Squibs and Discussions

Do the Right Thing . . . but Expect the Unexpected

Jon Oberlander*
University of Edinburgh

1. Do the Right Thing

Dale and Reiter (1995) have recently discussed the nature of referring expression generation, focusing on the case of definite noun phrases. In particular, they consider Gricean approaches, whereby the speaker is supposed to take into account likely inferences by the hearer, in accord with Gricean maxims (Grice 1989), and select the generated NP accordingly, so as to avoid false or misleading inferences (Joshi 1982). They observe that previous accounts (including their own) have attempted to optimize the generated noun phrase, making it as brief as possible, within the constraints of accurately distinguishing the intended referent from any other candidate referents. For instance, consider a situation containing three animals: one small white cat and two dogs, one large and black, and the other small and white. It is usually assumed that an optimal description of the first dog is either *the large dog* or *the black dog*, whereas *the large black dog* will be suboptimal, since it contains two adjectives where one will do; it is longer than strictly necessary, and suffers from a degree of redundancy (Dale 1992; Reiter 1990).

However, Dale and Reiter argue that the previous algorithms proposed for this task are computationally inefficient, and that the task itself must be reconsidered. In particular, they suggest that there is substantial psycholinguistic evidence that *people* don't generate the shortest, most efficient NPs, and that this behavior is regarded as perfectly natural (see Levelt [1989] for a survey). Hence, generation algorithms need not optimize their descriptions either.

Dale and Reiter go further; they state that:

> One could even argue that an algorithm based on psycholinguistic observations of human speakers may in fact be superior to one that attempts to interpret the maxims as strictly as (computationally) possible. This would be justified if one believed that the Gricean maxims were simply an approximation to the general principle of "if a speaker utters an unexpected utterance, the hearer may try to infer a reason for the speaker's failure to use the expected utterance"; under this perspective, a system that imitated human behaviour would be more likely to generate 'expected' utterances than a system that simply tried to obey general principles such as brevity, relevance, and so on. (p. 253)

The primary point is that the behavior of human speakers involves the production of nonminimal utterances, and their hearers expect this behavior. Conversely, hearers do not expect speakers to produce optimal, minimized utterances; such an unexpected utterance would in fact provoke its hearer to search for reasons for its speaker's failure

* Human Communication Research Centre, University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, Scotland

to use an expected utterance. Both people and natural language generation systems should therefore strive to produce the most expected utterance, if they are to avoid unwanted implicatures. It has been suggested that this position can be encapsulated in a new high-level maxim:

Spike Lee's Maxim

Do the right thing.[1]

Gricean maxims have been discussed in considerable detail; by analogy, we can scrutinize the ideas underlying this "Spikean maxim," and the injunction to "do the right thing." The key question is whether we can make consistent sense of the notion of a generator attempting to generate an "expected" utterance.

2. What's Right?

So, let us equate generating expected utterances with doing the right thing; the failure to generate them counts as "doing the wrong thing," and leads to additional processing effort on the part of the hearer. But what is the "right" thing? What counts as "expected"? Given Dale and Reiter's evidence, we can immediately concede that the shortest, most efficient description need not be the most expected, and hence is not always the right thing. But is there any way of spelling out the injunction in more detail?

In fact, there are two ways of adding detail, and both seem to be intended by Dale and Reiter; we may thus consider them Spikean submaxims:

1. Do the human thing.
2. Do the simple thing.

The first of these trades on the obvious fact that human language users are sensitive to the conventional behavior of other human language users. We learn those conventions and, by generating in accord with them, produce the kinds of utterance expected under the circumstances. We thereby maximize our chances of being understood. So, at an engineering level, the best thing for natural language engineers to do is to build systems that emulate human behavior as faithfully as possible.

The second submaxim is slightly less obvious, but should come as a relief for the engineers. The right thing might after all be characterized as the simplest output—so long as the simplicity lies in the algorithm that produced it, instead of in the relative complexity of the output string itself. The nature of the algorithm is an empirical matter, and a good way of uncovering it is, of course, through observing human behavior.

Indeed, according to Dale and Reiter, the psycholinguistic evidence on definite noun phrases is that doing the human thing and doing the simple thing go hand

1 The suggested formulation is Robert Frederking's, and it was proposed at the 1996 AAAI Spring Symposium on Computational Implicature, chaired by Barbara Di Eugenio and Nancy Green. The name of the maxim was settled by popular consensus, and is a reference to Spike Lee's 1989 film *Do the Right Thing*. As we shall see, this maxim mainly provides a convenient label, bringing together the more detailed claims made by Dale and Reiter.

in hand; people use a simple algorithm, which doesn't waste excessive resources on computing possible misinterpretations:

> The principle that has emerged from our study of the referring expression generation task is that a simple and nonliteral interpretation of the Gricean maxims is to be preferred . . . Perhaps it may some day be possible to make a very general statement such as "human speakers in general use very simple (in computational terms) interpretations of the maxims of conversational implicature, and hence computer natural language generation systems should also use such interpretations." (p. 262)

It is easy to see how one might be tempted to assume that the human thing is the simple thing, and that this grounds out the notion of an expected utterance. For one thing, if speakers follow simple algorithms, then since most hearers are themselves speakers, they could conceivably predict speakers' behavior by unconsciously anticipating what they themselves would do. But of course, this only stays simple—and avoids infinite recursion—so long as the predicted speaker behavior doesn't involve consultation of a sophisticated model of the hearer. However, a more convincing reason might lie in the conventional behavior of language communities. In a community of language users, if speakers follow simple algorithms, then simply generated utterances will provide the corpus of observed human language use from which any conventions will arise, and from there guide future behavior. The simple algorithm for referring expressions could involve run-time computation of the set of predicates to use, or it could involve selection from precompiled items in a phrasal lexicon; but either way, behavior in accord with the simple algorithm becomes what is expected by other speakers.

To summarize this position: human speakers can do the right thing, and hence observe the Spikean maxim, by following simple algorithms. Emulating human generation behavior, via the use of a simple, empirically discovered algorithm, delivers a system that generates the expected utterances.

3. What's Wrong with This?

Unfortunately, this position might be plausible in the case of definite noun phrase generation, but it cannot be correct in general. The difficulties hinge on the notion of expectation. Dale and Reiter's argument relies on psycholinguistic findings on the generation of definites to help reveal what people regard as expected or unexpected. But there are other empirical results, concerning the generation and interpretation of pronouns, which do not fit into this picture.

A good deal of work has been carried out on human behavior with respect to the processing of pronominal expressions. In particular, one strand of research has examined the psychological plausibility of Grosz, Joshi, and Weinstein's (1983, 1995) Centering Theory. Hudson D'Zmura (1988), for instance, is one of several researchers to have shown that pronouns in subject position that specify the highest-ranked *Cf* (forward-looking center) of the preceding utterance are interpreted more rapidly than repeated names in subject position. This is attributed to an expectancy effect because the subject position is the preferred site of the *Cb* (backward-looking center), which is normally a pronoun specifying the highest ranked *Cf*. A related strand of research, by Stevenson and her collaborators (Stevenson, Crawley, and Kleinman 1994; Stevenson and Urbanowicz 1995), connects this work on centering to other possible influences on processing, including preferences concerning thematic roles (Dowty 1991), and the effects of connective expressions in multiclause sentences.

Stevenson and her collaborators have pursued two main types of empirical study: continuation tasks, and reading time tasks. In continuation tasks, a subject is typically presented with a sentence or sentence fragment, and asked to continue it. When the fragment contains two sentences (or clauses), the first mentioning two entities, and the second either empty, or containing merely an initial pronoun, the completions can be categorized on the basis of which entity functions as antecedent, and the results analyzed to reveal preferences for particular patterns of anaphoric reference. In reading time studies, a subject is presented with a complex sentence, or pair of sentences, and the time they take to read it is measured and analyzed.

Both types of study are highly relevant to the current issue. Continuation studies examine the kinds of (written) utterances that people prefer to generate in given, carefully controlled contexts. Reading time studies examine the relative ease with which people interpret the (written) utterances they are presented with in carefully controlled contexts. So, on the one side, we would predict that continuation preferences will reveal the output that speakers are most likely to generate in a given context; on the other side, we would predict that reading times will reveal which inputs hearers expect most strongly in a given context. It is worth reiterating the latter point: a good guide as to whether an utterance is expected or not is the amount of processing it gives rise to; certainly, Dale and Reiter are not alone in assuming that an unexpected utterance will give rise to additional inferences on the part of its hearer. In the case of written text, the amount of processing is operationally detected by measuring reading times: unexpected sentence fragments in a given context will take longer to read than expected fragments in the same context.

Now, Stevenson's own hypothesis is that preferences due to centering constraints interact with those due to the thematic roles of the entities referred to. On this view, centering primarily influences *how* an entity introduced in one sentence will be referred to in the next (by pronoun, or by name, for instance); thematic roles influence *which* entities will be subsequently referred to (the Agent, or the Patient, from the first sentence, for instance). In particular, centering tells us to expect a pronoun in subject position to specify the highest ranked *Cf* from the previous sentence. On the other hand, thematic role information tells us to expect that the subject of the current sentence is more likely to specify an entity associated with the consequences of the event introduced in the previous sentence; thus, if the verb in the previous sentence introduced roles for Goal and Source, then the subject of the current sentence is most likely to be the Goal from the previous sentence.

To illustrate, analysis of continuation data confirms that people prefer to use a pronoun to refer to the entity in initial position and to use a repeated name for the entity in second position. This effect is independent of who gets referred to, which depends on the thematic role of each referent. Compare examples (1) and (2):

(1) John gave the book to Bill and ...

(2) Bill took the book from John and ...

People continue the fragment with a pronoun when they want to refer to John in (1) and Bill in (2), but they are more likely to repeat the name when they want to refer to Bill in (1) and John in (2). This is despite the fact that in both sentences Bill (the Goal) is the person they are most likely to refer to.

Now, there are some very interesting apparent discrepancies between certain results from the continuation studies and the reading time studies. Take an example of the form in (3). Look at four variants, all using the connective *so*:

(3) a. John gave the book to Bill so he ... [*he* = *John*]

b. John gave the book to Bill so John ...

c. John gave the book to Bill so he ... [*he* = *Bill*]

d. John gave the book to Bill so Bill ...

First, consider the results of continuation studies for such examples; the evidence here is from Stevenson, Crawley and Kleinman's (1994) third experiment, the results of which confirm the findings from two other continuation experiments reported there. The materials contained three initial fragment types: goal-source; experiencer-stimulus; and agent-patient. Each fragment ended in a connective. The design manipulated two factors: order of thematic roles in the initial fragment (for instance, source-goal, as in (3a), or goal-source); and connective (*so* versus *because*).

It was found that, if a person generated the pronoun *he* as the first word in their continuation, it was almost always used to refer to John (pp. 537–39). It seems that the effects of centering swamp the effects of thematic role—the generator here prefers to write about John. Thus, (a)-type continuations would be preferred to (c)-type continuations.

Now, compare the results of reading time studies for such examples; the evidence here is from Stevenson and Urbanowicz (1995). The materials consisted of two-clause sentences, such as (4):

(4) Malcolm won some money from Stuart because he was very good at poker.

Each sentence was presented a clause at a time, followed by a comprehension question, which probed the correct resolution of any anaphor that appeared. The design manipulated four factors: type of anaphor in second clause (pronoun, such as *he*, versus repeated name, such as *Malcolm*); order of thematic roles in the initial fragment (source-goal versus goal-source); connective (*so* versus *because*); and antecedent (Goal or Source).

It was found that, when encountered, (c)-type sentences were read significantly faster than (a)-type sentences (p. 331). That is, with *so* and *he*, complex sentences where the antecedent is the Goal-in-second-position prove faster to read than those in which the antecedent is Source-in-initial-position. It seems that effects of thematic role win out—the interpreter expects to read about Bill.

Bringing the results of these studies together, there is an apparent asymmetry between interpretive and generative behavior. At a thematic level, both generators and interpreters prefer to talk about goals, and so (c)-type and (d)-type sentences are most likely to be generated, and are most expected by interpreters. However, at the syntactic realization level, generators will only rarely produce (c)-type utterances. Given a fragment ending with *so*, a generator that next outputs *he* will usually go on to produce an (a)-type sentence. Yet, given that reading time is an indication of expectedness, it seems that, on the contrary, (c)-type sentences are *more expected*, and easier to process, than (a)-types.

To put it another way, given a certain type of prior context, generators will systematically fail to produce the utterance that is easiest to process. And to put it yet another way, this is evidence *against* the hypothesis that people do the right thing, in the sense of generating the most expected utterance.

In fact, this body of evidence simultaneously undermines the idea that the output of a simple, highly plausible algorithm will be what is most expected, and the

idea that the emulation of human behavior will invariably produce the most expected utterances. The first idea falls because a simple centering-based generator would invariably generate (a)-type continuations—and these are less expected than (c)-type continuations. Now, it could be suggested that we replace the simple centering algorithm with one that also includes thematic role preferences. Unfortunately, even if this algorithm correctly emulates human behavior, it will still generate (a)-type continuations in preference to (c)-types—because that is what people do, too. And for this reason, the second idea also falls: emulating human behavior here cannot hope to produce the most expected utterances, for the good reason that *people* don't produce the most expected utterances in the first place.

To summarize this argument: if you emulate human behavior, you must—as a matter of course—expect to generate unexpected utterances. Thus, if doing the right thing is doing the simple thing, or even just doing the human thing, it is here doomed to fail to generate the expected utterance.

4. Expect the Unexpected

Now, how serious is this problem for Dale and Reiter? If a psycholinguistically inspired algorithm for pronoun generation leads us to expect unexpected output, should we conclude that doing the right thing—the human thing, the simple thing—will always get the wrong results?

Obviously not. Dale and Reiter's algorithm was for generating definite noun phrases, not pronouns. It might be thought that by comparing continuation and reading time studies on definites, we could undermine their position more directly. However, there is a sense in which this would be immaterial; their algorithm clearly achieves reasonable output much of the time. Thus, it is tempting to conclude that Spike Lee's maxim simply has limited applicability: people (and machines) should do the right thing, but only on some—not all—generation tasks.

However, this does not seem quite right, either. Rather, the key lesson from the work on pronoun generation and interpretation is that we must develop a more sophisticated view of "expectation." In all the examples like (3), Bill is the person a generator is expected to talk about—and for whom a pronoun would thus make good sense—but this thematically based expectation must be played off against the centering-based expectation; there are multiple factors underlying any expectation. The existence of such multiple factors is the fundamental reason why speakers can always be put into a context in which they will generate utterances they themselves find relatively hard to interpret. The existence of multiple factors also carries a more general lesson for computational linguists. If we hope to trade in psycholinguistic findings for new algorithms, we must resist the temptation to be overly selective concerning the results we pick on.[2]

Thus, Dale and Reiter could be strategically correct: careful study of human generative behavior may well reveal that it is less complex than we have come to believe, and that emulating this simple behavior is the right policy in general. Doing the right thing, emulating human performance *is* generally better than slavishly following literal interpretations of the maxims.

Nonetheless, it remains true that, because there are multiple factors underlying any expectation, a human-inspired algorithm for generating the expected utterance can in fact produce unexpected utterances. However simple or complex the algorithm

2 This paper has itself obviously been selective; but it is not intended to be perniciously so.

for choosing what to talk about next, and how to refer to it, people will still say things they would find surprising. Human performance thus embraces a mild paradox, which means that even the best emulator will always generate unexpected utterances, which violate Spike Lee's maxim, and cause unwanted implicatures.

Some would argue that the semblance of paradox is simply evidence that there are limits to the experimental paradigm. That there are such limits is not in doubt. But there remains something odd about the fact that not only would I myself generate the unexpected utterance, but everyone else would too. Thus, over time, I really should come to expect this unexpected utterance—and it should therefore stop being unexpected. Perhaps I am exposed to too few instances of the paradoxical behavior in my lifetime to become attuned to it, but the language community as a whole has already had plenty of time to adjust its expectations. It has not done so: the asymmetry in behavior appears to be stable. Thus, individual language users will continue to be surprised by the behavior of their language community. By contrast, researchers in natural language generation should not be surprised at such asymmetries in behavior. Once we understand better both the behavior, and the interplay of expectations which underlies it, analysts and engineers alike will know exactly when they should expect the unexpected.

Acknowledgments

The support of the Economic and Social Research Council for HCRC is gratefully acknowledged. The author is supported by an EPSRC Advanced Fellowship. This paper was inspired by discussions with participants at the AAAI Spring Symposium on Computational Implicature, held at Stanford, in March 1996; my thanks also to Rosemary Stevenson and Keith Stenning, and to the paper's anonymous reviewers, for very helpful comments.

References

Dale, Robert. 1992. *Generating Referring Expressions: Building Descriptions in a Domain of Objects and Processes*. MIT Press, Cambridge, MA.

Dale, Robert and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19: 233–263.

Dowty, David R. 1991. Thematic proto-roles and argument selection. *Language*, 67: 547–619.

Grice, H. Paul. 1989. *Studies in the Way of Words*. Harvard University Press, Cambridge, MA.

Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein. 1983. Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st Annual Meeting*, pages 44–50. Association for Computational Linguistics.

Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21: 202–225.

Hudson D'Zmura, Susan B. 1988. *The Structure of Discourse and Anaphor Resolution: The Discourse Center and the Roles of Nouns and Pronouns*. Ph.D. thesis, University of Rochester.

Joshi, Aravind K. 1982. Mutual beliefs in question answering systems. In Neilson V. Smith, editor, *Mutual Knowledge*. Academic Press, New York.

Levelt, William. 1989. *Speaking: From Intention to Articulation*. MIT Press, Cambridge, MA.

Reiter, Ehud. 1990. The computational complexity of avoiding unwanted computational implicatures. In *Proceedings of the 28th Annual Meeting*, pages 97–104. Association for Computational Linguistics.

Stevenson, Rosemary J., Rosalind A. Crawley, and David Kleinman. 1994. Thematic roles, focus and the representation of events. *Language and Cognitive Processes*, 9:519–548.

Stevenson, Rosemary J. and Agnieszka J. Urbanowicz. 1995. Structural focusing, thematic role focusing and the comprehension of pronouns. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, pages 328–332.