Published in final edited form as:
Psychol Sci. 2006 April ; 17(4): 273–277. doi:10.1111/j.1467-9280.2006.01697.x.

Don't Talk About Pink Elephants! : Speakers' Control Over Leaking Private Information During Language Production

Liane Wardlow Lane, **Michelle Groisman**, and **Victor S. Ferreira**
University of California, San Diego

Abstract

Speakers' descriptions sometimes inappropriately refer to information known only to them, thereby "leaking" knowledge of that private information. We evaluated whether speakers can explicitly control such leakage in light of its communicative consequences. Speakers described mutually known objects (e.g., a triangle) that had size-contrasting matches that were privileged to the speakers (e.g., a larger triangle visible to the speakers only), so that use of a contrasting adjective (e.g., *small*) involved referring to the privileged information. Half the time, speakers were instructed to conceal the identity of the privileged object. If speakers can control their leaked references to privileged information, this conceal instruction should make such references less likely. Surprisingly, the conceal instruction caused speakers to refer to privileged objects more than they did in the baseline condition. Thus, not only do speakers have difficulty not leaking privileged information, but attempts to avoid such leakage only make it more likely.

Though people nearly always bring their own perspectives to any given situation, sometimes they behave as though they fail to appreciate that fact. This egocentrism has been explored experimentally in tasks like that illustrated in Figure 1. In this example, four objects are positioned between two people. One person can see three of them: a triangle, circle, and heart. The other person can additionally see a fourth object—a larger triangle. If the second person is asked to identify the mutually visible triangle so that the first person can pick it out, he or she ought to say "triangle," just as "circle" would describe the sole circle. Yet sometimes speakers in this circumstance say "small triangle" instead (Horton & Keysar, 1996;Nadig & Sedivy, 2002;Wardlow & Ferreira, 2003), as if they fail to appreciate their unique perspectives.

Why might speakers produce such seemingly erroneous utterances, like "small triangle," in these situations? One possibility is that low-level factors might compel speakers to pay more or less attention to the shape that only they can see (hereafter, the *hidden* shape). (For accounts of how factors like these might operate, see Horton & Keysar, 1996, and Nadig & Sedivy, 2002.) For example, too much attention to the hidden shape may boost its salience, overwhelming the knowledge that it is hidden, and leading speakers to refer to it when labeling the to-be-described (*target*) shape. To the extent that such low-level factors are influential, utterances like "small triangle" are like "Simon says" errors; undue attention to the hidden shape compels speakers to refer to it, even though they should not and may not intend to.

Another factor that might affect the likelihood that speakers will disregard their knowledge of perspective differences is knowledge of the high-level communicative consequences of producing such errors (Clark, 1996; Schober & Brennan, 2003): When a speaker says "small

Address correspondence to Liane Wardlow Lane, Department of Psychology 0109, University of California, San Diego, La Jolla, CA 92093-0109, e-mail: lwardlow@ucsd.edu..

triangle" instead of "triangle," he or she not only has communicated which shape the addressee ought to pick out, but also has potentially leaked *implicit* information. In particular, the addressee can infer that the speaker can probably see another triangle, which is likely to be the hidden shape. In most situations, leaked information is unlikely to harm speakers' communicative goals (Nadig & Sedivy, 2002); if speakers aim to convey which triangle addressees should select, "small triangle" works about as well as "triangle" (addressees can see only one triangle), and the leaked information is largely irrelevant. Indeed, by communicating more information with fewer words, use of utterances like "small triangle" might be generally adaptive.

But what happens when leaked information conflicts with speakers' goals? Assume that in the situation illustrated in Figure 1, speakers are instructed not only to name the target shape, but also to conceal the hidden shape. In this case, speakers should avoid describing the target as "small triangle," because the leaked information might cue addressees to the identity of the hidden shape. Can speakers' high-level communicative intentions (to name the target and conceal the hidden shape) overcome their basic tendency to sometimes violate their knowledge of perspective differences? Or are the low-level factors (e.g., salience) that compel speakers to produce utterances like "small triangle" not under speakers' intentional control?

Ironic-processes theory (Wegner, 1994) suggests another possibility: Speakers may be more, rather than less, likely to refer to a hidden object precisely because of an intention to conceal it. Ironic-processes theory is a dual-process account of performance according to which an *operator* process attempts to perform a desired action, while a *monitor* process checks for signs of failure. Critically, monitoring can bring failure conditions into awareness, thereby ironically causing "precisely counterintentional" (Wegner, 1994) behaviors, especially when task conditions are challenging. For example, subjects attempting to hold a pendulum steady while counting backward by 3s will swing it along a particular axis more when that axis is forbidden than when it is not (Wegner, Ansfield, & Pilloff, 1998). Analogously, instructions to conceal the hidden object in situations like the one illustrated in Figure 1 could engage an ironic-process monitor, making counterintentional behaviors (e.g., saying "small triangle") more likely.

The present experiment used a referential communication task (Hanna, Tanenhaus, & Trueswell, 2003; Keysar, Barr, Balin, & Brauner, 2000) like that illustrated in Figure 1. Speakers described to addressees mutually visible shapes on *target* cards while trying to ignore hidden shapes on *foil* cards. On *critical* trials, the object on the target card was medium-sized (see Fig. 2). On half the critical trials (*test* trials), foils and targets were the same shape, but contrasted in size. Thus, test trials were designed to elicit utterances that included modifiers that contrasted the target with the hidden shape. On the other half of critical trials (*control* trials), the foil was a different shape from the target. Control trials thus assessed how often utterances included modifiers irrespective of the contrast to the hidden shape.

Speakers were tested in two blocks that were presented in counterbalanced order. In *baseline* blocks, speakers were instructed to identify each target so that addressees could select it from the mutually visible set. Separate scores were kept for speakers and addressees, each receiving 1 point whenever addressees selected the target. In *conceal* blocks, participants were given additional instructions encouraging speakers to hide the foil's identity when identifying the target. Specifically, after addressees selected a target, they were allowed to guess the identity of the foil. A point was added to addressees' scores for each correct guess, and a point was subtracted for each incorrect guess. Speakers were instructed not to allow addressees to gain additional points. Therefore, speakers should have avoided behavior that

might cue the foil's identity (e.g., producing utterances like "small triangle"), because the modifying adjective could cue the identity of the foil.

Performance in the conceal condition should show whether speakers can control leaking information in light of communicative consequences. If speakers have such control, instruction to conceal the hidden shape should reduce the mention of foil-contrasting modifiers relative to their frequency when no conceal instructions are given (i.e., in the baseline block). However, if information is leaked as an uncontrollable consequence of low-level factors such as attention increasing the salience of the hidden shape, then instruction to conceal that hidden shape should not decrease production of modifiers. Finally, according to an ironic-processes account, the conceal instruction should cause speakers to use modifiers even more in the conceal block than in the baseline block.

METHOD

Subjects

Participants were 88 undergraduates at the University of California, San Diego. Forty-four served as speakers, and 44 as addressees. All participants were native speakers of English.

Materials and Design

Participants were tested with 288 cards. Each displayed one simple line drawing of a familiar object. The objects varied in actual size across and within trials such that the size of a given object relative to the size of the other objects on the same trial could be large on some trials and small on others. The target and foil on each test trial differed in size. Each object type was used only on one trial, and no object ever occurred with more than one other object of the same type.

Two manipulations were used: contrast type (test vs. control) and instruction (conceal vs. baseline). On test trials, the foil contrasted in size with the to-be-named mutually visible object. On control trials, the to-be-named mutually visible object was unique. On conceal trials, speakers were told not to provide addressees with any information about the hidden shape. On baseline trials, speakers were not given any special instructions regarding the hidden shape.

Four experimental conditions were assigned to each target object by crossing the levels of contrast type and instruction. Both factors were manipulated within speakers and items in counterbalanced fashion. Subjects were presented with 36 critical trials, half test and half control, and, factorially, half in the conceal condition and half in the baseline condition. Conceal and baseline trials were blocked, with the order of blocks counterbalanced across subjects. Thirty-six additional filler trials, half composed of two pairs of cards and half composed of one pair and two unique cards, were administered. Speakers were asked to identify a mutually visible shape on all trials.

Procedure

A coin toss randomly assigned participants to the roles of speaker and addressee. Participants sat at opposite sides of a table; the speaker could see a computer monitor, but the addressee could not. At the beginning of each trial, the addressee closed his or her eyes while the experimenter placed four cards on the table. The speaker then looked at the computer monitor, which displayed a schematic of the four blank cards, one of which had an arrow above it with the instruction, "Block this card." The speaker blocked the corresponding actual card by positioning an occluder between the card and the addressee so that the addressee could not see the card. Next, the speaker looked back at the computer

screen, which showed the four-card schematic with an arrow pointing at a different card with the instruction, "Identify this card." The speaker was instructed to describe the corresponding card with just enough information so that the addressee could identify it. Upon hearing the speaker's description, the addressee opened his or her eyes and attempted to identify the target card. On conceal trials, the addressee was told that after trying to identify the described card, he or she could guess the identity of the hidden shape.

RESULTS AND DISCUSSION

On critical trials, the description of the target was transcribed and coded for whether it included a designated modifier (e.g., the description of the medium-sized triangle was coded for whether it was described as "small triangle" on both test and control trials). For each experimental condition, the percentage of targets described with such modifiers was computed for each subject. These percentages were submitted to repeated measures analyses of variance (ANOVAs) using subjects (F_1) and items (F_2) as random factors. (Analyses carried out using arcsine-transformed proportions yielded the same pattern of significance as reported here.) The ANOVA design was 2 × 2, with the factors of contrast type and instruction. Planned comparisons assessed performance on test versus control trials separately under each instruction condition. All significant effects achieved the .05 level unless otherwise specified. We report variability with repeated measures 95% confidence-interval half-widths (CIs) based on single-degree-of-freedom comparisons (Loftus & Masson, 1994).

Figure 3 shows the mean percentages of target descriptions that included specified modifiers (e.g., "small triangle") as a function of contrast type and instruction. As expected, speakers produced more modifiers overall on test trials (10%) than on control trials (0.9%), $F_1(1, 43) = 16.3$, CI = ±4.5%, $\eta_p^2 = .275$; $F_2(1, 35) = 35.9$, CI = ±3.0%, $\eta_p^2 = .507$; this result shows that modifiers were used specifically in response to the size contrast. Speakers also produced more modifiers overall in conceal blocks (7.9%) than baseline blocks (3.0%), $F_1(1, 43) = 6.5$, CI = ±3.9%, $\eta_p^2 = .131$; $F_2(1, 35) = 10.1$, CI = ±3.3%, $\eta_p^2 = .224$. In fact, speakers produced 13% more adjectives on test trials (14.4%) than on control trials (1.4%) in the conceal block, but only 4.9% more adjectives on test trials (5.4%) than on control trials (0.5%) in the baseline block, leading to a significant interaction, $F_1(1, 43) = 5.3$, CI = ±5.1%, $\eta_p^2 = .109$; $F_2(1, 35) = 6.4$, CI = ±4.4%, $\eta_p^2 = .156$. Planned comparisons revealed the difference between test and control trials to be significant in the conceal condition, $F_1(1, 43) = 26.9$ and $F_2(1, 35) = 34.1$; marginally significant by speakers in the baseline condition, $F_1(1, 43) = 3.8$, $p < .06$; and significant by items in the baseline condition, $F_2(1, 35) = 5.0$. In short, speakers tended to modify target descriptions with respect to hidden information, but they did so even more when instructed to conceal the hidden information than when not so instructed.

Thus, at least under these task conditions, speakers were unable to control whether they leaked hidden information, and did not reduce such leakage when it had negative consequences. In fact, the opposite was observed: When speakers were provided with instructions and incentives not to leak information about a hidden shape, they were even more likely to do so. These results support the idea that when speakers fail to account for their unique perspectives, it is because relatively autonomous, low-level processes cause privileged information to be unintentionally incorporated into their descriptions. This finding carries implications for theories of language production. Generally, production models distinguish conceptual processing and grammatical encoding (Bock, 1982; Levelt, 1989). It is at the level of conceptual processing that speakers encode their communicative intentions and define a to-be-conveyed message. That message consists of those conceptual features that are accessible and whose production will allow speakers to achieve their

communicative objective. However, additional conceptual features that are not needed to convey the intended message may also be accessible (e.g., contextually activated but unimportant information, or relevant but private information). The question is whether grammatical encoding processes encode only those conceptual features that make up the intended message, or whether they can also encode features that, though accessible, are not intended to be expressed. Given the present results, we suggest that being part of a communicative intention is not a necessary condition for an accessible conceptual feature to influence grammatical encoding.

Furthermore, the direction of the observed difference between the two instruction conditions can be accounted for with ironic-processes theory. Specifically, the conceal instruction may have exaggerated the influence of a monitor process tasked with checking for failure. Note that in the present experiment, unlike in previous demonstrations of ironic-processes effects (see Wegner, 1994), counterintentional behaviors did not arise as a function of increased cognitive load, perhaps because the demands of production are already inherently taxing.

Indeed, if ironic-processes mechanisms are responsible for the outcome reported here, these results extend the practical implications of ironic-processes effects into the communicative domain. Consider the relation between the present results and the well-known observation that when directed not to think of a pink elephant, people inevitably do just that. The latter observation illustrates that people do not have total explicit control over what thoughts come to mind. The present results incorporate this observation, as the instruction to conceal the hidden shape evidently only made that shape more salient. However, the present results go further, by showing that when directed to conceal the hidden shape, speakers were more likely not only to think of it, but also to refer to it. The fact that private information was sometimes leaked despite explicit attempts to avoid doing so suggests not only that leaked information may sometimes be information speakers might want to keep private, but also that attempts to conceal that private information might make its leakage even more likely. If so, these results are likely to be relevant to many kinds of interactions, ranging from interpersonal interactions to adversarial negotiation.

Acknowledgments

This work was presented at the 13th annual Architectures and Mechanisms for Language Processing conference, September 2005, Ghent, Belgium. This research was supported by National Institutes of Health Grant R01 MH64733. We thank Kristy Lawson for assistance with data collection, Bob Slevc and Tamar Gollan for helpful discussions, and an anonymous reviewer for helpful comments.

References

Bock JK. Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. Psychological Review. 1982; 89:1–47.

Clark, HH. Using language. Cambridge, England: Cambridge University Press; 1996.

Hanna JE, Tanenhaus MK, Trueswell JC. The effects of common ground and perspective on domains of referential interpretation. Journal of Memory and Language. 2003; 49:43–61.

Horton WS, Keysar B. When do speakers take into account common ground? Cognition. 1996; 59:91–117. [PubMed: 8857472]

Keysar B, Barr DJ, Balin JA, Brauner JS. Taking perspective in conversation: The role of mutual knowledge in comprehension. Psychological Science. 2000; 11:32–38. [PubMed: 11228840]

Levelt, WJM. Speaking: From intention to articulation. Cambridge, MA: MIT Press; 1989.

Loftus GR, Masson MEJ. Using confidence intervals in within-subject designs. Psychonomic Bulletin & Review. 1994; 1:476–490.

Nadig AS, Sedivy JC. Evidence of perspective-taking constraints in children's online reference resolution. Psychological Science. 2002; 13:329–336. [PubMed: 12137135]

Schober, MF.; Brennan, SE. Processes of interactive spoken discourse: The role of the partner. In: Graesser, AC.; Gernsbacher, MA., editors. Handbook of discourse processes. Mahwah, NJ: Erlbaum; 2003. p. 123-164.

Wardlow, L.; Ferreira, VS. Finding common ground: How do speakers block privileged information?; Poster presented at the Architectures and Mechanisms for Language Processing conference; Glasgow, Scotland. 2003 September.

Wegner DM. Ironic processes of mental control. Psychological Review. 1994; 101:34–52. [PubMed: 8121959]

Wegner DM, Ansfield M, Pilloff D. The putt and the pendulum: Ironic effects of the mental control of action. Psychological Science. 1998; 9:196–199.

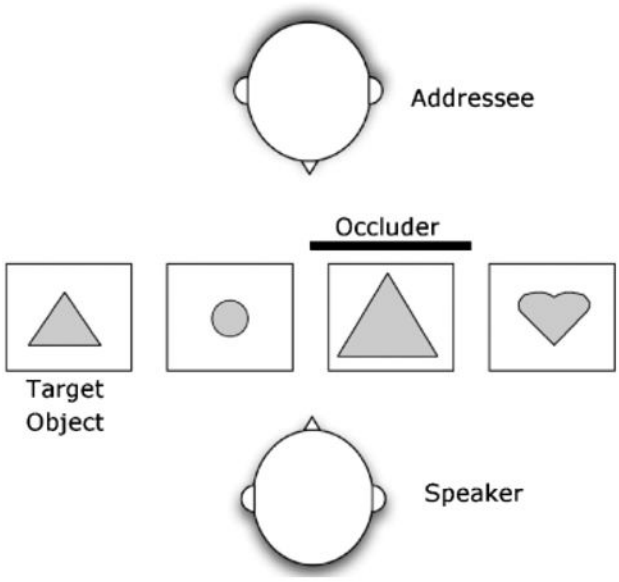

Fig. 1.
Example of the expergimental setup.

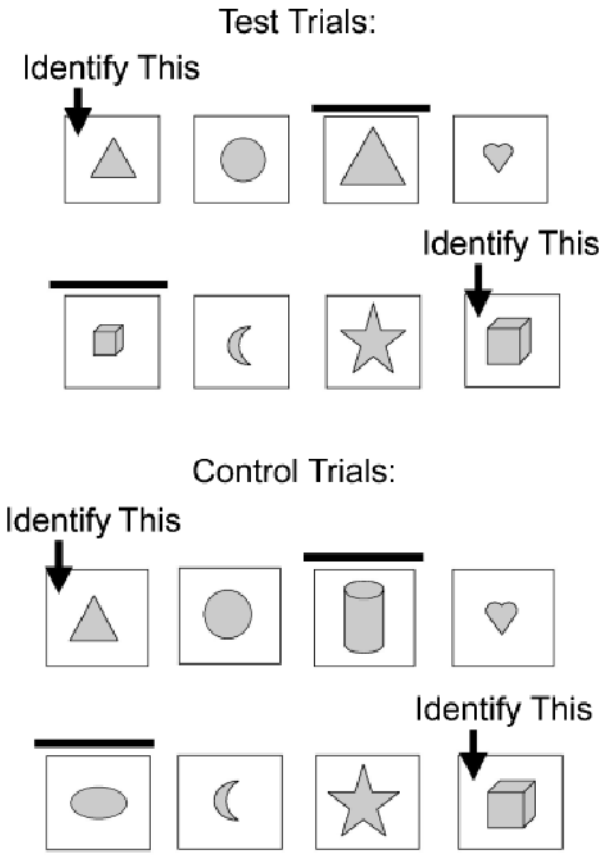

Fig. 2.
Examples of the stimuli used on test (top) and control (bottom) trials.

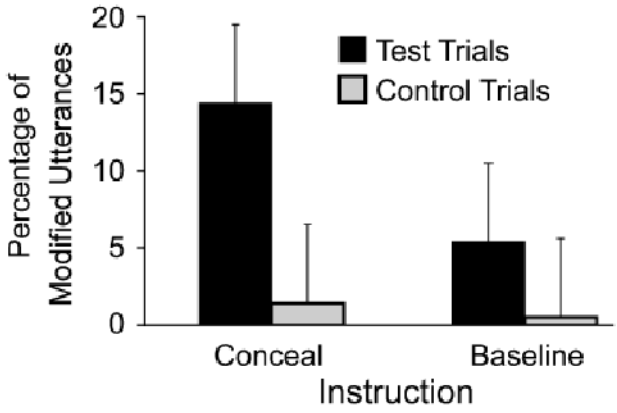

Fig. 3.

Percentage of target descriptions including foil-contrasting modifiers as a function of whether the foil and target were the same shape (test trials) or different shapes (control trials) and whether speakers were given instructions to conceal the hidden shape (conceal trials) or not (baseline trials). Error bars illustrate 95% confidence intervals of the interaction by speakers.