ExpLing TermProject

진행 상황	0%
상태	시작 전

Current Research Plan Summary

Topic

Does hate speech affect sentence processing, memory retention, and reproduction?

Main Research Questions

1. Does hate speech **impair semantic processing**?
2. Does hate speech **enhance memory retention** due to emotional salience?
3. Does hate speech **increase the likelihood of negative content reproduction** in participants' own language?

Research Background

- Ding et al. (2016)
    - **Negative emotional verb → attention narrowing → shallow semantic processing**
        - Negative emotion verb: large N400 → requires more attention (attention narrowing)
        - Incongruent noun + negative verb: less N400, P600 → lack of attention
- Kissler et al. (2006)
    - https://pubmed.ncbi.nlm.nih.gov/17015079/
    - **Emotional word → elevate cortical response → stronger processing**

- Kensinger et al. (2006)
    - http://sciencedirect.com/science/article/abs/pii/S0749596X06000635
    - **Emotional word → enhanced main, impaired surrounding memory → narrowing**
        - Main: occurs emotion, Surrounding: others
        - Condition: Negative, neutral, positive images
        - Results
            1. Recognition test: increased hit rate for main, impaired for peripheral
            2. Recall test: more memory for main, impaired for peripheral
            3. False Memory test: increased false alarm for peripheral, not for main
            4. RT: Increased in peripheral
- Schindler et al. (2023)
    - https://www.nature.com/articles/s41598-023-43794-4
    - **Effect by emotion → when task & attention directs to emotion**
        - EPN, LPP: increased significantly in emotional task
- Conclusion
    - Narrowing effect of emotional word is widely accepted
        1. But, explicit focus on 'hate speech' is not suggested
        2. It's impact on processing and memorizing relevant element is not investigated
        3. Relation with reproduction is not suggested
        4. Neurological evidence of such effect is not suggested

Predictions

- **Simple cognitive model**: hate speech → online processing, memory, and reproduction

0. Reading hate speech elicits negative affect and perceived group-based threat
1. Hate speech will **narrow attention toward** hate element.
2. Hate speech will **impair semantic integration and reanalysis** of subsequent information
    a. As attentional resources are focused on hate-consistent content
3. Hate speech's attentional narrowing will **yield a biased encoding pattern**.
    a. Hate-consistent "central" elements are over-encoded
    b. Neutral or contextual "peripheral" elements are under-encoded
4. Eventually, hate speech will **increase the likelihood of reproducing hate-consistent** implications in later descriptions or judgments.
    a. As biased encoding will lead to a distorted mental representation of the group.

Hypotheses

H1 (Attention capture)

- Hate-modifier sentences will show **longer reading times at the hate modifier** than at neutral modifiers.
    - This reflects affect-driven attentional capture.

H2 (Attention narrowing & shallow integration)

- At the critical noun and spillover regions, neutral-modifier sentences will show a clear plausibility effect (implausible > plausible RT)
- Whereas **in the hate condition** this **plausibility effect will be reduced**,
    - This indicates shallower integration of subsequent content under attentional narrowing.

H3 (Biased memory / "trade-off + distortion")

- Relative to neutral context, hate context will lead to

    (a) **Lower accuracy** for neutral/factual statements

    (b) **Higher false alarm rates** for hate-consistent lures

    - This reflects a biased encoding of central hate-consistent elements and reduced encoding of peripheral factual details.

H4 (Encoding bias in reproduction)

- Free descriptions after hate context will contain

  (a) a **higher proportion of hate-consistent propositions** and negative adjectives

  (b) **fewer neutral background details**, compared to descriptions after neutral context

    - This indicating a **biased and partially distorted reproduction** of the target group.

Experimental Design (2×2 factorial design)

Factor 1	Factor 2	Example	Description
Modifier (Hate vs. Neutral)	Plausibility (Plausible or not)	a. hate + plausible b. neutral + plausible c. hate + implausible d. neutral + implausible	Crosses emotional tone and plausibility

Participant tasks:

1️⃣ Word-by-word presentation for fine-grained RT (self-paced reading / paced RSVP)

- **Design**
    - Use **self-paced reading (SPR)** with trial structure
        - context
        - modifier (hate/neutral) → VP/Noun (plausible vs implausible or true vs false) → spillover region(s).
    - Record RTs at **critical word(s)** and **spillover**; predefine which word carries each manipulation.
    - Include **filler trials** to prevent strategy.
- **Timing & interface**
    - SPR: space-bar to reveal each word; mask previous word; set a maximum trial duration; instruct natural reading.

  - RSVP: 250–350 ms/word with 100–200 ms ISI for critical words; adjust in piloting.
- **Controls**
  - Match frequency, length, bigram probability
  - **Manipulation checks**: independent **plausibility** ratings and **valence/arousal** ratings collected in a separate norming sample.
- **Analysis**
  - **Linear mixed-effects**
    - by-subject and by-item random intercepts
    - random slopes for factors when possible
  - Primary tests: Hate × Plausibility at **critical region** and **spillover**.

2️⃣ **Judgment task (semantic/plausibility or truth)**

- **What to ask**
  - **Plausibility/congruency**: Multiple choice questions
- **Why it's useful**
  - Provides a **trial-wise explicit measure** that complements implicit RT effects and creates a bridge to **P600 (controlled reanalysis)** predictions.
- **Design cautions**
  - Place judgments **immediately after** each sentence to anchor trial-level linkage, but avoid long on-screen text that re-exposes hateful content.
  - Include **catch trials** with obvious answers to maintain engagement.
- **Analysis**
  - **Logistic mixed-effects** for accuracy
  - **linear mixed-effects** for judgment RT
  - Divergences can signal **conflict/reanalysis** consistent with P600 theories
    - e.g., long RT but correct judgment

3️⃣ **Free description / evaluation task (reproduction & attitude)**

- **Prompting**
    - Neutral prompts (e.g., “Briefly describe the ZZ group in your own words.”; “Write 2–3 sentences.”).
    - Counterbalance **task order**: consider running free description **before** explicit judgment on a **separate block** or **separate day** to reduce demand characteristics; if same session, run **SPR → free description → judgment** or **counterbalance 2↔3** across participants.
- **Coding plan**
    - Pre-register a **codebook**: negativity, hate-lexicon presence, moral/emotional tone, certainty/hedging, reproduction of specific claims.
    - Use A**utomatic NLP** (sentiment, toxicity lexicons) for reliability
        - report **Cohen’s κ** / **ICC**.
- **Outcome links**
    - Test whether hate condition increases **negative descriptors** and **verbatim reproduction** of critical claims; examine correlations with RT and judgment outcomes (mediation: online cost → explicit judgment → reproduction).

Metrics

1. **Primary metrics**

Self-paced reading time at the modifier, critical noun, and spillover regions

- Behavioral index of **semantic processing difficulty** (N400-like).

Judgment accuracy (plausible vs implausible / true vs false)

- Outcome of successful vs failed **reanalysis** (P600-related).

Sentiment and hate-consistency of free descriptions

- Downstream attitudinal and reproduction bias.

2. **Secondary metrics**

Judgment RT

- Additional index of reprocessing cost

    - Interpreted cautiously because it reflects multiple processes
    - memory, reanalysis, uncertainty

Any "accuracy" coding in free writing

- Treated as exploratory, as free writing primarily targets attitudes rather than memory.

Experiment Example

0) Background passage (shown once at block start)

자알(ZZ)족은 높은 산간 계곡에 거주하며, 의식 때 **양털로 만든 겹옷**을 입고, **오리구이**를 함께 나누는 전통이 있습니다. 의식 전에는 **허브 차**를 마시고, **노래 의식**을 치르며, 주식으로 **발효 곡물 음식**을 먹고, **흙집**을 짓습니다.

- suggested with an image created by AI

> Participants are told all later sentences refer to this fictional group.

---

1) Step-1 — Word-by-Word Reading (Self-Paced Reading / RSVP)

Factorial 2×2 (Modifier × Plausibility), critical region in **bold**:

- **Hate + Plausible**

  "ZZ족은 **열등한** 민족으로 **의식을 위해 전통 의상을 입는다.**"

    - Tokens: `["ZZ족은","열등한","민족으로","의식을","위해","전통","의상을","입는다"]`

- **Neutral + Plausible**

  "ZZ족은 **고립된** 민족으로 **의식을 위해 전통 의상을 입는다.**"

- **Hate + Implausible**

  "ZZ족은 **열등한** 민족으로 **모임 전 돌을 먹는다.**"

- **Neutral + Implausible**

  "ZZ족은 **고립된** 민족으로 **모임 전 돌을 먹는다.**"

- **Filler (Neutral + Plausible + True/False)**

  ZZ족은 산길 을 따라 천천히 이동 한다 . → True

  ZZ족은 강 바닥 을 매일 청소 한다 . → False

Presentation & logging

- SPR: one word at a time; previous word masked; spacebar to advance.
- RSVP alternative: 275 ms/word, 125 ms ISI; critical word extended to 325 ms.
- Log **RT per word**, with special flags for:
  - `region = modifier` (hate/neutral adjective/phrase)
  - `region = verb` (integration onset)
  - `region = critical_noun` (plausibility locus; e.g., *garments* vs *stones*)
  - `region = spillover_postposition` , `spillover_verb`
- **Comprehension probe (33% of trials)** after sentence:

  "Do ZZ wear layered wool garments during ceremonies?"

  Expected answer: **Yes** (based on background).

---

2) Step-2 — Judgment Task (immediate, trial-wise)

Option A — Plausibility (recommended for ERP alignment)

- Prompt: *"Given the background, does this continuation make sense?"*

  Response: **Yes / No** (record RT too)

  Answer key for the four variants:

  - Hate+Plausible → **Yes**
  - Neutral+Plausible → **Yes**
  - Hate+Implausible → **No**
  - Neutral+Implausible → **No**
- Example
  - "ZZ족은 의식 때 양털로 만든 겹옷을 입는다." → True
  - "ZZ족은 바다 근처에 산다." → False

---

3) Step-3 — Free Description & Evaluation

Free description prompt (2–3 sentences)

- *“In your own words, briefly describe what the ZZ group is like.”*

Example participant responses

- Neutral condition (typical):

  “그들은 높은 산속에 살며 전통 의상을 입고 의식을 치른다. 오리구이를 함께 먹는다.”

- Hate condition (risk of reproduction):

  “그들은 열등하고 더러운 옷을 입는다. 모임 전 이상한 음식을 먹는다고 들었다.”

Evaluation scales (7-point Likert; reverse-score as needed)

- *Warmth* (1=cold, 7=warm)
- *Competence* (1=incompetent, 7=competent)
- *Trust* (1=untrustworthy, 7=trustworthy)
- *Willingness to interact* (1=avoid, 7=approach)
  - Automatic NLP: sentiment polarity, toxicity lexicon hits, bigram overlap with stimuli.

Preprocessing & Exclusions (TBA)

- Exclude outliers:
  - Trim RTs per subject×region: exclude `<150 ms` or `>3000 ms`; then **±2.5 SD** winsorization or exclusion.
- Data Quality:
  - Drop trials with comprehension-probe errors on that item (optional).
  - Exclude participants with **<75%** probe accuracy or **>25%** trimmed trials.

Analyses

A) Word-by-word RT (Linear Mixed-Effects)

- Model (R `lme4` style):

```
rt_ms ~ Mod * Plaus + Region + (1 + Mod*Plaus | subj) + (1 + Mod*Plaus | item)
```

  - Primary tests at **critical_noun** and **spillover1** (subset or interaction with `Region`).
  - Planned contrasts: Hate vs Neutral within **Plausible** and within **Implausible**.

B) Judgment accuracy & RT

- Accuracy (logistic mixed model):

  ```
  correct ~ Mod * Plaus + (1 + Mod*Plaus | subj) + (1 + Mod*Plaus | item)
  ```

- Judgment RT (linear mixed model as in A).

C) Free description & evaluation

- Sentiment/toxicity/reproduction scores:

  ```
  outcome ~ Mod * Plaus + (1 + Mod*Plaus | subj) + (1 + Mod*Plaus | item)
  ```

- **Exploratory linkage**: correlate per-subject **RT cost** (Hate–Neutral at critical/spillover) with **negativity/reproduction** indices (Pearson/Spearman); optional mediation (RT → judgment → reproduction).

Reporting

- Fixed-effect estimates (β), **95% CIs**, p-values (Satterthwaite df or parametric bootstrap), effect sizes (standardized β).

Minimal item set (ready to pilot; 8 items total)

- Create **24–32 items** for final study; for a pilot, use **8 items** (2 per cell) + 8 fillers.
- Balance frequency/length; pre-rate **plausibility** and **valence/arousal** in a separate sample.

Expected Outcomes & Theoretical Significance

- Hate speech is expected to:
  - Increase **semantic processing load** (longer RTs - potential N400 modulation).

    - Strengthen **memory retention** through emotional salience.
    - Promote **linguistic reproduction** of negative or hateful content.
- The results could clarify **how emotionally charged language spreads cognitively and socially —** linking linguistics, cognition, and ethics.

Long-Term Plan

- This behavioral mini-experiment will serve as a **foundation for the BA thesis**.
- Later integration with **EEG measurements** (N400/P600) to model real-time semantic and emotional effects.
- Aim: a publishable study bridging **experimental linguistics and computational cognitive neuroscience**.