JBLiMP: Japanese Benchmark of Linguistic Minimal Pairs

Taiga Someya and **Yohei Oseki**
The University of Tokyo
{taiga98-0809,oseki}@g.ecc.u-tokyo.ac.jp

Abstract

In this paper, we introduce **JBLiMP** (Japanese Benchmark of Linguistic Minimal Pairs), a novel dataset for targeted syntactic evaluations of language models in Japanese. JBLiMP consists of 331 minimal pairs, which are created based on acceptability judgments extracted from journal articles in theoretical linguistics. These minimal pairs are grouped into 11 categories, each covering a different linguistic phenomenon. JBLiMP is unique in that it combines two important features independently observed in existing datasets: (i) coverage of complex linguistic phenomena (cf. CoLA) and (ii) presentation of sentences as minimal pairs (cf. BLiMP). In addition, JBLiMP is the first dataset for targeted syntactic evaluations of language models in Japanese, thus allowing the comparison of syntactic knowledge of language models across different languages. We then evaluate the syntactic knowledge of several language models on JBLiMP: GPT-2, LSTM, and *n*-gram language models. The results demonstrated that all the architectures achieved comparable overall accuracies around 75%. Error analyses by linguistic phenomenon further revealed that these language models successfully captured local dependencies like nominal structures, but not long-distance dependencies such as verbal agreement and binding.

1 Introduction

The past few years have seen a remarkable success of neural language models, and some language models based on Transformer (Vaswani et al., 2017) have achieved the state-of-the-art performance in various natural language processing (NLP) tasks (Wang et al., 2018, 2019). In fact, recent neural language models are extremely successful in solving a variety of downstream tasks, but it remains to be understood how well these neural language models understand the syntax of natural languages. In order to address this question, some studies investigated the syntactic knowledge of language models with a specially designed dataset for targeted syntactic evaluations (e.g., Linzen et al., 2016; Marvin and Linzen, 2018; Wilcox et al., 2018; Gulordava et al., 2018; Futrell et al., 2019; Chaves, 2020). However, most of these studies have focused on English and other European languages, and only few studies extended this investigation to non-European languages (Gulordava et al., 2018; Ravfogel et al., 2018). Importantly for the purpose here, even fewer studies have dealt with a wide variety of linguistic phenomena in non-English languages (Xiang et al., 2021; Trotta et al., 2021).

In this paper, we introduce **JBLiMP** (Japanese Benchmark of Linguistic Minimal Pairs), a novel dataset for targeted syntactic evaluations of language models in Japanese.[1] JBLiMP consists of 331 minimal pairs, which are created based on acceptability judgments extracted from journal articles in theoretical linguistics. These minimal pairs are grouped into 11 categories, each covering a different linguistic phenomenon. JBLiMP is unique in that it successfully combines two important features independently observed in existing datasets: (i) coverage of complex linguistic phenomena (cf. CoLA; Warstadt et al., 2019) and (ii) presentation of sentences as minimal pairs (cf. BLiMP; Warstadt et al., 2020). We evaluate the syntactic knowledge of several language models on JBLiMP: GPT-2 (Radford et al., 2019), LSTM (Hochreiter and Schmidhuber, 1997) and *n*-gram language models. The results demonstrated that all the architectures achieved comparable overall accuracies around 75%. Error analyses by linguistic phenomenon further revealed that these language models successfully captured local dependencies like nominal structures, but not long-distance dependencies such as verbal agreement and binding.

[1] JBLiMP is available at https://github.com/osekilab/JBLiMP.

Findings of the Association for Computational Linguistics: EACL 2023, pages 1581–1594
May 2-6, 2023

Language	Linguistic Phenomenon			
	Subject-verb agreement	Filler-gap	Anaphor/binding	Argument structure
English	Linzen et al. (2016); Gulordava et al. (2018); Marvin and Linzen (2018); Warstadt et al. (2019)	Wilcox et al. (2018); Futrell et al. (2019); Chaves (2020); Da Costa and Chaves (2020); Warstadt et al. (2019)	Marvin and Linzen (2018); Warstadt et al. (2019); Futrell et al. (2019)	Warstadt et al. (2019); Kann et al. (2019); Chowdhury and Zamparelli (2019)
French	Gulordava et al. (2018); Mueller et al. (2020); An et al. (2019)			
Italian	Gulordava et al. (2018); Mueller et al. (2020); Trotta et al. (2021)	Trotta et al. (2021)	Trotta et al. (2021)	
Russian	Gulordava et al. (2018); Mueller et al. (2020)			
German	Mueller et al. (2020)			
Basque	Ravfogel et al. (2018)			
Hebrew	Gulordava et al. (2018); Mueller et al. (2020)			
Chinese	Xiang et al. (2021)	Xiang et al. (2021)	Xiang et al. (2021)	Xiang et al. (2021)
Japanese	This work			

Table 1: Related work organized by language and linguistic phenomenon

2 Related Work

Evaluation of language models has been mainly performed by computing metrics such as perplexity. This gives us an objective standard of the performance of language models, but doesn't provide insight into their performance on specific downstream tasks. While recent large-scale benchmarks like GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) are informative in this respect, many recent studies have sought to provide evidence that language models have learned the syntax of natural languages. In a pioneering work by Linzen et al. (2016), minimal pairs were employed to investigate whether language models are sensitive to subject-verb agreement in English. For instance, they tested whether language models assign a higher probability to *are* than *is* in (1).

(1) a. The keys to the cabinet are on the table.

b. *The keys to the cabinet is on the table.

Their results suggested that LSTM language models are fairly sensitive to English subject-verb agreement. However, this and related studies (e.g., Marvin and Linzen, 2018; Futrell et al., 2019) only covered a limited range of linguistic phenomena like subject-verb agreement.

In order to tackle this problem, more recent studies have introduced large-scale datasets for comprehensive syntactic evaluations (Warstadt et al., 2019, 2020). One such dataset is CoLA (Corpus of Linguistic Acceptability; Warstadt et al., 2019), which consists of 10,000 sentences with binary acceptability labels extracted from linguistics journals and textbooks. CoLA is incorporated into GLUE benchmark (Wang et al., 2018) and has been used to evaluate the sensitivity of language models to the syntax of natural languages. While CoLA has enabled the comprehensive syntactic evaluations of language models, this dataset is not without its limitation, as noted by Warstadt et al. (2019) themselves. The limitation lies in the need to train a supervised classifier on CoLA for evaluation. In short, CoLA is designed for binary classification of acceptability judgements, but there is no clear way to map the probability of the sentence estimated by language models to binary acceptability judgements. Unfortunately, "the use of supervision prevents making strong conclusions about the sentence encoding component, since it is not possible to distinguish what the encoder knows from what is learned through supervised training on acceptability data" (Warstadt et al., 2019).

Dataset	Linguistics Journal	Minimal Pairs
CoLA (Warstadt et al., 2019)	✓	
ItaCoLA (Trotta et al., 2021)	✓	
BLiMP (Warstadt et al., 2020)		✓
CLiMP (Xiang et al., 2021)		✓
JBLiMP	✓	✓

Table 2: Comparison of JBLiMP and other existing datasets

With this limitation in mind, BLiMP (Benchmark of Linguistic Minimal Pairs; Warstadt et al.,

2020) is developed, which includes 67 datasets automatically generated from grammar templates created by linguists. These 67 datasets are grouped into 12 categories based on linguistic phenomenon, each containing 1,000 minimal pairs. Note that each pair has one acceptable sentence and one unacceptable sentence. Importantly, this dataset has overcome an aforementioned problem, because sentences are not presented as binary classification problems, but as minimal pairs: the evaluation can be readily performed by comparing the probabilities of an acceptable sentence and an unacceptable sentence. Nevertheless, BLiMP also has its limitation to overcome. Namely, since minimal pairs are automatically generated with template grammars and vocabularies, BLiMP doesn't necessarily cover complex and important linguistic phenomena (cf. Class Ⅲ judgement, see Marantz 2005; Linzen and Oseki 2018), compared to those datasets which are created by extracting sentences from linguistics journals.

There is also a general problem with the datasets for targeted syntactic evaluations of language models as a whole: imbalance in target languages and linguistic phenomena (cf. Table 1). In fact, most of the existing datasets have focused on English. Although some studies have extended the scope of their research to other languages (Gulordava et al., 2018; An et al., 2019; Ravfogel et al., 2018; Mueller et al., 2020), only few studies have covered a wide range of syntactic phenomena and focused on languages other than English (Xiang et al., 2021; Trotta et al., 2021).

3 JBLiMP

In order to overcome all the limitations mentioned above, we introduce JBLiMP (Japanese Benchmark of Linguistic Minimal Pairs), a novel dataset for targeted syntactic evaluations of language models in Japanese. JBLiMP is unique in that it successfully combines two important features independently observed in existing datasets (Table 2): (i) coverage of complex linguistic phenomena (cf. CoLA; Warstadt et al., 2019) and (ii) presentation of sentences as minimal pairs (cf. BLiMP; Warstadt et al., 2020). In addition, JBLiMP is the first dataset for targeted syntactic evaluations of language models in Japanese, thus alleviating the imbalance in target languages and allowing the comparison of syntactic knowledge of language models across different languages.

3.1 Data Collection

JBLiMP consists of acceptability judgments from journal articles on Japanese syntax published in JEAL (Journal of East Asian Linguistics): one of the prestigious journals in theoretical linguistics. Specifically, we examined all the articles published in JEAL between 2006 and 2015 (133 papers in total), and extracted 2,323 acceptability judgments from 28 papers on Japanese syntax (cf. Table 3). Acceptability judgments include sentences in appendices and footnotes, but not sentences presented for analyses of syntactic structures (e.g. sentences with brackets to show their syntactic structures).

3.2 Categorization by linguistic phenomenon

We categorized the extracted sentences into different groups to enable detailed analyses of results by linguistic phenomenon. The categorization mostly followed that of BLiMP (Warstadt et al., 2020) and was conducted at three levels of granularity: `type`, `phenomenon` and `paradigm`.

3.2.1 Type

First, the extracted sentences were categorized based on the type of acceptability judgements and how those sentences were presented in the articles. This level of categorization has 8 different types. These categories are mutually exclusive, meaning that no further typing is done for sentences in footnotes or appendices.

Acceptability: acceptability judgements that do not depend on a specific context or interpretation.

Interpretation: acceptability judgements that depend on a specific context or interpretation.

Coreference: acceptability judgements that depend on a specfic interpretation of coreference.

Lexical: acceptability judgements that depend on a specific lexical item.

Footnote: acceptability judgements presented in footnotes.

Appendix: acceptability judgements presented in appendices.

Repeat: acceptability judgements repeated by the authors.

Variation: acceptability judgements that only differ in unimportant elements for theory construction. For example, (2b) below is categorized into

`variation` because the difference between *da* 'is' and *desu*, a polite form of 'is', is not relevant for theory construction.

(2) a. Taro-ga atta no-wa Hanako-ni da
Taro-Nom saw that-Top Hanako-Dat is
'It was Hanako that Taroo saw.'

b. Taro-ga atta no-wa Hanako-ni desu
Taro-Nom saw that-Top Hanako-Dat is
'It was Hanako that Taroo saw.'

Source	# Sentences
Takahashi (2006)	60
Oshima (2006)	34
Tenny (2006)	70
Bobaljik and Wurmbrand (2007)	18
Ivana and Sakai (2007)	51
Kishimoto (2008)	254
Saito et al. (2008)	46
Takita (2009)	13
Hayashishita (2009)	73
Miyamoto (2009)	36
Tomioka (2009)	27
Asano and Ura (2010)	144
Watanabe (2010)	40
Grosu (2010)	43
Takahashi (2010)	77
Tsujioka (2011)	226
Abe (2011)	53
Takano (2011)	81
Kishimoto (2012)	120
Grosu and Landman (2012)	28
Kishida and Sato (2012)	98
Yoon (2013)	55
Sawada (2013)	81
Watanabe (2013)	118
Nishigauchi (2014)	115
Shimoyama (2014)	63
Sudo (2015)	184
Shibata (2015)	115
Total	2,323

Table 3: Number of extracted sentences by source

3.2.2 Phenomenon

Second, the extracted sentences were further categorized based on linguistic phenomena. Phenomenon basically corresponds to that in BLiMP, but some modifications were applied to make the categorization more suitable for Japanese.

Argument Structure: acceptability judgements based on the order of arguments and case marking.

(3) a. Taroo-ga Hanako-**ni** au.
Taroo-Nom Hanako-Dat see.
'Taroo sees Hanako.'

b. *Taroo-ga Hanako-**o** au.
Taroo-Nom Hanako-Acc see.
'Taroo sees Hanako.'

Binding: acceptability judgements based on the binding of noun phrases. For instance, this includes the coreference resolution of anaphors.

(4) a. Hazimete **soitu**-ni atta
for-the-first-time him-Dat saw
hito-ga **Taroo**-o kenasita
person-Nom Taroo-Acc criticized
'The person who saw him for the first time criticized Taroo.'

b. *Hazimete **soitu**-ni atta
for-the-first-time him-Dat saw
hito-ga **daremo**-o kenasita
person-Nom everyone-Acc criticized
'The person who saw him for the first time criticized everyone.'

Control/Raising: acceptability judgements based on predicates that are categorized as control or raising.

(5) a. **Taroo**-ga korobi sokoneta.
tumbler.doll-Nom tumble failed.
'Taroo failed to tumble.'

b. ***Daruma**-ga korobi sokoneta.
tumbler.doll-Nom tumble failed.
'Tumbler doll failed to tumble.'

Ellipsis: acceptability judgements based on the possibility of omitting elements in the sentences. For instance, this includes nominal and verbal ellipsis.

(6) a. Hare-no-hi-ha yoi ga
clear-NO-day-Top good though
ame-no-hi-ha otikomu.
rain-NO-day-Top feel.depressed.
'Clear days are OK, but I feel depressed on rainy days.'

b. *Hare-no-hi-ha yoi ga
clear-NO-day-Top good though
ame-no-ha otikomu.
rain-NO-Top feel.depressed.
'Clear days are OK, but I feel depressed on rainy days.'

Filler-gap: acceptability judgements based on the dependency between the moved element and the gap. For instance, this includes wh-movements and cleft sentences.

(7) a. **Nani-o** **daremo** yom-ana-katta-no.
What-Acc anyone read-neg-past-Q.
'What did no one read?'

b. *__Daremo__ **nani-o** yom-ana-katta-no.
anyone What-Acc read-neg-past-Q.
'What did no one read?'

Island effects: acceptability judgements based on the restrictions on filler-gap dependencies such as wh-movements.

(8) a. Taroo-ha Hanako-ga naze kare-no
Taroo-Top Hanako-Nom why he-Gen
tegami-o suteta **to** omotteiru no.
letter-Acc discarded C think Q
'Why is Taro angry because Hanako discarded his letters?'

b. *Taroo-ha Hanako-ga naze kare-no
Taroo-Top Hanako-Nom why he-Gen
tegami-o suteta **kara** okotteiru no.
letter-Acc discarded because be.angry Q
'Why is Taro angry because Hanako discarded his letters?'

Morphology: acceptability judgements based on the morphology. BLiMP has `irregular forms` category for the conjugation of past tenses, but we adopted this category instead to incorporate minimal pairs on morphology in general.

(9) a. sore-wa keesoku
that-Top measurement
kanoo-**na** ryuusi-da
possibility-Cop.Adnom particle-Cop
'That is a measurable particle'

b. *sore-wa keesoku
that-Top measurement
kanoo-**da** ryuusi-da
possibility-Cop.Fin particle-Cop
'That is a measurable particle'

Nominal Structure: acceptability judgements based on the internal structure of noun phrases. BLiMP has `determiner-noun agreement` category, but we adopted this category instead, because Japanese doesn't have explicit determiner-noun agreements.

(10) a. Watashi-ga kinoo **mita** **hito-wa**
I-Nom yesterday saw person-Top
suteki datta
beautiful was
'The person I saw yesterday was beautiful'

b. *Watashi-ga kinoo **mita** **no**
I-Nom yesterday saw *no*
hito-wa suteki datta
person-Top beautiful was
'The person I saw yesterday was beautiful'

NPI Licensing: acceptability judgements based on the restrictions on where negative polarity items (NPIs) can appear. For instance, NPIs include *nan-imo*, a Japanese counterpart of 'any'.

(11) a. *John-ga moshi **nani-ka**
John-Nom if something
nusun-dara, taihos-areru daroo.
steal-COND arrest-PASS be.will
'If John steals anything, he will be arrested.'

b. *John-ga moshi **nani-mo**
John-Nom if what-MO
nusun-dara, taihos-areru daroo.
steal-COND arrest-PASS be.will
'If John steals anything, he will be arrested.'

Quantifiers: acceptability judgements based on the distribution of quantifiers such as floating quantifiers.

(12) a. Taroo-ga tomodati-ni **huta-ri** CD-o
Taroo-Nom friend-Dat 2-CL CD-Acc
okutta.
sent.
'Taro sent two friends a package.'

b. *Taroo-ga CD-o tomodati-ni **huta-ri**
Taroo-Nom CD-Acc friend-Dat 2-CL
okutta.
sent.
'Taro sent two friends a package.'

Verbal Agreement: acceptability judgements based on the dependency between subjects and verbs. Japanese doesn't have the same kind of subject-verb agreement as in English. Instead, this includes the linguistic phenomena such as subject honorification where the social status of subjects are reflected in the morphology of verbs.

(13) a. **Ito-sensei-ga** Mary-o
Ito-teacher-Nom Mary-Acc
o-home-**ni-nat**-ta
Hon-praise-Lv-Past
'Prof. Ito praised Mary.'

b. *__Watashi-ga__ Mary-o
I-Nom Mary-Acc
o-home-**ni-nat**-ta
Hon-praise-Lv-Past
'I praised Mary.'

3.2.3 Paradigm

Finally, the extracted sentences are further categorized into 39 more fine-grained types named `paradigm`. Paradigm also corresponds to that in BLiMP and is basically sub-categorization of `phenomenon`.

3.3 Minimal pairs

For direct evaluation of language models through the probabilities assigned by these language models, we created minimal pairs using the sentences categorized above. First, we selected all the sentences that satisfy the following conditions:

- The sentences are presented as unacceptable examples (marked with '?' or '*', for example). Exceptions are those sentences that are presented as acceptable examples, but marked with '?' or '%'.
- Type is not one of `variation`, `repeat`, `footnote` or `appendix`.
- The sentences are grouped into one of the 11 `phenomena`.

We deduplicated the selected unacceptable examples and removed those unacceptable examples whose (un)acceptability depends on the context. Second, since we are concerned with sentence-level acceptability judgements, we augmented incomplete sentences, replacing, for example, (14a) with (14b).

Phenomenon	# Minimal pairs
ARGUMENT STRUCTURE	140
VERBAL AGREEMENT	61
MORPHOLOGY	35
NOMINAL STRUCTURE	23
ELLIPSIS	19
QUANTIFIERS	14
BINDING	13
ISLAND EFFECTS	11
FILLER-GAP	9
NPI LICENSING	4
CONTROL/RAISING	2
Total	331

Table 4: Number of minimal pairs by phenomenon

(14) a. *Sono futari gakusei
that two-CL student
'those two students'

b. *Taroo-ha sono futari gakusei-ni atta
Taroo-Top that two-CL student-Dat saw
'Taroo saw those two students.'

Finally, we created minimal pairs based on the selected unacceptable sentences, on the assumption that all the unacceptable sentences for theory construction generally have their acceptable counterparts to demonstrate the contrasts in acceptability (Sprouse et al., 2013). Specifically, for each unacceptable example, we either found an appropriate acceptable example from the extracted sentences, or created a corresponding acceptable example. When creating acceptable sentences, we read the relevant papers to understand the authors' intent to present the corresponding unacceptable sentences.

3.4 Data Validation

In order to validate the quality of minimal pairs in JBLiMP, we conducted an acceptability judgement experiment with Lancers, a Japanese crowdsourcing platform.[2] For each minimal pair, 15 native speakers of Japanese completed a forced-choice task which reflects the evaluation procedure of language models. Specifically, annotators are asked to select the more grammatical of the two sentences, following the experimental design in Sprouse et al. (2013). To minimize the burden on annotators, we split 367 minimal pairs into 16 different groups:

[2] https://www.lancers.jp

15 groups of 23 minimal pairs and 1 group of 22 minimal pairs. Each annotator completes 22 or 23 acceptability judgements and is compensated 150 yen ($\simeq$ \$ 1.2). The order of minimal pairs and the vertical order of acceptable and unacceptable examples within a minimal pair was randomized. Majority vote is taken to determine human-annotated acceptable sentences. For each minimal pair, if the annotation of JBLiMP and the majority vote of human annotations do not match, that minimal pair is removed from JBLiMP. In this way, 36 minimal pairs were removed, resulting in 331 minimal pairs in total (Table 4). In addition, we calculated human baseline accuracy, dividing the number of human annotations that match JBLiMP's judgements by the total number of annotations. As a result, the human baseline accuracy was 90.90% as reported in Table 5.

4 Experiment

4.1 Models

In this paper, we evaluate language models trained by Kuribayashi et al. (2021) with JBLiMP.

GPT-2 GPT-2 (Radford et al., 2019) is one of the large-scale language models based on Transformer architectures (Vaswani et al., 2017). We evaluate two different sizes of GPT-2 models (`Trans-LG`, `Trans-SM`). `Trans-LG` has 24 layers, 16 attention heads, and 1024 embedding dimensions. `Trans-SM` has 8 layers, 6 attention heads, and 384 embedding dimensions.

LSTM LSTM (Hochreiter and Schmidhuber, 1997) is a language model based on RNN architectures (Elman, 1990), which is known to achieve a better language modeling performance than vanilla RNN language models (Sundermeyer et al., 2012). We evaluate a 2-layer LSTM language model with 1024 hidden layer dimensions and 400 embedding dimensions.

***n*-gram** We also evaluate a 5-gram language model as a baseline. This model is implemented by KenLM (Heafield et al., 2013).

Training settings (Kuribayashi et al., 2021) Training data was approximately 5M sentences extracted from news and Japanese Wikipedia. Each sentence in training data was first segmented by MeCab and then segmented into subwords by BPE (Byte-Pair Encoding).[3] All the neural language models (`Trans-LG`, `Trans-SM` and `LSTM`) were trained with the data of three different sizes: LG (full training data), MD (1/10 training data), SM (1/100 training data). These language models were trained with three different random seeds, and saved at four different points in the training: 100, 1,000, 10,000, 100,000 training steps.

4.2 Evaluation metrics

The probability assigned to a sentence can be mapped into acceptability judgements in multiple ways (Lau et al., 2017). In this work, we employ SLOR (Lau et al., 2017) as a mapping function, which mitigates the confounding effects of sentence lengths and lexical frequencies. SLOR score for a sentence X is defined as follows:

$$SLOR(X) = \frac{\log p_m(X) - \log p_u(X)}{|X|}$$

where $p_m(X)$ is the probability of a sentence given by a language model, and $p_u(X) = \prod_{w \in X} p_u(w)$ is the unigram probability of a sentence. Unigram probabilities are estimated via maximum likelihood estimation for each subword in the training corpus. For each minimal pair, we examine whether language models assign a higher probability/acceptability to an acceptable sentence than an unacceptable one.

5 Results and Discussion

5.1 Overall accuracy

Overall accuracy of each language model on JBLiMP is reported in Table 5. While Trans-LG achieves the best accuracy of 77.95%, all the models notably achieve the comparable accuracy and fall short of human accuracy by a wide margin, which may suggest that language models can't necessarily recognize complex linguistic phenomena.

5.2 Accuracy by linguistic phenomenon

For each language model, we calculate accuracy by linguistic phenomenon on JBLiMP, as reported in Table 5. Analysis by linguistic phenomenon reveals that the performance of language models drastically differs depending on linguistic phenomenon. Language models achieve a relatively high accuracy on

[3] Vocabulary size was set to 100,000 and character coverage to 0.9995. Implementation by SentencePiece (Kudo and Richardson, 2018) was employed.

Model	Overall	Argument Structure	Verbal Agr.	Morph.	Nominal Structure	Ellipsis	Quant.	Binding	Island Effects	Filer Gap	NPI Licensing	Control Raising
Trans-LG	77.95	89.05	53.55	82.86	**95.65**	85.96	73.81	58.97	75.76	55.56	50.00	16.67
Trans-SM	76.54	89.05	44.26	82.86	**97.10**	89.47	71.43	46.15	84.85	55.56	75.00	0.00
LSTM	75.73	86.67	46.99	83.81	**95.65**	91.23	66.67	41.03	87.88	44.44	66.67	50.00
5-gram	74.02	78.57	57.38	82.86	86.96	**89.47**	78.57	53.85	72.73	66.67	50.00	0.00
Human	90.90	92.19	89.62	94.86	**97.68**	87.37	85.71	82.05	92.12	78.52	90.00	70.00
Model Ave.	76.06	85.76	50.55	83.10	**93.84**	89.03	72.62	50.00	80.31	55.56	60.42	16.67

Table 5: Accuracy of each language model and human by phenomenon. Accuracy is averaged over 3 different random seeds except 5-gram and human. All the language models are trained for 100,000 steps on full training corpus (LG). The number in bold indicates the best score within a model, while the number with underscore indicates the worst score.

phenomena like `nominal structure`. This phenomenon includes minimal pairs with relatively local dependencies, as exemplified in (15).

(15) Nominal structure

a. Watashi-ga kinoo **mita** **hito-wa** suteki datta
I-Nom yesterday saw person-Top beautiful was
'The person I saw yesterday was beautiful'

b. *Watashi-ga kinoo **mita** **no** **hito-wa** suteki datta
I-Nom yesterday saw *no* person-Top beautiful was
'The person I saw yesterday was beautiful'

In sharp contrast, language models suffer a sharp drop in accuracy on linguistic phenomena such as `verbal agreement` and `binding`. (Here, `control/raising` is taken out of consideration because its data size is small compared to the other phenomena.) These phenomena generally involve relatively long dependencies: `verbal agreement` involves dependency between the subject and the verb of the sentence as exemplified in (16), while `binding` involves dependency between anaphors and their antecedents as illustrated in (17).

(16) Verbal agreement

a. **Ito-sensei-ga** Mary-o **o**-home-**ni-nat**-ta
Ito-teacher-Nom Mary-Acc Hon-praise-Lv-Past
'Prof. Ito praised Mary.'

b. ***Watashi-ga** Mary-o **o**-home-**ni-nat**-ta
I-Nom Mary-Acc Hon-praise-Lv-Past
'I praised Mary.'

(17) Binding

a. Hazimete **soitu**-ni atta hito-ga **Taroo**-o kenasita
for-the-first-time him-Dat saw person-Nom Taroo-Acc criticized
'The person who saw him for the first time criticized Taroo.'

b. *Hazimete **soitu**-ni atta hito-ga **daremo**-o kenasita
for-the-first-time him-Dat saw person-Nom everyone-Acc criticized
'The person who saw him for the first time criticized everyone.'

Lower accuracy in these kinds of minimal pairs suggests that language models are less sensitive to long-distance dependencies. These results are compatible with the previous results that RNN-based language models cannot capture long-distance dependencies without explicit supervision (Linzen et al., 2016), but are not necessarily consistent with the results that Transformer-based language models can successfully capture long-distance dependencies (Goldberg, 2019).

5.3 Human confidence and model confidence

Figure 1 shows the relationship between model confidence and human confidence. Each model's confidence on a minimal pair is defined as the difference of the SLOR scores between the acceptable and unacceptable sentence: $SLOR(X_{pos}) - SLOR(X_{neg})$ where X_{pos} is an acceptable sentence and X_{neg} is an unacceptable sentence. Human confidence on a minimal pair is defined as the number of annotators who had the same annotation as the JBLiMP. While the language models are able to make predictions with relatively high confidence for sentences with high human confidence, the confidence of the language models is low for sentences with low human confidence, i.e., for which there

are fluctuations in acceptability judgments among humans. Furthermore, many of the language models have negative confidence for the sentences with low human confidence. These results may suggest that language models have successfully captured the gradience in human acceptability judgements, whose existence was suggested in Lau et al. (2017).

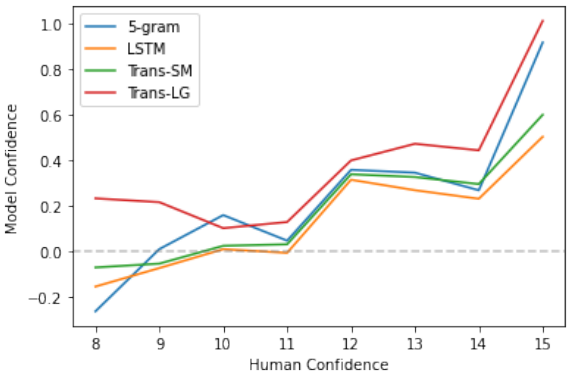

Figure 1: The relationship between model confidence and human confidence. All the neural language models are trained for 100,000 steps on full training corpus (LG).

5.4 Effects of perplexity on accuracy

We investigate the relationship between the perplexity, which is widely used as an evaluation metric of language models' performance, and the accuracy on JBLiMP for each language model.The perplexity is calculated on the validation data in Kuribayashi et al. (2021). Figure 2 shows the language models' accuracy on JBLiMP as a function of perplexity. In contrast to the results in Kuribayashi et al. (2021) that lower perplexity does not necessarily ensure better psychometric predictive power of language models, our results suggest that language models with lower perplexity will generally achieve better syntactic performance. Note incidentally that language models with particularly high perplexity ($> 3 \times 10^4$), represented as the points to the right of the black dashed line in Figure 2, are trained for more than 10,000 steps with relatively small data (SD or MD). These language models seem to be overfitted to the training data, and thus were taken out of consideration in this discussion.

6 Conclusion

In this paper, we introduced JBLiMP (Japanese Benchmark of Linguistic Minimal Pairs), a novel dataset for targeted syntactic evaluations of language models in Japanese. JBLiMP consists of

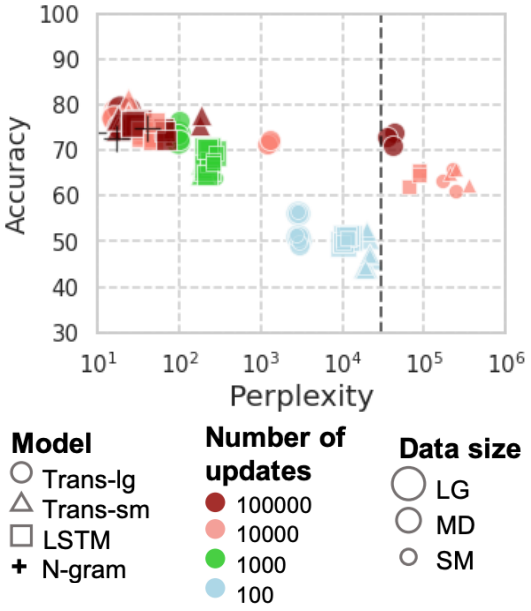

Figure 2: Models' accuracy on JBLiMP as a function of perplexity. The perplexity is calculated on the validation data in Kuribayashi et al. (2021). The vertical dashed line in black indicates the perplexity of 3×10^4.

331 minimal pairs, which are created based on acceptability judgments extracted from journal articles in theoretical linguistics. These minimal pairs are grouped into 11 categories, each covering a different linguistic phenomenon. JBLiMP is unique in that it successfully combines two important features independently observed in existing datasets: (i) coverage of complex linguistic phenomena (cf. CoLA) and (ii) presentation of sentences as minimal pairs (cf. BLiMP). In addition, JBLiMP is the first dataset for targeted syntactic evaluations of language models in Japanese, thus allowing the comparison of syntactic knowledge of language models across different languages. We then evaluated the syntactic knowledge of several language models: GPT-2, LSTM and *n*-gram language models. The results demonstrated that all the architectures achieved comparable overall accuracies around 75%. Error analyses by linguistic phenomenon further revealed that these language models successfully captured local dependencies like nominal structures, but not long-distance dependencies such as verbal agreement and binding. Finally, these detailed analyses of language models' knowledge on complex linguistic phenomena using minimal pairs are only possible with the unique design of JBLiMP. This paper will hopefully encourage the development of the datasets with JBLiMP's two important features in other languages.

Limitations

All the example sentences in JBLiMP were manually transcribed from linguistic journals. While this method of data collection has enabled it to cover complex linguistic phenomena, it also made it difficult to increase the size of the dataset. Additionally, the quantity of minimal pairs on a specific linguistic phenomenon is directly influenced by how often that phenomenon is discussed in linguistic journals, hence the imbalanced distribution of minimal pairs across different linguistic phenomena in JBLiMP. These problems could be overcome by collecting additional examples from linguists (if possible, the authors of the source linguistic journals in JBLiMP).

Acknowledgements

This work was supported by JST PRESTO Grant Number JPMJPR21C2, Japan. We are also grateful for the anonymous reviewers and area chairs for their detailed and helpful feedback.

References

Jun Abe. 2011. Real parasitic gaps in japanese. *J. East Asian Ling.*, 20(3):195–218.

Aixiu An, Peng Qian, Ethan Wilcox, and Roger Levy. 2019. Representation of constituents in neural language models: Coordination phrase as a case study. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2888–2899, Hong Kong, China. Association for Computational Linguistics.

Shin'ya Asano and Hiroyuki Ura. 2010. Mood and case: with special reference to genitive case conversion in kansai japanese. *J. East Asian Ling.*, 19(1):37–59.

Jonathan D Bobaljik and Susi Wurmbrand. 2007. Complex predicates, aspect, and anti-reconstruction. *J. East Asian Ling.*, 16(1):27–42.

Rui P Chaves. 2020. What don't RNN language models learn about Filler-Gap dependencies? *Proceedings of the Society for Computation in Linguistics*, 3(1):20–30.

Shammur Absar Chowdhury and Roberto Zamparelli. 2019. An LSTM adaptation study of (un)grammaticality. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 204–212, Florence, Italy. Association for Computational Linguistics.

Jillian Da Costa and Rui Chaves. 2020. Assessing the ability of transformer-based neural models to represent structurally unbounded dependencies. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 12–21, New York, New York. Association for Computational Linguistics.

Jeffrey L Elman. 1990. Finding structure in time. *Cogn. Sci.*, 14(2):179–211.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities.

Alexander Grosu. 2010. The status of the internally-headed relatives of Japanese/Korean within the typology of "definite" relatives. *J. East Asian Ling.*, 19(3):231–274.

Alexander Grosu and Fred Landman. 2012. A quantificational disclosure approach to japanese and korean internally headed relatives. *J. East Asian Ling.*, 21(2):159–196.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

J-R Hayashishita. 2009. Yori-Comparatives: A reply to beck et al. (2004). *J. East Asian Ling.*, 18(2):65–100.

Kenneth Heafield, Ivan Pouzyrevsky, J. Clark, and Philipp Koehn. 2013. Scalable modified kneser-ney language model estimation. In *ACL*.

S Hochreiter and J Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Adrian Ivana and Hiromu Sakai. 2007. Honorification and light verbs in japanese. *J. East Asian Ling.*, 16(3):171–191.

Katharina Kann, Alex Warstadt, Adina Williams, and Samuel R Bowman. 2019. Verb argument structure alternations in word and sentence embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 287–297.

Maki Kishida and Yosuke Sato. 2012. On the argument structure of zi-verbs in japanese: reply to tsujimura and aikawa (1999). *J. East Asian Ling.*, 21(2):197–218.

Hideki Kishimoto. 2008. Ditransitive idioms and argument structure. *J. East Asian Ling.*, 17(2):141–179.

Hideki Kishimoto. 2012. Subject honorification and the position of subjects in japanese. *J. East Asian Ling.*, 21(1):1–41.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower perplexity is not always Human-Like. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5203–5217, Online. Association for Computational Linguistics.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cogn. Sci.*, 41(5):1202–1241.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn Syntax-Sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Tal Linzen and Yohei Oseki. 2018. The reliability of acceptability judgments across languages. *Glossa: a journal of general linguistics*, 3(1).

A Marantz. 2005. Generative linguistics within the cognitive neuroscience of language. *The Linguistic Review*.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Yoichi Miyamoto. 2009. On the Nominal-Internal distributive interpretation in japanese. *J. East Asian Ling.*, 18(3):233–251.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Taisuke Nishigauchi. 2014. Reflexive binding: awareness and empathy from a syntactic point of view. *J. East Asian Ling.*, 23(2):157–206.

David Y Oshima. 2006. Adversity and Korean/Japanese passives: Constructional analogy. *J. East Asian Ling.*, 15(2):137–166.

A Radford, J Wu, R Child, D Luan, D Amodei, and others. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Shauli Ravfogel, Yoav Goldberg, and Francis Tyers. 2018. Can LSTM learn to capture agreement? the case of basque. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 98–107, Brussels, Belgium. Association for Computational Linguistics.

Mamoru Saito, T-H Jonah Lin, and Keiko Murasugi. 2008. N'-Ellipsis and the structure of noun phrases in chinese and japanese. *J. East Asian Ling.*, 17(3):247–271.

Osamu Sawada. 2013. The comparative morpheme in modern japanese: looking at the core from 'outside'. *J. East Asian Ling.*, 22(3):217–260.

Yoshiyuki Shibata. 2015. Negative structure and object movement in japanese. *J. East Asian Ling.*, 24(3):217–269.

Junko Shimoyama. 2014. The size of noun modifiers and degree quantifier movement. *J. East Asian Ling.*, 23(3):307–331.

J Sprouse, C T Schütze, and D Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001-2010. *Lingua*, 134:219–248.

Yasutada Sudo. 2015. Hidden nominal structures in japanese clausal comparatives. *J. East Asian Ling.*, 24(1):1–51.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.

Daiko Takahashi. 2006. Apparent parasitic gaps and null arguments in japanese. *J. East Asian Ling.*, 15(1):1–35.

Masahiko Takahashi. 2010. Case, phases, and Nominative/Accusative conversion in japanese. *J. East Asian Ling.*, 19(4):319–355.

Yuji Takano. 2011. Double complement unaccusatives in japanese: puzzles and implications. *J. East Asian Ling.*, 20(3):229–254.

Kensuke Takita. 2009. If chinese is Head-Initial, japanese cannot be. *J. East Asian Ling.*, 18(1):41–61.

Carol L Tenny. 2006. Evidentiality, experiencers, and the syntax of sentience in japanese. *J. East Asian Ling.*, 15(3):245–288.

S Tomioka. 2009. Why questions, presuppositions, and intervention effects. *J. East Asian Ling.*, 18(4):253–271.

Daniela Trotta, Raffaele Guarasci, Elisa Leonardelli, and Sara Tonelli. 2021. Monolingual and cross-lingual acceptability judgments with the Italian CoLA corpus. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2929–2940, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Takae Tsujioka. 2011. Idioms, mixed marking in nominalization, and the basegeneration hypothesis for ditransitives in japanese. *J. East Asian Ling.*, 20(2):117–143.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł Ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

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.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Akira Watanabe. 2010. Notes on nominal ellipsis and the nature of no and classifiers in japanese. *J. East Asian Ling.*, 19(1):61–74.

Akira Watanabe. 2013. Non-neutral interpretation of adjectives under measure phrase modification. *J. East Asian Ling.*, 22(3):261–301.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about Filler–Gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online. Association for Computational Linguistics.

Suwon Yoon. 2013. Parametric variation in subordinate evaluative negation: Korean/Japanese versus others. *J. East Asian Ling.*, 22(2):133–166.

A Examples of minimal pairs in JBLiMP

Phenomenon	Paradigm	Unacceptable	Acceptable
Argument structure	case	太郎が その 本に 読んだ。 Taroo-Nom that book-Dat read	太郎が その 本を 読んだ。 Taroo-Nom that book-Dat read
	passive	家が 大工に 建てさせられた。 House-Nom carpenter-Dat build-cause-pass-past	大工が 家を 建てさせられた。 carpenter-Dat house-Dat build-cause-pass-past
	scrambling	最も 太郎が 面白かった 人を 取材した。 most Taroo-Nom funny-Cop-Past person-Acc interviewed	太郎が 最も 面白かった 人を 取材した。 Taroo-Nom most funny-Cop-Past person-Acc interviewed
	animacy	ジョンには お金が 居る。 John-Dat-Top money-Nom have.Animate-Pres	ジョンには 兄弟が 居る。 John-Dat-Top brother-Nom have.Animate-Pres
	aspect	太郎が プールで 1 時間で 泳いだ。 Taroo-Nom pool-in 1-hour-in swim-Past	太郎が プールで 1 時間 泳いだ。 Taroo-Nom pool-in 1-hour swim-Past
	internal argument	ジョンが 息子を 自殺した。 John-Nom son-Acc self-killing-do-Past	ジョンが 息子を 自慢した。 John-Nom son-Acc self-boast-do-Past
Verbal agreement	subject honorification	健が 山田先生に お会いになった。 Ken-Nom Yamada-teacher-Dat met.Hon	健に 山田先生が お会いになった。 Ken-Dat Yamada-teacher-Nom met.Hon
	person constraint	あなたは 寒い です。 You-Top cold Cop-Pres	私は 寒い です。 I-Top cold Cop-Pres
Binding	weak crossover	初めて そいつに 会う 人が 貶す のは 誰を です か？ for-the-first-time him-Dat see person-Nom criticize that-Top who-Acc Cop Q	初めて 会う 人が 貶す のは 誰を です か？ for-the-first-time see person-Nom criticize that-Top who-Acc Cop Q
	variable binding	花子が そいつが 書いた 論文を 修正させた のは 誰に です か？ Hanako-Nom him-Nom wrote paper-Acc revise-made NL-Top who-Dat Cop Q	花子が 誰に そいつが 書いた 論文を 修正させた の です か？ Hanako-Nom who-Dat him-Nom wrote paper-Acc revise-made NL Cop Q
	anaphor	自分$_i$の 先生$_i$には 学生が わかる。 self-Gen teacher-Dat-Top student-Nom recognize-Pres	先生$_i$には 自分$_i$の 学生が わかる。 teacher-Dat-Top self-Gen student-Nom recognize-Pres
	reciprocal	お互い$_i$の 母親から 彼ら$_i$に そのことを 伝えた。 each.other-Gen mother-from they-Dat that-fact-Acc tell-Past	彼ら$_i$に お互い$_i$の 母親から そのことを 伝えた。 they-Dat each.other-Gen mother-from that-fact-Acc tell-Past
Ellipsis	nominal ellipsis	晴れ の 日は 良い が、 雨 の は 落ち込む。 clear *no* day-Top good though rain *no* Top feel-depressed	晴れ の 日は 良い が、 雨 の 日は 落ち込む。 clear *no* day-Top good though rain *no* day-Top feel-depressed
	adjunct ellipsis	太郎が その理由で 解雇された 後、 花子も 解雇された。 Taroo-Nom that-reason-for was-fired after Hanako-aslo was-fired	太郎が その理由で 解雇された 後、 花子も その理由で 解雇された。 Taroo-Nom that-reason-for was-fired after Hanako-also that-reason-for was-fired
	parasitic-gap	初めて 会う 人が 誰を けなしますか？ for-the-first-time see person-Nom who-Acc criticize-Q	初めて 会う 人が けなす のは 誰を ですか？ for-the-first-time see person-Nom criticize that-Top who-Acc is-Q
morphology	part of speech	子供 そう。 child seem	美味し そう。 tasty seem
	idiom	太郎の 忠告は 花子には 糠にも 釘 だった。 Taroo-Gen advise-Top Hanako-Dat-Top bran-Dat-also nail was	太郎の 忠告は 花子には 糠に 釘 だった。 Taroo-Gen advise-Top Hanako-Dat-Top bran-Dat nail was
	reflexive	強い 地震のため 建物が 自壊を した。 strong earthquake-for building-Nom self-collapse-Acc do-Past	強い 地震のため 建物が 自壊 した。 strong earthquake-for building-Nom self-collapse do-Past
	inflection	それは 計測 可能だ 粒子だ。 that-Top measurement possibility-Cop.Fin particle-Cop	それは 計測 可能な 粒子だ。 that-Top measurement possibility-Cop.Adnom particle-Cop
	nominalization	原稿に 手の 入れ方は 人それぞれだ。 draft-Dat hand-Gen put.in-way-Top person-each-Cop	原稿への 手の 入れ方は 人それぞれだ。 draft-to-Gen hand-Gen put.in-way-Top person-each-Cop
	honorification	伊藤先生から そのことを 話して おいでになる。 Ito-teacher-from that-fact-Acc tell-Te Hon-be-Lv-Pres	伊藤先生から そのことを お話になって いる。 Ito-teacher-from that-fact-Acc Hon-tell-Lv-Te be-Pres

Phenomenon	Paradigm	Unacceptable	Acceptable
Quantifiers	floating quantifiers	学生が 家を 4 人 買った。 student-Nom house-Acc four-CL buy-Past	学生が 4 人 家を 買った。 student four-CL house-Acc buy-Past
	universal quantifiers	みんながみんな 大学へ 行かない。 everyone-Nom-everyone university-to go-Neg-Pres	みんながみんな 大学へ 行く 訳では ない。 everyone-Nom-everyone university-to go-Pres reason-Cop-Top Neg-Pres
	classifier	太郎は 3 本ずつの その 鉛筆を 買った。 Taroo-Top three-Cl-Dist-Gen that pencil-Acc buy-Past	太郎は その 3 本ずつの 鉛筆を 買った。 Taroo-Top that three-Cl-Dist-Gen pencil-Acc buy-Past
	negation	ジョンは メアリーが 賢い 以上に 賢くない。 John-Top Mary-Nom smart more smart-Neg	ジョンは メアリーが 賢い 以上に 賢い。 John-Top Mary-Nom smart more smart
Island effects	complex-NP island	太郎が 昨日 会った 人を 探している のは 花子に だ。 Taroo-Nom yesterday saw person-Acc looks-for that-Top Hanako-Dat is	太郎が 昨日 花子に 会った 人を 探している のだ。 Taroo-Nom yesterday Hanako-Dat saw person-Acc looks-for that-is
	adjunct island	太郎が 読んだ から 花子が 怒った のは その 本を だ。 Taroo-Nom read because Hanako-Nom got-angry that-Top that book-Acc is	太郎が その 本を 読んだ から 花子が 怒った のだ。 Taroo-Nom that book-Acc read because Hanako-Nom got-angry that-is
	specificity island	ジョンは その メアリーより 高い 指輪を 買った。 John-Top that Mary-than expensive ring-Acc bought	ジョンは メアリーより 高い 指輪を 買った。 John-Top Mary-than expensive ring-Acc bought
	negative island	ジョンは メアリーが 雇わなかったより 賢い 人を 見つけた。 John-Top Mary-Nom hire-Neg-Past-than smart person-Acc found	ジョンは メアリーが 雇ったより 賢い 人を 見つけた。 John-Top Mary-Nom hire-Past-than smart person-Acc found
	factive island	メアリーが ジョンが 自分の 学生が 新しい 仮説を 提案した と 知っていたのの 欠陥を 指摘した。 Mary-Nom John-Nom self-Gen student-Nom new hypothesis-Acc proposed Czer know-had-no-Gen defect-Acc pointed-out	メアリーが ジョンが 自分の 学生が 新しい 仮説を 提案した と 言っていたのの 欠陥を 指摘した。 Mary-Nom John-Nom self-Gen student-Nom new hypothesis-Acc proposed Czer say-had-no-Gen defect-Acc pointed-out
Filler-gap	intervention effects	誰も 何を 読まなかったの？ anyone what-Acc read-Neg-Past-Q	何を 誰も 読まなかったの？ what-Acc anyone read-Neg-Past-Q
	relative clause	山田先生は この本を なった ことは お読み だ。 Yamada-teacher-Top this-book-Ac become-Past fact-Top Hon-read-Ren Cop	山田先生は この本を お読みに なった。 Yamada-teacher-Top this-book-Ac Hon-read-Ren-Obl become-Past
	cleft	山田先生が なった Nz-Top この本の お読み だ。 Yamada-teacher-Nom become-Pastのは this-book-Gen Hon-read-Ren Cop	山田先生が この本を お読みに なった。 Yamada-teacher-Nom this-book-Acc Hon-read-Ren-Oblbecome-Past
	resumptive pronoun	トムが それらを 食べた ことが 明らかな 芋は 大きかった。 Tom-Nom these-Acc ate fact-Nom clear potato-Top big-Past	トムが 食べた ことが 明らかな 芋は 大きかった。 Tom-Nom ate fact-Nom clear potato-Top big-Past
NPI licensing	NPI	今回は 誰が 寄付を 呼びかけも しなかった。 this-time-Top anyone-Nom donation-Acc call.for-Q do-Neg-Past	今回は 誰から 寄付を 呼びかけも しなかった。 this-time-Top anyone-from donation-Acc call.for-Q do-Neg-Past
	NCI	ジョンが もし 何も 盗んだら、 逮捕される だろう。 John-Nom if what-MO stead-Cond arrest-Pass be.will	ジョンが もし 何か 盗んだら、 逮捕される だろう。 John-Nom if what-Q stead-Cond arrest-Pass be.will
Nominal structure	modifier	私が 昨日 見たの人は 素敵だった。 I-Nom yesterday saw-NO-person-Top beautiful-Past	私が 昨日 見た人は 素敵だった。 I-Nom yesterday saw-person-Top beautiful-Past
	measure phrase	このビルは 低さ 20 メートル ある。 this-building-Top shortness 20-meter is	このビルは 高さ 20 メートル ある。 this-building-Top height 20-meter is
Control/Raising	subject control	だるまが 転び損ねた。 Dharma-Nom tumble failed	太郎が 転び損ねた。 Taroo-Nom tumble failed