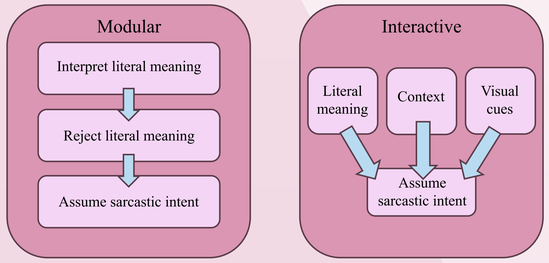
An Exploratory Eye-Tracking Study into the Time Course of Sincere and Sarcastic Speech Recognition

Isabella Noor Warner, Sueah Kim, Thomas Zhao, Ryuki Matsuura, Seth Wiener

Language Acquisition, Processing, and Pedagogy Lab, Carnegie Mellon University

Background

- Speakers use changes in prosody to denote sarcasm.
- In English, reductions in F0 and harmonics-to-noise ratio (HNR) are the most prominent sarcasm signalers.
- There are two schools of thought on sarcastic processing:
  - Modular theory: follows Gricean principles and suggests that processing time for sarcasm will always take longer.
  - Interactive theory: suggests that visual cues and context can be integrated into the interpretation for faster processing.



Research Question

- Do prosodic cues and added visual stimulation provide the necessary supporting context for participants to analyze the communicative intent of the utterance in an interactive manner, or is processing still modular?

Methods

46 participants

- 100 participants initially recruited via Profilic
- 54 removed for failing dichotic pitch task, eye-tracking calibration or having a framerate < 5 Hz
- Mean participant age = 37.8 years old
- English was a first, primary, and fluent language for all participants.

Stimuli Creation

- Audio stimuli recorded in the form of statement-response pairs.
- Responses contained 21 or 22 syllables and a 440ms 3-syllable carrier phrase.
- Each pair was recorded once with a sarcastic answer and once with a sincere answer.
- 20 pairs resulted in 40 unique audio recordings.
- The keyphrase contained one to three words in which the prosody was manipulated.
- Remaining audio retained similar prosody regardless of the sincere or sarcastic intended meaning.
- Visual stimuli were generated using ChatGPT 5.0 and Google AI Studio Gemini 2.5 Flash Image.
- F0 and HNR extracted automatically using a deep neural network-based Forced Alignment tool implemented with wav2vec 2.0.

Eye-tracking Task

- The task consisted of 20 trials with no fillers. Half of these trials were sincere and half were sarcastic counterbalanced with two lists.
- Eye movements were continuously recorded using webgazer in Gorilla.
- Presentation: Fixation cross, statement audio, 1000ms preview, response audio.
- Participants were asked to click on the image that represented their perceived interpretation.

Statement: “That party was crazy”

Response: “Last night was so wild that my parents would probably approve of it if they were there.”

Sincere Interpretation

Sarcastic Interpretation

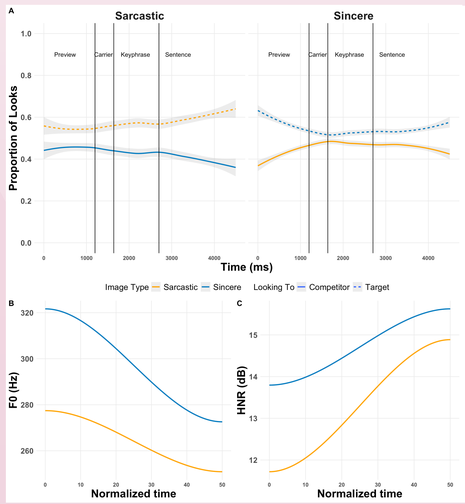
Statement: “I look sharp in this jacket”

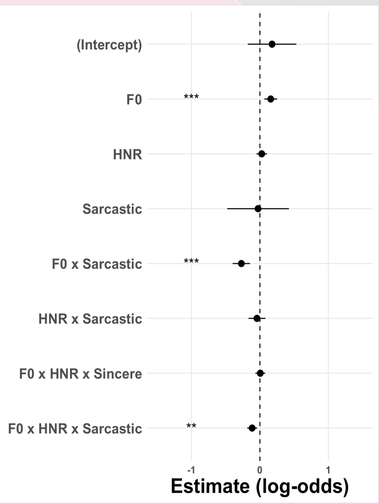
Response: “I think that no one is going to be able to look away from it when we go out.”

Sincere Interpretation

Sarcastic Interpretation

Results





Fixation Analysis
1000 ms ~ 11 consecutive 50 ms time windows

Generalized linear mixed model

glmer(looks to target ~ scaled F0 + scaled HNR + prosody + sF0:prosody + sHNR:prosody + sF0:sHNR:prosody + (1 | Participant) + (1 | Item), family = "binomial")

Model fit: Conditional $R^2 = 0.20$

- F0: β = .16
- F0 x Sarcastic: β = –.27
- F0 x HNR x Sarcastic: β = –.11

Conclusions

- An increase in F0 led to more looks to sincere targets.
- An increase in F0 alone and an increase in F0 and HNR led to fewer looks to sarcastic targets.
- The addition of congruent (or incongruent) visual stimuli potentially aided interpretation of prosody.
- Broader conclusions in regards to the interactive or modular theories of sarcastic processing cannot be drawn at this time.

Limitations

- Audio stimuli design inherently reduced the naturalness of speech in favor of uniformity; all prosodic variations in keyphrase, rather than entirety of utterance.
- Eye-tracking data revealed an initial bias towards sincere targets for an unknown reason.
- “Sarcastic” and “sincere” are not absolute terms.

Overview

- We paired audio stimuli with visual representations of utterance context to determine the time course of sarcastic processing.
- We believe this is the first use of web-based eye-tracking in this manner and context.

Future Directions

- Broadening the number of productions to include humorous, positive, neutral, or other contexts.
- Exploring these interactions across languages.

A Note on AI Usage

We used generative AI tools to generate visual stimuli for this experiment. This decision was informed by a growing consensus that AI-driven tools can benefit linguistics research, and the acknowledgement that human creation of the necessary stimuli would also be difficult. However, we found that using AI to create these stimuli was somewhat problematic. Iterative prompt engineering would be necessary to make the best use of these new tools.

References

- H. S. Cheang and M. D. Pell, “The sound of sarcasm,” *Speech Comm.*, vol. 50, no. 5, pp. 366–381, 2008.
- H. S. Cheang and M. D. Pell, “Acoustic markers of sarcasm in Cantonese and English,” *J. Acoust. Soc. Amer.*, vol. 126, no. 3, pp. 1394–1405, 2009.
- A. Ţurcan and R. Filik, “An eye-tracking investigation of written sarcasm comprehension: The roles of familiarity and context,” *J. Exp. Psychol.: Learn. Mem. Cogn.*, vol. 42, no. 12, pp. 1867–1893, 2016.
- H. P. Grice, “Logic and conversation,” in *Syntax and Semantics, vol. 3: Speech Acts*, J. Morgan and P. Cole, Eds. New York, NY: Academic Press, 1975, pp. 41–58.
- S. Suzuki, H. Takatsu, R. Matsuura, M. Koyama, M. Saeki, and Y. Matsuyama, “Feedforwarding diagnostic language assessment: Artificial intelligence-(AI-) driven weakness identification and contextualised feedback for second language speaking,” *Lang. Testing*, vol *42*, no.4, pp. 476–507, 2025.
- A. Bramlett and S. Wiener, “The Art of Wrangling: Working with Web-based Visual World Paradigm Eye-tracking Data in Language Research,” *Ling. Approaches Biling,* vol. 15, no. 4, pp. 538-570, 2025.

Acknowledgements

We thank Jay Dasilva for his help creating the stimuli and the anonymous reviewers for providing critical feedback.

Contact us

Full Paper