouldn't fall off.'

The results showed that the ambiguous two-referent condition triggered a sustained negative deflection, broadly distributed over the scalp but with a pronounced frontal component, and starting at about 200 ms post word onset but without a well-defined peak. This effect has come to be known as the 'Nref effect' and has been taken to indicate difficulties identifying a unique discourse reference for definite descriptions and pronouns to refer back to (Nieuwland et al. 2007; Van Berkum et al. 2007). The original data did not establish whether the Nref effect was related to the lack of salience of a unique discourse referent or would emerge as long as two discourse referents had been introduced in the previous context. However, Nieuwland et al. (2007) subsequently looked at whether the Nref effect would disappear if the ambiguity was eliminated prior to the occurrence of the target definite description. Participants were presented with stories in which one of the potential discourse referents of a definite description was made more salient by having one individual leave the scene of the protagonist, for example:

(21) David had asked the two girls to clean up their room before lunchtime. But one of the girls had been sunbathing in the front yard all morning, and the other had actually just driven off in his car for some serious downtown shopping. As he gazed at the empty driveway, David told the girl . . .

The Nref effect was found to disappear when one discourse referent was made salient. In (21), for example, there is only one salient discourse referent available for the definite description 'the girl' to refer back to. This shows that the Nref effect is not triggered simply by the initial availability of two potential discourse referents but is triggered when there is not a unique salient discourse referent for it to refer back to.

The data from neuroscience thus give us some reason to think that definite descriptions trigger both a familiarity and a uniqueness presupposition. Speakers accommodate plausible familiarity presuppositions that are not already satisfied by the previous discourse. When a presupposition is implausible given the existing discourse representation, it is more difficult for the reader to accommodate it. At this point the speaker

may accommodate the presupposition or choose not to integrate the new information into the existing discourse representation. Difficulties integrating a definite description can also arise because of lack of saliency of the discourse referent, which provides some evidence for the hypothesis that definite descriptions trigger both a familiarity and a uniqueness presupposition.

17.6 The Reference of Proper Names

In philosophy the majority of debates about reference have centered around the reference of proper names. Until the 1970s the dominant theory was the descriptive theory, according to which proper names refer indirectly via descriptions or concepts. Saul Kripke's (1980) lectures "Naming and Necessity" won many philosophers over on the side of direct referentialism. On the 'new' standard view, proper names refer directly to an individual by virtue of a causal-historical connection, to their uses, and an original baptismal event. One of the most influential arguments for direct reference theories of proper names is the modal argument. Consider the following two sentences:

(22) a. Aristotle is Aristotle
b. Aristotle is the teacher of Alexander the Great

(22a) and (22b) have different modal profiles. As Aristotle could not have failed to be Aristotle, (22a) is metaphysically necessary. By contrast, since Aristotle might not have been the teacher of Alexander the Great, (22b) is metaphysically contingent. As (22a) and (22b) have different modal profiles, they cannot be semantically equivalent. So, 'Aristotle' and 'the teacher of Alexander the Great' cannot be semantically equivalent. But the same argument can be made, regardless of which descriptions we propose are semantically equivalent to 'Aristotle'. So, 'Aristotle' is not semantically equivalent to any description. Or so the argument goes.

There have been many subsequent replies to this argument on behalf of descriptive theories of reference. For example, identifying proper names with rigidified descriptions (e.g., 'the actual teacher of Alexander the Great') or clusters of descriptions seems to avoid this particular problem (Reimer 2010; Abbott 2011). But it is fair to say that none of the replies to Kripke's modal argument has gained widespread support. There is, indeed, still a considerable number of philosophers who adhere to a direct reference theory of proper names (http://philpapers.org/surveys/metaresults.pl).

In DRT, proper names are treated as predicates (Kamp and Reyle 1993).[5] The first occurrence of a proper name either introduces a discourse referent x and a condition N(x) or refers back to a previously introduced discourse referent, as in 'They elected a

[5] In Kamp (1981), proper names are treated as individual constants and introduce the condition x is identical to N.

new chair. His name was "Kelvin". Kelvin chaired the committee for two years'. Additional occurrences of the name refer back to the discourse referent, which refers to an actual individual just in case that individual satisfies all the conditions introduced by co-referring expressions. Despite the semantic differences between the direct reference theories and DRT, the condition introduced by names, N(x), might turn out to map onto an individual only if there is a causal-historical relation between the occurrence of 'N' and that individual (this is indeed an implication of Kamp's notion of anchoring). However, DRT inflicts the further requirement that the individual must satisfy any other constraints imposed by co-referring expressions. For example, a discourse referent introduced by a proper name (e.g., 'Obama') and subsequently referred to by a definite description (e.g., 'the president') can have a particular value only if the constraining conditions of definite description (e.g., being the president) are satisfied. This requirement imposed by DRT has the virtue that even if a name turns out to have no actual referent, it still has a meaning contributed by the name-qua-predicate and other constraining conditions. In traditional formal semantics, proper names are individual constants. So, if they have no actual referent, they are 'empty' and therefore meaningless.

Whether a causal-historical relation between a name and an individual is required in order for the name to map onto the individual is an open question but not one that has been directly investigated from the perspective of neuroscience. Other empirical approaches have been used to shed light on this question. In the standard paradigm of experimental philosophy, Machery et al. (2004) presented a version of Kripke's Gödel case to forty undergraduates at Rutgers University and forty undergraduates from the University of Hong Kong. The specific case they used was the following:

> Suppose that John has learned in college that Gödel is the man who proved an important mathematical theorem, called the incompleteness of arithmetic. John is quite good at mathematics and he can give an accurate statement of the incompleteness theorem, which he attributes to Gödel as the discoverer. But this is the only thing that he has heard about Gödel. Now suppose that Gödel was not the author of this theorem. A man called 'Schmidt', whose body was found in Vienna under mysterious circumstances many years ago, actually did the work in question. His friend Gödel somehow got hold of the manuscript and claimed credit for the work, which was thereafter attributed to Gödel. Thus, he has been known as the man who proved the incompleteness of arithmetic. Most people who have heard the name 'Gödel' are like John; the claim that Gödel discovered the incompleteness theorem is the only thing they have ever heard about Gödel. When John uses the name 'Gödel', is he talking about:
>
> (A) the person who really discovered the incompleteness of arithmetic? or
> (B) the person who got hold of the manuscript and claimed credit for the work?
>
> Machery et al. (2004: B6)

The team found that intuitions varied across culture. The Westerners were significantly more likely than the Chinese to give causal-historical responses (response A).

The researchers speculate that the difference in intuitions may be grounded in differences in cognition between Westerners and Chinese.

The research casts doubt on a certain methodology used in the philosophy of language, known as 'intuitions about cases', but it does not provide any obvious starting point for neuroscience to investigate the reference of proper names. It remains an open question how most speakers in John's situation who learned in college that Gödel is the man who proved the incompleteness theorem and who then subsequently learn that a man named 'Schmidt' discovered the incompleteness theorem would update the existing DRS (i.e., the mental representation) to reflect this fact. If speakers know nothing else about Gödel, it is plausible that they would integrate the name 'Schmidt' into the DRS as another name for 'Gödel'. In that case, their uses of 'Gödel' and 'Schmidt' would both fail to refer to any actual individuals, as Gödel and Schmidt are different people. So, answering this sort of question about brain processing would not tell us anything useful about the reference of proper names. If participants learned that the man who was baptized 'Gödel' didn't discover the incompleteness theorem but that a man named 'Schmidt' discovered it, they would in all likelihood reject the information they learned in college but this sort of discourse revision would merely reflect their knowledge of how we name people. This type of revision would be consistent with the condition 'Gödel(x)' mapping onto an individual only if he is named 'Gödel'. It wouldn't demonstrate a required historical-causal relation between 'Gödel' and that individual.

The most promising neuroscientific approach to the reference of proper names may be to look at how updating takes place with proper names in the scope of modals (see Roberts 1989; and Kamp et al. 2011 for some testable predictions). This type of research should look not only at proper names in alethic modal contexts but also embedded under epistemic operators such as 'belief'. Kripke's causal-historical requirement is considerably less plausible for names embedded in epistemic modal contexts than in alethic modal contexts. For example, 'I told John that Allen Stewart Konigsberg is Woody Allen but he doesn't believe me. He thinks Allen Stewart Konigsberg is Woody Allen's father' cannot introduce the same discourse referent for all occurrences of 'Allen Stewart Konigsberg' and 'Woody Allen'. Occurrences of names embedded in non-alethic modal contexts may thus require their own discourse referents (Roberts 1989; Kamp et al. 2011).

17.7 Methodological Considerations

One of the main complaints about electrophysiological studies of semantics has been that they don't study linguistic information. Pylkkänen et al. (2011), for example, point out that the traditional way of studying linguistic phenomena has been to identify grammatically ill-formed sentences on the basis of the judgments of native speakers and then generate linguistic theories to explain why sentences have the grammar they do. The problem with the electrophysiological studies that claim to be studying

semantics, Pylkkänen et al. argue, is that the alleged semantically ill-formed stimuli are, in fact, not ill-formed in the formal sense of linguistic theory. They do not violate syntactic rules but simply conflict with world knowledge. To back up this claim, they point to the original studies performed by Kutas and Hillyard (1980). "He spread the warm bread with socks" is not grammatically ill-formed. It's a perfectly grammatical sentence. The reason it gives rise to an N400 effect is that the participants know that socks are neither edible nor spreadable but this fact is a fact about the world, not a fact about language. They add that they are unaware of any semantic theory that would claim that typical N400 sentences are semantically ill-formed. The idea that the N400 effect is a semantics-related ERP, they say, is not conducive to the project of generating theoretically grounded models of how the brain processes language.

However, this criticism seems to presuppose that a model of how the brain processes language must be grounded only in syntax or knowledge of truth-conditions. But such a model would provide a rather limited picture of language processing, and would have little to do with semantics, more broadly speaking. It is widely agreed that perfectly grammatical sentences can be semantically ill-formed, as witnessed by Chomsky's 'colorless green ideas sleep furiously'. Likewise, sentences that have possible truth-conditions may be meaningless to speakers (e.g., "he spread the warm bread with socks"). In DRT a sentence is semantically ill-formed when it cannot coherently be added to the existing discourse representation. Outside of the realm of fiction, there are few contexts that would be able to accommodate a sentence like "he spread the warm bread with socks", which ordinarily makes the sentence semantically ill-formed.

Now, it is true that in the original ERP studies, unlike in later studies, only single sentences were presented in the experimental setting. But the participants used in the study were not blank slates. They came into the experimental setting with existing discourse representation structures, and DRT predicts that sentences are analyzed and integrated relative to existing representations. Furthermore, in DRT the discourse representation is supposed to serve not only as a mental model of the discourse but also as a mental model of the world when supplemented by discourse-independent information. So when seen against the background of DRT or other dynamic theories, the N400 effect can indeed be reliably interpreted as a semantics-related ERP.

That said, no one is claiming that the N400 effect is exclusively semantic. As Van Berkum et al. (2007) point out, there are no ERP effects that correspond uniquely to the exact level of language representation manipulated in experiments (syntax, semantics, reference, pragmatics). Many regularities have been unearthed, for example, difficulties of semantic and syntactic integration tend to be manifested as N400 and P600 effects respectively. However, language processing occurs at multiple levels of language representation, and what experimental researchers define as processing at one level (e.g., semantic) can affect processing at a different level (e.g., syntactic) and may even sometimes be manifested only at that level.

A different methodological question is that of whether the standard distinction between a level of semantics and a level of pragmatics can be maintained within dynamic semantic theories. There are, of course, numerous linguistic phenomena

that depend on pragmatic factors for their interpretation. But the question is whether there are good grounds for thinking that the brain processes semantic and pragmatic information by the same processing sites and in a particular order. There are studies that suggest that the brain processes semantic and pragmatic information differently (Hunt et al. 2013; Politzer-Ahles et al. 2013). Hunt et al. (2013), for example, had participants read existentially quantified sentences that had both a semantic meaning ('at least one') and a pragmatic meaning ('not all') presented next to an image with content related to the sentence. Sentences like 'The boy cut some of the steaks in this story' were presented together with images that made both the semantic and pragmatic interpretations true (when the boy cut some but not all), images that made neither true (when the boy cut none), or images that made the semantic interpretation true but the pragmatic one false (e.g., when the boy cut all the steaks). The largest N400 effects were found when the sentence was false both on its semantic interpretation and its pragmatic interpretation. When the sentence was false on its pragmatic interpretation but true on the semantic interpretation, there was an intermediary N400 amplitude. The smallest amplitude was observed for pictures that made both the semantic and the pragmatic interpretation true. Using a similar paradigm Politzer-Ahles et al. (2013) further found that pragmatically false but semantically true sentences elicited a sustained posterior negative component that was distinct from the N400 effect. The researchers propose that the sustained negativity reflects that the integration requires cancellation of the pragmatic inference and retrieval of the semantic meaning. While these results indicate that the brain processes semantic and pragmatic aspects of meaning differently, the pragmatic aspects appear to be calculated fast enough to affect subsequent interpretation.

The findings suggest that although the brain is capable of distinguishing pragmatic and semantic interpretations, new information that is integrated into the existing DRS may be the result of pragmatic inferences and the fact that the pragmatically inferred information is sometimes preferred, unless this information is inconsistent with the existing discourse representation. This is consistent with the basic tenet of DRT, which does not necessitate a rigid distinction between semantics and pragmatics. This is so insofar as it captures how we process what the speaker appears to attempt to convey to us rather than the semantic properties of language per se (cf. Recanati 2004, 2012).

17.8 Conclusion

Discourse Representation Theory (DRT) and other dynamic semantic theories have been considerably more successful than traditional semantics in accommodating linguistic phenomena, such as anaphora, presupposition, and event order. DRT predicts that discourse gives rise to a mental representation that is ever growing and continually revised. On this view, sentences have meanings only derivatively, depending on their capacity to change the existing discourse representation. The mental

representations, not sentences, have a truth-conditional impact. A discourse representation consists of discourse referents and conditions introduced by linguistic expressions. New linguistic information either introduces new discourse referents into the discourse representation or refers back to discourse referents in the existing representation. The discourse referents refer to entities in the external world when these entities satisfy the referent and the conditions imposed on it throughout the entire discourse representation.

Because DRT takes discourse representation to be the most central notion in semantics, it is suitable as an object of investigation for neuroscience. Since Kutas and Hillyard (1980) discovered that difficulty of integrating new linguistic information into an existing discourse context has its own ERP signature—the N400 effect, a large body of research has looked at the ERP signatures associated with different linguistic phenomena. The majority of neuroscientific studies relating to reference have been concerned with anaphoric reference and related phenomena. The results of these studies lend evidence to the general framework of DRT. DRT implies that a discourse referent introduced by a referring expression can refer to a single individual only if that individual satisfies all the conditions imposed by co-referential expressions.

Although the focus here has primarily been on DRT and discourse referents, however, it is quite plausible that speakers very often manage to refer to extra-mental entities even when most of the constraining conditions in the discourse representation are not satisfied. To borrow the classical example of Donnellan (1966), a speaker may succeed in picking out a man who is drinking water using the description 'the man with the martini'. This sort of speaker reference is useful in contexts in which the purpose of the description is not to describe an entity but to pick it out for further conversation (Kripke 1977). The bulk of empirical work in this area has been in developmental psychology and computer science looking at how speakers generate referring expressions for singling out entities in a visual scene (see e.g., Dale and Reiter 1995). However, the ERP paradigms for studying discourse could easily be adjusted to look at event-related brain potentials in individuals generating referring expressions for picking out objects in a visual scene.

Acknowledgments

I am grateful to Barbara Abbott and Jeanette Gundel for invaluable comments on an earlier draft of this paper.

THE OXFORD HANDBOOK OF

REFERENCE

Edited by

JEANETTE GUNDEL

and

BARBARA ABBOTT

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First Edition published in 2019

Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2018943672

ISBN 978-0-19-968730-5

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY