To Drop or Not to Drop? Predicting Argument Ellipsis Judgments: A Case Study in Japanese

Yukiko Ishizuki[1,2] **Tatsuki Kuribayashi**[3] **Yuichiroh Matsubayashi**[1,2]
Ryohei Sasano[4,2] **Kentaro Inui**[3,1,2]
[1]Tohoku University, [2]RIKEN, [3]MBZUAI, [4]Nagoya University
yukiko.ishizuki.p7@dc.tohoku.ac.jp, y.m@tohoku.ac.jp
sasano@i.nagoya-u.ac.jp, {tatsuki.kuribayashi, kentaro.inui}@mbzuai.ac.ae

Abstract

Speakers sometimes omit certain arguments of a predicate in a sentence; such omission is especially frequent in pro-drop languages. This study addresses a question about ellipsis—what can explain the native speakers' ellipsis decisions?—motivated by the interest in human discourse processing and writing assistance for this choice. To this end, we first collect large-scale human annotations of *whether* and *why* a particular argument should be omitted across over 2,000 data points in the balanced corpus of Japanese, a prototypical pro-drop language. The data indicate that native speakers overall share common criteria for such judgments and further clarify their quantitative characteristics, e.g., the distribution of related linguistic factors in the balanced corpus. Furthermore, the performance of the language model–based argument ellipsis judgment model is examined, and the gap between the systems' prediction and human judgments in specific linguistic aspects is revealed. We hope our fundamental resource encourages further studies on natural human ellipsis judgment.

Keywords: Argument Ellipsis, Language Resource, Discourse Processing, Japanese

1. Introduction

In pro-drop languages, arguments are often omitted from sentences, contrasting markedly with English where a subject is typically mandated (Craenenbroeck and Temmerman, 2019). For example, our corpus analysis revealed that 37% of arguments, such as subjects and objects in Japanese, are omitted (§3.1). This frequent occurrence of *argument ellipsis* or *omission*[1] has raised several challenges in natural language processing (NLP), such as recovering the omitted information to address the textual ambiguity (Sasano and Kurohashi, 2011; Wang et al., 2018).

In this study, we tackle a new question toward this phenomenon—*when is such a context-dependent ellipsis preferred by native speakers and which linguistic factor is associated with this decision?* The identification of human consensus on argument ellipsis judgments and its factors will contribute to linguistics and education: it can clarify the underlying consensus on the judgments among native speakers and be a helpful resource for writing assistance. Furthermore, it may be possible to explore applications such as an automatic text proofreading system can be explored based on the findings.

This paper presents the first large-scale study wherein the aforementioned question was investigated within a balanced corpus of Japanese. Japanese is a prototypical pro-drop language, which is typically adopted to study the ellipsis-related phenomena (Iida et al., 2007; Shibata and Kurohashi, 2018; Konno et al., 2021). In contrast, existing studies have typically investigated similar phenomena within specific constructions, e.g., verb phrase ellipsis (Schäfer et al., 2021) and relativizer (Jaeger and Levy, 2006) in English. We specifically explored two key questions: (i) To what extent can native speakers agree on ellipsis judgments and their rationale in a natural, balanced sample of text? (ii) How accurately can human ellipsis judgments be predicted by recent NLP approaches?

To achieve these objectives, we first created a dataset that determines whether and why a particular argument should be omitted.[2] This dataset was built with over 2,000 data points extracted from a large, balanced Japanese corpus (Maekawa et al., 2014) and our newly designed annotation protocol and tool (Figure 1). Through creating our dataset, we found a general consensus in human judgments among the annotators, and the degree of this agreement varied depending on the associated linguistic factors. Furthermore, our annotations exhibited a clear parallel with the ellipsis judgments in the original corpus, suggesting a shared consensus about these judgments between the author (of the corpus text) and the readers (i.e., workers).

To examine question (ii), we benchmarked the performance of language models (LMs) on the ellipsis judgment task, including BERT, Japanese LMs, and GPT-4 (Devlin et al., 2019; OpenAI, 2023). Experimental results showed that the models strug-

---

[1] We use both terms interchangeably in this paper.

[2] Our data and codes are available at `https://github.com/cl-tohoku/JAOJ`.

LREC-COLING 2024, pages 16198–16210
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

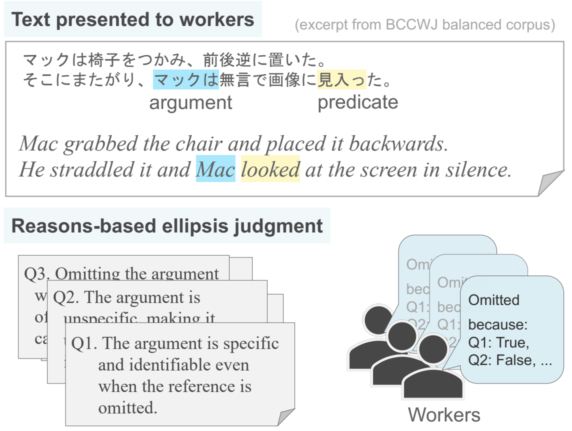

Text presented to workers
(excerpt from BCCWJ balanced corpus)
マックは椅子をつかみ、前後逆に置いた。
そこにまたがり、マックは無言で画像に見入った。
argument
predicate
Mac grabbed the chair and placed it backwards.
He straddled it and Mac looked at the screen in silence.
Reasons-based ellipsis judgment
Q3. Omitting the argument
Q2. The argument is unspecific, making it
Q1. The argument is specific and identifiable even when the reference is omitted.
Omitted
because:
Q1: True,
Q2: False, ...
Workers

Figure 1: Illustration of the argument ellipsis judgment annotation. In addition to asking for the final judgments, we also asked workers to answer the questions corresponding to the linguistic factors that potentially influence their judgments.

gled with certain types of ellipsis judgments, highlighting the human-LMs gap in discourse processing in the realm of ellipsis.

2. Japanese Argument Ellipsis

In Japanese, every argument of a predicate can potentially be omitted depending on the context:

(1) *Kare-wa mô kare-no ringo-wo tabe-ta.*
He-TOP already his_apple-ACC ate.

'He already ate his apple.'

(2) *Kare-no ringo-wa nai.*
His apple-TOP there_is_not.
(Kare-wa) mô (kare-no ringo-wo)
(He–TOP) already (his_apple-ACC)
tabe-ta-node.
eat-PAST-REASON.

'He doesn't have his apple. The reason is that (he) already ate (his apple).'

For example, in Example (1), the subject (*he*) and object (*his apple*) of the predicate (*eat*) are hardly omitted without context, but these would be omitted with a particular context as in Example (2) considering the redundancy. The question is the decision criteria behind such ellipsis judgments. Note that an argument ellipsis occurs in a chunk-level structure. Thus, when the phrase *ringo-wo* (*apple*) is dropped in Example (2), its modifier *kare-no* (*his*) should also be omitted; this rule is applied to our dataset throughout this study.

Notably, our scope is beyond the typical focus of computational linguistics, for example, a syntactic reduction of relativizer (i.e., "*that*") (Jaeger and Levy, 2006), in the sense that we commit more to the strategy of deciding what semantic information, i.e., who did what for whom, to say (or not to say) beyond the omission of supplemental, functional words/markers (Lee, 2006). This broader scope makes it necessary for us to first break down the problem according to the associated linguistic factors (§3.3), and hopefully, our systematic analysis will pose an elaborated question of which type of ellipsis can be explained by a particular approach, e.g., information-theoretic measurement (Jaeger and Levy, 2006; Meister et al., 2021).

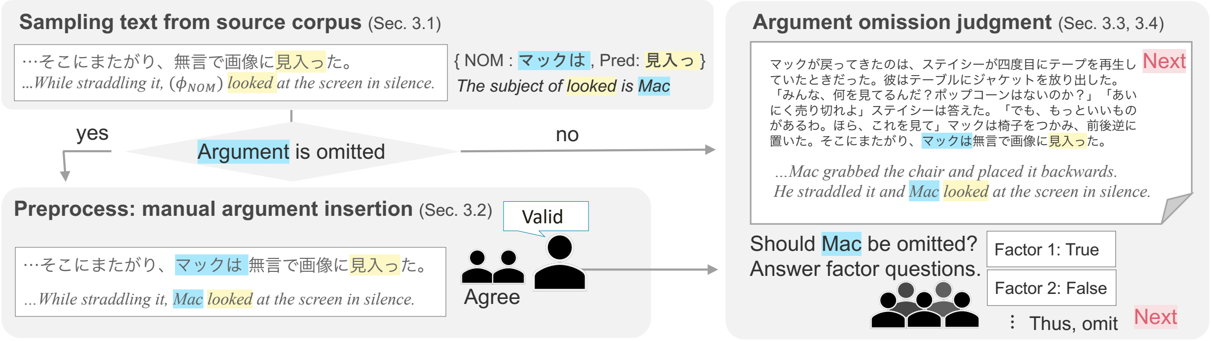
3. Annotation Task Design

Annotation Overview: Figure 2 shows data preprocessing (left part) and the annotation task (right part). We first collected sentences potentially involving argument ellipsis from the balanced corpus (§3.1 and §3.2). Then, workers were provided with a sentence along with the information of which predicate (*looked*) and argument (*Mac*) are targeted and the preceding document-level context. Here, the workers were asked to infer whether the specified argument should be omitted or not along with its potential reason (§3.3 and §3.4). The text was annotated using our newly designed tool for this task (§3.5).

3.1. Source Texts

Corpus: BCCWJ-DepParaPas (Maekawa et al., 2014), a widely-used Japanese balanced corpus with a predicate-argument structure (PAS) annotation, was used to create instances for asking ellipsis preferences. We executed annotation using 32 documents from the book domain of this corpus. The PAS annotation enabled us to identify the originally omitted arguments in the corpus; these can be used for collecting data points potentially favoring omission. We targeted nominative (NOM), accusative (ACC), and dative (DAT) arguments that were annotated in this corpus. Among the 17,481 arguments, including those omitted in the corpus, we finally used 2,373 arguments given the one-argument-per-sentence restriction (§3.5). The average length of the document-level contexts for each argument is 2,990 characters, and the average length of arguments is 6.25 characters. Further details of the data sampling procedure are provided in Appendix A.

Omission Statistics: Table 1 shows the frequency of argument omission in the corpus. The "Original" denotes the statistics of BCCWJ-DepParaPas, and the "Sampled" denotes the arguments targeted in this study. Omission is not a rare phenomenon in a natural corpus; specifically, NOM argument is omitted as frequently as around


Sampling text from source corpus (Sec. 3.1)
{ NOM : マックは , Pred: 見入っ }
The subject of looked is Mac
yes
Argument is omitted
no
Preprocess: manual argument insertion (Sec. 3.2)
Valid
Agree
Argument omission judgment (Sec. 3.3, 3.4)
Next
Should Mac be omitted?
Answer factor questions.
Factor 1: True
Factor 2: False
⋮ Thus, omit
Next

…そこにまたがり、無言で画像に見入った。
…*While straddling it,* (ϕ_{NOM}) *looked at the screen in silence.*

…そこにまたがり、マックは無言で画像に見入った。
…*While straddling it, Mac looked at the screen in silence.*

マックが戻ってきたのは、ステイシーが四度目にテープを再生していたときだった。彼はテーブルにジャケットを放り出した。「みんな、何を見てるんだ？ポップコーンはないのか？」「あいにく売り切れよ」ステイシーは答えた。「でも、もっといいものがあるわ。ほら、これを見て」マックは椅子をつかみ、前後逆に置いた。そこにまたがり、マックは無言で画像に見入った。

…*Mac grabbed the chair and placed it backwards. He straddled it and Mac looked at the screen in silence.*

Figure 2: Annotation procedure for our ellipsis judgment task.

	Original			Sampled		
Case	#Args	Omit	Insert	#Args	Omit	Insert
NOM	10,537	46.2%	53.8%	1,868	50.6%	49.4%
ACC	4,576	22.7%	77.3%	348	18.4%	81.6%
DAT	2,368	27.6%	72.4%	157	24.2%	75.8%
Total	17,481	37.5%	62.5%	2,373	44.1%	55.9%

Table 1: Omission rate of each grammatical case in the original and sampled corpus.

50%. The sampled arguments also reflect similar statistics to those in the original corpus.

Data Presented to Workers: The right part of Figure 2 shows the data provided to the workers. The arguments were either present or omitted in the original sentence, and we asked the workers to make an ellipsis judgment—*to drop or not to drop*—by showing a sentence with the target argument always **inserted**. In other words, the workers were not provided with information regarding whether or not the argument was initially omitted in the corpus. This protocol was employed to avoid any potential bias in their decision-making process. An additional preprocessing was required to naturally insert the originally omitted arguments into the sentence (§3.2). Further details on the data annotation interface are explained in §3.5.

3.2. Preprocess: Argument Insertion

The left part of Figure 2 shows an overview of the argument insertion procedure, which is needed in advance to implement our annotation scheme. To insert the originally omitted arguments into the sentence, a pair of workers, who were linguistics graduate students and not involved in the next ellipsis judgment task (§3.3), were asked to determine both the appropriate surface form and the position of the argument to be filled in the sentence. Such a task of positioning is necessary due to the flexible nature of word-order in Japanese sentences. If the two workers provided different answers, they were required to decide on the most plausible one. Subsequently, an additional worker further checked the data to exclude apparently collapsed sentences due to this insertion step.

3.3. Factor-level Questions

One of our objectives is to explore the underlying rationale influencing human judgments of argument ellipsis. To achieve this aim, instead of enquiring about the ellipsis decision directly, we asked workers a series of questions based on the potential linguistic factors related to ellipsis judgment. The final judgment of a worker was subsequently derived from the responses to these questions. The preprocessing and ellipsis judgment parts were performed by different annotators.

Factors: To investigate the linguistic factors associated with ellipsis judgments, the authors of this study conducted a trial annotation with small hold-out data. Then, we listed the following five factors by exploring potential reasons for cases in which the authors' choices aligned based on Japanese grammar (Tsujimura, 2013). Table 2 shows examples for each factor.

1. **Identifiability:** If a referent of the argument cannot be uniquely identified due to omission, it can potentially necessitate its explicit mention. This can also be seen as the difficulty of zero-anaphora resolution (Iida et al., 2008).
2. **Specificity:** If an unidentifiable argument needs to be specified to comprehend the text, it should be realized. Otherwise, if a context does not require uniquely specifying the argument entity, one can leave it unspecified by omitting it (Dipper et al., 2007).
3. **Connotation:** Affixed particles to the argument can serve a diverse range of connotative functions, such as topicalization and emphasizing contrasts or exclusivity. The argument must be

Factor and Question	Example	Answer
1. **Identifiability**: The argument is specific and identifiable even when the reference is omitted.	*Ringo-to mikan-ga aru. Saru-wa (ryôhô-wo) taberu-koto-ga dekiru.* Apple-AND orange-NOM be. Monky-TOP (both-ACC) eating-NOM can_do. ‘There are apples and oranges. The monkey can eat (both).’	False
2. **Specificity**: The unidentifiable argument should be specified (True) or unnecessary to specify in the given context (False).	*Seifu-wa (seifu-ga) taisaku-wo kento-suru-tame-ni iinkai-wo* government-TOP (government-NOM) measure-ACC consider-FOR committee-ACC *settisi-ta。* set_up-PAST. ‘The government has set up a committee as (the government) consider measures.’ (government vs. comittee for the subject of *consider*)	False
3. **Connotation**: *a*. Omitting the argument would result in different connotations.	*(Watashi-ga) choshoku-wo tabere-ba yokat-ta。* (I-NOM) breakfast-ACC eat-IF was_good. ‘(No one else but I) should have eaten the breakfast.’ (*Only the author (I) appears in this context)	True
(Only when 3*a* is true) *b*. Connotation incurred by omission should be plausible given the context.[3]		True*
4. **Grammaticality**: *a*. Omitting the argument would render the sentence ungrammatical or unnatural.	*Jiken-no kizi-wo sakusei-suru-tame, higaisha-ni (hanasi-wo) kî-ta。* incident-OF article-ACC create-FOR, victim-DAT (story-ACC) heard. ‘I heard (the story) from the victim to create an article about the incident.’	True
b. Including the argument would render the sentence ungrammatical or redundant.	*Mac-wa isu-wo tsukami, (Mac-wa) gyaku-ni oi-ta.* Mac-TOP chair-ACC grab-AND, (Mac-TOP) backwards put-PAST. ‘Mac grabbed a chair and (he) set it backwards.’	True
5. **Miscellaneous preferences**: Omitting/Inserting the argument enhances the naturalness of the sentence, although not necessary.	*(Watashi-wa) wakai-koro, America-ni ryûgaku-shi-te-ita。* (I-TOP) when_young, the U.S.-DAT studied_abroad. ‘(I) studied abroad in the U.S. when (I was) young.’	Could go either way

Table 2: Five factors found in the preliminary observation. We created the inquiry corresponding to each of these factors to ask workers the reason for their judgments. The *Answer* on the right side indicates whether the example sentence is True or False in response to the question about that specific factor.

retained if its omission can lead to the irrevocable loss of subtle connotations (Kenkyūkai, 2008).[3]

4. **Grammaticality:** If the realization or omission of the argument considerably disrupts the syntactic fluency of the sentence, it is undesirable to do so.
5. **Miscellaneous preferences:** When the aforementioned factors are inadequate for judgment, a more favorable one is selected based on naturalness.

Constraints and Preferences: We presumptively associated Factors 1-4 with a type of linguistic constraint and Factor 5 with preferences, respectively. In other words, different ellipsis judgments associated with Factors 1-4 can directly affect the information contents communicated between the writer and readers (Factors 1-3) or the accuracy of communication (Factor 4). In this sense, these factors were constrained to make the communication disambiguated and clear. In contrast, Factor 5 can be a type of subjective preference related to the naturalness of the text (e.g., preference toward salience and the efficacy/redundancy of communication). We tentatively introduced this categorization in the analyses and experiments (§5 and §6). Specifically, we referred to the ellipsis judgments associated with Factors 1-4 as “*hard-omission and insertion*” (HO and HI) and those with Factor 5 as “*soft-omission and insertion*” (SO and SI), respectively (see §3.4 for details). This distinction will generally be tied with the bi-dimensional pressures underlying communication: informativity (constraints) and complexity (preferences) (i Cancho and Solé 2003; Piantadosi et al. 2012; Frank and Goodman 2012; Kirby et al. 2015; Gibson et al. 2019; Xu et al. 2020; Hahn et al. 2020; *inter alia*).

---

[3] Practically, there might be a case wherein a worker cannot guess the preferable decision only from the context when Factor 3 is true (ellipsis affects the connotation). At least during our annotation, such a case rarely occurs; therefore, we just ignored these instances herein.

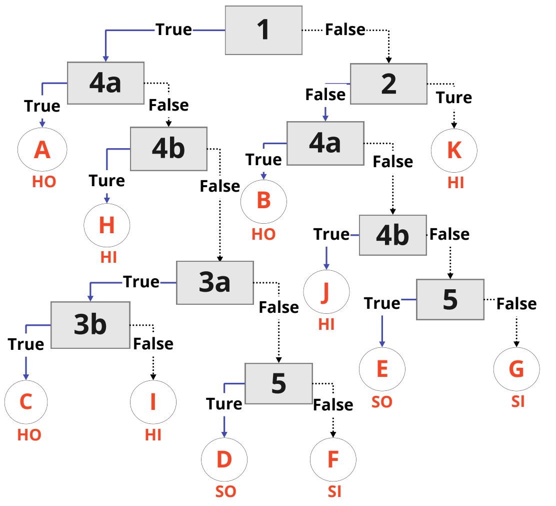


Figure 3: The decision tree used in the annotation process. Each number denotes the factor ID. The validity of the tree was confirmed during the preliminary tasks among the authors and the instructions provided to the workers.

3.4. From Factor-level Answers to Ellipsis Judgments

The workers were presented with the factor-level questions first and were asked to arrive at a decision on whether to omit the argument based on their answers. Specifically, the answers to the factor-level questions were mapped to the final ellipsis decision based on the decision tree shown in Figure 3. The node label in Figure 3 corresponds to the factor ID in Table 2. For example, if a worker answers *False* for *Factor 1* (identifiability) and *True* for *Factor 2* (specificity), the argument should be inserted as it is necessary to comprehend the text and it cannot be complemented with the context information.

To reduce the load on the workers, the questions were asked in the top-to-bottom order shown in Figure 3, i.e., Factor 1 is the first question. The worker finished answering when the answers could be mapped to the final ellipsis decision—whether it should be omitted or not.

3.5. Annotation Interface

Interface: To make the annotation efficient and controlled, we developed a dedicated annotation tool. Within this tool, the targeted argument and predicate are highlighted, and the worker answers the factor-level questions (§3.3) on the targeted argument with the displayed context. When workers press the "Next" button on the screen (right side of Figure 2), the subsequent target predicate-argument pair and its context are appended after the current text. In other words, the targeted arguments are shifted in the order in which they appear from the beginning of the text, resulting in the workers reading the entire text incrementally as they proceed with the task. This process is repeated until the end of one document. Note that to simplify the task, a single argument-predicate pair is targeted in each sentence, as described in §3.1. That is, when a sentence contains more than two arguments in a sentence, one to be worked on is randomly determined in advance.

		Omitted		Inserted	
Case	#Args	Hard	Soft	Soft	Hard
NOM	1,868	35.2%	14.9%	11.5%	38.4%
ACC	348	13.0%	8.9%	18.0%	60.0%
DAT	157	13.0%	5.7%	8.6%	73.0%
Total	2,373	30.4%	13.2%	11.5%	44.9%

Table 3: Distribution of aggregated human ellipsis judgment labels for each grammatical case.

Original Ellipsis Decision is not Leaked: As described in §3.1 and §3.2, the workers were not provided with information regarding whether the argument was initially omitted or not in the original corpus. To ensure this but prevent the incorrect interpretation of context, after the workers answered the questions and pressed the "Next" button, the original sentence was disclosed as the actual prior context for the forthcoming target instances. Therefore, we strictly prohibited the workers from revising their preceding judgments.

4. Data Creation

4.1. Workers

The ellipsis judgment task was conducted by five native Japanese speakers who were university students. Prior to the task, they participated in an instruction and training session where we ensured that they could achieve the expected judgments, which were predetermined by the authors of this study, in a hold-out small dataset.

4.2. Data Statistics

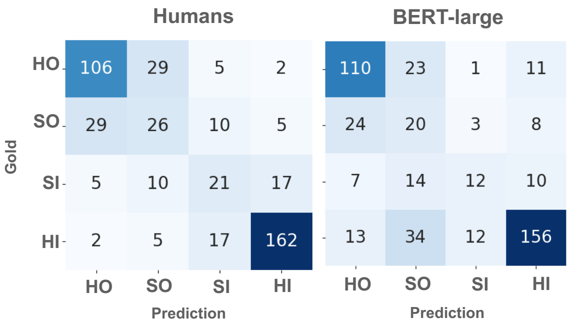
Label Distributions: To show the overall statistics, we aggregate the human judgment labels by the five workers for each data point as their median. Here, we assume the original scale of HO$\succ$SO$\succ$SI$\succ$HI (see §3.3 for these categorizations). Table 3 shows the label distribution in the collected data, where an overall trend that hard omissions/insertions more frequently occur than soft ones can be observed. Moreover, the distributions varied depending on the grammatical case types; `NOM` arguments were omitted more often (50.1%=35.2%+14.9%) than the other cases. Note that such an omission rate was similar to that in the corpus (Table 1).

Inter-worker Agreement: We calculated Krippendorff's alpha, the agreement measure for ordinal labels, for our five workers, resulting in 0.87. This indicates that the workers' judgments were generally consistent.[4] To investigate the label-wise inter-worker agreement, we calculate a one-vs-one F_1 score for every pair of two workers and consider an average of the scores. The first line in Table 5 shows the resulting scores. As suggested in §3.3, the constraint-based judgments (Factors 1–4) show comparatively higher scores of 74.6% for HO and 87.2% for HI. The scores of preference-based judgments (Factor 5) were considerably low (35.9% and 39.3%), suggesting that these categories have some degree of freedom in the choice.

Inter-worker Confusion Matrix: The left part of Figure 4 shows the confusion matrix on the one-vs-one evaluation. The number of instances was averaged by the number of trials. We observed that the SO/SI judgments had a larger variance than the HO/HI judgments. Moreover, the different label choices among the workers basically occurred between adjacent labels (e.g., HO and SO, or SO and SI) and rarely in between distant labels (e.g., HO and HI, or SO and HI).

Agreement with the Original Corpus: We also compared the workers' judgments with those made in the original corpus by converting four categories of HO, HI, SO, and HI into the binary decisions between omission and insertion. The agreement score was calculated in terms of the accuracy of these binary judgments in the original text and the aggregated label, resulting in 97.0%. This score suggests that the judgments for argument omission were consistent between a writer (authors of the corpus texts) and readers (annotators).

Factor-level Statistics: Table 4 shows the distribution of linguistic factors associated with the worker judgments. We confirmed that no specific linguistic factor dominated the worker judgments, but the associated factors spanned the listed multiple factors. Note that the associated factor for each argument was aggregated on a majority voting basis (see §6).



Humans	HO	SO	SI	HI
HO	106	29	5	2
SO	29	26	10	5
SI	5	10	21	17
HI	2	5	17	162

BERT-large	HO	SO	SI	HI
HO	110	23	1	11
SO	24	20	3	8
SI	7	14	12	10
HI	13	34	12	156

Figure 4: Confusion matrix of humans (left side) and BERT-large (right side) in validation data. Note that the numbers of instances are based on the average of one-vs-one evaluations for humans and based on the comparison between the system's predictions and the aggregated labels of humans for BERT_L.

	#Instances				
Label	1	2	4*a*	4*b*	5
True	1,419	736	119	25	295
False	940	704	1,322	914	273

Table 4: Statistics of factor-level questions.

5. Experiment: Predicting Human Ellipsis Judgments

We benchmark ellipsis judgment prediction models powered by standard natural language processing (NLP) approaches: (i) zero-shot prompting with large LMs, such as GPT-4 (OpenAI, 2023) and the Japanese LLM, Swallow (§5.2), and (ii) finetuning with BERT (Devlin et al., 2019). Note that pursuing a high-performance model for this task is not the objective of this paper; rather, we use these straightforward baselines to introduce our task.

5.1. Task

Task Formulation: We formulate the task involving the prediction of the aggregated judgment label of five workers (computed as in §4.2), i.e., one of the following four categories: HO, SO, SI, and HI. We employ the categorical classification as our initial foray. In a future study, we will introduce an ordinal scale to the models and metrics will be our future work. We used F_1 as the evaluation metric herein.

Data Split: We partitioned the data points into the training (1,459 arguments), validation (456), and test (458) sets. This partitioning was conducted at the document level, thus ensuring that data points across different folds do not share the context from

---

[4] A value greater than 0.677 is considered a high agreement (Krippendorff, 2004) .

Instruction for factor-level binary classification

task description = " You are a native Japanese speaker. You are requested to read the provided text and answer the following questions. Please note that the target predicate and its argument are surrounded by <predicate> and <argument>, respectively. \n
Question: Is the argument specific and identifiable even when the reference is omitted? "

prompt = f" Texts:{**text**} \n
Please output 'YES' or 'NO'."

Target text

text = "マックが戻ってきたのは、ステイシーが四度目にテープを再生していたときだった。彼はテーブルにジャケットを放り出した。「みんな、何を見てるんだ？ポップコーンはないのか？」「あいにく売り切れよ」ステイシーは答えた。「でも、もっといいものがあるわ。ほら、これを見て」マックは椅子をつかみ、前後逆に置いた。そこにまたがり、 <argument> マックが <argument> 無言で画像に<predicate> 見入っ<predicate>た。"

Figure 5: Prompt example for the GPT models. The targeted predicate and argument are surrounded by special tags (`<argument>` or `<predicate>`). An explanation of the targeted factor and an example illustrated in the annotation manual are also provided to the model.

the same document. We confirmed that the label distribution for each split was comparable.

5.2. Models

GPT-3.5 and GPT-4: GPT-3.5 and GPT-4 are multilingual models and have sufficient ability to use Japanese (Kasai et al., 2023). We obtain their argument ellipsis judgments by posing factor-level questions to them, utilizing a virtually identical protocol to that designed for human workers (§3), which enables a direct comparison between humans and models. Figure 5 shows the prompt template for a particular factor-level question, where the description of "Question" is changed depending on the corresponding factor-level question. The target context and predicate-argument pair are fed into the models, following the question.[5] The final ellipsis judgment is induced from the responses to all factor-level questions in the same manner as explained in §3.

Japanese large LMs: We also test Japanese large LMs, specifically, instruction-tuned versions of Swallow LMs with 13B (Swallow$_\mathrm{S}$)[6] and 70B (Swallow$_\mathrm{L}$)[7] parameters, which are created by further pre-training of Llama 2 (Touvron et al., 2023) with Japanese data. We use the same prompting setting as GPT-3.5 and GPT-4 experiments.

	Avg.	Omitted		Inserted	
Model	Macro	Hard	Soft	Soft	Hard
Humans	59.3$_{\pm 2.8}$	74.6$_{\pm 2.5}$	35.9$_{\pm 9.7}$	39.3$_{\pm 6.0}$	87.2 $_{\pm 2.2}$
GPT-3.5	21.4	17.5	4.2	15.1	48.9
GPT-4	25.2	32.7	11.6	9.2	47.2
Swallow$_\mathrm{S}$	21.7	40.8	7.0	3.8	35.2
Swallow$_\mathrm{L}$	22.7	36.7	10.5	7.6	35.9
BERT$_\mathrm{S}$	51.0$_{\pm 0.6}$	69.5$_{\pm 0.5}$	24.0$_{\pm 2.0}$	32.3$_{\pm 2.3}$	78.1$_{\pm 0.6}$
BERT$_\mathrm{L}$	50.2$_{\pm 4.3}$	76.0$_{\pm 4.7}$	29.6$_{\pm 3.7}$	10.9$_{\pm 5.0}$	84.2$_{\pm 4.2}$
Rand	20.2	28.2	11.4	10.8	30.3
Rand+	27.2	27.0	22.5	11.8	47.7

Table 5: Performance of human workers and the tested models. The values of human and BERT models are the average and standard deviation of five workers/models' results.

BERT: As another simple approach, we fine-tune BERTs, where the input is the "target text" shown in Figure 5 alone, and the training objective is simply handled as the four-class classification task. We use BERT-base-japanese (BERT$_\mathrm{S}$)[8] and BERT-large-japanese (BERT$_\mathrm{L}$).[9] We fine-tune models with five different seeds and report the standard deviation as well as the average of their performance. More detailed settings of input sequences and hyperparameters are described in Appendix C.

Baselines: We also examine two simple baselines: (i) *Rand* that predicts labels randomly, and (ii) *Rand+* that assigns the HI label for the `DAT` and `ACC` arguments, which are the majority label for these grammatical cases, and randomly assigns the decision label for the `NOM` argument.

5.3. Results

Table 5 shows the results of the four-class classification. As explained in §4.2, the human performance is the average F_1 score calculated through one-vs-one comparisons of each worker pair's predictions. We observed that the GPT and Swallow models only achieved baseline-level performance, indicating that they cannot replicate human-like Japanese ellipsis judgments, at least within this prompting setting.

---

[5] The context is limited to 512 subwords, considering the BERT's maximum input length. We randomly chose the answer label when the models did not yield a specific answer.

[6] `https://huggingface.co/tokyotech-llm/Swallow-13b-instruct-hf`

[7] `https://huggingface.co/tokyotech-llm/Swallow-70b-instruct-hf`

[8] `https://huggingface.co/cl-tohoku/bert-base-japanese-v2`

[9] `https://huggingface.co/cl-tohoku/bert-large-japanese`

The fine-tuned BERTs exhibited better performance than the prompt-based models, although this difference itself was not surprising since BERTs were trained with human annotations, whereas GPTs and Swallows were under the zero-shot prompting setting. BERT's performance was still inferior to humans in some aspects; predicting the soft preferences (SO/IO) was particularly challenging for the fine-tuned BERT compared to the human upper bound. The right side of Figure 4 shows the confusion matrix of the BERT$_{\mathrm{L}}$ model, which incorrectly predicted the distant labels than humans (e.g., HI v.s. SO). In general, the scores for HO/HI consistently surpass those for SO/SI. This tendency was also observed in human judgments, although the performance gap is more significant for LM-based models. In summary, all the tested models could not fully simulate human-like ellipsis judgments, which is a challenge for LMs.

6. Model Analysis

In this section, we analyze the ability of LMs for ellipsis judgments from the factor-level perspective.

Task: To evaluate the difficulty faced by LMs in answering the factor-level questions, we report the binary classification performance for each factor listed in §3.3. Here, a model is provided with the same instance as the ellipsis judgment task, but the task is to predict whether the answer is *True* or *False* for each factor-level question.

Models: As in §5, we examine zero-shot prompting with the GPT and Swallow models and finetuning with BERTs, but again the task is the factor-level binary classifications. Note that the answers of GPTs and Swallows to the factor-level questions are exactly those used in §5 to induce the final ellipsis judgment. Models are evaluated in terms of the macro F_1 of *True*/*False* labels. For human performance, the same one-vs-one evaluation method described in §4.2 is used.

Data: We partitioned each data point into binary classification tasks of *True*/*False*, each of which corresponds to the question of a specific factor. Here, we only used the data points where the gold labels could be determined via majority votes among the workers; this resulted in the creation of a different train/valid/eval split in a 3:1:1 ratio with the main experiment (§5). A very small number of samples were associated with Factor 3 (Table 4), thus this factor was excluded from the analysis. These procedures are detailed in Appendix D.

Model	Fac.1	Fac.2	Fac.4*a*	Fac.4*b*	Fac.5
Human	86.3$_{\pm 1.3}$	64.2$_{\pm 2.1}$	80.5$_{\pm 0.7}$	71.1$_{\pm 2.4}$	70.6$_{\pm 2.4}$
GPT-3.5	44.0	39.4	49.8	43.9	35.3
GPT-4	53.2	29.5	51.2	33.8	48.0
Swallow$_{\mathrm{S}}$	49.5	26.1	48.9	31.8	46.4
Swallow$_{\mathrm{L}}$	48.8	31.3	49.1	36.3	51.4
BERT$_{\mathrm{S}}$	85.5$_{\pm 0.4}$	53.3$_{\pm 1.6}$	76.8$_{\pm 1.6}$	61.6$_{\pm 1.8}$	58.8$_{\pm 2.1}$
BERT$_{\mathrm{L}}$	84.6$_{\pm 0.5}$	52.5$_{\pm 1.8}$	70.0$_{\pm 4.4}$	56.0$_{\pm 3.0}$	58.8$_{\pm 2.7}$
Rand	47.9	49.0	37.8	37.7	48.4

Table 6: Macro F_1 scores of the factor-level binary classifications. The values of human and BERT models are the average and standard deviation of five workers/models' results. The values of Rand are influenced by the imbalance in the distribution of the labels in Table 4.

Results: Results are shown in Table 6. Some parallels were observed in the performance differences across the factors between humans and BERTs. Factors 1 and 4a were relatively easy to solve, and Factors 2, 4b, and 5 were difficult for all humans and BERTs. Specifically, BERTs exhibited nearly human-upper bound performance in Factor 1 (identifiability), whereas a relatively large BERT–human performance gap was observed in Factor 5 (preference). These findings clarify that particular linguistic factors are associated with ellipsis judgments that are difficult for the models to predict. Notably, the GPT and Swallow models still yielded nearly random baseline performance in almost all factors except Factor 4a.

7. Related Work

7.1. Ellipsis in Computational Linguistics

The scope of ellipsis studies spans various fields, including theoretical syntax, semantics, discourse, and psycholinguistics (Merchant et al., 2001; Craenenbroeck and Temmerman, 2019). Our investigation will largely be aligned with psycholinguistic interests in ellipsis, aiming to unveil the biases associated with the ellipsis judgment or coreference resolution (Carminati, 2005; Michaelov and Bergen, 2022) during language processing; other perspectives are also involved in our listed linguistic factors though (§3.3). Some studies have tested specific hypotheses on ellipsis judgments, such as uniform information density (Jaeger and Levy, 2006; Schäfer et al., 2021) hypothesis and centering theory (Walker et al., 1994; Grosz et al., 1995), using controlled examples. This study, in contrast, employs a complementary empirical approach of analyzing the types of ellipsis that occur frequently in the balanced corpus and predicting them from both linguistic and engineering perspectives. We hope that our created resource facilitates further studies toward ellipsis judgment.

7.2. Ellipsis in NLP

In NLP, the zero anaphora resolution that automatically recovers the omitted argument/antecedent from its context has been actively explored (Sasano and Kurohashi, 2011; Poesio, 2010; Konno et al., 2021). Complementary to such studies, our study focuses on a different problem—when the argument should be omitted and why based on its context. This perspective is also important when writing/generating natural texts, achieving writing assistance, and computationally modeling human discourse processing. Notably, the Japanese language we focused on is a popular choice in such ellipsis studies (Iida et al., 2007; Shibata and Kurohashi, 2018; Konno et al., 2021).

7.3. Analyzing Discourse Processing of NLP Models

Discourse is a challenging linguistic aspect to process computationally. NLP models have been tested/probed from various discourse perspectives, and their hindered or non-human-like abilities have been typically reported (Bender and Koller, 2020; Sakaguchi et al., 2021; Upadhye et al., 2020; Koto et al., 2021; Schuster and Linzen, 2022; Fujihara et al., 2022). The experiments conducted herein were also aimed in this direction and exhibited that predicting the soft preferences was particularly challenging for the models.

Additionally, the linguistic competence of large LMs, e.g., the ability to judge the acceptability of a sentence (Dentella et al., 2023), has been tested via meta-linguistic prompting (Hu and Levy, 2023). We also employed such a prompting approach in evaluating large LMs to align the evaluation scheme with humans and facilitate their direct performance comparison. Nevertheless, Hu and Levy (2023) pointed out that the prompt-based method tends to underestimate the model's linguistic knowledge, and the probability-based evaluation is indeed another standard approach for quantifying the (context-dependent) sentence acceptability judgment ability of LMs (Warstadt et al., 2020; Hu et al., 2020; Sinha et al., 2023). Thus, benchmarking the model's ellipsis judgment preference through generative, probabilistic perspectives should be future work, although simple probability comparison might pose another potential limitation of simply avoiding a sentence containing new information (not identifiable in Factor 1, though) with low probability.

8. Conclusion

In this study, we examined the degree of consensus among native speakers on ellipsis judgments and their rationale within a naturally balanced text sample, motivated by the interest in human discourse processing and potential application to writing assistance. To this end, we have collected human annotations of whether and why a particular argument should be omitted across over 2,000 data points in a large balanced corpus. The collected data varied based on their associated linguistic factors. We also examined the performance of the LM-based argument ellipsis judgment model and quantified the gap between the system's and human's judgments. We hope that the annotation data we created serves as a foundational resource for future research toward argument omission.

9. Limitations

The annotation task was somewhat artificially designed compared with the human's real reading/writing activities. Another approach, such as directly measuring human reading behavior (e.g., reading time), will be complementary to our study. In addition, the number of workers involved in our annotation task was limited to five; thus, there might be some annotator biases in human gold judgment, although they were carefully instructed in trial sessions and showed high agreement, and we attempted to avoid these biases in workers' majors, academic years, and genders. We also suspect that some data points, such as those with "soft" decisions, may particularly reflect workers' biases, such as their reading ability. Analyzing such biases themselves may provide additional insights into human language processing, and thus, this will be an interesting future research.

The adopted linguistic factors 1–5 and the course categorization of SO/HO/HI/SI were mainly based on the trial examinations with the authors of this paper and previous linguistic studies. Quantitative support for this decision should have been preferred. Furthermore, Factor 5 (miscellaneous preferences) was a somewhat vague category and required further categorization. Nevertheless, this study isolated the data points with respect to potential linguistic factors. This will enable researchers to conduct studies by focusing more on specific aspects (e.g., preference) of ellipsis. For example, contrasting information-theoretic statistics with Japanese ellipsis preference will provide support for a particular hypothesis from a non-English perspective.

As for the experimental designs, the input for the models was limited to a particular token length (512 tokens). This setting might have underestimated the model performance, although in most cases, the context was less than 512 tokens in length. We also observed that GPT-3.5, GPT-4, and Swallow models yielded chance-level performances, which may be further improved by refining prompt designs or examining probability measurements. Regardless, the experiments in this paper are positioned as a demonstration aimed at predicting human ellipsis judgments. Consequently, the improvement of the performance will be the subject of a future investigation.

10. Ethics Statement

We ensured the collected data did not contain any information that could identify an individual worker. The hourly wage of each worker was the standard salary of university students in their country.

11. Acknowledgement

This work was supported by JST CREST Grant Number JPMJCR20D2.

12. Bibliographical References

Emily M Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198. Association for Computational Linguistics.

Maria Nella Carminati. 2005. Processing reflexes of the feature hierarchy (person > number > gender) and implications for linguistic theory. *Lingua*, 115(3):259–285.

Jeroen van Craenenbroeck and Tanja Temmerman. 2019. *The Oxford Handbook of Ellipsis*. Oxford University Press.

Vittoria Dentella, Elliot Murphy, G Marcus, and Evelina Leivada. 2023. Testing AI performance on less frequent aspects of language reveals insensitivity to underlying meaning. *arXiv*, abs/2302.12313.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Stefanie Dipper, Michael Goetze, and Stavros Skopeteas. 2007. *Information structure in cross-linguistic corpora: Annotation guidelines for phonology, morphology, syntax, semantics and information structure*. Universität Potsdam.

Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998.

Riki Fujihara, Tatsuki Kuribayashi, Kaori Abe, Ryoko Tokuhisa, and Kentaro Inui. 2022. Topicalization in language models: A case study on Japanese. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 851–862. International Committee on Computational Linguistics.

Edward Gibson, Richard Futrell, Steven P Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. How efficiency shapes human language. *Trends Cogn. Sci.*, 23(5):389–407.

Barbara J Grosz, Aravind K Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Comput. Linguist.*, 21(2):203–225.

Michael Hahn, Dan Jurafsky, and Richard Futrell. 2020. Universals of word order reflect optimization of grammars for efficient communication. *Proc. Natl. Acad. Sci. U. S. A.*, 117(5):2347–2353.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. *arXiv*, abs/2305.13264.

Ramon Ferrer i Cancho and Ricard V Solé. 2003. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3):788–791.

Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2008. Zero-anaphora resolution by learning rich syntactic pattern features. *ACM Transactions on Asian Language Information Processing*, 6(4).

Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. 2007. Annotating a Japanese text corpus with predicate-argument and coreference relations. In *Proceedings of the Linguistic Annotation Workshop*, pages 132–139. Association for Computational Linguistics.

T. Jaeger and Roger Levy. 2006. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. 2023. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations. *arXiv*, abs/2303.18027.

Nihongo Kijutsu Bunpō Kenkyūkai. 2008. Gendai nihongo bunpō vol. 5 [toritate, topic].

Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. 2015. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102.

Ryuto Konno, Shun Kiyono, Yuichiroh Matsubayashi, Hiroki Ouchi, and Kentaro Inui. 2021. Pseudo zero pronoun resolution improves zero anaphora resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3790–3806. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Discourse probing of pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3849–3864. Association for Computational Linguistics.

K Krippendorff. 2004. *Content analysis: An introduction to its methodology*. Sage publications.

Hanjung Lee. 2006. Parallel optimization in case systems: Evidence from case ellipsis in korean. *J. East Asian Ling.*, 15(1):69–96.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written japanese. *Language resources and evaluation*, 48(2):345–371.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the Uniform Information Density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980. Association for Computational Linguistics.

Jason Merchant et al. 2001. *The syntax of silence: Sluicing, islands, and the theory of ellipsis*, volume 1. Oxford University Press on Demand.

James A. Michaelov and Benjamin K. Bergen. 2022. Do language models make human-like predictions about the coreferents of Italian anaphoric zero pronouns? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1–14, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

OpenAI. 2023. GPT-4 technical report.

Steven T Piantadosi, Harry Tily, and Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

Massimo Poesio. 2010. Computational models of anaphora resolution: A survey.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Ryohei Sasano and Sadao Kurohashi. 2011. A discriminative approach to Japanese zero anaphora resolution with large-scale lexicalized case frames. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 758–766. Asian Federation of Natural Language Processing.

Lisa Schäfer, Robin Lemke, Heiner Drenhaus, and Ingo Reich. 2021. The role of UID for the usage of verb phrase ellipsis: Psycholinguistic evidence from length and context effects. *Front. Psychol.*, 12:661087.

Sebastian Schuster and Tal Linzen. 2022. When a sentence does not introduce a discourse entity, transformer-based models still sometimes refer to it. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 969–982. Association for Computational Linguistics.

Tomohide Shibata and Sadao Kurohashi. 2018. Entity-centric joint modeling of Japanese coreference resolution and predicate argument structure analysis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 579–589. Association for Computational Linguistics.

Koustuv Sinha, Jon Gauthier, Aaron Mueller, Kanishka Misra, Keren Fuentes, Roger Levy, and Adina Williams. 2023. Language model acceptability judgements are not always robust to context. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6043–6063, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv*, abs/2307.09288.

Natsuko Tsujimura. 2013. *An introduction to Japanese linguistics*. John Wiley & Sons.

Shiva Upadhye, Leon Bergen, and Andrew Kehler. 2020. Predicting reference: What do language models learn about discourse models? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 977–982. Association for Computational Linguistics.

Marilyn Walker, Masayo Iida, and Sharon Cote. 1994. Japanese discourse and the process of centering. *Computational Linguistics*, 20(2):193–231.

Longyue Wang, Zhaopeng Tu, Shuming Shi, Tong Zhang, Yvette Graham, and Qun Liu. 2018. Translating pro-drop languages with reconstruction models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4937–4945.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Yang Xu, Emmy Liu, and Terry Regier. 2020. Numeral systems across languages support efficient communication: From approximate numerosity to recursion. *Open Mind (Camb)*, 4:57–70.

		Omit		Insert	
Split	#Args.	Hard	Soft	Soft	Hard
Train	1,459	29.5%	15.2%	11.4%	43.9%
Valid	456	31.6%	8.1%	14.3%	46.1%
Test	458	31.6%	11.8%	9.6%	47.1%
All	2,373	30.3%	13.2%	11.6%	44.9%

Table 7: Label distribution of the dataset.

		Factors				
Archtecture	Params	1	2	4*a*	4*b*	5
BERT_S	batch size	8	8	16	8	8
	learning rate	5e-5	5e-7	3e-5	2e-5	3e-5
BERT_L	batch size	16	8	16	16	16
	learning rate	2e-5	5e-6	2e-5	5e-5	2e-5

Table 8: Batch sizes and learning rates for §6.

Appendix

A. Sampling Arguments from Corpus

In this study, we focused on arguments corresponding to `NOM`, `DAT`, and `ACC` cases among the arguments appearing in the source corpus. We also excluded the predicates that primarily serve a functional or grammatical role rather than conveying specific semantic content such as “aru” (be), “naru” (become), and “yaru” (do) in advance.

The original corpus sometimes contained multiple predicates within a single sentence. In cases wherein a sentence had multiple predicates, the predicate-argument pair to be annotated was determined randomly to ensure that each annotated sentence had only one target argument.

B. Label Distribution of Dataset

The label distribution of the dataset used in the experiment in §5 is shown in Table 7.

C. Hyperparameter and Model Settings

All BERT-based classification models were trained on NVIDIA RTX A6000. Early stopping was applied based on the loss of training data. All other parameters not specified in this section followed the default values of the `TrainingArguments` class in the Hugging Face Transformers library.

Settings for §5: For BERT-base-japanese, a batch size of 16 and a learning rate of 3e-05 were used, whereas for BERT-large-japanese, a batch size of 8 and a learning rate of 5e-05 were used.

Settings for §6: Table 8 shows the hyperparameters for BERTs. Eventually, we built five factor-level classifiers in total.

D. Finetuning Procedure in §6

Regarding the BERT-based models, we finetuned them to solve the classification tasks.

Dataset: In our human annotation process using the decision tree, it is not obligatory for each instance of annotation to be responded to all questions at the factor level. Moreover, each worker might follow a distinct decision-making path for the same annotation instance. Therefore, the number of answers to each factor-level question may vary.

Consequently, for each set of responses gathered for each factor, where the number of collected responses, denoted by n, lies in the range of 0 to 5, we treated it as a binary classification instance only under the condition that the majority label within these n responses can be determined; this approach resulted in a reduction of the dataset size as depict in Table 4. Eventually, a very small number of samples remained for Factor 3, and therefore, Factor 3 was excluded from the experiment.

Evaluation: In order to align the evaluation of BERT with that of human workers, which was measured by the average macro-F_1 score of five workers (§4.2), we computed the average F_1 score of five fine-tuned models.