Language and Linguistics Compass 1/1–2 (2007): 84–99, 10.1111/j.1749-818x.2007.00009.x

The Form of Referential Expressions in Discourse

Amit Almor* and Veena A. Nair
University of South Carolina

Abstract

Most instances of real-life language use involve discourses in which several sentences or utterances are coherently linked through the use of repeated references. Repeated reference can take many forms, and the choice of referential form has been the focus of much research in several related fields. In this article we distinguish between three main approaches: one that addresses the 'why' question – why are certain forms used in certain contexts; one that addresses the 'how' question – how are different forms processed; and one that aims to answer both questions by seriously considering both the discourse function of referential expressions, and the cognitive mechanisms that underlie their processing cost. We argue that only the latter approach is capable of providing a complete view of referential processing, and that in so doing it may also answer a more profound 'why' question – why does language offer multiple referential forms.

Coherent discourse typically involves repeated references to previously mentioned referents, and these references can be made with different forms. For example, a person mentioned in discourse can be referred to by a proper name (e.g., *Bill*), a definite description (e.g., *the waiter*), or a pronoun (e.g., *he*). When repeated reference is made to a referent that was mentioned in the same sentence, the choice and processing of referential form may be governed by syntactic constraints such as binding principles (Chomsky 1981). However, in many cases of repeated reference to a referent that was mentioned in the same sentence, and in all cases of repeated reference across sentences, the choice and processing of referential form reflects regular patterns and preferences rather than strong syntactic constraints. The present article focuses on the factors that underlie these patterns. Considerable research in several disciplines has aimed to explain how speakers and writers choose which form they should use to refer to objects and events in discourse, and how listeners and readers process different referential forms (e.g., Chafe 1976; Clark & Wilkes 1986; Kintsch 1988; Gernsbacher 1989; Ariel 1990; Gordon, Grosz & Gilliom 1993; Gundel, Hedberg & Zacharski 1993; Garrod & Sanford 1994; Gordon & Hendrick 1998; Almor 1999; Cowles & Garnham 2005). One of the central observations in this research is that there exists an inverse relation between the specificity of the referential expression and the salience of the referent in the context of the discourse (Givon 1976; Ariel 1990; Gundel et al. 1993). For example, pronouns, which convey only limited information about their referent, are primarily used for referents that are salient in the discourse. In contrast, proper names and definite descriptions, which in most cases unambiguously identify their referent, are often used to introduce new referents to the discourse or to make reference to discourse referents that are not salient (Ariel 1990; Gundel et al. 1993; Garrod & Sanford 1994; Gordon & Hendrick 1998). The seemingly complementary distribution of different referential forms has led many researchers to postulate that different forms are processed differently and serve distinct discourse functions, but views of what are these differences in processing and function vary between theories and between disciplines. The different views can be roughly divided by whether they address the 'why' question (why do certain forms tend to be used in certain contexts?) or the 'how' question (how are different forms processed?). In the rest of this article, we describe how these two types of views apply to the study of reference and then show how they can be profitably combined into a detailed view of referential form as reflecting a balance between discourse function and processing cost (Almor 1999). The combined view we advocate considers the multiplicity of referential forms as a solution to a problem created by memory constraints and the requirements of sequential communication. We conclude by discussing the implications of this view for a different and arguably a more profound 'why' question – why does language offer multiple referential forms?

Why Do Certain Forms Tend to Be Used in Certain Contexts?

Views that attempt to address the 'why' question mostly consider language as a collaborative activity in which participants use linguistic devices to achieve a common goal, which is a successful linguistic exchange. Referential form can be considered as being one such device and some researchers working in this tradition consequently view referential form as providing an additional channel of communication between speakers and listeners and between writers and readers (e.g., Ariel 1990; Vonk, Hustinx & Simons 1992; Gundel et al. 1993). In this view, speakers and writers use this channel to convey helpful information about the structure of the discourse, such as common ground and topic shifts, and listeners and readers rely on this channel for updating and maintaining their representation of the discourse structure (Clark & Wilkes 1986; Vonk et al. 1992; Clark 1996). For example, the use of a pronoun can be interpreted as indicating topic continuity (Gordon et al. 1993), and the use of a fuller reference can be interpreted as indicating the speaker's or writer's intention to shift topic, which will result in a different entity becoming the most salient in readers' discourse representation (Vonk et al. 1992).

In this view, referential form is seen as a product of cooperation between speakers and listeners, with speakers choosing the form that they think would maximize the likelihood of successful communication, and listeners likewise assuming that speakers would always choose a form that would maximize communicative success (Clark & Wilkes 1986). This view is closely tied to work in pragmatics, where referential form has often been argued to constitute a marker of the referent's memory accessibility (Ariel 1990), or its 'cognitive status' (Gundel et al. 1993). According to Ariel's Accessibility Theory (Ariel 1990), each possible referential form is associated with a unique value along an 'accessibility marking scale' such that higher positions on the scale correspond to greater referent accessibility. Full names and definite descriptions are on the low end of this scale, demonstrative expressions (*this*, *that*) are on the middle, and pronouns (*he*, *she*) are on the high end. Referent accessibility at each point during discourse is affected by four factors: (i) distance from the last mention of the referent, (ii) competition with other possible referents, (iii) salience of the referent in the context, and (iv) unity – whether the previous mention of the referent is in the same or previous sentence, or in the same or previous paragraph. According to this theory, speakers and writers choose the form of referential expressions so as to encode the accessibility of the referent and thus aid listeners and readers to successfully identify that referent.

A similar approach was proposed by Gundel et al. (1993) who argued that it is the extent to which referents are mutually identifiable to discourse participants, and not just memory accessibility, that affects the choice and processing of referential form. The degree to which a referent is identifiable constitutes its cognitive status and each status is necessary and sufficient for the appropriate use of a specific form. The different cognitive statuses form a hierarchy such that a higher status on the hierarchy (meaning more easy to identify referents) entails all the lower statuses as well. For example, a referent with the most restrictive 'referent in focus' status, which is associated with the use of unstressed pronouns, also inherits all the less restrictive cognitive statuses such as 'familiar' (associated with demonstratives) and 'type identifiable referent' (the least restrictive status, which is associated with indefinite nouns).

In both the accessibility and cognitive status approaches, the way referential form is used to convey the accessibility or cognitive status of the referent is tied to the use of pragmatic principles. One such principle is the Gricean 'maxim of quantity' according to which speakers and writers should choose the form of reference that is sufficiently informative for their purpose but not more informative than is necessary (Grice 1975). This principle explains why the most informative expression that is the lowest in the hierarchy is not always used. However, this principle in and of itself does not explain why the accessibility scale or the cognitive status hierarchy exist to begin with, and why are they organized the way they are. A different approach that has become more prevalent in recent years is based on Relevance Theory (Sperber & Wilson 1995), which reduces all pragmatic principles to a single 'be relevant' principle that is based on balancing processing cost with 'cognitive effect' or communicative function. This approach goes one step further than the Gricean approach in that it tackles the question of why the accessibility scale and the cognitive hierarchy exist by tying each position on the scale or hierarchy to a point of balance between communicative function and the processing cost associated with achieving this function. However, work in this area has not provided an independent and cognitively based notion of processing cost and in this sense has made a circular argument where cost is defined on the basis of the scale or hierarchy, which it is supposed to explain. Moreover, work in this area has typically not addressed issues related to how people actually process referential form during production and comprehension. Indeed, much of the empirical evidence considered in this work comes from corpus analyses or observations of written and spoken discourse, which provide a good measure of the outcome of processing, and of the possible function of referential expressions, but do not provide much information about the actual course of processing and the cost associated with it. Fortunately, there is considerable work in psycholinguistics that has examined referential processing using empirical methods geared towards answering the 'how' question – how are referential expressions processed? Importantly, this work allows the notion of cost to be independently construed. We now turn to discuss some of this work but will then argue that a complete picture of how referential expressions are processed can only be drawn on the basis of considering both cost and function and how they are balanced.

How Are Different Forms Processed?

Some approaches to referential processing rely on general psycholinguistic frameworks in language production and comprehension. For example, Arnold's Expectancy hypothesis (Arnold 1998, 2001; Arnold, Fagnano & Tanenhaus 2003) distinguishes between the production and comprehension of referential expressions, emphasizing factors that are known to affect language production such as the availability of information, and processes that are known to occur in comprehension such as the statistical weighing of different cues. In this view, many factors can affect the comprehension of referential expressions by modulating the likelihood of reference to different referents, and as a matter of convention, the more likely a referent is to be mentioned, the more reduced will the form of the referential expression be. This view has been supported by some elegant experiments showing that listeners quickly adjust their referential expectations based on speakers' disfluencies, presumably because listeners interpret these disfluencies to indicate the retrieval of new information that is associated with a topic shift (Arnold et al. 2003, 2004). In line with connectionist constraint-based views of language comprehension (MacDonald, Pearlmutter & Seidenberg 1994; Trueswell & Tanenhaus 1994), the Expectancy hypothesis leads to a notion of processing cost that is based on predictability – less predictable references require more information and therefore tend to be referred to with fuller references. However, as is often the case with statistically based approaches, it is not clear to what extent the statistical regularities underlie the processing of referential expressions, or simply reflect patterns generated by other constraints. Moreover, as we shall see later, there is good evidence for the role of constraints that are related to memory function in the comprehension of referential expressions and these constraints cannot be easily explained by predictability of reference. Thus, while this work makes a convincing argument that the comprehension of referential expressions is sensitive to expectations constructed from available cues, it nevertheless leaves open the question of what are the underlying mechanisms.

The question of how referential expressions are processed has also been considered as part of general theories of discourse processing and representation. For example, in Gernsbacher's Structure Building theory (Gernsbacher 1990), full repeated reference is argued to be a useful way to enhance the memory representation of referents. This theory emphasizes the role of referential expressions as memory operators, mostly in enhancing the activation of referents through repeated reference. In Sanford and Garrod's Memory Focus model (Sanford & Garrod 1981), pronouns are immediately associated with referents in *explicit focus*, a special memory store for recently activated information, and fuller references such as names and definite descriptions lead to a search of *implicit focus*, which is a memory store for less active information that was mentioned earlier in the discourse or that is closely associated with the current topic. This theory thus postulates that pronouns and full names are associated with different memory systems and different memory processes – pronouns are associated with a random-access-like memory system, and fuller references are associated with a searchable list-like memory system. In McKoon and Ratcliff's Minimalist framework (McKoon & Ratcliff 1992), pronouns and full repeated references are processed at different times with the reference resolution of pronouns viewed as a controlled process that in some cases can be postponed to the end of utterances, and full repeated references viewed as an automatic process that occurs immediately upon encountering the referential expression (Greene, McKoon & Ratcliff 1992; McKoon, Gerrig & Greene 1996). These theories are clearly useful in that they attempt to place referential processing in the broader context of discourse processing and general cognitive processes and representations. However, as it turns out, referential processing does not always operate according to established and expected memory principles and in some cases shows effects that appear contrary to commonly observed priming effects related to repetition and semantic overlap (Gordon, Hendrick & Foster 2000).

Most strikingly, in contrast to Gernsbacher's view of fully specified referential expressions as memory enhancers (Gernsbacher 1989), pronouns, despite their inherent ambiguity and low informative value, are sometimes processed faster than fuller referential expressions. For example, Gordon, Grosz and Gilliom (1993) discovered that when reference is made to the most salient referent in the discourse, pronouns are read faster than repeated proper name references. In their study, Gordon, Grosz and Gilliom used a self-paced reading task with short discourse items similar to the 'Bruno the bully' text below, and found longer reading times for the last sentence in the 'repeated name' version than in the 'pronoun' version, an effect which they termed the 'repeated name penalty':

<u>'Bruno the bully' sample text – repeated name version</u>
Bruno was the bully of the neighborhood.
Bruno chased Tommy all the way home from school one day.
Bruno watched Tommy hide behind a big tree and start to cry.

<u>'Bruno the bully' sample text – pronoun version</u>
Bruno was the bully of the neighborhood.
He chased Tommy all the way home from school one day.
He watched Tommy hide behind a big tree and start to cry.

The work of Gordon et al. (1993) established that this effect occurred in reference to proper names, and later research (Almor 1999) extended the repeated name penalty to definite descriptions. This effect has now been demonstrated with written short texts (e.g., Gordon et al. 1993; Almor 1999), with longer textual materials (Shapiro & Milkes 2004), and recently also with texts that refer to depicted objects (Peters & Almor 2006). This effect was also demonstrated in spoken language comprehension where it is reflected in worse memory for discourses with repeated references to salient referents than for discourses in which repeated reference was made to non-salient referents (Almor & Eimas 2007). Recently, this effect has also been claimed to be reflected in event-related potentials (ERP), which are physiological measures of readers' brain electric activity. Specifically, the repeated name penalty has been argued to be reflected by increased N400 in response to repeated name references to salient referents than to non-salient referents (Swaab, Camblin & Gordon 2004). The N400 is a negative waveform, which typically peaks 400 milliseconds after a critical word and is commonly interpreted as a marker of the difficulty of integrating the word with its preceding discourse context. The repeated name penalty has clearly become an important finding that requires theorists to consider the special issues that are involved in referential processing. We now turn to review some of these theories.

When it was originally reported, the repeated name penalty was used to support the claims of Centering theory (Grosz, Joshi & Weinstein 1995), which is a framework developed in computational linguistics as an effective algorithm for reference resolution. Centering theory postulates that all the referents in the discourse are ordered by their relative salience in the discourse context, and that each utterance includes one special reference that connects the utterance to previous discourse, and which is termed 'the backward looking center'. For example, in the 'Bruno the bully' text above, Bruno is the backward looking center in sentences 2 and 3. Centering theory specifies special purpose rules that determine which referential form should be used on the basis of the relative salience of the referents in the previous sentence. For example, in its original version, Centering theory stated that if the backward looking center refers to the most salient referent in the previous sentence, then the backward looking center must be referred to by a pronoun. Gordon et al. (1993) explained the repeated name penalty as the consequence of violating this rule.

In later research, Gordon and Hendricks (1998) developed the Discourse Prominence theory that aimed to explain both inter- and intrasentence co-reference relations on the basis of formal construction rules adapted from Discourse Representation theory (Kamp & Reyle 1993), and Centering theory's notion of graded referent prominence (Grosz et al. 1995). In Discourse Prominence theory, referential processing is governed by a set of construction rules that are applied as part of building and maintaining a discourse representation. According to Discourse Prominence theory, when a proper name is encountered, a construction rule generates a new representation in the discourse model, but when a pronoun is used, a different construction rule searches for a matching referent in the list of previously mentioned referents in decreasing order of prominence. This model explains the repeated name penalty as reflecting the application of a special 'equivalence' construction rule that reconciles the representation of the new referent generated by the proper name construction rule, and the representation of the referent already in memory. This rule searches the list of referents in ascending prominence order, thus involving longer searches for prominent referents. Both Centering theory and Discourse Prominence theory emphasize special purpose syntactic procedural rules that operate solely on the basis of the form of referential expressions. Although the detailed specification of computational operations in these two approaches forms an appealing model of how reference is processed and what factors affect processing cost, these approaches do not consider the discourse function of referential expressions, and as we shall see, are unable to account for differences in the processing of referential expressions that are not based on word class (e.g., definite descriptions *vs*. pronouns).

Bringing the 'Why' and 'How' Together

In contrast to approaches that emphasize formal rules, the Informational Load hypothesis (ILH; Almor 1999, 2000, 2004) attributes the repeated name penalty to a combination of pragmatic principles and the architecture of working memory. In the spirit of Relevance theory (Sperber & Wilson 1995) and Accessibility theory (Ariel 1990), the ILH argues that referential processing is governed by balancing discourse function with processing cost. The discourse function of referential expressions is identifying the referent but possibly also adding new information to the discourse about the referent or about the speaker's or writer's attitude toward the referent. The ILH further attributes the processing cost of referential expressions to the architecture of the working memory system that is used for discourse representation. At some stage during reference resolution, both the representation of a new reference (or some part of it) and the representation of the previous discourse and the referent are simultaneously active before integration can take place. According to the ILH, this stage in processing is prone to interference that is affected by the semantic overlap between the representation of the referential expression and the representation of the discourse and the referent in memory, as well as by the amount of activation in memory of both representations. Consequently, while semantic overlap and memory activation may initially aid the lexical processing of the referential expression and may also help identifying the referent, they might also accrue processing cost during integration. The working memory resources used for the construction and updating of the discourse model play an important part in this theory and in how it construes processing cost.

Although in its original formulation (Almor 1999) the ILH did not specify how the processing cost related to semantic overlap could be calculated, later work (Almor 2004) proposed a specific method for calculating this cost. This method incorporated several assumptions. First, this method assumed distributed representations such that referents and referential expressions are represented as vectors in a multi-dimensional semantic space. Second, this method assumed a memory model in which capacity is determined by not only the number of stored items but also by their activation (e.g., Just & Carpenter 1992) such that higher activation can result in more competition and therefore in higher cost. Third, overall activation was assumed to be affected by not only memory activation but also by the amount of semantic detail in the representation. This is because representations that are rich in semantic detail require bigger chunks of the capacity limited representational space than representations that involve little semantic detail.

Based on these assumptions, processing cost was expressed as a function of two quantities. The first was the overlap between the semantic representations of the referential expression and the pre-existing representation of the referent. The second was the amount of activation of both representations. The usefulness of this quantification was tested in a series of simulations reported in Almor (2004), in which feature lists generated by human participants were used to model the various empirical findings reported in Almor (1999) and in Almor et al. (1999).

The work described so far attributed processing cost to a direct interference between the representation of the referential expression and the prior representation of the discourse and referent. However, the results of a recent memory study (Almor 2005), which examined the effect of semantic overlap on word list recall, suggest that semantic overlap might in fact enhance the pre-existing representation of the referent, but at the expense of the representation of the other information in the discourse, thus making overall integration difficult. By this alternative interpretation, interference is caused through the suppression of the memory representation of the discourse information surrounding the repeated or semantically overlapping referential expression (Gernsbacher 1989), but not the representation of the referent itself. In support of this interpretation, research on the relation between referential form and memory accessibility has shown that repeated reference, which represents high semantic overlap, reduces the accessibility of the other information in the discourse beside the referent in comparison to pronominal reference, which represents minimal semantic overlap (Vonk et al. 1992).

This revised notion of interference can also explain the disparity in the literature between studies of reference that employed a probe recognition task and studies that employed reading time-based measures. In the probe recognition task, participants indicate whether a target word appeared in a sentence they just read. Several studies that used this task to examine referential processing found facilitation associated with a semantically overlapping reference to a memory salient referent (e.g., Gernsbacher 1989; Greene et al. 1992). In contrast, several studies that used reading time based measures found slower processing associated with a semantically overlapping reference to a memory-salient referent (Gordon et al. 1993; Gordon & Chan 1995; Gordon & Scearce 1995; Almor 1999). Although Gordon, Hendrick and Ledoux (2000) argued that the disparity between the results of the two tasks is due to inherent flaws in the probe recognition task, the present interpretation suggests that probe recognition and self-paced reading simply measure different stages in processing. The probe recognition task likely taps the initial processing of the referential expression, whereas many reading-based measures tap the integrative processing that follows. While the initial processing may be facilitated by repetition and semantic overlap, integrative processing can be hindered.

According to the ILH, the repeated name penalty results from an imbalance between function and cost during the integration of the referential expression into the representation of the discourse. Referential expressions that carry more information than necessary are more difficult to process when the referent is already salient in the discourse than when it is not because the extra information in these expressions serves a discourse function only when the referent is not focused and therefore has to be properly identified and possibly reactivated in working memory. In support of this argument, Almor (1999) showed that repetitive NP anaphors (e.g., '*the bird*' in '*The bird seemed very satisfied*') are read slower when their antecedent is focused (e.g., following the sentence '*It was the bird that ate the fruit*') than when it is not (e.g., following the sentence '*What the bird ate was the fruit*'), arguably because the information in the repetitive anaphors only serves a discourse function when the antecedent is not focused and has to be reactivated. In contrast, referential expressions that carry a low amount of information, such as pronouns, would be easier to process when the referent is salient than when it is not, because they do not pose a processing burden that has to be balanced by serving some special function such as reactivating the referent. In this theory, the amount of semantic overlap between the referential expression and the prior representation of the referent matters because it leads to memory interference. Thus, according to the ILH, the inverse relation between reference specificity and the memory salience of referents is not a matter of an arbitrary cooperative discourse convention, but a direct outcome of the architecture of the underlying working memory system.

The balance between processing cost and discourse function provides a systematic way to explain the effects of many factors on referential processing. Several of these factors appear to affect referential processing by modulating the prior activation of the referents. For example, recency of mention, topicality, and focus affect the activation of referents in memory. In ILH terms, references to such prominent referents incur high processing cost due to the high level of memory interference associated with these referents. Fuller references to such referents are justified only if they serve some discourse function beyond identifying the referent. Other factors are more likely to affect processing through the pre-activation of semantic information. For example, the semantic fit between general discourse context, the thematic roles associated with the most recent verb, and the voice of the utterance are likely to yield some semantic activation that will affect the cost of processing a subsequent referential expression. In line with Arnold's Expectancy hypothesis (Arnold 1998), some factors may simply make reference to a certain referent more likely and therefore only justify little cost, leading to an overall preference for general expressions. This can explain the effect of factors such as parallelism, namely, the finding that references to referents that appear in the same grammatical position in the preceding clause tend to be more reduced (Arnold 1998; Chambers & Smyth 1998). By this explanation, parallelism may be driven by statistical regularity rather than by some underlying memory constraint.

Some factors affect the processing of referential expressions by modulating the functionality of referential expressions. For example, the existence of several referents that share grammatical features that are encoded in reduced expressions such as pronouns (e.g., gender and number), may justify the use of an expression with a higher cost so as to avoid ambiguity. In other cases a special discourse function justifies the use of a fuller expression. Full noun phrase references are justified when used for style, for adding emphasis, or for adding new information. In such cases, despite the high initial processing cost, overall processing will not suffer because this cost will be balanced with discourse function (Almor et al. 2007).

There are several empirical findings that support the ILH. First is the finding that the repeated name penalty occurs for only repeated references but not for definite noun phrase references that are more general than their referent, such as category anaphors (Almor 1999). Like pronouns, and unlike repeated references, category anaphors are read faster when their referent is focused than when it is not focused. This finding supports the ILH's attribution of the repeated name penalty to semantic overlap rather than referential form per se.

A second finding that supports the ILH is the non-intuitive detrimental effect semantic overlap has on the processing of reference to a salient discourse entity. Almor (2004) reported an inverse typicality effect that was reflected in faster readings of category noun phrase anaphors (e.g., '*the bird*') with a focused antecedent when the antecedent was an atypical member of the category (e.g., '*ostrich*') than when it was a typical member (e.g., '*robin*'). Cowles and Garnham (2005) showed a similar inverse conceptual distance effect reflected in faster reading of superordinate anaphors (e.g., '*the creature*') when the focused antecedent was a remote subordinate (e.g., '*robin*') than when it was a close subordinate (e.g., '*bird*'). This finding further supports the role the ILH assigns to semantic representations in the cost of processing referential expressions. These findings show that the processing of definite noun references presents heterogeneity compatible with their cost and function as argued by the ILH. Moreover, using a category reference with an atypical referent is analogous to using a pronoun, whereas using a category reference with a typical referent is analogous to using a repeated reference. Clearly, theories that appeal to only the word class of referential expressions cannot explain the consistent heterogeneity within the class of definite noun phrases that mirrors the distinction between repeated names and pronouns.

A third finding that supports the ILH's notion of balance between cost and function and the role of working memory in that balance comes from recent experiments on the processing of metaphorical anaphors (Almor et al. 2007). Participants in these experiments were divided into low-span and high-span readers based on a verbal working memory task and then read discourses with metaphorical anaphors. High span participants did not take longer to read sentences with metaphoric anaphors (e.g., '*the creampuff*' as a reference to a weak boxer in a boxing match) than comparable sentences with literal anaphors (e.g., using the expression '*the fighter*' as reference) when the context preceding the metaphor provided enough information to facilitate reference resolution, but not too much information so as to obviate the function of the metaphor. Low-span participants always read sentences with metaphor references slower than sentences with literal anaphors. These results support the ILH in showing that metaphoric anaphors are processed similarly to other anaphors as a matter of balancing processing cost and discourse function, and that this balance is affected by working memory performance.

A fourth finding that provides more evidence for the role of working memory in referential processing is the deficits shown by patients with Alzheimer's disease (AD) in producing and comprehending referential expressions. These patients, who suffer from a working memory impairment, show a tendency to overuse pronouns in production but their ability to comprehend pronouns is significantly compromised, and they are better able to access information about the referent when a noun phrase anaphor is used (Almor et al. 1999). Furthermore, these aspects of reference production and comprehension are correlated with working memory performance. According to the ILH, the working memory impairment in AD shifts the balance of cost and function. In production, AD patients' representation of referents in working memory is degraded, leading to the loss of some distinguishing semantic features. For example, the representation of '*robin*' might become more similar to the representation of '*bird*'. According to the ILH, this loss of specific information about the referent causes an increase in the processing cost of all possible referential expressions. This is because, according to the ILH, cost is a matter of the semantic relation between the representations of the referential expression and the referent. For example, although the expression 'the bird' has only little processing cost with respect to the referent 'the robin', it has a higher processing cost with respect to the referent 'the bird' (because of the greater amount of repeatedly activated semantic features in the latter case than in the former.) Therefore, when semantic detail is lost in a referent's representation, a more general and less costly expression, such as a pronoun, is likely to be produced. In comprehension, the rapid loss of semantic information from the representation of the discourse in working memory renders fuller references such as repeated full noun phrases more functional than reduced expressions such as pronouns. In other words, for AD patients, fuller references can help reactivate information that would otherwise be lost. These findings are important not only in demonstrating the importance of working memory in referential processing, but also in that they demonstrate that the balance between cost and function is not a matter of simple convention but rather reflects the operation of the underlying mechanisms. Changes to the cost and function associated with these mechanisms result in changes to the point of balance.

A fifth group of findings supports the ILH's view of the repeated name penalty as reflecting integrative processing. Like other theories, the ILH assumes that the resolution of anaphoric expressions involves multiple stages, with the initial processing of these expressions followed by integration into the representation of the discourse. However, the ILH specifically attributes the repeated name penalty to the integrative stage and not to the initial processing of the referential expression, which can in fact be facilitated by repetition. In line with this view, the repeated name penalty has not been observed in reading paradigms in which stimuli were presented word by word but only when stimuli were presented as larger units of sentence fragments or whole sentences (Nair & Almor 2006). Nair and Almor (2006) interpreted this as indicating that the referential processes underlying the repeated name penalty may be delayed and spill over to the processing of words following the referential expression. Similar findings have also been observed in studies of spoken language where a repeated name penalty was observed in a delayed recall task but not in lexical decision latencies during processing (Almor & Eimas 2007). In a recent study using the visual world paradigm, Almor and Phillips (2006) tracked listeners' eye movements as they heard discourses about visual displays and showed that focus and semantic overlap can facilitate initial processing but impede later processing.

Further evidence for the delayed nature of these processes comes from eye-tracking studies of reading that found the repeated name penalty (Camblin et al. forthcoming) and the inverse typicality effect (van Gompel, Liversedge & Pearson 2004) only in regressive eye movements but not in initial fixation times or gaze duration. Cook, Myers and O'Brien (2005) also found that readers were as fast or faster to comprehend a question with a category anaphor when the exemplar antecedent was not mentioned in the preceding text than when it was. Cook, Myers and O'Brien interpreted this result as suggesting that reference resolution involves a primary stage in which memory is searched in a fast and automatic process. If a relevant antecedent is found, this antecedent is reinstated and integrated with previous discourse in a secondary stage. If no relevant antecedent is found, processing is terminated after only one stage, thus allowing the reader to move on to subsequent text.

Although there is much evidence for the existence of multiple processing stages in referential processing, it is important to stress that we are not claiming that these processing stages cannot overlap under certain conditions or even occur simultaneously, but rather that they do not always do.

Summary and Future Directions

The multiplicity of possible referential forms has been traditionally viewed as a problem that has to be solved during language production and that may call for special strategies during language comprehension. Much existing research has focused on either why certain forms are used in certain contexts, or on how different forms are processed. In this article we described an approach that addresses both questions together and that views the multiplicity of referential form not as a problem but as the solution language offers to a problem created by the constraints of serial communication and the architecture of the memory system that is used for representing discourse. Our view shares its emphasis on the balance between cost and function with theories that aim to answer the 'why' question (e.g., Ariel 1990; Gundel et al. 1993). However, unlike these theories, the present view espouses a clear and independently motivated view of computational cost that affects the integrative stage of referential processing. The present view also shares its appeal to memory mechanisms and processing stages with many theories that address the 'how' question. However, in contrast to many of these theories, the present view emphasizes the role of semantic representation in working memory and views differences between word classes such as full names and pronouns as driven by these semantic factors and mirroring semantically driven differences between expressions within the same word class (e.g., definite descriptions of varying levels of specificity). The description of different word classes in terms of semantic overlap that also apply to words within a single word class may suggest why language offers multiple forms and in particular why language has pronouns. We argue that the memory system that is used for representing discourse is prone to interference and that the existence of reduced expressions such as pronouns provides an optimized solution for repeated reference with minimal memory interference. In our view, pronouns are the solution language developed to the problem posed by the need to communicate sequentially using limited size informational units that have to be coherently linked, within a memory system that is prone to interferences.

Short Biography

Amit Almor is an Associate Professor of Psychology and Linguistics at the University of South Carolina, Columbia, SC, USA. He is broadly interested in the relation between language and memory in both healthy and impaired populations. Much of his work has focused on reference processing in discourse. The theoretical foundation for his work in this area was laid out in a 1999 Psychological Review article titled '*Noun-phrase anaphora and focus: The informational load hypothesis*'. Since the publication of this article, he has been working on further developing this framework through computational modeling, behavioral investigations of reading and spoken language comprehension, and, more recently, neuroimaging techniques. He holds a BSc in Mathematics and Computer Science from Tel Aviv University, Tel Aviv, Israel, and a PhD in Cognitive Science from Brown University, Providence, RI, USA.

Veena A. Nair is a graduate student in the Experimental Psychology program at the University of South Carolina, Columbia, SC, USA. She is working with Dr. Almor on reference processing in reading using self paced reading, ERP, and functional magnetic resonance imaging methodologies. She holds a Bachelor of Engineering degree from Government College of Engineering, Pune, India, and MS (Psychology) from the Old Dominion University, Norfolk,VA, USA.

Endnote

* Correspondence address: Amit Almor, University of South Carolina, Psychology Department, University of South Carolina, 1512 Pendleton Street, Columbia, SC 29208, USA.

Works Cited

Almor, A. 1999. Noun-phrase anaphora and focus: the informational load hypothesis. Psychological Review 106.748–765.

——. 2000. Constraints and mechanisms in theories of anaphor processing. Architectures and mechanisms for language processing, ed. by M.W. Crocker, M. Pickering and C. Clifton, Jr., 341–54. New York: Cambridge University Press.

——. 2004. A Computational Investigation of reference in production and comprehension. Approaches to studying world-situated language use: bridging the language-as-product and language-as-action traditions, ed. by J. C. Trueswell & M. K. Tanenhaus, 285–301. Cambridge, MA.: MIT Press.

——. 2005. Phonological and semantic interference in serial recall: bringing in the lexicon. Paper presented at the 2005 Annual Psychonomics Society Meeting, Toronto, Canada.

Almor, A., and P. Eimas. 2007. Focus and noun phrase anaphors in spoken language comprehension. Manuscript submitted for publication.

Almor, A., and M. Phillips. 2006. Category NP anaphors in spoken language comprehension. Paper presented at the 2006 Annual Psychonomics Society Meeting, Houston, TX.

Almor, A., S. Arunachalam, and B. W. Strickland. 2007. When the creampuff beat the boxer: working memory, cost, and function in reading metaphoric reference. Metaphor & Symbol 22.169–193.

Almor, A., D. Kempler, M. C. MacDonald, E. S. Andersen, and L. K. Tyler. 1999. Why do Alzheimer patients have difficulty with pronouns? Working memory, semantics, and reference in comprehension and production in Alzheimer's disease. Brain & Language 67.202–27.

Ariel, M. 1990. Accessing noun-phrase antecedents. London: Routledge.

Arnold, J. E. 1998. Reference form and discourse patterns. Unpublished PhD Dissertation, Stanford University.

——. 2001. The effect of thematic roles on pronoun use and frequency of reference continuation. Discourse Processes 31.137–62.

Arnold, J. E., M. Fagnano, and M. K. Tanenhaus. 2003. Disfluences signal theee, um, new information. Journal of Psycholinguistic Research 32.25–36.

Arnold, J. E., M. K. Tanenhaus, R. J. Altmann, and M. Fagnano. 2004. The old and thee, uh, new: disfluency and reference resolution. Psychological Science 15.578–82.

Camblin, C. C., K. Ledoux, M. Boudewyn, P. C. Gordon, and T. Y. Swaab. (in press). Processing new and repeated names: effects of coreference on repetition priming with speech and fast RSVP. Brain Research.

Chafe, W. L. 1976. Giveness, contrastiveness, definiteness, subjects, topics, and point of view. Subject and Topic, ed. by C. N. Li, 25–55. New York: Academic Press.

Chambers, C. G., and R. Smyth. 1998. Structural parallelism and discourse coherence: a test of centering theory. Journal of Memory and Language 39.593–608.

Chomsky, N. 1981. Lectures on government and binding. Dordrecht, The Netherlands: Foris Publications.

Clark, H. H. 1996. Using language. Cambridge, UK: Cambridge University Press.

Clark, H. H., and G.-D. Wilkes. 1986. Referring as a collaborative process. Cognition 22.1–39.

Cook, A. E., J. L. Myers, and E. J. O'Brien. 2005. Processing an anaphor when there is no antecedent. Discourse Processes 39.101–20.

Cowles, H. W., and A. Garnham. 2005. Antecedent focus and conceptual distance effects in category noun-phrase anaphora. Language and Cognitive Processes 20.725–50.

Garrod, S. C., and A. J. Sanford. 1994. Resolving sentences in discourse context: how discourse representation affects language understanding. Handbook of psycholinguistics, ed. by M. A. Gernsbacher, 675–98. New York: Academic Press.

Gernsbacher, M. A. 1989. Mechanisms that improve referential access. Cognition 32.99–156.

——. 1990. Language comprehension as structure building. Mahwah, NJ: Lawrence Erlbaum Associates.

Givon, T. 1976. Topic, pronoun and grammatical agreement. Subject and Topic, ed. by C. N. Li. New York: Academic Press.

Gordon, P. C., and D. Chan. 1995. Pronouns, passives, and discourse coherence. Journal of Memory and Language 34.216–31.

Gordon, P. C., and R. Hendrick. 1998. The representation and processing of coreference in discourse. Cognitive Science 22.389–424.

Gordon, P. C., and K. A. Scearce. 1995. Pronominalization and discourse coherence, discourse structure and pronoun interpretation. Memory and Cognition 23.313–23.

Gordon, P. C., B. J. Grosz, and L. A. Gilliom. 1993. Pronouns, names, and the centering of attention in discourse. Cognitive Science 17.311–47.

Gordon, P. C., R. Hendrick, and K. Ledoux Foster. 2000. Language comprehension and probe-list memory. Journal of Experimental Psychology: Learning, Memory, and Cognition 26.766–75.

Gordon, P. C., R. Hendrick, and K. Ledoux. 2000. Language comprehension and probe-list memory. Journal of Experimental Psychology: Learning, Memory, and Cognition 26.766–75.

Greene, S. B., G. McKoon, and R. Ratcliff. 1992. Pronoun resolution and discourse models. Journal of Experimental Psychology: Learning, Memory and Cognition 18.266–83.

Grice, H. P. 1975. Logic and Conversation. Syntax and semantics III: Speech acts, ed. by P. Cole and J. Morgan, 41–58. New York: Academic Press.

Grosz, B., A. K. Joshi, and S. Weinstein. 1995. Centering: a framework for modeling the local coherence of discourse. Computational Linguistics 21.203–25.

Gundel, J., N. Hedberg, and R. Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. Language & Cognitive Processes 69.274–307.

Just, M. A., and P. A. Carpenter. 1992. A capacity theory of comprehension: individual differences in working memory. Psychological Review 99.122–49.

Kamp, H., and U. Reyle. 1993. From discourse to logic: Introduction to model theoretic semantics of natural language, formal logic and discourse representation theory. Dordrecht, The Netherlands: Kluwer Academic.

Kintsch, W. 1988. The role of knowledge in discourse comprehension: a construction-integration model. Psychological Review 95.163–82.

MacDonald, M. C., N. J. Pearlmutter, and M. S. Seidenberg. 1994. The lexical nature of syntactic ambiguity resolution. Psychological Review 101.676–703.

McKoon, G., and R. Ratcliff. 1992. Inference during reading. Psychological Review 99.440–66.

McKoon, G., R. J. Gerrig, and S. B. Greene. 1996. Pronoun resolution without pronouns: some consequences of memory-based text processing. Journal of Experimental Psychology: Learning, Memory, and Cognition 22.919–32.

Nair, V., and A. Almor. 2006. Referential processing in word by word reading. Poster presented at the 2006 Annual CUNY Conference on Sentence Processing, New York, NY.

Peters, S. A., and A. Almor. 2006. The repeated name penalty in reference to concrete objects. Poster presented at the 2006 Annual CUNY Conference on Sentence Processing, New York, NY.

Sanford, A. J., and S. C. Garrod. 1981. Understanding written language. Chichester, UK: John Wiley & Sons.

Shapiro, A., and A. Milkes. 2004. Skilled readers make better use of anaphora: a study of the repeated-name penalty on text comprehension. Electronic Journal of Research in Educational Psychology 2.161–80.

Sperber, D., and D. Wilson. 1995. Relevance: Communication and cognition, 2nd edn. Oxford, UK: Blackwell.

Swaab, T. Y., C. C. Camblin, and P. C. Gordon. 2004. Electrophysiological evidence for reversed lexical repetition effects in language processing. Journal of Cognitive Neuroscience 16.715–26.

Trueswell, J. C., and M. K. Tanenhaus. 1994. Toward a lexicalist framework of constraint-based syntactic ambiguity resolution. Perspectives on sentence processing, ed. by C. Clifton, Jr., L. Frazier and K. Rayner, 155–79. Hillsdake, NJ: Lawrence Earlbaum Associates.

van Gompel, R. P. G., S. P. Liversedge, and J. Pearson. 2004. Antecedent typicality effects in the processing of noun phrase anaphors. The on-line study of sentence comprehension: Eyetracking, ERPs, and beyond, ed. by J. M. Carreiras and C. Clifton, 119–37. Brighton, UK: Psychology Press.

Vonk, W., L. G. M. M. Hustinx, and W. H. G. Simons. 1992. The use of referential expressions in structuring discourse. Language and Cognitive Processes 7.301–33.