JCoLA: Japanese Corpus of Linguistic Acceptability

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

Abstract

Neural language models have exhibited outstanding performance in a range of downstream tasks. However, there is limited understanding regarding the extent to which these models internalize syntactic knowledge, so that various datasets have recently been constructed to facilitate syntactic evaluation of language models across languages. In this paper, we introduce **JCoLA** (Japanese Corpus of Linguistic Acceptability), which consists of 10,020 sentences annotated with binary acceptability judgments. Specifically, those sentences are manually extracted from linguistics textbooks, handbooks and journal articles, and split into in-domain data (86 %; relatively simple acceptability judgments extracted from textbooks and handbooks) and out-of-domain data (14 %; theoretically significant acceptability judgments extracted from journal articles), the latter of which is categorized by 12 linguistic phenomena. We then evaluate the syntactic knowledge of 9 different types of Japanese and multilingual language models on JCoLA. The results demonstrated that several models could surpass human performance for the in-domain data, while no models were able to exceed human performance for the out-of-domain data. Error analyses by linguistic phenomena further revealed that although neural language models are adept at handling local syntactic dependencies like argument structure, their performance wanes when confronted with long-distance syntactic dependencies like verbal agreement and NPI licensing.

Keywords: linguistics, targeted syntactic evaluation, language resource

1. Introduction

Neural language models, especially Transformer-based language models (Vaswani et al., 2017), have exhibited outstanding performance in a range of downstream tasks (Wang et al., 2018, 2019), yet there is limited understanding regarding the extent of linguistic knowledge these models have internalized. Several studies have explored the syntactic competence of language models through acceptability judgment tasks (e.g., Linzen et al., 2016; Marvin and Linzen, 2018). These and other related studies are critical as they mark the beginning of syntactic evaluations of language models, but they were limited in the scope of linguistic phenomena. In more recent times, researchers have constructed extensive datasets to facilitate more comprehensive syntactic evaluations (Warstadt et al., 2019, 2020; Xiang et al., 2021; Trotta et al., 2021; Mikhailov et al., 2022). Nonetheless, the majority of these investigations have centered around English and other European languages (Gulordava et al., 2018; Warstadt et al., 2019, 2020; Wilcox et al., 2018), with only a handful expanding their scope to encompass non-European languages (Gulordava et al., 2018; Ravfogel et al., 2018). Notably, an even smaller number of studies have addressed a broad spectrum of linguistic phenomena in languages other than English (Trotta et al., 2021; Xiang et al., 2021; Mikhailov et al., 2022).

In this paper, we introduce JCoLA (Japanese Corpus of Linguistic Acceptability) [1], which consists of 10,020 sentences with acceptability judgments by linguists. Specifically, those sentences are manually extracted from linguistics textbooks, handbooks and journal articles, and split into in-domain data (86 %; relatively simple acceptability judgments extracted from textbooks and handbooks) and out-of-domain data (14 %; theoretically significant acceptability judgments extracted from journal articles), the latter of which is categorized by 12 linguistic phenomena. We then evaluate the syntactic knowledge of 9 different types of Japanese and multilingual language models on JCoLA. The results demonstrated that several models could surpass human performance for the in-domain data, while no models were able to exceed human performance for the out-of-domain data. Error analyses by linguistic phenomena further revealed that although neural language models are adept at handling local syntactic dependencies like argument structure, their performance wanes when confronted with long-distance syntactic dependencies like verbal agreement and NPI licensing.

2. Related Work

Acceptability judgment is a crucial aspect of human linguistic competence. It refers to the innate ability of individuals to differentiate between sen-

[1] Our dataset, JCoLA, is publicly available at https://github.com/osekilab/JCoLA. JCoLA is adopted as one of six tasks of JGLUE (Kurihara et al., 2022), a benchmark for natural language understanding (NLU) in Japanese.

LREC-COLING 2024, pages 9477–9488
20-25 May, 2024.

Language	Binary Acceptability Judgment	Minimal Pairs
English	CoLA (Warstadt et al., 2019)	BLiMP (Warstadt et al., 2020)
Italian	ItaCoLA (Trotta et al., 2021)	
Chinese		CLiMP (Xiang et al., 2021)
Russian	RuCoLA (Mikhailov et al., 2022)	
Japanese	JCoLA (This work)	JBLiMP (Someya and Oseki, 2023)

Table 1: Comparison of JCoLA and other existing datasets. As of now, there are no languages other than English for which both CoLA-style and BLiMP-style datasets are available.

tences that are grammatically correct and those that are not, even without any explicit training in grammar. For instance, when presented with two sentences, individuals can intuitively recognize which one is more acceptable or natural-sounding. Such judgments are considered the primary behavioral measure used by generative linguists to study the underlying structure of language in humans (Chomsky, 1957). By examining acceptability judgments, linguists can gain insights into the rules that govern language and how these rules are applied by speakers of a particular language.

Historically, the evaluation of language models has been conducted using metrics such as perplexity, or based on how well the models perform on specific downstream tasks, as seen in benchmarks like GLUE (Wang et al., 2018). However, in recent years, there have been efforts to assess the syntactic knowledge of language models through acceptability judgment tasks.

Linzen et al. (2016) first employed minimal pairs to examine how well LSTM language models could capture subject-verb agreement in English.

(1) The key is on the table.

(2) * The key are on the table.

This and other related studies are critical as they mark the beginning of syntactic evaluations of language models. However, they were limited in the scope of linguistic phenomena considered (e.g., Marvin and Linzen, 2018; Futrell et al., 2019; Gulordava et al., 2018).

In light of this, more recent approaches introduced large-scale acceptability judgment corpora for targeted syntactic evaluations of language models (Warstadt et al., 2019, 2020). Similar to Linzen et al. (2016), Warstadt et al. (2020) constructed BLiMP (Benchmark of Linguistic Minimal Pairs) as a dataset employing minimal pairs. BLiMP consists of 67,000 minimal pairs automatically generated across 12 types of linguistic phenomena. This enables the evaluation of language models on a wide range of linguistic phenomena, not limited to subject-verb agreement. Furthermore, similar datasets have been developed for languages other than English, allowing for comparable evaluations across various languages (Xiang et al., 2021; Someya and Oseki, 2023).

Concurrently, there is also an approach to targeted syntactic evaluations of language models that does not rely on minimal pairs but instead evaluates language models with binary classification tasks based on acceptability. CoLA (Corpus of Linguistic Acceptability; Warstadt et al. (2019)) is the first corpus that achieves this, a dataset built by collecting sentences from syntax textbooks, handbooks, and linguistics journals. Similar datasets to CoLA have also been emerging for languages other than English (Trotta et al., 2021; Mikhailov et al., 2022), though none exist for Japanese as of yet (cf. Table 1).

3. JCoLA

In this study, we introduce JCoLA (Japanese Corpus of Linguistic Acceptability), which will be the first large-scale acceptability judgment task dataset focusing on Japanese. JCoLA consists of sentences from textbooks and handbooks on Japanese syntax, as well as from journal articles on Japanese syntax that are published in JEAL (Journal of East Asian Linguistics), one of the prestigious journals in theoretical linguistics.

3.1. Data Collection

Sentences in JCoLA were collected from prominent textbooks and handbooks focusing on Japanese syntax. In addition to the main text, example sentences included in the footnotes were also considered for collection. We also collected 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 2,252 acceptability judgments from 26 papers on Japanese syntax (Table 2). Acceptability judgments include sentences in appendices and footnotes, but not sentences presented for anal-

yses of syntactic structures (e.g. sentences with brackets to show their syntactic structures). As a result, a total of 11,984 example sentences were collected. Using this as a basis, JCoLA was constructed through the methodology explained in the following sections.

Source	N	%
Gunji (1987)	301	88.0
Inoue (1976a,b)	1805	86.2
Kuno (1973)	1553	78.0
Kuroda (1965)	332	91.6
Kuroda (1992)	681	85.5
Miyagawa (2008)	591	82.7
Shibatani (1976)	2209	83.3
Shibatani (1990)	387	90.2
Tsujimura (1999)	531	75.9
Tsujimura (2013)	259	81.1
In-Domain	8649	83.4
Abe (2011)	15	53.3
Asano and Ura (2010)	92	63.0
Bobaljik and Wurmbrand (2007)	11	72.7
Grosu (2010)	11	18.2
Grosu and Landman (2012)	8	62.5
Hayashishita (2009)	34	76.5
Ivana and Sakai (2007)	38	73.7
Kishida and Sato (2012)	81	77.8
Kishimoto (2008)	204	71.1
Kishimoto (2012)	90	61.1
Miyamoto (2009)	17	94.1
Nishigauchi (2014)	68	94.1
Oshima (2006)	25	96.0
Saito et al. (2008)	32	78.1
Sawada (2013)	40	95.0
Shibata (2015)	72	80.6
Shimoyama (2014)	51	92.2
Sudo (2015)	133	65.4
Takahashi (2006)	26	57.7
Takahashi (2010)	29	79.3
Takano (2011)	41	90.2
Takita (2009)	6	16.7
Tenny (2006)	45	93.3
Tomioka (2009)	15	60.0
Tsujioka (2011)	67	56.7
Watanabe (2010)	27	81.5
Watanabe (2013)	93	64.5
Out-of-Domain	1371	73.2
Total	10,020	82.0

Table 2: The number of sentences in JCoLA by source. *N* is the number of sentences in a source. *%* is the percent of the acceptable sentences in a source. While *In-Domain* sources are textbooks and handbooks on Japanese syntax, while all the sources listed above as *Out-of-Domain* are journal articles published in JEAL.

3.2. Data Preparation

3.2.1. Data Preprocessing

Among the sentences extracted through the above method, there were sentences that were not appropriate for JCoLA, a binary classification dataset based on single-sentence acceptability judgments. We either remove or modify these sentences in preprocessing. First, sentences labeled with '?', '#', '%', or '(?)' were removed. Additionally, sentences that did not have such labels but were noted to have variable acceptability depending on the speaker were also removed. Furthermore, duplicates, examples that were not single-sentence acceptability judgments, those containing inappropriate vocabulary, and examples whose unacceptability depends on the context were eliminated. Lastly, some sentences were found to be incomplete. In these cases, they were supplemented to form complete sentences, ensuring that the acceptability did not change. (e.g., John's book -> John's book is red.)

3.2.2. Categorization

A part of the data is annotated based on linguistic phenomena in order to analyze each phenomenon in detail. We categorize the 12 phenomena in JCoLA as follows (Table 3):

Phenomenon	# Sentences
Argument structure	545
Filler-gap	257
Morphology	159
Nominal structure	150
Quantifier	127
Verbal agreement	105
Binding	101
Ellipsis	44
Island effects	19
NPI/NCI	12
Control/raising	11
Simple	71

Table 3: Number of sentences by phenomenon in out-of-domain data. Note that the examples in JCoLA could be categorized into multiple phenomena.

Argument Structure: acceptability judgements based on the order of arguments (3a) and case marking (3b).

(3) a. Ken-ni tegami-ga todoita.
Ken-DAT letter-NOM reached
'A letter reached Ken.'

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 reflexive binding (4a) and the coreference resolution of anaphors (4b).

(4) a. Ken-ga zibun-no heya-ni
Ken-NOM self-GEN room-DAT
modotta
returned
'Ken returned to his room.'

b. ?* Hazimete soitu-ni au
for-the-first-time him-DAT see
hito-ga kenasu no-wa
person-NOM criticize that-TOP
dare-o desu ka?
who-ACC is Q
'?*Who is it that people who see him for the first time criticize?'

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

(5) John-wa ie-o tukuri-sokoneta
John-TOP house-ACC make-to-fail-PAST
'John failed to make a house.'

Ellipsis: acceptability judgements based on the possibility of omitting elements in the sentences. For instance, this includes nominal (6a) and adjunct ellipsis (6b).

(6) a. Taroo-ga zibun-o
Taroo-NOM self-ACC
hihansita-ra Hanako-wa
criticized-when Hanako-TOP
hometa.
praised
'When Taroo criticized himself, Hanako praised.'

b. * Taroo-ga sono riyuu de
Taroo-NOM that reason for
kaikosareta atode, Hanako-mo
was-fired after Hanako-also
kaikosareta.
was-fired
'*After Taroo was fired for that reason, Hanako was fired too.'

Filler-gap: acceptability judgements based on the dependency between the moved element and the gap. For instance, this includes comparatives (7a) and cleft sentences (7b).

(7) a. Mary-wa John-ga kaita yori nagai
Mary-TOP John-NOM wrote than long
ronbun-o kaita.
paper-ACC wrote
'Mary wrote a longer paper than John wrote'

b. Taroo-ga atta no-wa Hanako-ni
Taroo-NOM saw that-TOP Hanako-DAT
da.
is
'It was Hanako that Taroo saw.'

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

(8) * Taroo-wa Hanako-ga naze kare-no
Taroo-TOP Hanako-NOM why he-GEN
tegami-o suteta kara
letter-ACC discarded because
okotteiru no?
be.angry C
'*Why is Taro angry because Hanako discarded his letter?'

Morphology: acceptability judgements based on the morphology. For instance, it includes idioms.

(9) Taroo-no kotoba-wa hi-ni abura-o
Taroo-GEN words-TOP fire-DAT oil-ACC
sosoida.
pour
'Taroo's words made the situation worse'

Nominal Structure: acceptability judgements based on the internal structure of noun phrases.

(10) amen-no hi-wa kiraida
rainy day-TOP hate.be
'I hate rainy days.'

NPI/NCI: acceptability judgements based on the restrictions on where negative polarity/concord items (NPIs/NCIs) can appear. For instance, NCIs include *daremo*.

(11) Daremo monku-o iw-anakat-ta.
who-MO complaint-ACC say-NEG-PAST
'Nobody complained.'

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

(12) John-wa hon-o san-satsu katta.
John-TOP book-ACC three-CL bought
'John bought three books.'

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. * Mary-ga Ito-sensei-o
Mary-NOM Ito-teacher-ACC
o-home-ni-nat-ta.
HON-praise-LV-PAST
'Mary praised Prof. Ito.'

Simple: acceptability judgements that do not have marked syntactic structures. For instance, it includes a simple transitive sentence.

(14) John-ga hon-o yonda
John-NOM book-ACC read
'John read a book.'

Sentences that do not fall into these 12 phenomena were deleted.

Note that the examples in JCoLA could be categorized in multiple phenomena. For example, the following sentence includes a classifier *mit-tu* 'three', which is a quantifier-binder and a variable *soko* 'it', which gets a bound variable interpretation. Thus, this is a combination of binding and quantifier phenomena.

(15) Mit-tu-izyoo-no kaisya-o soko-no
3-CL-or.more-GEN company-ACC it-GEN
syain-ga hihansi-ta
employee-NOM criticized-PAST
'Three companies, its employee(s) criticized.'

3.3. Data Validation

As a reference for the upper limit of accuracy in JCoLA, human acceptability judgment experiments were conducted on Lancers[2] with a subset of the JCoLA data. Specifically, we conducted acceptability judgment experiments on 200 sentences sampled from the in-domain data and all the sentences in the out-of-domain data, making a total of 1,951 sentences. To reduce the burden on each annotator, the sentences were divided into 38 groups of 50 sentences and one group of 51 sentences. Each annotator performed a forced-choice binary acceptability judgment task on 50 or 51 sentences. For the out-of-domain data, if the results of the acceptability judgment experiment did not match between the human majority vote and the JCoLA annotation, that data was removed. As a result, 380 instances were deleted, leaving 1,371 instances in the out-of-domain data. The results showed that for the in-domain data, the individual agreement with JCoLA was 75.9%, and the majority vote agreement with JCoLA was 79.5%. For the out-of-domain data, the individual agreement with JCoLA was 85.4%, and the majority vote agreement with JCoLA was 100.0% (due to the aforementioned data removal).

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

3.4. Data Split

While CoLA includes out-of-domain data in addition to the standard train/dev/test splits to assess whether overfitting occurs to specific sources or linguistic phenomena within the training data, JCoLA will also incorporate out-of-domain data. However, in JCoLA, the data collected from journal articles in JEAL are designated as out-of-domain. This is because JCoLA aims to evaluate whether language models can generalize to more complex linguistic phenomena (cf. Class Ⅲ judgement, see Marantz 2005; Linzen and Oseki 2018) after learning relatively simple grammatical rules (Class Ⅱ judgement). The in-domain data is split into training data (6,919 instances), development data (865 instances), and test data (865 instances). On the other hand, the out-of-domain data is only used for evaluation, and divided into development data (685 instances) and test data (686 instances).

4. Experiments

4.1. Models

In this paper, we evaluate some pretrained Japanese and multilingual neural language models on JCoLA. Specifically, we evaluate nine different neural language models provided by different organizations, which are different in size, method of morphological analysis and tokenization, and training corpus. Additionally, to provide a benchmark for state-of-the-art language models, we also conduct evaluations on GPT-4 and GPT-3.5-turbo.

BERT We evaluate three different types of BERT language models provided by Tohoku University NLP group[3]: Tohoku $\text{BERT}_{\text{BASE}}$[4], Tohoku $\text{BERT-char}_{\text{BASE}}$[5] and Tohoku $\text{BERT}_{\text{LARGE}}$[6]. These models are trained on the Japanese version of Wikipedia. The texts are first tokenized by MeCab (Kudo et al., 2004) and then split into subwords by BPE (Sennrich et al., 2016).[7] Tohoku

[3] https://github.com/cl-tohoku

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

[5] https://huggingface.co/cl-tohoku/bert-base-japanese-char-v2

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

[7] For Tohoku $\text{BERT-char}_{\text{BASE}}$, the texts are segmented into characters.

$\text{BERT}_{\text{BASE}}$ and Tohoku BERT-$\text{char}_{\text{BASE}}$ have 12 layers, 12 attention heads, and 768-dimensional hidden states, while Tohoku $\text{BERT}_{\text{LARGE}}$ has 24 layers, 16 attention heads, and 1024-dimensional hidden states.

In addition, we evaluate a BERT language model provided by NICT (NICT $\text{BERT}_{\text{BASE}}$).[8] The model configuration is the same as Tohoku $\text{BERT}_{\text{BASE}}$ and Tohoku BERT-$\text{char}_{\text{BASE}}$.

Japanese RoBERTa We also evaluate three variants of RoBERTa language models provided by Kawahara Lab. at Waseda University[9]: Waseda $\text{RoBERTa}_{\text{BASE}}$[10], Waseda RoBERTa-$\text{seq128}_{\text{LARGE}}$[11] and Waseda RoBERTa-$\text{seq512}_{\text{LARGE}}$[12]. These models are trained on the Japanese version of Wikipedia and the Japanese portion of CC-100. The texts are first tokenized by Juman++ (Morita et al., 2015) and then split into subwords using Sentence Piece (Kudo and Richardson, 2018) with a unigram language model (Kudo, 2018). Waseda $\text{RoBERTa}_{\text{BASE}}$ has 12 layers, 12 attention heads, and 768-dimensional hidden states. Waseda RoBERTa-$\text{seq128}_{\text{LARGE}}$ and Waseda RoBERTa-$\text{seq512}_{\text{LARGE}}$ both have 24 layers, 16 attention heads, and 1024-dimensional hidden states, but are trained with the maximum sequence length of 128 and 512, respectively.

XLM-RoBERTa To compare the performance of monolingual and multilingual language models on JCoLA, we also evaluate two multilingual language models with different parameter sizes: XLM-$\text{RoBERTa}_{\text{BASE}}$[13] and XLM-$\text{RoBERTa}_{\text{LARGE}}$[14]. These models are trained on multilingual Common Crawl (Wenzek et al., 2020) and the train texts are directly tokenized using Sentence Piece (Kudo and Richardson, 2018) with a unigram language model (Kudo, 2018). XLM-$\text{RoBERTa}_{\text{BASE}}$ has 12 layers, 12 attention heads and 768-dimensional hidden states. XLM-$\text{RoBERTa}_{\text{LARGE}}$ has 24 layers, 16 attention heads and 1024-dimensional hidden states.

[8] https://direct.nict.go.jp/

[9] https://nlp-waseda.jp/en/

[10] https://huggingface.co/nlp-waseda/roberta-base-japanese

[11] https://huggingface.co/nlp-waseda/roberta-large-japanese

[12] https://huggingface.co/nlp-waseda/roberta-large-japanese-seq512

[13] https://huggingface.co/xlm-roberta-base

[14] https://huggingface.co/xlm-roberta-large

GPT-4 and GPT-3.5-turbo In addition to the models previously mentioned, we extend our evaluation to include large-scale language models developed by OpenAI, namely GPT-4 (gpt-4) and GPT-3.5-turbo (gpt-3.5-turbo). These models represent some of the most advanced developments in the field of natural language processing and are publicly available through OpenAI's platform. For our experiments, we employed the code provided by LLM-jp,[15] ensuring our prompts were consistent with the standards set within this framework. The generation configurations were adopted as per the default settings provided by the OpenAI API, which are designed to optimize the performance of these models under a variety of tasks. To ensure the reliability of our results, we conducted each experiment three times, calculating the mean and variance of the outcomes to present a comprehensive view of the model's performance in our context.

4.2. Training Settings

Each language model is trained for five epochs with AdamW optimizer (Loshchilov and Hutter, 2019) and linear warmup with a warmup ratio of 0.1. In addition, the language models are trained using three different learning rates (5e-5, 3e-5, and 2e-5) and we evaluate models which achieved the highest Matthews Correlation Coefficient (MCC; Matthews (1975)) on the development data. This evaluation metric is an evaluation metric suitable for unbalanced binary classifiers also used in Warstadt et al. (2019). For each configuration, we trained 20 models with different random seeds to mitigate the effect of randomness. The score for each language model is calculated as the average across 20 different random seeds, but we ignore those results where the models achieved less than zero MCC score on the development set, as in Warstadt and Bowman (2020).

5. Results and Discussion

5.1. Overall performance

Table 4 presents the Matthews Correlation Coefficient (MCC) and accuracy of various models on the in-domain and out-of-domain data, along with human performance. In the in-domain data, several models demonstrate performance surpassing that of human individuals. However, in the case of out-of-domain data, none of the models were able to exceed human performance (MCC). This suggests that the language models may not necessarily capture the complex linguistic phenomena addressed in theoretical linguistics (Class Ⅲ

[15] https://github.com/llm-jp/llm-jp-eval

Model	In-domain		Out-of-domain	
	Acc.	MCC	Acc.	MCC
Tohoku BERT base	0.838 ± 0.008	0.357 ± 0.032	0.758 ± 0.007	0.264 ± 0.033
Tohoku BERT base (char)	0.813 ± 0.010	0.225 ± 0.024	0.739 ± 0.006	0.175 ± 0.037
Tohoku BERT large	0.836 ± 0.008	0.352 ± 0.036	0.771 ± 0.009	0.320 ± 0.038
NICT BERT base	0.843 ± 0.006	0.374 ± 0.025	0.773 ± 0.010	0.327 ± 0.040
Waseda RoBERTa base	0.859 ± 0.009	0.410 ± 0.046	0.788 ± 0.015	0.385 ± 0.059
Waseda RoBERTa large (s128)	**0.864 ± 0.007**	**0.466 ± 0.028**	0.822 ± 0.014	0.506 ± 0.044
Waseda RoBERTa large (s512)	0.859 ± 0.017	0.416 ± 0.104	0.799 ± 0.025	0.421 ± 0.098
XLM RoBERTa base	0.824 ± 0.007	0.152 ± 0.084	0.748 ± 0.018	0.166 ± 0.135
XLM RoBERTa large	0.833 ± 0.010	0.242 ± 0.089	0.759 ± 0.019	0.230 ± 0.155
GPT-3.5-Turbo	0.625 ± 0.016	0.185 ± 0.030	0.701 ± 0.022	0.398 ± 0.040
GPT-4	0.794 ± 0.005	0.295 ± 0.010	**0.855 ± 0.002**	**0.629 ± 0.005**
Human (Individual)	0.760	0.384	0.854	0.653
Human (Majority vote)	0.795	0.437	1.000	1.000

Table 4: Performance of each language model on JCoLA out-of-domain test set. The score for each language model is calculated as the average across 20 different random seeds, but we ignore those results where the models achieved less than zero MCC score on the development set, as in Warstadt and Bowman (2020). The best performance across models is indicated in bold.

judgement). However, while the majority of models have lower performance on out-of-domain data compared to in-domain data, some models perform better on out-of-domain data. These models appear to be generalizing the linguistic phenomena observed in in-domain data correctly and are somewhat able to judge acceptability even for more complex linguistic phenomena.[16]

5.2. Performance by phenomenon

Figure 1 shows the Matthews Correlation Coefficient (MCC) values for each linguistic phenomenon in the out-of-domain test set across different models. Notably, almost all models demonstrate high accuracy in the Simple category, which suggests that they are capable of accurately capturing this linguistic phenomenon, even with sentences from sources not seen during training. However, for other phenomena, the performance is generally lower than that for Simple. In fact, the average MCC across linguistic phenomena, excluding Simple, is 0.248, which is significantly lower than the 0.599 observed for Simple. This suggests that while language models can effectively learn relatively simple linguistic phenomena (Class Ⅱ judgement) as presented in textbooks and handbooks of syntactic theory, they may not necessarily be able to generalize to more complex linguistic phenomena (Class Ⅲ judgement).

Furthermore, upon examining the performance of language models on different phenomena, it becomes apparent that language models perform relatively well on certain linguistic phenomena, such as binding, argument structure, and filler-gap, but struggle with others. Relatively high performance in Binding could be attributed to the fact that the proportion of positive examples for Binding is 93.1%, significantly higher than the overall 73.2% for the out-of-domain data. For Argument Structure, many sentences only require capturing relatively local dependencies related to the order of arguments and/or case marking, as in (16).

(16) John-ga hon-o/*-ni yonda
John-NOM book-ACC/*DAT read
'John read a book.'

Regarding filler-gap, even though it generally involves complex linguistic phenomena such as wh-movement, the presence of a relatively large number of sentences involving simpler comparison phenomena could be contributing to the higher accuracy.

(17) Mary-wa John-ga kaita yori nagai
Mary-TOP John-NOM wrote than long
ronbun-o kaita.
paper-ACC wrote
'Mary wrote a longer paper than John wrote'

On the other hand, language models show lower accuracy on linguistic phenomena such as NPI/NCI and verbal agreement. This could be

[16] Interestingly, the models that exhibited higher performance on out-of-domain data all utilized Sentence Piece with a unigram language model for tokenization, indicating the possibility that this choice of tokenization method may have contributed in some way to their performance.

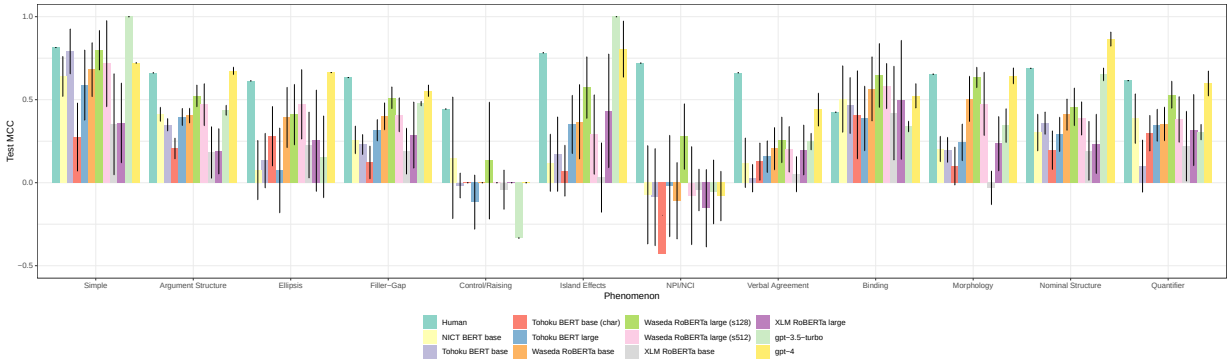

Figure 1: Performance of each language model on JCoLA out-of-domain test set by phenomenon. The MCC score for each language model is calculated as the average across 20 different random seeds, but we ignore those results where the models achieved less than zero MCC score on the development set, as in Warstadt and Bowman (2020). Error bars mark the mean ± 1 SD.

because NPI/NCI and verbal agreement often require capturing relatively long-distance dependencies, as seen in the examples below.[17]

(18) 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. * **Mary-ga** Ito-sensei-o
Mary-NOM Ito-teacher-ACC
o-home-**ni-nat**-ta.
HON-praise-LV-PAST
'Mary praised Prof. Ito.'

(19) **Daremo** monku-o iw-**anakat**-ta.
who-MO complaint-ACC say-NEG-PAST
'Nobody complained.'

Overall, the analysis by linguistic phenomenon highlights the strengths and limitations of language models in capturing various linguistic phenomena. While they are adept at handling simpler structures, their performance wanes when confronted with more complex linguistic phenomena, especially those requiring long-distance dependencies.

6. Conclusion

In this paper, we introduced JCoLA (Japanese Corpus of Linguistic Acceptability), which consists of 10,020 sentences annotated with binary acceptability judgments. Specifically, those sentences were manually extracted from linguistics textbooks, handbooks and journal articles, and split into in-domain data (86 %; relatively simple acceptability judgments extracted from textbooks and handbooks) and out-of-domain data (14 %; theoretically significant acceptability judgments extracted from linguistics journals), the latter of which was categorized by 12 linguistic phenomena. We then evaluated the syntactic knowledge of 9 different types of Japanese and multilingual language models on JCoLA. The results demonstrated that several models could surpass human performance for the in-domain data, while no models were able to exceed human performance for the out-of-domain data. Error analyses by linguistic phenomena further revealed that although neural language models are adept at handling local syntactic dependencies like argument structure, their performance wanes when confronted with long-distance syntactic dependencies like verbal agreement and NPI licensing.

Limitations

All the sentences included in JCoLA have been extracted from textbooks, handbooks and journal articles on theoretical syntax. Therefore, those sentences are guaranteed to be theoretically meaningful, making JCoLA a challenging dataset. However, the distribution of linguistic phenomena directly reflects that of the source literature and thus turns out to be extremely skewed. Indeed, as can be seen in Table 3, while the number of sentences exceeds 100 for most linguistic phenomena, there are several linguistic phenomena for which there are only about 10 sentences. In addition, since it is difficult to force language models to interpret sentences given specific contexts, those sentences whose unacceptability depends on contexts were inevitably removed from JCoLA. This removal process resulted in the deletion of unacceptable sentences from some linguistic phenomena (such as ellipsis), consequently skewing the balance between acceptable and unacceptable sen-

[17] The results for control/raising were not considered to be reliable due to the small sample size, and they were excluded from the analysis.

tences (with a higher proportion of acceptable sentences).

Acknowledgements

This work was supported by JST PRESTO Grant Number JPMJPR21C2, Japan.

7. Bibliographical References

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

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.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton, The Hague.

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.

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.

Takao Gunji. 1987. *Japanese Phrase Structure Grammar*, volume 8 of *Studies in Natural Language and Linguistic Theory*. Springer Netherlands, Dordrecht.

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

Kazuko Inoue. 1976a. *[Transformational Grammar and Japanese Vol.1] Henkei bumpou to nihongo (in Japanese)*. TAISHUKAN Publishing Co., Ltd.

Kazuko Inoue. 1976b. *[Transformational Grammar and Japanese Vol.2] Henkei bumpou to nihongo (in Japanese)*. TAISHUKAN Publishing Co., Ltd.

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

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. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

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.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain. Association for Computational Linguistics.

Susumu Kuno. 1973. The Structure of the Japanese Language.

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. JGLUE: Japanese general language understanding evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957–2966, Marseille, France. European Language Resources Association.

S.-Y. Kuroda. 1965. *Generative grammatical studies in the Japanese language.* Thesis, Massachusetts Institute of Technology. Accepted: 2006-08-09T19:22:05Z.

S.-Y. Kuroda. 1992. *Japanese Syntax and Semantics*, volume 27 of *Studies in Natural Language and Linguistic Theory*. Springer Netherlands, Dordrecht.

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).

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. ArXiv:1711.05101 [cs, math].

Alec 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.

B.W. Matthews. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451.

Vladislav Mikhailov, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov, and Ekaterina Artemova. 2022. RuCoLA: Russian corpus of linguistic acceptability. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5207–5227, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shigeru Miyagawa. 2008. *The Oxford Handbook of Japanese Linguistics*. Oxford University Press.

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

Hajime Morita, Daisuke Kawahara, and Sadao Kurohashi. 2015. Morphological analysis for unsegmented languages using recurrent neural network language model. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2297, Lisbon, Portugal. 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.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

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

Masayoshi Shibatani. 1976. *Japanese Generative Grammar*. Academic Press, Inc.

Masayoshi Shibatani. 1990. *The Languages of Japan*. Cambridge University Press.

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

Taiga Someya and Yohei Oseki. 2023. JBLiMP: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, Dubrovnik, Croatia. Association for Computational Linguistics.

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

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.

Satoshi 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.

Natsuko Tsujimura. 1999. *The Handbook of Japanese Linguistics*. Wiley-Blackwell.

Natsuko Tsujimura. 2013. *An Introduction to Japanese Linguistics*. Wiley-Blackwell.

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 and Samuel R. Bowman. 2020. Linguistic analysis of pretrained sentence encoders with acceptability judgments.

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.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

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.