Subject Omission in Japanese Across Different Sources and Medical Context

Shusuke Hashimoto
Indiana University Bloomington, MS, Department of Linguistics

INDIANA UNIVERSITY
Department of Linguistics

INTRODUCTION

This study explores subject omission in Japanese within the medical context, focusing on three different types of sources.

- Forum - informal discussions where individuals share experiences and seek advice in a conversational manner.
- Tweets - brief and spontaneous expressions often reflecting immediate thoughts or reactions.
- Reports - formal and structured documentation, presenting information in an organized and professional format.

The Japanese language is characterized by pro-drop features, which often manifest in the flexibility of subject appearance.

Research Questions:

1. How does subject omission behave differently in the medical context?
2. How does subject omission differ across different sources, such as forums, reports, and tweets?

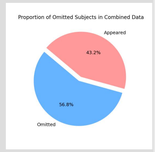

Chart 1.
Omitted or appeared subjects in combined data

METHODS AND MATERIALS

Data Collection

- The Japanese texts were sourced from Raithel *et al.* (2024), who presented their research in *"A Dataset for Pharmacovigilance in German, French, and Japanese: Annotating Adverse Drug Reactions Across Languages."*
- The forum texts were collected from *Yahoo! Japan Chiebukuro* (YJQA), while the tweet data were gathered from X (formerly Twitter).
- All texts focus on adverse drug reactions (ADR).

Data Cleaning

- Since the Japanese language does not include spaces between words, the *fugashi* tokenizer was implemented to segment the texts.
- Manual annotation was conducted to mark instances of subject omission using the placeholder "[0]" where a subject would be syntactically required.
- Additionally, emojis and extraneous symbols were removed from the YJQA and X texts.

Example:
Original: しょうがない😥と説明されています！
Cleaned and Annotated: *[0]* しょうがないと*[0]* 説明されています

Annotation

- The annotation process was carried out using the software ELAN.
- Annotation guidelines were developed based on prior studies, including *"Annotation Manual for the NPCMJ" (Nagasaki et al., 2019) and "Constructing Web-Accessible Semantic Role Labels and Frames for Japanese as Additions to the NPCMJ Parsed Corpus" (Takeuchi et al., 2020).*

Key Annotation Tiers

- Anaphora - how omitted or explicit elements relate to prior discourse.
  [Features] new / explicit / zero / exophoric
- Pronoun - based on their grammatical number and person
  [Features] "sing / plur=1 / 2 / 3" Ex) あなた("You") => sing=1
- Voice - whether the subject performs the action or receives the action
  [Features] active / passive
- Medical - whether a token is related to the medical domain
  [feature] medical

RESULTS

Subject Omission for Each Source (see chart 2)
Tweets show the highest rate of subject omission at approximately 60%, followed closely by Forums, which have a similar rate. Reports exhibit the lowest percentage, around 50%, indicating that subject omission occurs less frequently in this source compared to the others. This suggests that written reports may require more explicit subject use, while informal communication like Tweets and Forums allows for more subject omission.

Subject Omission in The Medical Context (see chart 3)
Reports have the highest percentage, around 15%, indicating that subject omission related to medical topics is most common in formal reports. Tweets show a smaller percentage, roughly 7%, while Forums have the lowest proportion, close to 2%. This suggests that medical-related subject omission occurs more frequently in structured, formal sources like reports compared to informal platforms such as forums and tweets.

Thematic Role for All Subjects vs Omitted Subjects (see chart 4 and 5)
The first chart shows that theme (45.7%) is the most common role for subjects, followed by agent (27.8%) and experiencer (16.2%). In the second chart, agent (46.1%) dominates among omitted subjects, while theme decreases to 18.5%. The experiencer role remains relatively stable, with smaller roles like patient and others showing minor proportions in both cases. This suggests that agents are more likely to be omitted, while themes are less frequently omitted.

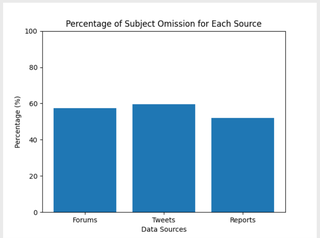

Chart 2.
Percentage of subject omission for each source

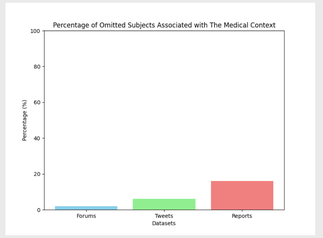

Chart 3.
Percentage of subject omission in the medical context

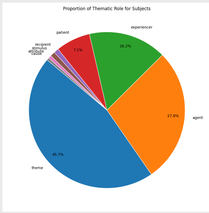

Chart 4.
Proportion of thematic role for subjects

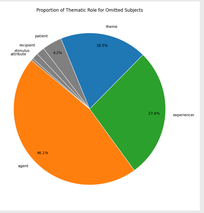

Chart 5.
Proportion of thematic role for omitted subjects

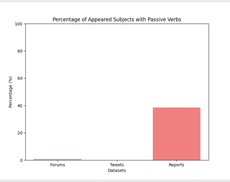

Chart 6.
Percentage of appeared subjects with passive verbs

Subject Appearance with Passive Verbs (see chart 6)
Reports stand out with a significantly higher percentage, around 38%, indicating that passive verbs with explicit subjects are more common in formal, structured texts. In contrast, Forums show a minimal percentage close to 1%, and Tweets register at 0%, suggesting that passive constructions are rarely used in these informal and conversational sources. This highlights a clear contrast in verb usage patterns between formal and informal contexts.

DISCUSSION

- Subject omission appears to be more acceptable in informal contexts (Forums and Tweets), where conversational style and brevity are prioritized, while Reports demand clearer subject presence to maintain explicitness and formality.
- In medical contexts, subject omission may be more flexible in Reports due to professional audiences relying on shared knowledge, whereas informal platforms like Tweets and Forums require explicit subjects to avoid ambiguity.
- Agents are more likely to be omitted compared to themes, suggesting that the agent role is often assumed or understood in context, particularly in informal sources, while themes remain more explicit in formal communication.
- Passive verbs in Reports often retain explicit subjects to ensure clarity and professionalism, contrasting with informal sources like Forums and Tweets, where subjects are omitted more frequently due to conversational norms.

CONCLUSIONS

This study reveals that subject omission occurs most frequently in informal sources like Tweets and Forums, while formal Reports show a higher rate of explicit subject use, particularly with passive verbs. In the medical context, subject omission is more common in structured Reports, likely due to professional shared knowledge, whereas Tweets and Forums maintain greater explicitness to avoid ambiguity. The thematic role analysis shows that agents are most frequently omitted, while themes remain more explicit, reflecting a tendency for assumed agents in informal communication. Overall, the results highlight the interplay between context formality, medical topics, and linguistic flexibility in Japanese subject omission.

REFERENCES

Raithel, A., et al. (2024). A Dataset for Pharmacovigilance in German, French, and Japanese: Annotating Adverse Drug Reactions Across Languages. *Proceedings of the 2024 Conference on Language Resources and Evaluation (LREC)*.

Nagasaki, I., et al. (2019). Annotation Manual for the NPCMJ. *National Institute for Japanese Language and Linguistics (NINJAL)*.

Takeuchi, S., et al. (2020). Constructing Web-Accessible Semantic Role Labels and Frames for Japanese as Additions to the NPCMJ Parsed Corpus. *Proceedings of the 2020 Conference on Language Resources and Evaluation (LREC)*

Iida, R., et al. (2007). Annotating a Japanese Text Corpus with Predicate-Argument and Coreference Relations. *Proceedings of the Linguistic Annotation Workshop*, 132–139

Futagi, Y. (2004). Japanese Focus Particles at the Syntax-Semantics Interface (Doctoral Dissertation). Rutgers, The State University of New Jersey.

NOTE

*exophoric
something/someone not mentioned in previous text, but known from the context