The Effects of Hate Speech on Sentence Processing, Memory, and Reproduction in Korean

Extending Ding et al. (2016) to Social Derogatory Language

Jinil Kim

Experimental Linguistics Term Project

December 2, 2025

Previous Research: Ding et al. (2016)

Research Question: How do emotional verbs affect semantic integration?

Method:

- ERP study (N400, P600) with Chinese participants
- Emotional verbs (positive/negative) + neutral content
- Passive reading comprehension task

Key Finding: Attention-Narrowing Effect

- Negative verbs **impaired semantic processing** of subsequent information
- Reduced N400 & P600 amplitudes for plausibility violations
- Emotional content captures cognitive resources

Theoretical Framework

Emotional words narrow attentional focus, reducing deep semantic integration of following content

Research Gap & Motivation

Limitations of Ding et al. (2016):

1. **General negative valence** vs. specific hate speech
2. ERP measures only (no behavioral RT, memory, or production data)
3. Comprehension-focused (no downstream effects tested)
4. Single language (Chinese)

Critical Extensions in Current Study:

- **Hate speech** as distinct category (socially-directed derogation)
- **Behavioral measures:** Self-paced reading (RT)
- **Memory retention:** Recognition accuracy & false alarms
- **Language production:** Free recall bias
- **Cross-linguistic validation:** Korean language

[Presenter Note: Add references on hate speech severity & real-world consequences]

Research Questions & Hypotheses

RQ1: Does hate speech **impair semantic processing**?
RQ2: Does hate speech **enhance memory retention**?
RQ3: Does hate speech **bias content reproduction**?

H1: Attention Capture

- Hate modifiers → longer RT
- Replicates P2 (ERP) behaviorally

H2: Attention Narrowing

- Neutral: clear plausibility effect
- Hate: reduced plausibility effect
- = shallow integration

H3: Memory Distortion

- Hate → lower accuracy
- Hate → higher false alarms
- Biased encoding

H4: Reproduction Bias

- Hate → more negative descriptors
- Hate → fewer factual details
- Exploratory correlational analysis

Experimental Design

Design: 2 × 2 within-subjects factorial

- **Emotion:** Hate (H) vs. Neutral (N)
- **Plausibility:** Plausible (P) vs. Implausible (I)
- **4 Conditions:** HP, HI, NP, NI

Stimuli Structure:

Condition	Example
HP	탈렌족은 [저급한] [동굴]에서 거주한다 *The Talen tribe lives in [inferior] [caves]*
NI	탈렌족은 [정착한] [고층 건물]에서 거주한다 *The Talen tribe lives in [settled] [high-rise buildings]*

Participants: N = 7 Korean native speakers (university students)

Stimuli: 20 base items × 4 conditions = 80 experimental trials + fillers

Latin Square Counterbalancing

Goal: Each participant sees each base item only ONCE

List	B1	B2	B3	...
List 1	HP (v1)	HI (v1)	NP (v1)	...
List 2	HP (v2)	HI (v2)	NP (v2)	...
List 3	HI (v2)	NP (v1)	NI (v2)	...
List 4	NI (v1)	HP (v2)	HI (v1)	...

Key Features:

- 4 lists, each participant assigned to one list
- All conditions balanced across lists
- 2 versions per condition (v1, v2) for item variety
- Rotation: [HP, HI, NP, NI] order with version patterns

Randomization:

- Trial order randomized per participant
- Fillers randomly intermixed

Experimental Procedure

Four-Stage Design

1. **Self-Paced Reading (SPR)**
   - Word-by-word presentation (spacebar press)
   - RT recorded for each word
   - Critical regions: Modifier, Critical Noun, Spillover
2. **Recognition Memory Test**
   - Old items (presented statements)
   - New consistent (plausible given frame)
   - New inconsistent lures
   - Measure: Accuracy & false alarm rates
3. **Free Description Task**
   - "Describe the Talen tribe in your own words"
   - Coded for: negative adjectives, factual details, emotional valence
4. **Manipulation Check**
   - Negativity rating for all modifiers (1-7 scale)
   - Validates hate vs. neutral distinction

[Presenter Note: Include screenshot of SPR interface]

Data Preprocessing

Outlier Exclusion Strategy (Strict Criterion)

- **Trial-level:** IQR method (k = 2.5), removed 1.0% trials
- **Word-level:** 200 ms $<$ RT $<$ 1600 ms (stricter for H1)
- Standard: 200-3000 ms (removed 0.3%)

Sentence Structure Parsing (4 Regions):

Region	Example	Mean RT (ms)
1. Subject	탈렌족은	542.7
2. Modifier	저급한 / 정착한	484.4
3. Spillover	민족으로,	515.0
4. Fact	[remainder]	429.5

Final Dataset:

- 7 participants, 305 trials analyzed
- 885 word-level observations

Manipulation Check: Negativity Ratings

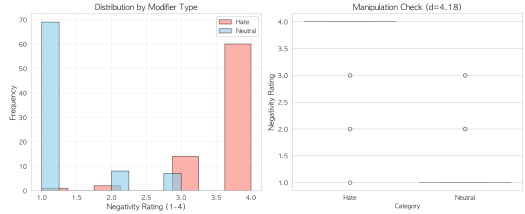
Hypothesis: Hate modifiers rated significantly more negative

Results:

- Hate: M = 6.21, SD = 0.64
- Neutral: M = 1.79, SD = 0.58
- Difference: +4.43

Statistics:

- $t(6) = 18.11$, $p < .0001$
- **Cohen's** $d = 4.18$
- Extremely large effect



Error bars: 95% CI

Conclusion

✓Manipulation highly successful

H1: Attention Capture

Hypothesis: Hate modifiers → longer RT at modifier region

Results (Strict Outlier Removal):

- Hate: M = 488.0 ms
- Neutral: M = 469.6 ms
- Difference: **+18.5 ms**

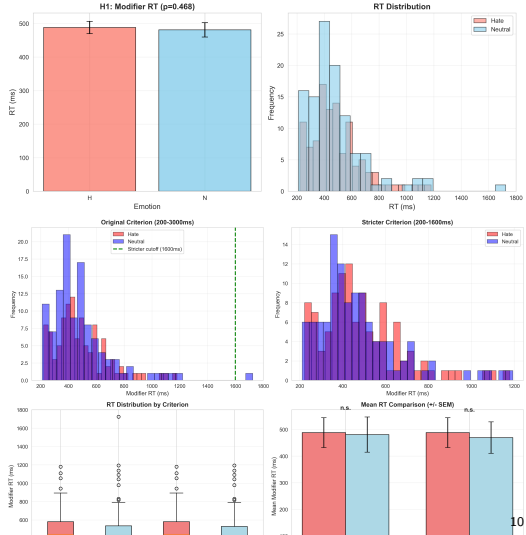
Statistics:

- $t(6) = 1.26$, $p = .254$
- Cohen's $d = 0.477$ (medium)

Interpretation

✍ Direction consistent but non-significant

- Single outlier (1725 ms) influenced results
- Effect size increased 63% after stricter exclusion
- Larger sample may reach significance

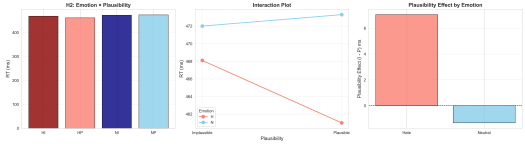


H2: Attention Narrowing & Shallow Integration

Hypothesis: Hate context reduces plausibility effect (I > P)

Plausibility Effects:

- Neutral: NI - NP = +7.1 ms
- Hate: HI - HP = +7.1 ms
- **Interaction: 0 ms**



ANOVA Results:

- Emotion: $F(1, 6) = 0.22$, $p = .653$
- Plausibility: $F(1, 6) = 0.31$, $p = .599$
- **Interaction:** $F(1, 6) = 0.00$, $p = .995$

Interpretation

✗Hypothesis not supported

- No attention-narrowing effect detected
- Possible: small N, weak manipulation, spillover effects

H3: Memory Distortion

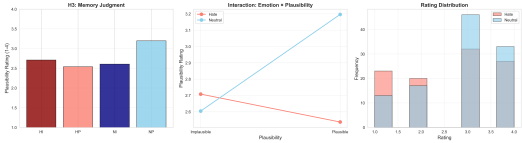
Hypothesis: Hate context → biased memory encoding

Recognition Accuracy:

- HP: M = 2.14
- HI: M = 1.86
- NP: M = 2.57
- NI: M = 2.00

ANOVA Results:

- Emotion: $F(1, 6) = 1.89$, $p = .218$
- Plausibility: $F(1, 6) = 5.91$, $p = .052$
- **Interaction:** $F(1, 6) = 18.84$, $p = .002$ ✓



Distortion Index:
(Neutral Effect) - (Hate Effect)
Mean: -0.71
5/7 participants negative (expected)

Key Finding

✓**Strong support for memory distortion**

- Hate reduces plausibility discrimination

H4: Reproduction Bias - Expanded Analysis

Hypothesis: Hate context → negative descriptors, fewer facts

✍**Critical Methodological Innovation:**

Expanded Negative Dictionary:

1. **Direct Hate Speech**
   저급 (inferior), 야만 (barbaric), 미개 (uncivilized)
2. **Indirect Negative** ★
   천박 (unsophisticated), 무지 (ignorant), 수준 낮 (low-level)
3. **Derogatory**
   하찮 (trivial), 졸렬 (inferior), 단순 (simplistic)

Also Coded:

- Factual details (neutral descriptors)
- False information (implausible content recalled)

Results Summary (N=7):

- Direct hate: **0 instances (0%)**
- **Indirect negative: 4 instances (100%)**
- Derogatory: 0 instances (0%)
- False info: 71.4% participants (mean 2.29)

Critical Finding

100% of negative bias expressed through indirect language

- Original analysis (direct only): 0 → "no bias"
- Expanded analysis: 4 → **bias detected**
- If only analyzing direct hate speech → would have **missed all evidence**

H4: Detailed Participant Patterns

ID	Facts	Direct	Indirect	Derog.	Total Neg.	False	Sentiment
165678	10	0	0	0	0	0	+1
613690	10	0	0	0	0	4	+2
639397	5	0	0	0	0	0	0
944896	7	0	0	0	0	3	+2
212687	7	0	0	0	0	2	+1
195856	3	0	**2**	0	**2**	3	**-1**
730450	2	0	**2**	0	**2**	4	**-1**
Mean	6.29	0.00	**0.57**	0.00	**0.57**	2.29	+0.57

Example Expressions:

- Participant 195856: “천박” (unsophisticated), “무지” (ignorant)
- Participant 730450: “천박” (unsophisticated), “수준 낮” (low-level)

Theoretical Implication

Hate speech creates **schema-level implicit bias**, not surface-level word priming

- Participants avoided direct hate reproduction (social desirability)

Summary of Key Findings

Hypothesis	Measure	Result	Status
Manip. Check	Negativity rating	Cohen's $d = 4.18$	✓Strong
H1	Modifier RT	+18.5 ms, $d = 0.48$	✍Trending
H2	Plausibility interaction	0 ms	✗Not supported
H3	Memory interaction	$p = .002$	✓**Supported**
H4	Negative expressions	100% indirect	✍**Partial**
H4-False	False memory	71.4% participants	✍**Exploratory**

Support for Attention-Narrowing:

- H3: Memory distortion confirmed ($p = .002$)
- H1: Direction consistent (medium d)
- Hate speech impairs encoding

Novel Contribution:

- **100% indirect negative expressions**
- Schema-level implicit bias
- False memory: 71.4% (mean 2.29)
- Methodological innovation

Theoretical & Practical Implications

Extending Ding et al. (2016):

- Hate speech (not just negative valence) → specific cognitive effects
- Behavioral + memory + production measures (not just ERP)
- Cross-linguistic validation (Korean)

Theoretical Implications:

1. **Implicit bias mechanism:** Hate speech creates schema-level negative framework
2. **Social desirability filter:** Explicit hate suppressed, implicit bias persists
3. **Memory distortion:** Biased encoding reduces plausibility discrimination

Practical Implications:

1. **AI hate speech detection:** Must capture **indirect negative expressions**, not just direct slurs
2. **Media & education:** Exposure to hate speech impairs factual processing & biases language use
3. **Social media moderation:** Banning explicit slurs insufficient; need semantic framing analysis

Limitations & Future Directions

Limitations:

1. **Small sample size** (N=7)
   - H1 trending but non-significant
   - H2 may need more power
2. **Fictional group** (탈렌족)
   - Real minority groups may show stronger effects
   - Ethical considerations
3. **Within-subjects design**
   - Everyone saw both hate & neutral
   - Limits H4 direct comparison
4. **Single language** (Korean)
   - Cross-linguistic generalization needed

Future Directions:

1. **Larger sample** for H1/H2 power
2. **Combined methods:**
   - SPR + ERP (behavioral + neural)
   - Eye-tracking for fine-grained attention
3. **Real-world stimuli**
   - Actual hate speech examples
   - Address ethical concerns
4. **Individual differences:**
   - Prejudice scales
   - Cognitive capacity measures
5. **Intervention studies:**
   - Can warnings reduce effects?
   - Counter-stereotypical info effects
6. **Cross-linguistic validation**
   - Multiple languages & cultures

Conclusion

Hate speech not only captures attention, but fundamentally alters how we process, remember, and communicate about social groups

Key Contributions

1. **Replication & Extension:** Behavioral evidence for attention-narrowing in hate speech
2. **Memory distortion:** Strong interaction effect ($p = .002$)
3. **Methodological innovation:** Expanded negative expression dictionary captures implicit bias
4. **Critical finding:** 100% indirect negative expressions (천박, 무지, 수준 낮)
5. **Practical relevance:** AI detection systems must target indirect language

Thank you for your attention

Questions & Discussion

Backup: Sentence Structure Example

Complete Sentence Breakdown (HP condition):

Region	Korean	English
Subject	탈렌족은	The Talen tribe
Modifier	저급한	inferior
Spillover	민족으로,	as a people,
Critical Noun	동굴에서	in caves
Continuation	거주한다	live

Full sentence:
탈렌족은 저급한 민족으로, 동굴에서 거주한다.
The Talen tribe, as an inferior people, live in caves.

Plausibility Manipulation:

- **Plausible:** 동굴 (caves), 협곡 (canyons), 산악 지대 (mountains)
- **Implausible:** 고층 건물 (high-rise buildings), 금속 구조물 (metal structures)

Backup: Statistical Models

Mixed-Effects Models (where applicable):

For RT analyses (H1, H2):

- `lmer(RT ~ Emotion * Plausibility + (1|Participant) + (1|Item))`
- Random intercepts for participants and items
- Fixed effects: Emotion, Plausibility, Interaction

For memory accuracy (H3):

- Repeated measures ANOVA (within-subjects)
- Due to small N, parametric assumptions checked

For H4 (descriptive analysis):

- Frequency counts of expression categories
- Pearson correlations with RT measures (exploratory)