Carl von Ossietzky
Universität
Oldenburg

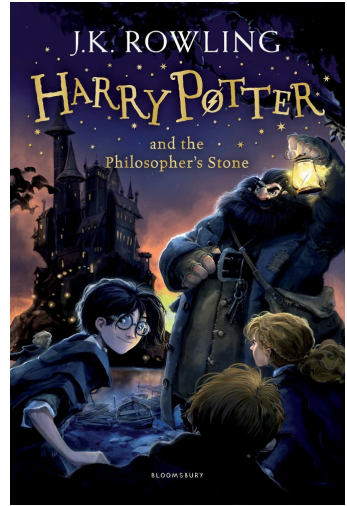
hhu Heinrich Heine
Universität
Düsseldorf

Semantics determines choices of writing styles in Japanese: A computational approach

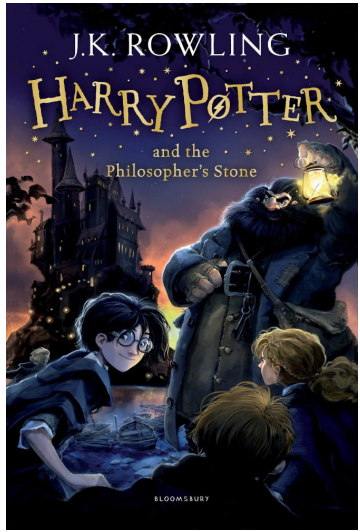
Motoki Saito & Ruben van de Vijver

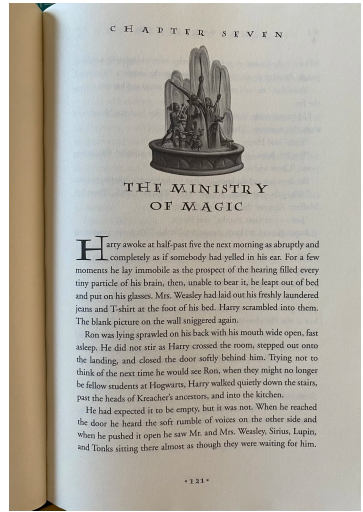
31.10.2025
International Symposium on Digital Humanities and AI Art
National Sun Yat-sen University, Taiwan

Harry Potter



Harry Potter





CHAPTER SEVEN

THE MINISTRY OF MAGIC

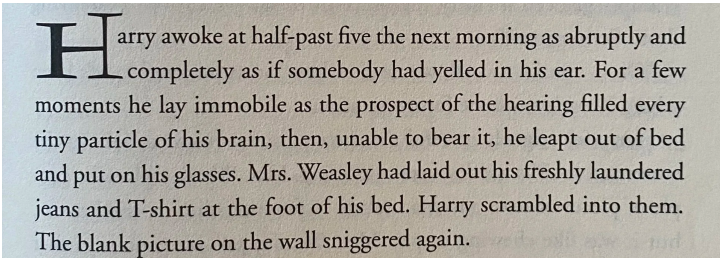
Harry awoke at half-past five the next morning as abruptly and completely as if somebody had yelled in his ear. For a few moments he lay immobile as the prospect of the hearing filled every tiny particle of his brain, then, unable to bear it, he leapt out of bed and put on his glasses. Mrs. Weasley had laid out his freshly laundered jeans and T-shirt at the foot of his bed. Harry scrambled into them. The blank picture on the wall sniggered again.

Ron was lying sprawled on his back with his mouth wide open, fast asleep. He did not stir as Harry crossed the room, stepped out onto the landing, and closed the door softly behind him. Trying not to think of the next time he would see Ron, when they might no longer be fellow students at Hogwarts, Harry walked quietly down the stairs, past the heads of Kreacher's ancestors, and into the kitchen.

He had expected it to be empty, but it was not. When he reached the door he heard the soft rumble of voices on the other side and when he pushed it open he saw Mr. and Mrs. Weasley, Sirius, Lupin, and Tonks sitting there almost as though they were waiting for him.

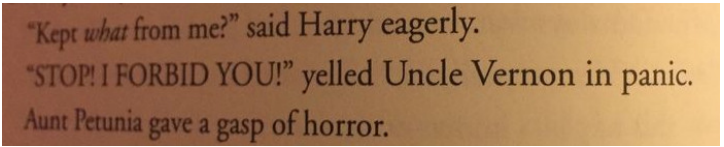
• 121 •

Harry Potter



Harry awoke at half-past five the next morning as abruptly and completely as if somebody had yelled in his ear. For a few moments he lay immobile as the prospect of the hearing filled every tiny particle of his brain, then, unable to bear it, he leapt out of bed and put on his glasses. Mrs. Weasley had laid out his freshly laundered jeans and T-shirt at the foot of his bed. Harry scrambled into them. The blank picture on the wall sniggered again.

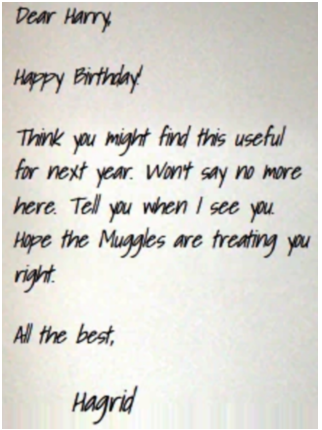
Uppercase


"Kept *what* from me?" said Harry eagerly.
"STOP! I FORBID YOU!" yelled Uncle Vernon in panic.
Aunt Petunia gave a gasp of horror.

Different fonts

Dear Professor Dumbledore,
Given Harry his letter.
Taking him to buy his things tomorrow.
Weather's horrible. Hope you're well.
Hagrid

Dear Hermione,
We lost. I'm allowed to bring him back to Hogwarts.
Execution date to be fixed.
Beaky has enjoyed London.
I won't forget all the help you gave us.
Hagrid



The same written words express different meanings

- Writers can choose what to express and how.

  e.g., Uppercase → Loud

  Compare “what?” vs “WHAT?”

The same written words express different meanings

- Writers can choose what to express and how.
  - e.g., Uppercase → Loud
    Compare “what?” vs “WHAT?”
  - e.g., Different fonts → Different feelings:
    Compare The cool brown fox. vs The chique brown fox.

One language, one writing system?

- How do you express different “feelings” for just *we*?
  e.g., “WE”? “we”?

One language, one writing system?

- How do you express different “feelings” for just *we*?
  e.g., “WE”? “we”?
- There is a limitation.
  - Because usually one language has one writing system.

One language, one writing system?

- How do you express different “feelings” for just *we*?
  e.g., “WE”? “we”?
- There is a limitation.
  - Because usually one language has one writing system.
  - Or having multiple writing systems, you can’t mix them.
    e.g., 我們 “we”
    e.g., *I們 “we”
    e.g., *me們 “we”

Japanese writing systems

- Japanese has three writing systems:

Japanese writing systems

- Japanese has three writing systems:
  - Hiragana
  - Katakana
  - Kanji (Chinese characters)

Japanese writing systems

- Japanese has three writing systems:
  - Hiragana

    e.g., わたしたち “we”
    → Casual, young, child, cute, etc.
  - Katakana
  - Kanji (Chinese characters)

Japanese writing systems

- Japanese has three writing systems:
  - Hiragana
    e.g., わたしたち “we”
    → Casual, young, child, cute, etc.
  - Katakana
    e.g., ワタシタチ “we”
    → Unnatural, robot, foerign, etc.
  - Kanji (Chinese characters)

Japanese writing systems

- Japanese has three writing systems:
  - Hiragana
    e.g., わたしたち “we”
    → Casual, young, child, cute, etc.
  - Katakana
    e.g., ワタシタチ “we”
    → Unnatural, robot, foerign, etc.
  - Kanji (Chinese characters)
    e.g., 私達 “we”
    → Formal, adult, official, etc.

Japanese writing systems

- Japanese has three writing systems:
  - Hiragana
    e.g., わたしたち “we”
    → Casual, young, child, cute, etc.
  - Katakana
    e.g., ワタシタチ “we”
    → Unnatural, robot, foerign, etc.
  - Kanji (Chinese characters)
    e.g., 私達 “we”
    → Formal, adult, official, etc.

- They can be mixed (relatively) freely:

Japanese writing systems

- Japanese has three writing systems:
  - Hiragana
    e.g., わたしたち “we”
    → Casual, young, child, cute, etc.
  - Katakana
    e.g., ワタシタチ “we”
    → Unnatural, robot, foerign, etc.
  - Kanji (Chinese characters)
    e.g., 私達 “we”
    → Formal, adult, official, etc.

- They can be mixed (relatively) freely:
  - Katakana + Hiragana
    e.g., ワタシたち “we”
    → Young women/teenagers/girls?

Japanese writing systems

- Japanese has three writing systems:
  - Hiragana
    e.g., わたしたち “we”
    → Casual, young, child, cute, etc.
  - Katakana
    e.g., ワタシタチ “we”
    → Unnatural, robot, foerign, etc.
  - Kanji (Chinese characters)
    e.g., 私達 “we”
    → Formal, adult, official, etc.

- They can be mixed (relatively) freely:
  - Katakana + Hiragana
    e.g., ワタシたち “we”
    → Young women/teenagers/girls?
  - Hiragana + Kanji
    e.g., わたし達 “we”
    → Sounds a bit older. Maybe young but adult women?

Japanese writing systems

- Japanese has three writing systems:
  - Hiragana
    e.g., わたしたち “we”
    → Casual, young, child, cute, etc.
  - Katakana
    e.g., ワタシタチ “we”
    → Unnatural, robot, foerign, etc.
  - Kanji (Chinese characters)
    e.g., 私達 “we”
    → Formal, adult, official, etc.

- They can be mixed (relatively) freely:
  - Katakana + Hiragana
    e.g., ワタシたち “we”
    → Young women/teenagers/girls?
  - Hiragana + Kanji
    e.g., わたし達 “we”
    → Sounds a bit older. Maybe young but adult women?
  - Katakana + Kanji
    e.g., ワタシ達 “we”
    → A particular way of talking?
  - ...etc.

Research question

- How does the writer choose a writing system or a combination of them?

Research question

- How does the writer choose a writing system or a combination of them?

⇓

- Intuitively: It is decided according to what you want to sound like.
  e.g., Should I sound like an adult person speaking in a formal situation?
  e.g., Should I sound cute like a little girl talking to her close friends?

Psycholinguistic theories

- Psycholinguistic theories do not say much about it (e.g., Dell, 1986; Levelt et al., 1999)

Psycholinguistic theories

- Psycholinguistic theories do not say much about it (e.g., Dell, 1986; Levelt et al., 1999)

↓

- This is because Japanese has a unique writing system, which has not been the subject of much psycholinguistic research.

Aim of the study

- Let's check if "what you want to sound like" really determines the choice of writing systems!

Discriminative Lexicon Model (DLM)

- A computational psycholinguistic model (e.g., Baayen et al., 2019)
- Conceptually, it is a model of an individual person.

Discriminative Lexicon Model (DLM)

- A computational psycholinguistic model (e.g., Baayen et al., 2019)
- Conceptually, it is a model of an individual person.
- We trained DLM to predict writing systems based on meanings.
  - Input: The meaning of the word.
  - Output: The writing systems of the word.

Semantic vectors

- You can think of a word's "meaning" as what context it may appear in (e.g., Landauer & Dumais, 1997)

  e.g., *dog* is similar in meaning to *cat*, compared to *universe*.

Semantic vectors

- You can think of a word's "meaning" as what context it may appear in (e.g., Landauer & Dumais, 1997)

  e.g., *dog* is similar in meaning to *cat*, compared to *universe*.

  Q. Why?

Semantic vectors

- You can think of a word's "meaning" as what context it may appear in (e.g., Landauer & Dumais, 1997)
  - e.g., *dog* is similar in meaning to *cat*, compared to *universe*.
    - Q. Why?
    - A. Because *dog* and *cat* are more likely to occur in the same context (similar sentences) than *universe*.

Semantic vectors

- You can think of a word's "meaning" as what context it may appear in (e.g., Landauer & Dumais, 1997)

  e.g., *dog* is similar in meaning to *cat*, compared to *universe*.

  Q. Why?

  A. Because *dog* and *cat* are more likely to occur in the same context (similar sentences) than *universe*.

- This "context" includes styles of writing:
  - Formal $\rightarrow$ A certain choice of words (e.g. good bye).
  - Foreign $\rightarrow$ Another choice of words (e.g. adios).
  - Casual $\rightarrow$ Yet another choice of words (e.g. see ya).

Semantic vectors

- You can think of a word's "meaning" as what context it may appear in (e.g., Landauer & Dumais, 1997)

  e.g., *dog* is similar in meaning to *cat*, compared to *universe*.

  Q. Why?

  A. Because *dog* and *cat* are more likely to occur in the same context (similar sentences) than *universe*.

- This "context" includes styles of writing:
  - Formal → A certain choice of words (e.g. good bye).
  - Foreign → Another choice of words (e.g. adios).
  - Casual → Yet another choice of words (e.g. see ya).

  ⇓

  - Do words in the formal register appear in different contexts than those in the foreign or casual registers?

Semantic vectors

- You can think of a word's "meaning" as what context it may appear in (e.g., Landauer & Dumais, 1997)
  - e.g., *dog* is similar in meaning to *cat*, compared to *universe*.
  - Q. Why?
  - A. Because *dog* and *cat* are more likely to occur in the same context (similar sentences) than *universe*.
- This "context" includes styles of writing:
  - Formal → A certain choice of words (e.g. good bye).
  - Foreign → Another choice of words (e.g. adios).
  - Casual → Yet another choice of words (e.g. see ya).

  ⇓
  - Do words in the formal register appear in different contexts than those in the foreign or casual registers?

  ↓
  - If yes, it would mean that we can predict writing systems based on semantics.

Calculating accuracies

- The trained DLM receives a word's meaning and prodicts which writing systems the word is written.
- We compared such a prediction against actual writing systems the word is really written in.

Results (1)

- Baseline accuracy: 16.67%

Results (1)

- Baseline accuracy: 16.67%
- Prediction accuracy: 99.79%

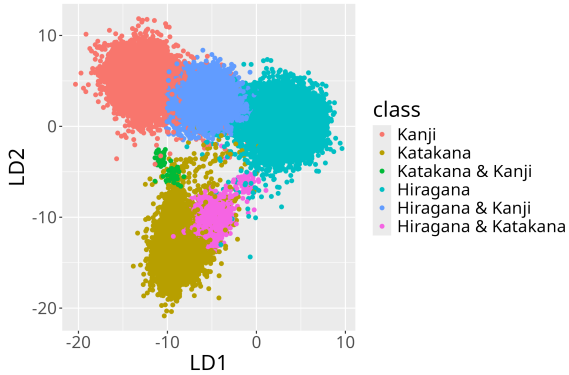
Results (1)

- Baseline accuracy: 16.67%
- Prediction accuracy: 99.79%

$\Downarrow$

- We can predict writing systems only by words' meanings.

Results (2)



- Writing systems are well-organized and separated in the semantic space.
  - LD1: Kanji (red) ↔ Hiragana (blue-green; turquoise)
  - LD2: Katakana (yellow) ↔ Others

Discussion

- Background:
  - Japanese has different ways of writing the same word.
  - How are writing systems chosen by the writer?

Discussion

- Background:
  - Japanese has different ways of writing the same word.
  - How are writing systems chosen by the writer?
- Results:
  - DLM could tell writing systems based only on meanings (accuracy 99.79%).
  - LDA analysis showed clear separation of writing systems in semantics.

Discussion

- Background:
  - Japanese has different ways of writing the same word.
  - How are writing systems chosen by the writer?
- Results:
  - DLM could tell writing systems based only on meanings (accuracy 99.79%).
  - LDA analysis showed clear separation of writing systems in semantics.
- Interpretation:
  - There is a tight relationship between a word's meaning and how the word is written.

Carl von Ossietzky
Universität
Oldenburg

hhu

Thank you very much!

ごせいちょうありがとうございました！

御清聴有難う御座いました！

ご清聴ありがとうございました！

ゴセイチョウアリガトウゴザイマシタ！

References I

[1] Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., & Blevins, J. P. (2019).The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. Complexity, 2019, 1–39. https://doi.org/10.1155/2019/4895891

[2] Dell, G. S. (1986).A spreading-activation theory of retrieval in sentence production. Psychological Review, 93(3), 283–321. https://doi.org/10.1037//0033-295x.93.3.283

[3] Landauer, T. K., & Dumais, S. T. (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.

[4] Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999).A theory of lexical access in speech production. Behavioral and Brain Sciences, 22, 1–75.