

## Using LSA to detect Irony

Aynat Rubinstein, Department of Linguistics, Tel Aviv University

### Abstract

In this work I propose a new model of verbal irony based on the notion of scales. The model, which stems from discourse theoretic accounts for irony, is then given computational concreteness based on Latent Semantic Analysis (LSA) [1]. Preliminary results are presented for automatically detecting irony in ironic headlines, a special type of irony which we argue is most fit for the LSA analysis.

### Irony on a Scale

From a discourse theoretic perspective, the model of scales suggests that understanding irony means perceiving the distance between two points on a scale [2].

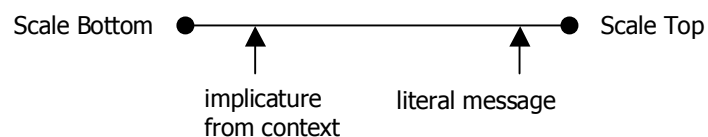


Figure 1: A scale for irony

A scale is the structure depicted in Figure 1 above. It is a line representing degrees, or cases, of the discourse topic alluded to in the utterance. The discourse topic defines the content of the scale and its edges, which represent extreme opposite cases of its realization. In understanding an ironic utterance, one point is conveyed by the literal meaning of the utterance, and the other is a relevant implicature extracted from context. The greater the difference between the two points, it is claimed, the better the resulting irony in terms of ease of perception and appropriateness.

For example, consider the following situation: your parents are away for the weekend, the house is totally at your disposal, it is Saturday afternoon and you have invited your boyfriend over. Just as the two of you are getting intimate on the sofa, your parents suddenly walk in. "What perfect timing!", you exclaim when you see them. Your boyfriend probably understands your ironic remark: the discourse topic being the nature of the timing of your parents' return, a scale is constructed that characterizes the timing in terms of degree of favorability. This is a scale ranging from good (very favorable) to bad (much unwanted). The literal message describes their timing as *perfect*, so one point is set on the scale close to the "Good" edge. In reality, as we remember, and as your boyfriend would readily admit, their

timing was quite *horrible*. A second point is then set at the “Bad” edge (see Figure 2). Once both points are set, the distance between them is computed. It is a significant distance, which licenses the ironic meaning.

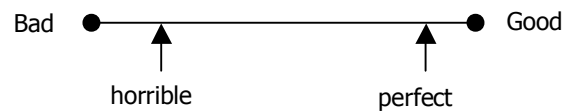


Figure 2: “What perfect timing!”

The present approach stems from the Indirect Negation theory of irony [3]. As such, it shares with classic pragmatic theory the classification of irony as a breach of a norm, but differs from it in its claim that the literal meaning is not discarded or rejected since it is crucial for computing the ironic meaning. Generalizing the Indirect Negation theory of irony, it is not required that the state of affairs designated by the ironic expression be an expected or desirable state of affairs. It makes no difference if the tone of the literal meaning is positive (as in ironic criticism) or negative (as in ironic praise), as long as it is located far enough from the implied meaning. Empirical experiments performed by Dukas [4] on visual irony in still and moving pictures corroborate our approach. Dukas has shown that the contrast between foreground (literal message) and background (implicated message) is more important in creating the irony than directionality, i.e. tone, of the two messages. Contrary to the predictions of the Indirect Negation theory of irony, he found that ironies in which the foreground was positive and the background was negative were not significantly easier to detect than ironies in which the messages were presented the other way around.

Scales provide additional insight in the account of ironic understatements and overstatements: an expression serving as an understatement in one context may function as an overstatement in an “opposite context”, one in which the contextual-point lays at the opposite edge of the scale.

The scale model accounts elegantly for the occurrence of the so-called “ironic cues” typical of ironic statements. Hyperboles, intensifying adverbs, and intonation all serve to widen the gap between the literal and contextual points on the scale. By driving the literal point closer to an edge the distance between the two points is increased, giving rise to better irony.

## **A computational model for irony using LSA**

From a computational point of view, the quantitative nature of the scale model suggests it can serve as a theoretical basis for a computational model of irony. The reasons for choosing LSA as the formal framework for our solution are a threefold: first, its latent contextual knowledge can be queried in order to extract relevant bits of information, namely the implicature. Furthermore, LSA provides a metric that can be utilized to calculate the distance between the implicature and the given literal utterance. Together, the contrast that lies in the heart of the irony can be computed. Lastly, the successful application of LSA as a model for metaphor [5] suggests it may play a role in a model for other types of non-literal language, and specifically for irony.

Latent Semantic Analysis, henceforth LSA, is a general theory of acquired similarity and knowledge representation. It is a bag-of-words model that ignores whatever linguistic structure is present in the text (morphological, syntactic, narrative, etc.) and is sensitive only to occurrences of words. The basic assumption of LSA is that words that have similar meanings tend to co-occur in texts. LSA's power lies in the fact that it is sensitive not only to direct co-occurrences, but can also infer indirect relations between words across texts. Similarity is measured in LSA as distance between vectors representing text items (or novel combinations of text items), defined as the cosine of the angle between them: the higher the cosine, the more similar the items.

In this work we attempt to utilize LSA for the task of automatic detection of irony. At first glance it seems to be the ideal model for irony: it has a metric for comparing sentences to sentences and words to words, it holds a representation of semantic relations between words, and it has shown "proof of concept" in many tasks that involve measures of similarity. However, LSA has its drawbacks. Two characteristics of LSA were taken into consideration before applying it to the task at hand. First, it is unable to distinguish synonyms from antonyms. Typical examples of irony that make use of antonyms ("Very funny", "What wonderful weather!") will go unnoticed. Second, it is ignorant of function words such as negation markers and intensifying adverbs, which are crucial clues in detecting irony. In light of these limitations, we decided to focus on a special type of irony, namely ironic headlines ("Afghanistan: a touristy leisure getaway"). These ironies can be expressed without negation markers and intensifying adverbs, and are typically based on the inappropriateness of concepts, not of antonyms.

Consider the following ironic headline:

Priorities<sup>1</sup>

“The most important thing in the world is eyebrow design”

Beauty queen and model Ilanit Levy (Yedioth Ahronoth)

Irony arises from the contrast between the meaning of the headline *Priorities* and the topic *eyebrow design*. Now suppose we replace *eyebrow design* by *buying a house* or *health*: the result is a literal and somewhat dull headline. We expect LSA to be sensitive to these differences.

The key idea in using LSA to detect irony is to look for dissimilarity and contrast, which in LSA means low similarity scores. Given a headline (*Priorities*) and a set of alternative topics (*eyebrow design*, *buying a house*, *health*), the model attempts to find the most ironic one by:

1. Computing the LSA similarity score between the headline and each of the alternative topics.
2. Ordering the alternatives according to their scores.
3. Outputting the pair headline-topic that received the lowest score as ironic.

We do not attempt in this work to cope with the more general detection task of judging for any arbitrary input utterance if it is ironic or not based on its content and the context, although it is definitely an interesting problem that should be addressed in the future.

In order to assess LSA’s applicability to the task of irony detection based on the model of scales, we performed a series of tests. The main question we set out to answer was whether the proposed computational model mimics humans’ behavior on tasks of irony detection.

## **Method**

### **Materials and Procedure**

Two irony detection tasks were presented to human subjects and to the computational model: two multiple choice tasks and a ranking task.

The multiple choice tasks consisted of 20 questions. Each question was presented as a set of alternative utterances, from which subjects were instructed to choose the most ironic one. In the main multiple choice task, 10 questions were presented with 4 alternatives each (Location items) - a total of 40 items. A second multiple choice task consisted of 10 questions with 2

---

<sup>1</sup> Appeared in the “Overheard” section of the Haaretz Magazine English edition, 28 February, 2003.

alternatives each (Government items) – a total of 20 items. An example of a Location question of this type with 4 alternatives (underline in the original):

- a. Iceland is really polluted.
- b. New York is really polluted.
- c. Goa is really polluted.
- d. Afghanistan is really polluted.

In the corresponding questions for the computational model, the alternatives were presented as pairs of the underlined elements in the questions presented to humans:

- a. (Iceland, polluted)
- b. (New York, polluted)
- c. (Goa, polluted)
- d. (Afghanistan, polluted)

The ranking task consisted of 7 questions, all based on real examples of ironic headlines from an Israeli newspaper<sup>2</sup>. The text under the headline included a blank, for which 3 alternative completions were given. Subjects were asked to rank the degree of irony for each alternative with respect to the headline on a scale of 1 (not ironic) to 10 (very ironic).

An example of a ranking question of this sort:

Priorities

“The most important thing in the world is \_\_\_\_\_”

- health
- eyebrow design
- buying a house

In the corresponding questions for the computational model, the alternatives were presented as pairs of the headline and each of the alternative completions:

- a. (priorities, health)
- b. (priorities, eyebrow design)
- c. (priorities, buying a house)

**Participants**

22 human subjects participated in the experiment: 52% female, 48% male, average age of 28.9 years. All were university students or graduates who volunteered to participate in the experiment.

---

<sup>2</sup> Examples were based on excerpts from sections of the Israeli Haaretz Magazine: “Kikar Ha-Medina” in the Hebrew edition (27 September, 2002; 28 February, 2003; 14 March 2003), and “Overheard” in the English edition (28 February, 2003). See Appendix A for the actual test items included in the analysis.

## Simulations

LSA simulations were performed using the online web-based LSA application One-To-Many Comparison<sup>3</sup>, on the General Reading up to 1<sup>st</sup> year college semantic space with 300 dimensions.

The questionnaire for human subjects was in Hebrew, and was translated to English for the evaluation of the computational model. It was verified that all words used in the questions existed in the corpus: if the word that appeared in the questionnaire for humans was not part of LSA's inventory, a near exact translation was used instead.

## Results

We now present the results of the computer simulations in comparison to humans' responses. Results for the multiple choice task are presented separately for the Location items (Table 1) and for the Government items (Table 2). For each alternative, LSA similarity scores are shown above the percentage of participants who chose it as most ironic. Shaded in dark gray are humans' and the model's first choices of the most ironic alternative for each question. Shaded in light gray are the model's second choices that match humans' first choices.

	New York	Goa	Iceland	Afghanistan
desert	0.07 59.09%	-0.06 4.55%	-0.06 36.36%	0.37 0%
tourism	0.13 0%	0.06 4.55%	0.19 0%	0.03 95.45%
highrise	0.13 4.76%	-0.03 33.33%	0.1 9.52%	-0.01 52.38%
polluted	0.06 0%	0.05 4.55%	-0.01 95.45%	-0.01 0%
island	0.24 18.18%	0.07 9.09%	0.57 0%	0.08 72.73%
desolate village	0.07 95.45%	0.15 4.55%	0.06 0%	0.08 0%
romantic atmosphere	0.08 0%	0.01 0%	0.03 0%	0 100%
over populated	0.32 0%	0.12 13.64%	0.16 77.27%	0.23 9.09%
bustling metropolis	0.26 0%	-0.03 31.82%	0.03 45.45%	0.03 22.73%
modern	0.18 0%	0.02 4.55%	0.05 4.55%	0.1 90.91%

Table 1: Results for Location items

<sup>3</sup> Available at <http://lsa.colorado.edu>.

As can be seen in Table 1 above, when comparing the model's first choice with humans' first choice (shaded dark gray), the model got only 3 items, 30%, correct (7% corrected for guessing by the formula  $[\text{correct-chance}/1\text{-chance}]^4$ ). However, on a more lax comparison taking into consideration the model's first and second choices (shaded light gray), the model got 9 items, 90%, correct (80% corrected for guessing).

	democracy	dictatorship
freedom of opinion	0.56 27.27%	0.31 72.73%
censorship	0.27 80.95%	0.26 19.05%
decentralization	0.43 4.55%	0.36 95.45%
secret police	0.13 100%	0.2 0%
human rights	0.3 4.55%	0.16 95.45%
centralization	0.41 95.45%	0.28 4.55%
political parties	0.66 9.52%	0.45 90.48%
free elections	0.51 9.09%	0.35 90.91%
rule of the people	0.43 13.64%	0.42 86.36%
violence	0.35 94.74%	0.33 5.26%

Table 2: Results for Government items

Comparing the model's first choice with humans' first choice (shaded dark gray) for Government items, the model got 7 items, 70%, correct (40% corrected for guessing).

Item analysis was performed to check the correlation between humans' judgment of irony and the model's judgment. Each pair of headline and topic received two scores: the number of participants that chose it as most ironic (out of 22), and a score of irony according to the model from 1 to 4, where 1 is least ironic and 4 is most ironic. Spearman correlation revealed a significant correlation between the two variables for the Location items: Spearman coefficient = 0.61,  $p=0.0001$ . In the Government items correlation was not found. In both

<sup>4</sup> Akin to the correction in Landauer & Dumais (1997) in evaluating LSA's success rate on the TOEFL synonymy test.

Location and Government items together a correlation of Spearman coefficient = 0.34, was found ( $p < 0.01$ ).

We turn now to the results of the ranking task. Recall that questions in this part were based on real examples of ironic headlines. Of the 7 questions in this part of the experiment, participants failed to detect the irony in one question, and it was not included in the comparison. An average of participants' rankings was calculated for each of the remaining six headline-completion pairs (10 - very ironic, 1 - not ironic). These averages are shown in Table 3, along with the LSA similarity scores for each pair.

priorities	eyebrow design	buying a house	health
	9.59	4.38	1.19
	0.13 ( <b>0.02</b> )	<b>0.08</b>	<b>0.15</b>
matchmaker	rapist	cashier	teacher
	9.36	5.19	3.19
	-0.05	0.01	0.06
Judaism	death	height	success
	8.68	4.38	1.19
	0.17 ( <b>0.04</b> )	-0.01 ( <b>0.06</b> )	0.01 ( <b>0.31</b> )
Shakespeare	soap opera	story	drama
	7.73	3.62	3.19
	0.28	0.13	0.84
nirvana	full volume	loud	quiet
	6.68	5.19	1.81
	-0.01	0	0.09
profession	son	driver	consultant
	8.91	2.95	1.81
	0.13 ( <b>0.08</b> )	0.03 ( <b>0.08</b> )	0.34 ( <b>0.10</b> )

Table 3: Results for ranking task

The model did not succeed in mimicking humans' rankings: only in 2 out of the 6 questions (*matchmaker*, *nirvana*: marked with gradual shading) did the model mimic the scale given by humans for the alternative topics. In three cases (*priorities*, *profession*, *Shakespeare*) it did not succeed in detecting the most ironic alternative (*eyebrow design*, *son*, *soap opera* respectively). In the remaining question (*Judaism*) the model ranked the alternatives totally opposite to the participants. However, carefully varying the items presented to LSA had a drastic effect on the results, as indicated by the figures in boldface. These effects are described and discussed in the next section.

## Discussion and conclusions

The results presented above provide supporting evidence for the model of scales, showing that humans' judgments of irony correlate with distances between concepts. However, they do not univocally support the viability of the proposed model of irony based on LSA. On the one hand, a significant substantial correlation was found between judgments of the model and humans in the main multiple choice task. On the other hand, in order to achieve a success rate of 80% on this task, we had to take into consideration both the model's first and second answers. Results in the second multiple choice task and in the ranking task were less encouraging.

However, we believe the model should be tested more thoroughly before a conclusion regarding its viability is reached. Firstly, varying the corpus on which LSA is trained may have a considerable effect on the results. For example, the real-life examples in the ranking task were taken from a contemporary Israeli newspaper. In order to fully appreciate these ironic headlines one must be knowledgeable about current political and social issues in today's Israel. The human participants were clearly knowledgeable in this respect, but the corpus LSA was trained on was not.

Secondly, we noticed that subtle changes in the input to LSA have drastic effects on the results (see figures in boldface in Table 3). Thus, using *job* instead of *profession* and *eyebrow* instead of *eyebrow design* in the ranking task resulted in different, and correct, rankings by the model ( $\text{cosine}(\textit{job}, \textit{son})=0.08$ ,  $\text{cosine}(\textit{job}, \textit{driver})=0.08$ ,  $\text{cosine}(\textit{job}, \textit{consultant})=0.10$ ,  $\text{cosine}(\textit{priorities}, \textit{eyebrow})=0.02$ ). Using *Atonement* instead of *Judaism* also brought out a correct scale:  $\text{cosine}(\textit{Atonement}, \textit{death})=0.04$ ,  $\text{cosine}(\textit{Atonement}, \textit{height})=0.06$ ,  $\text{cosine}(\textit{Atonement}, \textit{success})=0.31$ . This variability in similarity scores demonstrates that LSA's judgments do not always match our intuitions: while we would judge *job* and *profession* as near synonyms, they behave differently in the semantic space; while we know irony results from *eyebrow* and not from *design*, LSA should be told this explicitly.

In conclusion, based on our findings we believe LSA can serve as a basis for a working model of irony. However, its limitations should be understood, and it should be augmented by mechanisms that are sensitive to negation markers, to intensifiers, and to the distinction between synonyms and antonyms. Further research should also explore the effects of changing the training corpus and methods for detecting irony without relying on a set of pre-defined alternatives.

## References

- [1] Landauer, Thomas K. and Susan T. Dumais (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review* 104(2): 211-240.
- [2] Rubinstein, Aynat (2002). "Irony on a scale - A discourse theoretic account of irony", a talk given at GIM2002, German Israeli Minerva Summer School on Computational Linguistics.
- [3] Giora, Rachel (1995). On Irony and Negation. *Discourse Processes* 19: 239-264.
- [4] Dukas, Gideon (1997). On Aptness of Visual Irony: Testing Irony in Photography and Cinema. MA Thesis, Tel Aviv University.
- [5] Kintsch, Walter (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin & Review* 7: 257-266.

## Appendix A

Following are the test items used in the ranking task. Originally in Hebrew, they are all based on excerpts from the Israeli Haaretz Magazine (see footnote 2 for details):

### Priorities

"The most important thing in the world is \_\_\_\_\_"

- health
- eyebrow design
- buying a house

### Matchmaker

"I would have introduced her to John, who is a renown \_\_\_\_\_"

- rapist
- cashier
- teacher

### Judaism

"I wished that their kids \_\_\_\_\_, and I also prayed for it in the synagogue on The Day of Atonement"

- would die
- would be tall
- would succeed in life

### Shakespeare

"It's just a \_\_\_\_\_. After all, what is Romeo and Juliet?"

- drama
- story
- soap opera

### nirvana

"While he cooks he turns on the TV \_\_\_\_\_. He claims it calms him"

- loud
- quiet
- full volume

### Profession

"The prime minister's \_\_\_\_\_"

- consultant
- son
- driver