A Pragmatic, Cognitive and Computational Model of Verbal Irony

Akira Utsumi

Department of Computational Intelligence and Systems Science,
Tokyo Institute of Technology, Yokohama 226-8502, Japan
utsumi@uts.dis.titech.ac.jp

Introduction

Verbal irony is an intelligent, witty use of words found in many language activities. It has attracted the interest of linguists, philosophers, psychologists, rhetoricians and other linguistic scholars. However it appears to me that none of the previous irony theories can distinguish ironic utterances from nonironic ones completely, i.e., they do not provide a sufficient explanation of how people interpret irony. Furthermore, in the domains of computational linguistics and cognitive modeling, no attention has been given to ironic uses of language, although other nonliteral language such as metaphor has been a popular topic.

This paper presents irony research I have recently been engaged in: a comprehensive theory of irony, implicit display theory, which resolves several difficulties involved in previous irony theories, and a cognitive/computational mechanism of interpreting irony based on the implicit display theory.

Implicit Display Theory of Verbal Irony

The implicit display theory argues that verbal irony implicitly displays ironic environment. The main claim underlying the theory is threefold.

1. Verbal irony presupposes the proper situational setting, ironic environment. Ironic environment consists of the speaker's expectation, incongruity between the expectation and the reality, and the speaker's negative attitude toward the incongruity.

For example, consider a mother’s utterance (1a) given in Situation 1 and the same utterance given in Situation 2.

Situation 1. A mother asked her son to clean up his messy room, but he was lost in a comic book. After a while, she discovered that his room was still messy, and said to him:

(1) a. Your room is totally clean!

Hearers who have noticed Situation 1 have no problems understanding a mother’s ironic intension in (1a), but when the remark (1a) is given in Situation 2, it is no longer ironic. In terms of the implicit display theory, Situation 1 is surrounded by ironic environment since the ironist mother’s expectation that his room is clean has not been fulfilled and she is disappointed with or angry at the incongruity, whereas Situation 2 is not surrounded by ironic environment because it includes no apparent incongruity between her expectation and the reality.

Among the three events/states of ironic environment, the speaker’s expectation is the most essential component because the other two components cannot be identified unless the speaker’s expectation is known. Especially, all expectations that motivate irony must be attributed to, or possessed by, the speaker of irony. For example, in the following exchange, the ironist Peter has an expectation about the addressee David’s mental state (belief), i.e., Peter expects that David should know his opinion expressed by the preceding utterance is false, rather than an expectation about the state of affairs.

Situation 3. David said “I’d be promoted before you” to his colleague Peter. This elicited the following reply:

(2) Oh! you’d be promoted before me.

2. Verbal irony is an expression (utterance or statement) that implicitly displays ironic environment. Implicit display of ironic environment is, in the most prototypical cases, achieved by an utterance which alludes to the speaker’s expectation, includes pragmatic insincerity by violating one of pragmatic principles and is accompanied by a variety of verbal/nonverbal cues.

Allusion can be captured by coherence relations — e.g., Volitional-Cause, Enable — which hold between what is said and what is expected. Pragmatic principles violated by ironic utterances include the maxim of quality, felicity conditions for speech acts, politeness principle, and the maxim of quantity. Cues for indirect expression of the speaker’s negative attitude includes hyperbolic words and phrases (Kreuz and Roberts 1995), interjections, and prosodic cues like so-called ironic tone of voice.

In the example above, the utterance (1a) satisfies all these conditions and thereby implicitly displays ironic environment. First, it mentions, and thus alludes to, her expectation of the room being clean. Second, it is a literally false statement that violates the maxim of quality. Third, the hyperbolic word “totally” is used to exaggerate the ironic attitude. On the other hand, even if the following remark (1b) is made by a mother in Situation 1, it is unlikely to be ironic.

(1) b. Your room is totally messy!

In terms of the implicit display theory, it does not allude to the expectation and it is a pragmatically appropriate (i.e., sincere) utterance.

3. Verbal irony is a prototype-based category characterized by the notion of implicit display: in order for an utterance to be interpreted ironically, the utterance must be recognized to achieve implicit display to a certain degree (implicit display condition), and the situation must

1 Happé’s (1993) empirical finding on autistic people’s understanding of ironic language serves as empirical evidence for the importance of the speaker’s expectation in irony interpretation.
be identified as ironic environment through the process of checking or inferring the three components (ironic environment condition).

The prototype view of irony claims that people do not have to see all the three conditions for implicit display in order to interpret irony, nor do they notice ironic environment beforehand, as long as there is a possibility that the situation is surrounded by ironic environment. For example, people can perceive the remark (3) as ironic even though they are unaware of the events of her morning and/or her expectation and they recognize neither pragmatic insincerity nor allusion to the expectation (Barbe 1995).

**Situation 4.** You see your friend at work for the first time that day, and she says:

(3) I’ve had a great morning!

Rather, it is more appropriate to think that people judge whether an utterance is ironic by assessing the similarity between the prototype and the utterance. The notion of implicit display provides typicality conditions characterizing the prototype ‘verbal irony’, and as the prototype theory (e.g., Lakoff 1987) says, utterances with more properties of implicit display are perceived as being more ironic.

### A Cognitive Model of Irony Interpretation

According to the implicit display view, irony interpretation is to know that the discourse situation is surrounded by ironic environment by judging an utterance to be ironic, and it can be modeled as a process shown in Figure 1.

First, since most of the utterances in ordinary verbal communication are nonironic, clearly nonironic utterances are screened out by the implicit display condition without ironic environment being considered (Step 1). When an utterance satisfies the implicit display condition, the ironic environment condition is examined differently as the speaker’s expectation is known or unknown. If hearers readily recognize the allusion to the speaker’s expectation they know, they only examine the incongruity of the known expectation and the negative attitude (Step 3). On the other hand, when hearers do not know the speaker’s expectation, or do not recognize any allusion to the known expectation, the expectation must be inferred from the utterance and the discourse context by a process of hypothesis formation and evaluation (Step 2), and it is checked for whether it is incongruous with the situation and whether a negative attitude can be elicited (Step 3). Then, in both cases, if hearers are successful in recognizing ironic environment, they judge the utterance as ironic and become aware of the speaker’s ironic intention of drawing hearers’ attention to and conveying the fact that the three components for ironic environment hold in the current situation.

The implicit display condition is checked on the basis of the prototype view of irony. The similarity between an utterance and the prototype of irony is calculated as the degree of ironicalness $d(U)$ by the following formula.

$$d(U) = \begin{cases} 
    d_A + d_I + d_E & \text{(if the speaker's expectation that motivates irony is known)} \\
    d_D + d_I + d_E & \text{(otherwise)} 
\end{cases}$$

In the formula, $d_A$ denotes the degree of allusion, $d_I$ denotes the degree of pragmatic insincerity, $d_E$ denotes the degree of indirect expression of the attitude, and $d_D$ denotes the degree of desirability of the content of $U$.

The formula (4) means that when the speaker’s expectation that motivates irony is not known beforehand, sentence polarity ($d_D$) is used for a subcondition of the implicit display condition instead of allusion of the utterance ($d_A$), and positive utterances can facilitate ironic interpretation. On the other hand, sentence polarity does not affect the degree of implicit display when the expectation is known beforehand. Such formalization reflects the asymmetry of irony that many empirical studies (e.g., Kreuz and Glucksberg 1989; Kumon-Nakamura, Glucksberg, and Brown 1995) have pointed out: positive utterances are, in general, recognized to be more ironic than negative utterances. Such polarity effect is considerable when the speaker’s expectation is implicit, but when the expectation is explicit, there is no polarity effect (i.e., negative utterances such as “New York subways are dirty” uttered in a clean train can convey irony as appropriately as positive ones).

The formula also means that the implicit display condition is satisfied to the extent that its degree $d(U)$ is high. Given a certain threshold value $C$, an utterance does not satisfy the implicit display condition if a value of $d(U)$ is less than $C$. If every subcondition is either satisfied or not satisfied (e.g., $d_A$, $d_I$, $d_E$, $d_D$ take either 0 or 1), it is reasonable to assume that recognition of at least two of the three components for implicit display is enough for satisfaction of the implicit display condition (i.e., $C = 2$).

Empirical evidence to support the “2-of-3” criterion for implicit display is provided by the evaluation I conducted (Utsumi 1999c). In the evaluation, after reading 48 utterances with paragraph-length contexts which can be interpreted ironically, 48 subjects (graduate students) were asked to write down the speaker’s expectation, and to rate the degree of ironicalness and all the components for implicit display and ironic environment on 7-point scales (0–6). The result was that utterances judged to satisfy the “2-of-3” criterion were rated as significantly more ironic than
utterances judged not to satisfy, but there was no such difference between the group of utterances judged to satisfy all the three components and the group of other utterances.

Gibbs’s (1986) time-course study of irony can be seen as additional support for the proposed model. He demonstrated that subjects significantly took less time to understand ironic remarks in the explicit contexts (i.e., contexts that contained the statements motivating an explicit echoic mention of some expectation) than to understand the same remarks in the implicit contexts (i.e., contexts that contained no such statement). This finding can be explained by the proposed model. In the explicit contexts, the speaker’s expectations are quite manifest to the subjects, but in the implicit contexts they are not known beforehand. Thus the additional process of inferring the speaker’s expectation (i.e., Step 2) is required in the implicit contexts, and as a result, the ironic utterances in the implicit contexts take longer to process.

**Superiority of the Implicit Display Theory**

The implicit display theory can cope with several difficulties involved in previous irony theories, and thereby, it is much more comprehensive (Utsumi 1996, 1997). The main problem with previous irony studies is that they have attempted to provide necessary and/or sufficient properties for distinguishing irony from nonirony, but there appears to be no such properties shared by all ironic utterances. On the other hand, the prototype-based view does not need such common properties. Rather, it takes a comparative view that the property of irony, implicit display, is achieved to the extent that the degree of irrationalness is high.

Especially, the implicit display theory essentially differs from and is better than Wilson and Sperber’s (1992) echoic interpretation theory, which has been the dominant view of irony. The echoic interpretation theory argues that verbal irony is a variety of echoic interpretations of someone’s thought, utterance, expectation or general desires/norms, in which the speaker dissociates herself from the echoed materials with ridicule or scorn. For example, Peter’s echoic reply (2) in Situation 3 is a typical example of echoic irony. In the same way, the ironic utterance (1a) of Situation 1 echoes the mother’s expectation with negative attitude. The most important difference between allusion and echoic interpretation lies in what materials are echoed/alluded to by ironic utterances. On our view, ironic utterances are always motivated by, and thus, allude to the speaker’s expectations, while the echoic interpretation theory argues that irony echoes not only the speaker’s expectations, but also other materials. However, in the following exchange between Peter and his other colleague James, who did not know what David said,

(5) a. James: What did David said to you?
   b. Peter (with ridiculing aversion): He’d be promoted before me.

Peter’s utterance (5b) echoes David’s preceding utterance and he simultaneously dissociates himself from the David’s opinion echoed in the same way as (2), but no irony results. Hence, the echoic interpretation theory cannot distinguish irony from nonironic echoes completely. On the other hand, the implicit display view provides a consistent explanation of Peter’s utterances (2) and (5b). In the case of Situation 3, what really makes the utterance (2) ironic is not David’s preceding utterance itself, but the speaker Peter’s expectation that the addressee David should know that his opinion expressed by the preceding utterance is false. The reason that (5b) is not ironic is that the addressee James does not (and cannot) assume any irony-motivating expectation of the speaker relevant to the current exchange, and thereby, the discourse situation of (5b) is not surrounded by ironic environment.

Furthermore, the implicit display theory provides a consistent explanation of other empirical findings and properties of irony.

Victims of irony — The implicit display theory does not need different explanations for different victims. Victims of irony are persons who performed intentional actions because of which the ironist’s expectation had not been realized. Victimless irony has no victims because the ironist’s expected state of affairs accidentally failed without someone’s intentional actions.

Ironic cues — The implicit display theory is consistent with the empirical finding (e.g., Gibbs and O’Brien 1991; Barbe 1995) that ironic cues (e.g., ironic tone of voice), by themselves, are neither sufficient nor necessary conditions for irony. An utterance without such cues (i.e., $d_e = 0$) is ironic when values of other factors ($d_a, d_g, d_i$) are high enough for its irrationalness value to meet the implicit display condition, and an utterance accompanied by these cues (i.e., a value of $d_e$ high) is nonironic when values of other factors are so low that it does not satisfy the implicit display condition.

What irony conveys — Irony conveys the information about ironic environment. In the case of typical irony, since the hearer already knows that the three conditions for ironic environment hold in the situation, irony interpretation results in confirmation of the most uncertain information, i.e., the speaker’s negative attitude. That is why previous irony theories argue that irony communicates the speaker’s negative attitude. When the hearer does not recognize ironic environment beforehand, he/she also obtains new information that the unrecognized components hold in the current situation.

Time-course of ironic versus literal comprehension —

The implicit display theory predicts that more prototypical ironies or ironies given in the situation more easily identified as ironic environment are processed faster. Hence, sometimes ironic interpretation is processed slower than literal interpretation but sometimes ironic interpretation does not take longer to process than literal one. This prediction is consistent with the seemingly incompatible time-course findings on irony (Utsumi 1999b): ironic sentences did not take longer to interpret than their literal equivalent sentences (Gibbs 1986), but often took longer to interpret than the same sentences used literally (Giora 1995).

**Computational Implementation of the Model**

I have implemented the cognitive model shown in Figure 1 in a computer program (Utsumi 1999a), although it is not a full-fledged implementation of the model.² The

²Furthermore, I have been trying to integrate the model of irony interpretation presented here with the relevance-based computational framework (Utsumi and Sugeno 1994; Utsumi 1995).
the following path by a breadth-first search.

Figure 2: A sample discourse context for Situation 1

Causal Relations:
16. (→ (¬ (B & T0) (and (in ?X ?A) (free ?X)))
   clean-up(?X ?A) (! (B & T1) (clean ?A))
17. (→ (¬ (B & T) (do read(?X ?A)))
   (! (B & T) (not (free ?X)))

Speech Act Schemes:
18. (→ (¬ (S & T0) (and (do ?P) (praiseworthy ?P ?S)))
   Thank(?S ?T) ?P
   (! (B & T1) (intend ?S Convince(?S ?T, ! (S & T0) (grateful ?S ?T ?P))))

Emotion-Eliciting Rules:
19. (→ (and (! (B & T0))
   (hope ?X (! (B & T1 & T2 & T3) (not ?T)))
   (! (B & T1 & T2 & T3) ! (B & T1 & T2 & T3))
   (vsl-cause ?A (! (B & T1) ?A)))
   (! (B & T1) (angry-at ?X Agent(?A ?A)))

Figure 3: An example of the shared knowledge

program uses the situation-theoretic representation scheme (Utsumi, 1996), and all events and states are expressed as formulas F=¬(¬ (B T) 1), support relations between situations (B T) and contents. A situation consists of a belief space B and time T of the event/state. An utterance U which will be interpreted is expressed by the propositional content P represented by formulas which are not ascribed to anyone or by predicates (actions), together with the surface illocutionary act F. The discourse situation in which U is given is expressed as the hearer’s cognitive environment C=W ∪ K where the discourse context W is the set of hearer’s beliefs about events/states represented by formulas shown in Figure 2, and the shared knowledge K consists of causal relations as domain knowledge, speech act schemes and emotion-eliciting rules, some of which are shown in Figure 3.

The program answers whether an inputted utterance U is ironic, and when U’ is ironic it returns the speaker’s ironic intention. As an example, let us consider how the algorithm interprets (1c) given in Situation 1, assuming that W is the context of Figure 2.

(1c). Thank you for reading a comic book.

The algorithm selects Formula 13 as the speaker’s expectation, and judges (1c) to achieve implicit display at Step 1: its propositional content read(y,b) is found to be coherently related to Q=¬(¬ (B & T) 1 & T0) (clean a) by the following path by a breadth-first search.

read(y,b) -- prevent -- clean-up(y,a)
-- volitional-cause --> (! (B & T) (clean a))
-- unifiable --> (! (B & T) (T & T)(do read(?X ?A)))

Also, the violation of the precondition (! (B & T) 0 (praiseworthy read(y,b) x)) for the illocutionary act Thank is recognized since (ask (! (B & T) 0 (not (praiseworthy read(y,b) x)) C) returns yes. Since the known expectation is judged to be alluded to by (1c), it is checked for incongruity with the situation and for negative attitude at Step 3. In this case, (ask (! (¬ (¬ (B & T) 0 (T & T)(do read(?X ?A)))) (! (B & T) 0 (not (T & T)(do read(?X ?A))))) returns no (i.e., her son’s room is not clean), and (ask (! (¬ (¬ (B & T) 0 (T & T)(do read(?X ?A)))) (! (B & T) 0 (not (T & T)(do read(?X ?A))))) returns yes (i.e., the speaker’s negative emotion is elicited using the emotion-eliciting rule for anger, i.e., Formula 19). As a result, the algorithm judges (1c) to be ironic, and produces the following intention:
Inform(x,y, and (! (¬ (B & T0) (hope x (! (¬ (¬ (B & T1 & T2 & T3) (not ?T))))
   (¬ (¬ (B & T1 & T2 & T3) ! (B & T1 & T2 & T3))
   (vsl-cause ?A (! (B & T1) ?A)))
   (! (¬ (B & T1) (angry-at ?X Agent(?A ?A))))

References