Identity, Causality, and Pronoun Ambiguity

Eyal Sagi and Lance J. Rips

Psychology Department

Northwestern University

2029 Sheridan Road

Evanston, IL 60208

Email: rips@northwestern.edu

Keywords: pronoun resolution, anaphora, singular concepts, identity over time, causation, causal reasoning, ambiguity resolution

Send correspondence about this article to:

Lance Rips

Psychology Department

Northwestern University

2029 Sheridan Road

Evanston, IL 60208

Email: rips@northwestern.edu

Phone: 847-491-5947; Fax: 847-491-7859

Abstract

This article looks at the way people determine the antecedent of a pronoun in sentence pairs, such as: *Albert invited Ron to dinner. He spent hours cleaning the house*. The experiment reported here is motivated by the idea that such judgments depend on reasoning about identity (e.g., the identity of the *he* who cleaned the house). Because the identity of an individual over time depends on the causal-historical path connecting the stages of the individual, the correct antecedent will also depend on causal connections. The experiment varied how likely it is that the event of the first sentence (e.g., the invitation) would cause the event of the second (the house cleaning) for each of the two individuals (the likelihood that if Albert invited Ron to dinner, this would cause Albert to clean the house, vs. cause Ron to clean the house). Decisions about the antecedent followed causal likelihood. A mathematical model of causal identity accounted for most of the key aspects of the data from the individual sentence pairs.

Identity, Causality, and Pronoun Ambiguity

Readers and listeners who encounter a personal pronoun sometimes face a choice of antecedent. When the pronoun and its potential antecedents occur within the same sentence, syntactic rules (e.g., those of binding theory) can determine which antecedent is correct (see Fiengo & May, 1996). But especially when the pronoun and the antecedents span separate sentences, two or more antecedents can each produce a grammatical discourse, and people must resort to semantic and contextual clues to figure out the right one. Neither of the sentence pairs in (1) and (2), for example, provides overt grammatical clues about the antecedent of *he*, and yet readers believe that the antecedent is the subject noun (*Albert*) in (1) but the object noun (*Ron*) in (2), as we show later:

(1) Albert invited Ron to dinner. He spent hours cleaning the house. (Subject strong/Object weak)

(2) Albert invited Ron to dinner. He brought a small gift. (Subject weak/Object strong)

We suggest that the antecedent of a pronoun depends on the identity of the individual the pronoun describes. In (1), the intended referent of *he* is a male, *x*, who spent hours cleaning the house. Our goal as readers is to establish whether *x* is identical to Albert, to Ron, or to some third party, according to the information presented in the discourse. We take *x* to be Albert if the facts we've learned convince us that *x* is on the same biographical path as Albert, but we take *x* to be Ron if *x* is more likely to be on Ron's path. In the case of (1), we try to determine if the earlier invitation would extend naturally to a situation in which Albert did the house cleaning (where *x* = Albert), and we compare this possibility to one in which Ron did the house cleaning (where *x* = Ron). We show in this article that a cognitive model of how people trace the identity of physical objects in nonlinguistic contexts (Rips, Blok, & Newman, 2006) can predict decisions about antecedents in pairs such as (1) and (2).

Pronoun Resolution and Identity

Our approach to pronouns unites two issues that previous investigators have pursued in separate research streams. One issue, pursued in psycholinguistics, asks how people decide on an antecedent for a pronoun. The second issue, pursued in cognitive and developmental psychology, asks how people decide whether an individual at one time is the same as an individual at another. We suggest that a solution to the second issue is a solution to the first. Of course, not all pronouns are anaphoric. Anaphoric pronouns derive their referents through other referring expressions (e.g., in the way that *he* in (1) or (2) gets its referent from *Albert* or *Ron*). Some pronouns, however, obtain their referents from nonlinguistic context rather than from discourse. And among anaphoric pronouns, some don't refer to individuals but serve instead as variables bound to quantifiers (as in *No host is such that he would spend hours cleaning the house*). Our concern here is with anaphoric pronouns that purport to refer to individuals (and where overt syntactic constraints are consistent with each of the potential antecedents).

In bridging these issues, we appeal to two principles. The first asserts that assigning a pronoun from one sentence to an antecedent in another is a matter of establishing identity between two referents:

> **Coreference is Identity:** The antecedent of a pronoun "*p*" is the term "*t*" if the discourse represents *t* as identical to *p*.

According to this principle, judging the antecedent of a pronoun involves the same mental processes as deciding whether a person we just encountered at a concert is the same person we saw an hour ago in the library. We are not proposing a mere analogy between these kinds of decisions. Rather, we believe they are the same decision—*Is t = p?*—though sometimes drawing on different sources of evidence. In determining the antecedent of a pronoun, we need to consider how the discourse represents the relation between the two referring expressions. For example, if a confused English student asserts, *Nick Carraway is an important author; he wrote "The Great Gatsby,"* the antecedent of *he* is *Nick Carraway*, despite the fact that the individual who wrote *The Great Gatsby* is not Nick Carraway. The critical relation for determining the antecedent is *internal coreference*—how the discourse represents the situation—rather than *external coreference*—how matters actually stand (see, e.g., Lawlor, 2010; Recanati, 2012).

The second principle we rely on is that identity over time for a physical object (including a person) depends on the causal-historical path that connects the object's stages:

> **Causality Guides Identity:** An individual x_0 at one time is identical to an individual x_1 at a later time (i.e., $x_0 = x_1$) iff: (a) x_1 is causally close enough to x_0, and (b) x_1 is closer than any other causally-close-enough competitor.

This principle adapts a similar rule about identity from Nozick (1981), and it provides an account of explicit identity judgments about a variety of objects (Rips, Blok, & Newman, 2006). The library patron and the concertgoer are the same individual if and only if the concertgoer is on a causal path that is close enough to the library patron to qualify as the patron, and no one else is as close. [1] To be close enough, the later stage of an object must be a causal outgrowth of its earlier stages. The cluster of causal forces responsible for maintaining the object over time must extend from the earlier to the later time point.[2] Of course, discourse is typically not explicit about the causal factors responsible for identity. So a reader or hearer must make inferences about this causal background, including inferences about what the writer/speaker takes the relevant causes to be.

Putting these principles together, we propose that people believe causal connectedness, as represented in the relevant sentences, determines a pronoun's antecedent. We should assign *he* to *Albert* in (1) if Albert is represented as the closest of the causally close-enough individuals to the person who spent hours cleaning the house. In this article, we report an experiment showing that relative causal connectedness predicts pronoun assignments.

Causality, Coreference, and the Scope of Our Proposal

We maintain that the antecedent of a pronoun is a matter of identity. We also believe that causal-historical connections are responsible for identity, at least for ordinary physical objects, such as people, cats, and toasters (e.g., Nozick, 1981; Shoemaker, 1979). But we don't mean to imply that readers and listeners use only causal principles to determine a pronoun's antecedent. Sentences containing referential pronouns sometimes don't provide enough causal information to make such a connection. Take the

sentence pair *Carolyn thought about Penelope. Sharon caught a glimpse of her*. Here, the thinking-about event is probably not a cause or an effect of the glimpsing event. We take *her* to be Penelope rather than Carolyn because we understand the two sentences to be rhetorically parallel (Hobbs, 1979; Kehler, Kertz, Rohde, & Elman, 2008; Wolf, Gibson, & Desmet, 2004). Earlier studies also document biases to assign pronouns to the subject noun of the preceding clause (e.g., Crawley, Stevenson, & Kleinman, 1990) and to the noun sharing the same syntactic role (e.g., Chambers & Smyth, 1998; Grober, Beardsley, & Caramazza, 1978).

We would argue (though we cannot do so fully here) that causality is privileged in resolving pronouns because causality dominates other factors in determining identity (Blok et al., 2005, and Rips et al., 2006). Some factors, such as gender marking, help single out an antecedent by making some causal paths more likely than others. Substituting *Alberta* for *Albert* in (1) leaves *Ron* as the antecedent of *he* because people don't typically change from female to male during the events described.[3] Other factors, such as the prominence of subject nouns or nouns in parallel positions, indicate likely antecedents, but we easily discard these cues in the face of explicit causal facts, as in (2). The situation is again the same as judging identity in nonlinguistic situations: We often use surface properties like perceptual similarity to establish identity, but that's because similarity is an easily accessible (but fallible) indicator of deeper causal relations. In the present article, however, we do not attempt to pit causality against other influences, but settle for the more limited goal of showing that when causal information is available, it predicts the choice of antecedent. [4]

Experiment: Pronoun Disambiguation and Causality

One way to study the connection between pronoun resolution and causality is to vary the causal relation between events described by sentence pairs and examine the effect this relation has on choice of a pronoun's antecedent. The hypothesis is that readers will choose the antecedent that yields the strongest causal connection to the pronoun. In referring to the sentences within these pairs, we call the first the *head*

and the second the *tail*. One of these pairs is (1): *Albert invited Ron to dinner* (the head sentence). *He spent hours cleaning the house* (the tail sentence). Because inviting someone to dinner is more likely to motivate the host to clean the house than the guest, participants are apt to think *he* is Albert rather than Ron.

For some pairs in this experiment, the tail sentence is a more likely outcome when *he* refers to the subject noun of the head sentence than when it refers to the object noun, as in (1). We refer to such items as *Subject strong/ Object weak* pairs. For other pairs—the *Subject weak/Object strong* items—the outcome is more likely when *he* refers to the object noun. The pair in (2) provides an example. For still others—the *Neither strong* pairs—neither outcome is especially likely:

(3) Albert invited Ron to dinner. He went to a rock concert. (Neither strong)

Finally, for a fourth set of pairs—the *Both strong* items—the tail sentence could be a plausible result of the head sentence when either the subject noun or the object noun substitutes for *he*:

(4) Albert invited Ron to dinner. He bought an expensive bottle of wine. (Both strong)

Table 1 gives further examples of each type.

The current theory applies in a special way to the Both strong cases, such as (4) (Rips et al., 2006). According to the Causality Guides Identity principle, decisions about the identity of an object—which of two designated objects, y_1 or y_2, at one time is identical to an object x at another time—depend on two factors: First, in order for y_1 (or y_2) to be x, the causal connection to x must be above threshold. If y_1 or y_2 fails to be close enough to x to qualify as x's causal successor, then it can't be identical to x. Second, if two (or more) candidates are above threshold, then the closest of these alternatives counts as the identical item. The model predicts, then, that when both individuals from the head sentence are causally possible antecedents, participants must consider the difference in their causal strength. By contrast, when only one or neither of the individuals is causally plausible, participants can shortcut the comparison by eliminating candidates below threshold. In this experiment, we measure causal closeness by asking participants to evaluate the likelihood of the head-tail sentence pairs for each of the two named people (e.g., *Albert invited Ron…Albert bought an expensive bottle* vs. *Albert invited Ron…Ron bought*

an expensive bottle). Our prediction is that the difference between these ratings will correlate more highly with participants' choice of antecedent when both potential antecedents are causally plausible [as in (4)] than when only one or neither is [as in (1)-(3)].

Method

In one part of this experiment, participants read 16 sentence pairs one at a time on a computer screen. After reading a pair, the participants indicated their interpretation of a pronoun (*he*) that appeared in the tail sentence. In a second part of the experiment, participants rated the likelihood that the event described in the head sentence would cause the event described in the tail.

Materials. We composed the stimulus pairs in this experiment from 16 head sentences and 16 tail sentences. The head sentences each described an event using a transitive verb and two named people, a subject and object. Each tail sentence described a second event but referred to only one individual, using an initial pronoun (*he*). These sentences divided into four groups of four head and four tail sentences each. Within a group, we paired each head sentence with the four tail sentences to create four Subject strong/Object weak pairs, four Subject weak/Object strong pairs, four Both strong pairs, and four Neither strong pairs (see Table 1 for examples). We verify assignment of the pairs to these categories by means of causality ratings, described shortly. Each group thus included 16 sentence pairs—64 pairs in all.

For the disambiguation part of the experiment, we created four different presentation lists of 16 sentences each. A given head sentence appeared just once in each list, but across lists, each head sentence appeared with all four tails from the same group. Each list also included four sentence pairs from each of the four types (Subject strong/Object weak, Subject weak/Object strong, Both strong, and Neither strong).

For the causality ratings, we constructed two questions from each head-tail sentence pair by replacing the pronoun *he* with one of the two names from the head sentence. For example, the pair in (1) gave rise to: *If Albert invited Ron to dinner, how likely is that to cause Albert to spend hours cleaning the house?* and *If Albert invited Ron to dinner, how likely is that to cause Ron to spend hours cleaning the*

house? Each participant rated 32 such questions, which corresponded to the 16 head-tail pairs that the same participant had seen in the disambiguation portion of the experiment.

Procedure: pronoun disambiguation. For each sentence pair, participants chose the person from the first sentence that the pronoun in the second sentence referred to or chose "neither" if the pronoun was unlikely to refer to either character. The 16 sentence pairs appeared, one at a time, at the top left of a computer screen. To record the choice of antecedent, participants pressed one of three response keys.

Procedure: causality ratings. The instructions told participants that the questions were about how people judge the implications of events. Participants were to rate the likelihood of these implications on a 10-point scale by typing in the number. The scale appeared as a range of numerals from 0 (marked "Not at all likely") to 9 ("Extremely likely"). The 32 items appeared in a new random order for each participant.

Participants. Participants were Northwestern University students from an introductory psychology class. Twenty-eight participants did the disambiguation task first, and thirty-four did the causality ratings first.

Results and Discussion

The focus of this experiment is whether the strength of the causal relation between events predicts choice of an antecedent. We designed the sentence pairs so that the event of the head sentence could be a cause of the tail sentence when the pronoun in the tail referred to the subject of the head, the object of the head, both these arguments, or neither. Our prediction is that participants should disambiguate the pronoun in accord with these relations by determining which antecedent yields the stronger causal connection. In what follows, we first check our classification of the sentence pairs by examining participants' causality ratings. Then we evaluate the central predictions about choice of antecedent.

Causality ratings. Participants' ratings of causal likelihood fell in line with our assessments of the head and tail sentence pairs. Table 2 gives the means of these ratings for the four head-tail combinations.

When the question focused on the subject noun (column 1), we predicted that ratings would be high for the Subject strong/Object weak items (e.g., *If Albert invited Ron to dinner, how likely is that to cause Albert to spend hours cleaning the house?*) and Both items (e.g., *If Albert invited Ron to dinner, how likely is that to cause Albert to buy an expensive bottle of wine?*). The mean causality rating was 6.83 for these questions, whereas the mean was only 1.39 for the remaining items. When the question focused on the object noun (column 2), we predicted high ratings for the Subject weak/Object strong items (*If Albert invited Ron to dinner, how likely is that to cause Ron to bring a small gift?)* and the Both items. In fact, the mean rating was 6.16 for these questions, but 1.42 for the rest of the questions. The ratings suggest, then, that we were successful in choosing sentence pairs that matched the intended degrees of causal relatedness.

Disambiguations. Our predictions for the disambiguations follow from the idea that participants should choose the name that produces the strongest causal connection between the events described in the head and tail sentences. This idea implies first that participants should choose the subject noun as the referent of *he* if substituting the subject noun produces a causally strong head-to-tail link but substituting the object noun doesn't. This case will occur for the Subject strong/Object weak pairs. The last three columns in Table 2 list the proportion of subject choices, object choices, and "neither" choices for each of the types of sentence pairs. These data show that participants selected the subject noun on 92% of the Subject strong/Object weak trials. Second (and for similar reasons), if the choice of the object noun produces a strong head-tail connection, but the choice of the subject noun doesn't, participants should favor the object noun. This will be true for the Subject weak/Object strong items, and the proportion of object choices was 94% for these items.

Two further predictions follow for the Neither tails and the Both pairs. When neither the subject noun nor the object noun yields a sensible causal interpretation, participants should choose *neither*. This

applies to the Neither pairs, and participants in fact chose *neither* on 63% of these trials. Although this figure is much higher than for the other tail types (column 5 of Table 2), it is lower than we might expect, and we will return to this finding in discussing the individual sentence pairs in the next subsection. Finally, when both the subject noun and object noun make for a strong causal link, participants should choose between the two based on the precise difference in strength of the causal connections. The result should be a split decision between the subject and object nouns. For these Both pairs, participants picked the subject noun on 72% of trials and the object noun on 25%. The advantage for the subject noun may be due to a bias favoring the subject of the preceding sentence (e.g., Crawley et al., 1990) or a bias favoring the noun sharing the same syntactic role as the pronoun (e.g., Chambers & Smyth, 1998; Grober et al., 1978). However, we will outline a different way to account for these results in the General Discussion.

We can summarize the first, second, and fourth of the predictions mentioned in the two preceding paragraphs by saying that participants' choice of the subject noun should be highest for the Subject strong/Object weak pairs, intermediate for the Both pairs, and minimal for the Subject weak/Object strong and Neither pairs. Overall, the types of pairs differed significantly in the proportion of subject choices, according to a generalized linear mixed model for binomial data [$F(3,66) = 44.37$, $p < .001$, using the Satterthwaite approximation for degrees of freedom]. In accord with the hypothesis, subject noun choices were significantly higher for Subject strong/Object weak pairs than for Both pairs [$t(59) = 2.97$, $p = .02$, by a Bonferroni test] and significantly higher for Both pairs than for either Subject weak/Object strong or Neither pairs [$t(85) = 7.66$ and $t(45) = 6.24$, respectively, $p < .001$ in both cases].

The remaining prediction is that participants should choose *neither* more often for the Neither pairs than for the others. We've already noticed the large size of this difference, and it is significant in an analysis of the proportion of *neither* responses, similar to the analysis just described. For the overall difference in *neither* choices, $F(3,151) = 32.20$, $p < .001$. Bonferroni tests showed that these choices were greater for the Neither pairs than for the Subject strong/Object weak pairs, the Subject weak/Object strong pairs [$t(175) = 6.72$, $p < .001$, in both cases], and the Both pairs [$t(88) = 6.99$, $p < .001$].

Individual sentence pairs. As we noted earlier, the Both pairs in this experiment should be especially responsive to the difference between the causal strength of the two interpretations. For the other kinds of head-tail pairs, however, people can avoid this comparison by eliminating one or both candidates as falling below a threshold level of causal strength.

We can test this prediction using the causality ratings we collected in this study. We can compute, for each sentence pair, the difference between the rating when the object noun substituted for *he* and the rating when the subject noun substituted for *he*. The hypothesis is that the Both pairs will produce a high correlation between this causality difference and participants' choice of antecedent in the disambiguation task. The correlations should be lower, however, for the other head-tail combinations. The top panel in Figure 1 contains a scatter plot relevant to this prediction, with proportion of object choices from the disambiguation task on the *y*-axis and the difference in rated causal likelihood on the *x*-axis. Filled symbols in the figure represent results for the individual sentence pairs. (The unfilled symbols are predicted values from our model, which we will describe in the General Discussion.)

The correlations between object choice and causal difference support the prediction: This correlation is quite high for the Both pairs, symbolized by diamonds in Figure 1a [$r(14) = .84$, $p < .001$], whereas the correlations for the Subject strong/Object weak pairs (circles) and the Subject weak/Object strong pairs (squares) are near zero [$r(14) = .02$ for the former and $r(14) = .04$ for the latter, $p > .10$ in both cases]. Planned comparisons on the *z*-transformed coefficients show that the correlation for Both pairs is significantly higher than that of either the Subject strong/Object weak or the Subject weak/Object strong items ($p < .01$, two-tailed, in both cases). Only the correlation for the Neither items (triangles) approaches that for the Both pairs. For Neither pairs, $r(14) = .73$, $p = .001$, which does not differ significantly from the Both pairs' correlation ($p > .10$). We will consider some reasons for this latter effect when we fit our model to these data.

In general, however, the correlations show that the choice of antecedent is sensitive to fine-grained variation in causal strength. We find differences due not only to the split between strong and

weak connections, as shown in Table 2, but also to the *degree* of causal relatedness within the Both category.

General Discussion

We have looked at the way people decide on a pronoun's antecedent within causally related sentence pairs. Our guiding Coreference is Identity principle says that establishing the antecedent for a pronoun across sentences is establishing the identity of an individual. If you meet someone at a class reunion whom you believe might be either Albert or Ron, then the choice depends on whether this person is causally connected to other stages of one of these individuals. Likewise, if you read or hear the pronoun *he* and have to decide which of two previously mentioned people, Albert or Ron, *he* refers to, then the answer depends on whether the discourse portrays the individual denoted by *he* as tracing back to Albert or Ron.

According to the Causality Guides Identity principle, causal relations determine identity. In our sentence pairs, the causal relation between events establishes this connection. For the pair *Albert invited Ron to dinner. He spent hours cleaning the house*, the *he* of the cleaning is more likely to be on the host's biographical path than on the guest's. So the antecedent of *he* is more likely Albert than Ron.

The results confirmed the hypothesis that causal connectedness drives the choice of antecedent. Participants inferred that *he* referred to the subject (object) noun if substituting the subject (object) for *he* produced a stronger causal link; they typically responded *neither* if neither substitution produced a strong causal link; and they split their responses if both substitutions produced a strong link.

Model Fitting

To examine the theory in more detail, we fit a mathematical version of the model to the data in Figure 1. We suppose that participants consider the causal pathways connecting the individual denoted by

he to each of the potential antecedents in the head sentence. We will refer to these relations as the *antecedent paths* for the subject and object nouns.

The model assumes that the strength of each antecedent path is a normally distributed random variable, whose mean and standard deviation are given by the relevant ratings of causal likelihood. Second, we assume that people adopt a criterion *c* on this strength continuum. To determine the referent of a pronoun, participants sample a strength value for the antecedent path of the subject noun and a value for the antecedent path of the object noun. When the strength of a sampled value falls below *c*, then the corresponding antecedent is eliminated. If this is true for both the subject and the object paths, then participants will give a *neither* response. If only one of the paths has strength greater than *c*, then that path will furnish the antecedent. Finally, if both antecedent paths are greater than *c*, participants will choose the antecedent whose path has greater strength.[5]

We fit the model simultaneously to the proportion of object noun and *neither* responses, using a nonlinear least-squares procedure with one free parameter (*c*). The model's predicted values appear as unfilled symbols in Figure 1. Figure 1a contains the results for object responses, and Figure 1b, the *neither* responses. The shape of the unfilled symbols matches the shape of the corresponding type of sentence pair (circles for the Subject strong/Object weak pairs, triangles for the Neither pairs, etc.). These points show that the model captures the main trends in the results. Overall, R^2 for predicted versus observed values is .88, and the root mean square deviation is 0.12. The estimated value of *c* is 2.74 on the 0-to-9 point scale.

A comparison of the predicted and observed values for the object responses in Figure 1a brings out a number of facts about these data. For Both items, the model predicts the correlation between participants' choice of object noun and the difference in strength of the antecedent paths, as we noted earlier. This is because this difference comes into play when the strength of both paths is above criterion. The predicted and observed values in the figure (diamonds) show this sensitivity to the difference measure, increasing from left to right in the figure. We also noticed that on average these Both pairs gave rise to a subject interpretation more often than an object interpretation (72 vs. 25%). This tendency is

present in the predicted results as well, though to a slightly reduced degree: 65% of responses are subject nouns and 34% object nouns. The difference, then, is likely due to the fact that the average causal strength of the antecedent path for the subject noun is somewhat greater than that for the object noun, according to our measure of causal likelihood (see also Table 2). Although biases toward the subject noun or toward a parallel syntactic role may be a factor in some studies, as mentioned earlier, they may not be necessary to account for the results in this one.

The model also correctly predicts that the Subject strong/Object weak pairs (circles) and Subject weak/Object strong pairs (squares) will be less dependent on the difference in strength of the antecedent paths. However, the model underpredicts some of the Subject strong/Object weak pairs, as shown at the left of Figure 1a.[6] In our initial look at the data, we also noticed that the Neither pairs produced a moderate correlation between choice of object antecedents and the difference in causality ratings. Both antecedent paths for these Neither items should have low causal strength and should therefore produce many *neither* responses. However, if the Neither pairs come in beneath threshold, the model will never explicitly compare the strengths of these paths. So why the correlation between the difference in strength and the choice of the object noun? A glance at Figure 1 shows that the predicted values (unfilled triangles) for these Neither pairs also exhibits a positive correlation with object choices at a level nearly equal to that of the observed values (r = .68 for the predictions vs. r = .73 for the observed responses). The model predicts the correlation because the Neither pairs vary somewhat in the strengths of the subject and object paths in a way that lines up with participants' decisions. Because the sampled strengths of the antecedent paths are probabilistic, the object paths will sometimes exceed criterion and, for these stimulus items, do so more often than the subject paths. This produces the observed tendency to choose the object as antecedent.

Connections to Previous Research on Causality and Pronouns

Previous research on the role of causality in interpreting pronouns has focused on the effect verbs have in highlighting a potential antecedent—an effect called the *implicit causality of the verb*. Participants

in an event differ in the degree to which they are responsible for bringing it about. If people learn that Tom flattered Paul, for example, they typically believe this is because of Tom's obsequiousness rather than Paul's worthiness; but if they learn that Tom likes Paul, they believe this is the result of Paul's likeability rather than Tom's magnanimity (Brown & Fish, 1983). Similarly, people who are asked to complete sentences of the form *x verbs y because…* base their completions on the person, *x* or *y*, most responsible for the associated event, as befits a clause beginning with *because*. For instance, given the fragment in (5a), they expand on Tom's qualities, but given the fragment in (5b), they expand on Paul's (e.g., Au, 1986; Ferstl, Garnham, & Manouilidou, 2011; Garvey, Caramazza, & Yates, 1974):

(5) a. Tom flatters Paul because [he]…

b. Tom likes Paul because [he]…

c. Tom consoles Paul because [he]…

This difference in expectations also shows up in sentence processing (Caramazza, Grober, Garvey, & Yates, 1977; Featherstone & Sturt, 2010; Garnham, Oakhill, & Cruttenden, 1992; Koornneef & Van Berkum, 2006; Stewart, Pickering, & Sanford, 2000; and Vonk, 1985).

According to the present approach, effects of the "implicit causality of verbs" are part of a broader picture. People's preference for an antecedent in such experiments stems from their knowledge of cause-effect relations between events rather than simply from the verbs' thematic roles (see Rudolph & Försterling, 1997, for a comparison of verb taxonomies linked to implicit causality, and Pickering & Majid, 2007, for a discussion of the relation between thematic role and explicit causes and consequences). If Tom flatters Paul, for example, the cause of the flattery is likely to be some plan of Tom's to ingratiate himself. So we can predict that the fragment in (5a) will probably conclude with an explanation of Tom's plan, and participants will be faster in reading a completed version of this fragment if it contains consistent information about Tom.

This view of implicit causality comes with several advantages. First, it is congruent with contemporary theories that suggest that the observed bias is the result of the natural distribution of causes and consequences (see Rudolph & Försterling's, 1997, discussion of the covariation hypothesis). Second,

it is consistent with recent evidence demonstrating that causal information from the verb can quite rapidly highlight one of its arguments (e.g., *Paul* or *Tom* in (5)), even before the second clause is encountered (Pyykkonen & Jarvikivi, 2010). Readers or listeners of such sentences should benefit by attending to those individuals who are likely to play a role in further events that are causally associated with the initial one. Finally, this way of thinking about implicit causality goes along with the finding that even verbs that share the same thematic roles sometimes give rise to different choices of antecedents. For example, the fragment with *flatter* in (5a) is usually completed with something about Tom, but (5c) with *console* is completed with something about Paul (Ferstl et al., 2011), despite the fact that the thematic roles are the same in both sentences (Levin, 1993, classifies both *flatter* and *console* as *amuse*-type psych verbs).

Conclusion

Causal theories have recently proved useful in understanding the psychology of concepts, decision making, inductive inference, and counterfactual thinking, among others. We have suggested that causal reasoning also plays a role in a core psycholinguistic phenomenon: determining the antecedents of pronouns. Our account rests on the idea that the referent of a pronoun and the referent of an antecedent are individuals, with the identity between them depending on one being a causal outgrowth of the other. The experiment reported here supports this idea, allowing us to predict people's choice of antecedent from the causal connections tacitly expressed in successive sentences. This theory considerably expands the role causality plays in this domain. Although causal information is certainly not the only clue people rely on, it seems to provide an underlying basis for pronoun resolution.

Acknowledgments

We thank Dasom Kim, Sharon Paravastu, and Samantha Thompson for their help with these experiments, and Dan Bartels, Winston Chang, Jacob Dink, Brian Edwards, Joshua Hartshorne, Rumen Iliev, Samuel Johnson, Emily Morson, Linsey Smith, Reina Uchino, and participants at a workshop on Choice and Self in Noordwijk, the Netherlands, for their helpful comments on the research presented here.

References

Au, T. K. (1986). A verb is worth a thousand words: The causes and consequences of interpersonal events implicit in language. *Journal of Memory and Language, 25*, 104-122. doi: 10.1016/0749-596x(86)90024-0

Blok, S. V., Newman, G. E., & Rips, L. J. (2005). Individuals and their concepts. In W.-k. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman & P. Wolff (Eds.), *Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin* (pp. 127-149). Washington, DC: American Psychological Association.

Blok, S. V., Newman, G. E., & Rips, L. J. (2007). Out of sorts? Some remedies for theories of object concepts: A reply to Rhemtulla and Xu (2007). *Psychological Review, 114*, 1096-1102. doi: 10.1037/0033-295x.114.4.1096

Brown, R., & Fish, D. (1983). The psychological causality implicit in language. *Cognition, 14*, 237-273. doi: 10.1016/0010-0277(83)90006-9

Caramazza, A., Grober, E., Garvey, C., & Yates, J. (1977). Comprehension of anaphoric pronouns. *Journal of Verbal Learning & Verbal Behavior, 16*, 601-609. doi: 10.1016/s0022-5371(77)80022-4

Chambers, C. G., & Smyth, R. (1998). Structural parallelism and discourse coherence: A test of centering theory. *Journal of Memory and Language, 39*, 593-608. doi: 10.1006/jmla.1998.2575

Crawley, R. A., Stevenson, R. J., & Kleinman, D. (1990). The use of heuristic strategies in the interpretation of pronouns. *Journal of Psycholinguistic Research, 19*, 245-264. doi: 10.1007/bf01077259

Featherstone, C. R., & Sturt, P. (2010). Because there was a cause for concern: An investigation into a word-specific prediction account of the implicit-causality effect. *The Quarterly Journal of Experimental Psychology, 63*, 3-15. doi: 10.1080/17470210903134344

Ferstl, E. C., Garnham, A., & Manouilidou, C. (2011). Implicit causality bias in English: a corpus of 300 verbs [Supplemental material]. *Behavior Research Methods, 43*, 124-135. doi: 10.3758/s13428-010-0023-2

Fiengo, R., & May, R. (1996). Anaphora and identity. In S. Lappin (Ed.), *Handbook of contemporary semantic theory* (pp. 117-144). Oxford, UK: Blackwell.

Garnham, A., Oakhill, J., & Cruttenden, H. (1992). The role of implicit causality and gender cue in the interpretation of pronouns. *Language and Cognitive Processes, 7*, 231-255. doi: 10.1080/01690969208409386

Garvey, C., Caramazza, A., & Yates, J. (1974). Factors influencing assignment of pronoun antecedents. *Cognition, 3*, 227-243. doi: 10.1016/0010-0277(74)90010-9

Grober, E. H., Beardsley, W., & Caramazza, A. (1978). Parallel function strategy in pronoun assignment. *Cognition, 6*, 117-133. doi: 10.1016/0010-0277(78)90018-5

Hobbs, J. R. (1979). Coherence and coreference. *Cognitive Science, 3*, 67-90. doi: 10.1207/s15516709cog0301_4

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

Kennedy, C., & Boguraev, B. (1996). *Anaphora for everyone: Pronominal anaphora resolution without a parser.* Proceedings of the 16th conference on computational linguistics, Copenhagen.

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

Lappin, S., & Leass, H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics, 20*, 535-561.

Lawlor, K. (2010). Varieties of coreference. *Philosophy and Phenomenological Research, 81*, 485-495.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. Chicago: University of Chicago Press.

Nichols, S., & Bruno, M. (2010). Intuitions about personal identity: An empirical study. *Philosophical Psychology, 23*, 293-312.

Noonan, H. W. (1985). The closest continuer theory of identity. *Inquiry, 28*, 195-229.

Nozick, R. (1981). *Philosophical explanations.* Cambridge, MA: Harvard University Press.

Parfit, D. (1984). *Reasons and persons*. Oxford, UK: Oxford University Press.

Pickering, M. J., & Majid, A. (2007). What are implicit causality and consequentiality? *Language and Cognitive Processes, 22*, 780-788. doi: 10.1080/01690960601119876

Pyykkonen, P., & Jarvikivi, J. (2010). Activation and persistence of implicit causality information in spoken language comprehension. *Experimental Psychology, 57*, 5-16. doi: 10.1027/1618-3169/a000002

Récanati, F. (2012). *Mental files*. Oxford, UK: Oxford University Press.

Rips, L. J., Blok, S., & Newman, G. (2006). Tracing the identity of objects. *Psychological Review, 113*, 1-30. doi: 10.1037/0033-295x.113.1.1

Rudolph, U., & Försterling, F. (1997). The psychological causality implicit in verbs: A review. *Psychological Bulletin, 121*, 192-218. doi: 10.1037/0033-2909.121.2.192

Shafir, E., Simonson, I., & Tversky, A. (1993). Reason-based choice. *Cognition, 49*, 11-36.

Shoemaker, S. (1979). Identity, properties, and causality. *Midwest Studies in Philosophy, 4*, 321-342.

Stewart, A. J., Pickering, M. J., & Sanford, A. J. (2000). The time course of the influence of implicit causality information: Focusing versus integration accounts. *Journal of Memory and Language, 42*, 423-443.

Vonk, W. (1985). The immediacy of inferences in the understanding of pronouns. In G. Rickheit & H. Strohner (Eds.), *Inferences in text processing* (pp. 205-218). Amsterdam: North-Holland.

Williams, B. (1970). The self and the future. *Philosophical Review, 79*, 161-180.

Williams, B. (1982, February 18). Cosmic philosopher. *New York Review of Books, 29*, 32-34.

Wolf, F., Gibson, E., & Desmet, T. (2004). Discourse coherence and pronoun resolution. *Language and Cognitive Processes, 19*, 665-675. doi: 10.1080/01690960444000034

Footnotes

[1] Critics of Nozick's (1981) theory have pointed out that this formulation makes identity context-sensitive (Noonan, 1985; Williams, 1982), and the same is true of our own model. By clause (b) of the Causality Guides Identity principle, identity (e.g., whether $x = y$) depends on whether y is closer to x than other competitors, and this implies that an item could be identical relative to one set of competitors but nonidentical relative to another. The criticism is that identity should depend only on x and y and not on other items that happen to exist at the same time as x or y. But although this point raises a possible problem for the metaphysics of identity, it may be an advantage for a psychological theory of identity *judgments*. Many studies of judgment and decision making (e.g., Shafir, Simonson, & Tversky, 1993), testify to the context sensitivity of choice: People's choices depend on the range of options on offer. It would be surprising if judgments of identity were not subject to the same kind of context sensitivity, and in fact, evidence for such an effect appears in earlier experiments (see, e.g., Rips et al., 2006, Figure 5).

[2] The theory leaves room for debate about exactly which causal forces support specific types of objects over time. In the case of people, in particular, the critical forces may be bodily processes (e.g., Williams, 1970) or psychological processes, such as a person's memory for his or her earlier experiences (e.g., Parfit, 1984). For experimental evidence on whether people believe psychological or bodily processes are responsible for personal identity, see Blok, Newman, and Rips (2005) and Nichols and Bruno (2010). For further thoughts about the nature of the sustaining causal forces, see Blok, Newman, and Rips (2007)

[3] However, causal information can trump gender marking in pairs like *Ed performed a sex change operation on Fred*. *She became a better-adjusted person*. We interpret *she* to be Fred, despite the clash in grammatical gender and the lack of parallel syntactic roles.

[4] An account of coreference based on causality is more explanatory than accounts based on general properties, such as plausibility, salience, and prominence. Of course, the correct antecedent for a pronoun is the one that makes the discourse more plausible, but this generality does little more than re-state the problem of what explains the correct assignment. The claim that the right assignment is more plausible than the wrong ones seems to mean only that the former is a better interpretation than the latter, which verges on a tautology. The theoretical work to be done in explaining pronoun assignment is unpacking the components of "plausibility" and their weighting in people's decisions. The present analysis proposes that causal continuity is of first importance in this ranking, and it provides a reason why this should be the case by linking pronoun resolution to an independently motivated theory about object identity.

The characters and props of discourse vary in their salience, and readers and listeners register these variations. When encountering a pronoun, people may consider the discourse entities as potential antecedents in order of their relative salience, and salience-based algorithms have proved useful in systems for anaphora resolution in computational linguistics (e.g., Kennedy & Boguraev, 1996; Lappin & Leass, 1994). But salience is too vague a concept to produce a satisfactory theory since salience is a cover term for more cognitively and linguistically basic factors, such as recency, given-new structure, and many others. Nor is salience either necessary or sufficient for coreference. The sentence *It was an old lady that swallowed a fly* features the old lady as new information, which makes her more salient than her snack. Still, if the following sentence is *She caused a bad case of gastritis*, then *she* is the fly, not the lady.

[5] In more detail, let p_o be the probability that the strength of the object path is above the criterion, and p_s the probability that the strength of the subject path is above criterion for a given sentence pair. As just noted, we can compute p_o from the proportion of the normal distribution that falls above criterion c, where the distribution's mean μ_o and standard deviation σ are estimated from the causal ratings. Similarly, for p_s. The predicted probability of a *neither* response, P(neither), is then equal to the probability that both paths are below c:

$$P(\text{neither}) = (1 - p_o)(1 - p_s).$$

Participants will make an object response under two conditions: (a) The strength of the object path exceeds *c* but the strength of the subject path doesn't [with probability $p_o(1 - p_s)$], or (b) both strengths are greater than *c* but the strength of the object path is greater than that of the subject path. The latter probability is the proportion of the normal distribution greater than 0, where $\mu_o - \mu_s$ is the distribution's mean and $\sqrt{2}\sigma$ is its standard deviation. If $p_{o>s}$ stands for this latter probability, then the predicted probability of an object response is:

$$P(\text{object}) = p_o(1 - p_s) + p_o p_s p_{o>s}.$$

The predicted probability of a subject response, *P*(subject) can be obtained by subtraction:

$$P(\text{subject}) = 1 - P(\text{object}) - P(\text{neither}).$$

[6] Here, participants are making more decisions in favor of the *object* noun than the model predicts; so the deviations can't be put down to biases in favor of the subject noun, such as those mentioned in the preceding paragraph. We need to be cautious in examining these deviations, since they depend on a relatively small number of responses. But a possible explanation is a secondary object interpretation, one that isn't properly reflected in the causal strength ratings. An example is *Larry accepted a promotion from Ray. He said "Thanks."* Although the dominant interpretation was the intended one in which Larry said thanks for the promotion, 18% of participants thought Ray said thanks, presumably for Larry's good performance on the job. The causal ratings miss this last possibility. We asked participants to rate how likely it is that Larry's accepting a promotion from Ray would cause Ray to say thanks. However, it's not Larry's promotion but Larry's job performance that would trigger Ray's gratitude. The latter interpretation is still causal but based on an inferred common cause for the events of the two sentences.

Table 1

Sample Sentence Pairs

Pair Type	Head Sentence	Tail Sentence
Subject strong/	Albert invited Ron to dinner.	He spent hours cleaning the house.
Object weak	Julian beat Aaron in a boxing tournament.	He won first place.
	Charlie pursued Paul through the crowd.	He caught up a few minutes later.
	Kyle handed a prize to Bob.	He said "Congratulations."
Subject weak/	Albert invited Ron to dinner.	He brought a small gift.
Object strong	Julian beat Aaron in a boxing tournament.	He came in last.
	Charlie pursued Paul through the crowd.	He eventually got away.
	Kyle handed a prize to Bob.	He said "Thanks."
Both strong	Albert invited Ron to dinner.	He bought an expensive bottle of wine.
	Julian beat Aaron in a boxing tournament.	He was covered in sweat.
	Charlie pursued Paul through the crowd.	He ran out of breath.
	Kyle handed a prize to Bob.	He smiled.
Neither strong	Albert invited Ron to dinner.	He went to a rock concert.
	Julian beat Aaron in a boxing tournament.	He read a good book.
	Charlie pursued Paul through the crowd.	He ate a leisurely lunch.
	Kyle handed a prize to Bob.	He snored loudly.

Table 2

Mean Causality Ratings (on 1-to-9 scale) and Disambiguation Proportions (Standard Deviations Given in Parenthesis)

Pair Type	Mean Causal Likelihood Rating		Proportion of Choices for Antecedent		
	Subject Name in Tail	Object Name in Tail	Proportion of Subject Choices	Proportion of Object Choices	Proportion of "Neither" Choices
Subject strong/ Object weak	6.84 (1.18)	1.49 (0.66)	.92	.06	.02
Subject weak/ Object strong	1.54 (0.90)	6.28 (1.31)	.04	.94	.02
Both strong	6.82 (1.13)	6.04 (1.49)	.72	.25	.03
Neither strong	1.24 (0.69)	1.35 (0.85)	.15	.22	.63

Figure 1. Proportion of object noun (top panel) and *neither* choices (bottom panel). Filled circles denote Subject strong/Object weak pairs, filled diamonds Both pairs, filled triangles Neither pairs, and filled squares Subject weak/Object strong pairs. Unfilled symbols are predictions from the model described in the General Discussion, with the shape of the symbol corresponding to the type of sentences just described.

a.

b.