Subject Omission in Japanese Across Different Sources and Medical Context

Shusuke Hashimoto

Dept. of Linguistics, Indiana University Bloomington

Abstract

This study investigates patterns of subject omission in Japanese texts within medical contexts, focusing on three distinct sources: forums, tweets, and formal reports. Using annotated data, we examine the frequency and conditions under which subjects are omitted, emphasizing the impact of formality and thematic roles on omission patterns. Results reveal that informal platforms, such as forums and tweets, exhibit higher rates of subject omission compared to formal reports, where precision and explicitness are prioritized. Additionally, the study identifies the correlation between thematic roles, verb voice, and subject appearance, providing new insights into the linguistic strategies employed in professional and informal contexts.

1 Introduction

Subject omission, a defining feature of Japanese pro-drop syntax, allows speakers to omit grammatical subjects when their identity can be inferred from context. This flexibility is shaped by discourse, pragmatic factors, and the communicative context, making it a central focus in Japanese linguistics.

This study examines subject omission across three distinct sources—forums, tweets, and reports—representing informal and formal communication in the medical domain. Forums and tweets, characterized by conversational and concise expressions, contrast sharply with reports, which prioritize clarity, precision, and professionalism. By comparing these sources, the study addresses two research questions: How does subject omission behave differently in the medical context? How do patterns of subject omission differ across forums, tweets, and reports?

By analyzing subject omission across these sources, this study seeks to understand how subject realization is influenced by formality and communicative goals, offering insights into Japanese syntax and discourse patterns.

2 Methodology

2.1 Data Collection

The Japanese texts used in this study were sourced from Raithel et al. (2024), who presented A Dataset for Pharmacovigilance in German, French, and Japanese: Annotating Adverse Drug Reactions Across Languages. The data include forum texts collected from Yahoo! Japan Chiebukuro (YJQA) and tweets gathered from X (formerly Twitter). Both sets of texts focus on adverse drug reactions (ADR), providing a relevant medical context for analysis.

2.2 Data Cleaning

Since the Japanese language does not naturally include spaces between words, the fugashi tokenizer was implemented to segment the texts into appropriate word units. Manual annotation was conducted to mark instances of subject omission using the placeholder [0], which represents a syntactically required but omitted subject. For example:

Original:

	しょうがない	(´口`)	と	説明	
	shouganai		to	setsumei	
Lex. Gloss	unavoidable	emoji	that	explain	
	され	て	い	ます	！
	sare	te	i	masu	
Lex. Gloss	be$_{pass}$	and	-ing	-polite	excl

Free: (It) is explained that (it) cannot be helped.

Cleaned:

	[0]	しょうがない	と	[0]	
		shouganai	to		
Lex. Gloss	(omitted)	unavoidable	that	(omitted)	
	説明	され	て	い	ます
	setsumei	sare	te	i	masu
Lex. Gloss	explain	be$_{pass}$	and	-ing	-polite

All irrelevant emojis and extraneous symbols were removed from the forum and tweet datasets to ensure consistent and clean input for annotation.

2.3 Annotation

The annotation process was carried out using the ELAN software. Annotation guidelines were partially adapted from previous studies, including the *Annotation Manual for the NPCMJ* (Nagasaki et al., 2019) and *Constructing Web-Accessible Semantic Role Labels and Frames for Japanese* (Takeuchi et al., 2020). In total, 799 subjects were annotated, of which 454 were omitted and marked as [0]. The dataset includes 280 clauses from forums, 302 clauses from tweets, and 217 clauses from reports, with omitted subjects accounted for in each source. Below is the table summarizing the annotation tiers and their features to be marked.

Annotation Tiers	Features
Pronoun	• (person) • (number of pronoun)
Voice	• *active* • *passive*
Medical	• *medical*
Thematic Role	• (All possible thematic roles)
Animacy	• *animate* • *inanimate*
Discourse Type	• *focus* • *shift* • *continuity*
Politeness	• *polite*
Anaphora	• *new* • *explicit* • *zero* • *exophoric*
Clause	• *main* • *sub* (subordinate) • *rel* (relative)

Table 1: A table of expected performance scenarios and their corresponding datasets

The Pronoun tier annotates the person and number of pronouns, such as *sing=1* for singular first-person and *plur=3* for plural third-person. The Voice tier identifies whether the verb associated with the subject takes an active or passive form. The feature *active* is marked when the verb is in active form, and *passive* is marked when the verb is in passive form. The Medical tier identifies tokens related to medical terminology, marked with *medical*. The Thematic Role tier assigns semantic roles to subjects, such as *agent*, *theme*, *experiencer*, or *patient*, indicating their function in the sentence.

The Animacy tier distinguishes between subjects that are *animate*, referring to living beings, and *inanimate*, representing non-living entities. The Discourse Type tier identifies how subjects contribute to the flow of discourse, with features like *focus* for new information introduced in the sentence, *shift* for transitions in subject reference, and *continuity* for maintaining a consistent discourse. The Politeness tier focuses on the formality level, as often seen in formal or honorific language, marked with *polite*. Anaphora identifies how subjects, whether omitted or explicitly stated, relate to prior discourse within the text. The annotated features include *new*, which marks subjects that are mentioned but not referred back to, *explicit*, for subjects that are mentioned and referred back to, *zero*, indicating omitted subjects that are referred back to in context, and *exophoric*, for subjects that are not mentioned and not referred back to within the immediate discourse. Lastly, the Clause tier captures the syntactic structure of sentences by identifying subjects within *main*, subordinate (*sub*), or relative (*rel*) clauses. While the last four tiers—Discourse Type, Politeness, Animacy Anaphora, and Clause—are briefly introduced here, they are not further analyzed in this paper. These tiers will be explored in greater detail as part of future research.

3 Results

The results reveal distinct patterns in subject omission across tweets, forums, and reports, shaped by formality, thematic roles, and the medical context.

Figure 1 indicates that subject omission rates were highest in tweets, closely followed by forums, which exhibited similar levels. Reports demonstrated a significantly lower omission rate, around 50%. This suggests that informal sources like forums and tweets allow for greater subject omission, while reports, as formal documents, require more explicit subject use to ensure clarity.

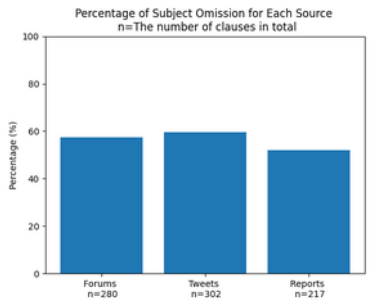

Figure 1: A diagram of the percentage of subject omission for each source

In Figure 2, subject omission related to medical topics appeared most frequently in formal reports, accounting for around 15%. Tweets showed a smaller proportion at roughly 7%, and forums exhibited the lowest rate, close to 2%. These results imply that professional audiences in formal medical contexts (e.g., reports) rely on shared knowledge, which allows for subject omission. In contrast, informal settings like tweets and forums require more explicit subjects to avoid ambiguity.

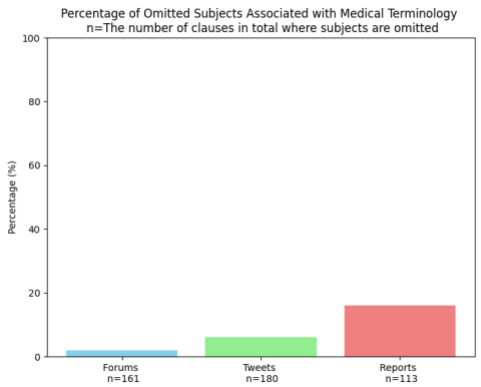

Figure 2: A diagram of the percentage of omitted subjects associated with medical terminology

When analyzing the thematic roles in Figure 3 and 4, theme was the most common role for explicit subjects (45.7%), followed by agent (27.8%) and experiencer (16.2%). However, for omitted subjects, agents dominated at 46.1%, while the proportion of themes dropped to 18.5%. This shift suggests that agents are more likely to be omitted due to contextual assumptions, whereas themes remain more explicitly stated.

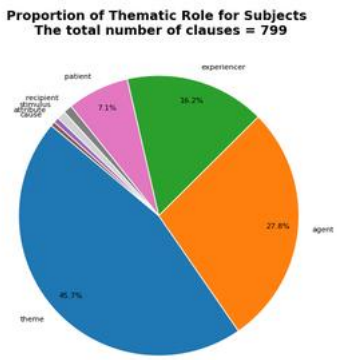

Figure 3: A diagram of the proportion of thematic roles for all subjects

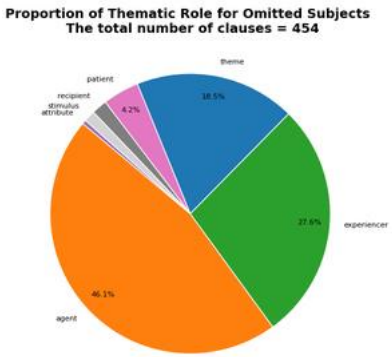

Figure 4: A diagram of the proportion of thematic roles for omitted subjects

In Figure 5, reports featured a notably high percentage of subjects appearing with passive verbs, around 38%. Forums, in contrast, showed minimal passive verb usage with explicit subjects, close to 1%, while tweets recorded 0%. These findings highlight a distinct pattern: passive constructions are more prevalent in formal and structured contexts like reports, where clarity and professionalism are prioritized.

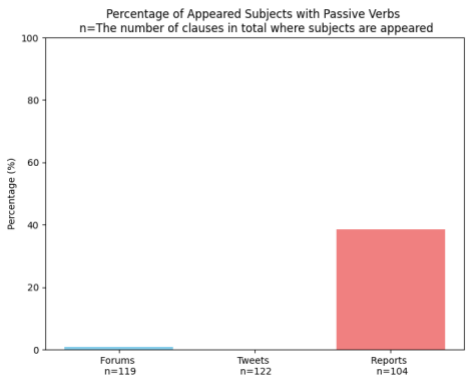

Figure 5: A diagram of the percentage of appeared subjects with passive verbs

4 Discussion

Subject omission exhibits distinct patterns when analyzed within the medical context and across different types of sources such as forums, tweets, and reports. In the medical context, subject omission is notably more frequent in formal reports. This tendency likely reflects the shared professional knowledge among readers, allowing for implicit understanding without the need for explicit subjects. Reports, as structured documents, maintain professional clarity even with omitted subjects due to their reliance on domain-specific language and context.

When comparing different sources, subject omission occurs most frequently in informal settings like forums and tweets, where *brevity and conversational tone are prioritized. In these contexts, the need for short and efficient communication encourages the omission of subjects, as interlocutors often rely on contextual cues to infer meaning. The casual and direct nature of these platforms allows for this economy of expression without significantly compromising clarity. In contrast, reports exhibit a lower rate of omission, as the professional and formal nature of this medium requires explicit subjects to ensure precision and avoid ambiguity for its specialized audience.

Thematic roles also differ between omitted and explicit subjects, which can explain the observed discrepancy in subject omission patterns. Agents, which tend to be animate entities, are more frequently omitted, particularly in informal contexts, as they are often assumed or inferred based on the surrounding context. This reflects a conversational or shared knowledge style where animate agents, such as patients and doctors, are naturally understood without explicit mention. Conversely, themes, which typically refer to medical materials or specific names of symptoms, are more likely to remain explicit because they provide critical and precise information. This is especially true in formal and structured texts like reports, where clarity and accuracy are prioritized to ensure the proper interpretation of medical content.

Passive constructions are notably more frequent with explicit subjects in formal reports, reflecting the emphasis on clarity and precision in professional medical communication. This can be attributed to the syntactic nature of the passive form, where the object of an active sentence is promoted to the subject position. As a result, passive constructions naturally require the subject to be overtly stated, reducing the likelihood of subject omission. In formal contexts like reports, this structure appears to function as a strategy to ensure unambiguous and explicit communication. Conversely, informal sources such as tweets and forums exhibit a preference for active voice and often leave subject positions unfilled, relying on contextual inference. This difference suggests a relationship between formality and subject omission, where formal writing, particularly in professional medical contexts, avoids omission more frequently, potentially through the use of passive constructions.

5 Conclusion

The findings of this study highlight the interplay between linguistic formality, context, and subject omission in Japanese texts. Informal sources, such as tweets and forums, prioritize brevity and conversational tone, resulting in higher rates of subject omission. In contrast, formal reports exhibit greater explicitness, often utilizing passive constructions to ensure clarity and precision. The analysis of thematic roles further underscores the contextual and semantic factors influencing omission patterns, with animate agents being more frequently omitted and inanimate themes being retained for their critical informational value.

These observations suggest a dynamic relationship between the communicative purpose of a text and its syntactic realization of subjects. While this study highlights Japanese subject omission in medical contexts through the analysis of thematic roles, verb voice, and source types, further investigation is necessary. Future research should incorporate the remaining annotation tiers—animacy, discourse type, politeness, clause, and medical features—alongside the current framework to deepen the understanding of subject omission patterns in Japanese, such as the correlation between politeness and subject omission. This comprehensive approach could also provide insights applicable to other languages and communicative domains.

6 References

Iida, Ryu, et al. "Annotating a Japanese text corpus with predicate-argument and coreference relations." *Proceedings of the linguistic annotation workshop*. 2007.

Horn, Stephen Wright, et al. "Annotation manual for the NPCMJ." (2019).

Raithel, Lisa, et al. "A Dataset for Pharmacovigilance in German, French, and Japanese: Annotating Adverse Drug Reactions across Languages." *arXiv preprint arXiv:2403.18336* (2024).

Takeuchi, Koichi, et al. "Constructing web-accessible semantic role labels and frames for Japanese as additions to the NPCMJ parsed corpus." *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 2020.

*Brevity, especially on platforms like Twitter with character limits, reflects the need for concise communication, encouraging subject omission to maximize efficiency.