日本語版 CoLA の構築

染谷 大河　大関 洋平
東京大学
{taiga98-0809, oseki}@g.ecc.u-tokyo.ac.jp

概要

近年、ニューラル言語モデルが自然言語の統語知識をどれほど有しているかを、容認性判断課題を通して検証する研究が行われてきている。しかし、このような言語モデルの統語的評価を行うためのデータセットは、主に英語を中心とした欧米の諸言語を対象に構築されてきた。本研究では、既存のデータセットの問題点を克服しつつ、このようなデータセットが構築されてこなかった日本語を対象とした初めてのデータセットである JCoLA (Japanese Corpus of Linguistic Acceptability) を構築した上で、それを用いた言語モデルの統語的評価を行った。

1　はじめに

近年、ニューラル言語モデルの成長は著しく、特に Transformer [1] をベースとしたモデルは、様々なタスクで高い精度を発揮している [2, 3]。一方、理論言語学の分野においては、伝統的に自然言語には一種の統語構造が存在しているということが主張されており [4, 5]、その構造の姿を解明すべく研究が進められている。確かに、先述のニューラル言語モデルは多くの自然言語処理タスクにおいて非常に高い精度を発揮することが確認されているものの、如何にしてこれらのタスクを解いているのか、特にこれらのニューラル言語モデルが自然言語の統語知識をどれほど有しているのかについては多くが分かっていない。

このような背景から、近年ではニューラル言語モデルが実際にどれほどの統語知識を獲得しているのかを検証する研究が盛んに行われている [6, 7, 8, 9, 10, 11]。しかし、その多くは英語を中心とした欧米の限られた言語を対象とした検証に終始し、性質の異なる言語に対象を拡張した検証は多くは行われておらず [12, 9]、中でも幅広い統語現象を扱った包括的な検証はごく限られた言語でしか行われていない [13, 14, 15, 16]。

本研究では、そのような状況に鑑みて、既存のデータセットの問題点を解決しつつ、幅広い統語現象について言語モデルの統語的評価を行う日本語データセットである JCoLA (Japanese Corpus of Linguistic Acceptability) を構築し、それを用いた言語モデルの性能検証を行う。

2　言語モデルの統語的評価

容認性判断課題とは、提示された文が容認可能（文法的）か容認不可能（非文法的）かの判断を下す課題である。近年、この容認性判断課題を用いた言語モデルの統語知識の検証が多く行われている。こうした研究の端緒となった Linzen et al. (2016) [6] では、言語モデルが英語の主述の一致 (subject-verb agreement) を捉えることができるかを以下のようなミニマルペアを用いて検証し、実際に LSTM 言語モデルが主述の一致を一定程度解くことができる統語知識を有していることが確認された。

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

このような文脈から、より広範囲の統語現象を扱い、かつデータの規模としても大きいデータセットの開発も進んでいる。Warstadt et al. (2019) [13] は、英語を対象とした大規模な言語理解ベンチマークである GLUE [2] に含まれるデータセットの一つとして、理論言語学のジャーナル論文や教科書から例文を抽出し、言語モデルの 2 値分類性能をテストする大規模データセットである CoLA (Corpus of Linguistic Acceptability) を構築したが、CoLA を用いて言語モデルの統語知識を検証するためには、言語モデルとは別に容認性の予測を出力するための分類器等を学習する必要があった。したがって、このような方法では得られた予測の良し悪しが言語モデルがテキストから学習した内部表現によるものなのか、分類器を教師ありで学習する際に獲得した表現によるものなのかが必ずしも明確ではないとい

う問題があることが指摘されている。この問題点を克服すべく、Warstadt et al. (2020) [14] は、島の制約 (island constraint) や動詞の項構造 (argument structure) をはじめとした 12 の現象を扱うミニマルペアを自動生成してまとめた大規模データセット BLiMP (The Benchmark of Linguistic Minimal Pairs for English) を構築した。これにより、言語モデルが正例に対して、負例よりも高い尤度を付与する確率分布を学習できていれば、言語モデルはそのミニマルペアの容認度の違いを正しく捉えているという仮定のもと、個別の分類器を学習することなく、言語モデルの出力を直接用いてその統語知識の検証を行うことが可能となった。

一方で、以上の言語モデルの統語知識を検証する試みは、その大多数が英語を対象としたものである。一部の研究で対象を英語以外にも拡張した検証 [9, 12, 17, 18] が行われてはいるが、主述の一致等の限られた統語現象を対象にした検証にとどまっており、幅広い統語現象を対象とし、かつ英語以外で検証を行った研究は非常に限られている [16, 15]。特に日本語においては自然言語処理の分野で広く使われる言語モデルの統語知識を評価するベンチマークとなるようなデータセットは存在していない。[1)] このような状況では、言語モデルが英語等の一部の言語だけではなく、自然言語一般の統語現象を捉えられているかどうかについての明確な証拠を得ることはできない。

本研究では、そのような現状に鑑みて JCoLA (Japanese Corpus of Linguistic Acceptability) を構築し、それを用いて既存言語モデルの統語的評価を行う。また、JCoLA は理論言語学のジャーナル論文から抽出した容認度と統語現象のアノテーションが付いた例文（2,323 文）と、それをもとに構築したミニマルペア（369 ペア）を収録したデータセットである。したがって複雑な統語現象を扱い、かつミニマルペアの形で提示できるという特徴を併せ持っており、既存のデータセットの問題点を克服したデータセットとなっている（表 1）。

表 1　既存の統語的評価用大規模データセットと JCoLA

データセット	言語学論文から抽出	ミニマルペア	データサイズ
CoLA [13]	✓		10,657（文）
ItaCoLA [16]	✓		9,722（文）
BLiMP [14]		✓	67,000（ペア）
CLiMP [15]		✓	16,000（ペア）
JCoLA（本研究）	✓	✓	369（ペア）

1）　Futrell et al. (2019)[10] と概ね同じタイトルである Futrell et al. (2018)[19] においては、日本語の否定極性項目についての議論が行われている。

3　JCoLA の構築

3.1　データ収集

言語モデルが理論言語学で重要とされている統語現象を捉えられているかを検証するため、本研究では東アジア・東南アジア言語の言語学のジャーナルとして著名な JEAL (Journal of East Asian Linguistics) に 10 年間（2006 年から 2015 年）で掲載された 133 本の論文の中で、特に日本語の統語論を扱っている論文 28 本を対象とし、その論文で提示されている全てのデータポイント（2,323 文）を抽出した。ここでの「全てのデータポイント」は、脚注や付録を含む本文の全ての日本語の例文の中で、構造分析のために提示された例文を除いたもののことである。

3.2　タイプ分類

単純な全データポイントに対する正解率による比較に終始することなく、個別の統語現象ごとのモデル評価を可能にするため、前節で抽出した例文を統語現象のタイプによって分類した。本研究では、全データポイントを 3 つの粒度で分類する。分類の名称については、BLiMP [14] を参考にした。

まず、大分類として各データポイントが問題としている容認性判断の性質や、本文中での提示のされ方に基づいて 8 種類に分類した。

Acceptability
特定の解釈・文脈に依存しない容認性判断。

Interpretation
特定の文脈・解釈に依存する容認性判断。

Coreference
指示詞や照応形の共参照に関する容認性判断。

Lexical
統語現象一般ではなく、特定の語彙項目に関する容認性判断。

Footnote
論文の脚注で提示されている例文。

Appendix
論文の付録で提示されている例文。

Repeat
既に提示された例文の繰り返しであることが、本文で明示的に断られている例文。

Variation
既に提示された例文と比べ、理論構築に無関係な要素のみにしか違いがない例文。

次に、中分類として各データポイントがどのような統語現象を扱っているのかに基づく分類を行った (*phenomenon*)。中分類は、基本的に BLiMP における 12 の現象に Others を加えたものであるが、今回対象とする日本語のデータに合わせてその一部を変更した（付録 A)。また、データポイントが二つ以上の現象に分類されうると判断された場合には、*phenomenon*-2 を用意して分類した。ただし、言語モデルの統語現象ごとの評価の際には *phenomenon* の分類が優先される。

最後に、小分類として中分類 (*phenomenon*) よりさらに粒度の細かい、個別の統語現象ごとに 39 種類の分類を行った (*paradigm*)。これにより、エラー分析の際により粒度の細かい分析が可能となる。

3.3 ミニマルペアの作成

まず、前節でタイプ分類が行われたデータセットのうち、以下の全ての条件を満たすものを抽出する。

- 非文として提示されている（?や*などのマーキングがされている）もの。ただし、?などのマーキングがされつつも、本文中で正例としてみなされているものは除く。
- 大分類が variation、repeat、footnote、appendix のいずれでもないもの。
- 中分類が others でないもの。

次に、言語学の論文において提示された全ての負例には、対応する正例が存在するという仮定のもと [20]、以上により抜き出された負例のぞれぞれに対応する正例を、論文の中から採用するか、本文の内容を確認しつつ筆者が作例することにより構築した。この際、解釈により容認度が変わる例は、JCoLA に含めない不適切な例として、ミニマルペア構築の対象外とした。また、重複している例文や、語彙項目が異なるのみで検証対象が同じである例文も除外した。以上の手順により、合計で 369 ペアのミニマルペアが作成された（表 2、付録 B)。

4 実験

4.1 言語モデル

以上で構築した JCoLA のミニマルペアを用いて、Kuribayashi et al. (2021) [21] で学習済みの言語モデルを評価する。モデルの種類は、2 つの異なるサイズの GPT-2 言語モデル（Trans-LG: 400M パラメータ、Trans-SM: 55M パラメータ）と LSTM 言語モデル、さらに 3-gram、4-gram、5-gram 言語モデルの合計で 6 種類である。それぞれの言語モデルは、学習データ量[2)]・学習ステップ数（100、1,000、10,000、100,000)・ランダムシード（3 種類）を変えた複数の設定で学習された（ただし、n-gram 言語モデルは学習データ全体で学習された)。また、学習データとしてはニュース記事と日本語版ウィキペディアに含まれる約 5,000,000 文からなるコーパスが用いられた。[3)]

表 2 中分類ごとのミニマルペアの数

phenomenon	ミニマルペア数
Argument structure	151
Verbal agreement	68
Morphology	38
Ellipsis	24
Nominal structure	24
Binding	16
Quantifiers	16
Filler-gap	13
Island effects	12
NPI licensing	4
Control/raising	3
総計	369

4.2 評価方法

ミニマルペアの正例に負例よりも高い尤度を付与できた場合に正解とみなし、各言語モデルの JCoLA における正解率を算出する。その際、系列長の違いを考慮するために、Lau et al. (2017) [24] で提案された正規化関数の一つである、MeanLP を用いて尤度を算出する（$|X|$ はサブワード単位の文の長さ)。

$$MeanLP = \frac{\log p(X)}{|X|}$$

5 結果

全学習データを用い、100,000 学習ステップの学習を行った 3 種類のニューラル言語モデルと 5-gram 言語モデルの全ミニマルペアに対する正解率を図 1 にまとめた。全ての言語モデルにおいて、正解率は 70% 程度に留まっている。したがって、JCoLA は複

2) 各モデルについて、学習データのうち 1/100(SM) を用いて学習されたもの, 1/10(MD) を用いて学習されたもの, 全て (LG) を用いて学習されたものの 3 種類が用意された。

3) 学習コーパス内の各文は、MeCab [22] により国語研短単位に分割された上で、Byte Pair Encoding(BPE)[23] を用いてサブワード単位に分割されている (vocab=100,000、character coverage=0.9995)。

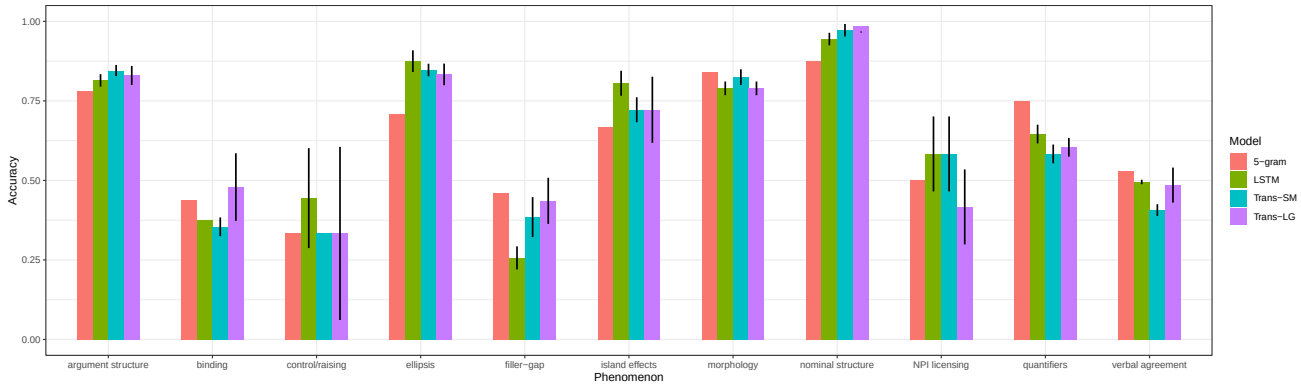

図 2　中分類ごとの正解率。LSTM・Trans-SM・Trans-LG は、全学習データ・100,000 学習ステップの設定で学習されたものである。エラーバーは 3 つのランダムシードの標準偏差である。

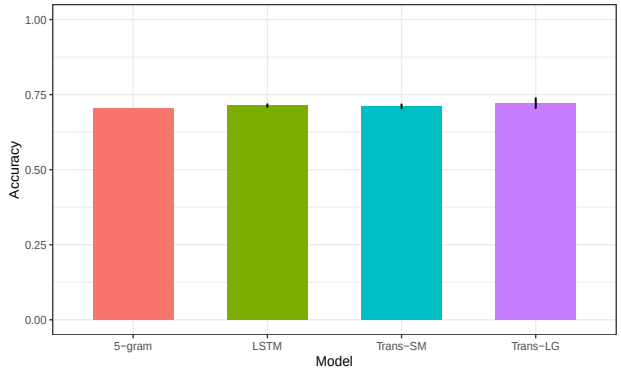

図 1　全ミニマルペアに対する正解率。LSTM・Trans-SM・Trans-LG は、全学習データ・100,000 学習ステップの設定で学習されたものである。エラーバーは 3 つのランダムシードの標準偏差である。

雑な統語現象の理解が求められ、ゆえに既存の言語モデルではその容認度の差を捉えるのが難しいミニマルペアを収録したデータセットであることが示唆される。

また、各中分類 (*phenomenon*) ごとの正解率を見ると、言語モデルが捉えられている統語現象と必ずしも捉えられていない統語現象が存在することが明らかになった（図 2)。まず、語の活用等を扱う morphology や名詞句内の構造を扱う nominal structure 等においては、5-gram 言語モデルを含め全ての言語モデルが高い精度を示したが、これは (2) のような比較的狭い範囲の情報のみを用いて正解することが可能な例が含まれていることが要因と考えられる。

(2)　a. 私が昨日**見た人**は素敵だった。
　　b:*私が昨日**見たの人**は素敵だった。

一方、filler-gap や verbal agreement では各モデル精度が低下しているが、これらは先ほどより長距離の依存関係を伴う統語現象である。verbal agreement の例としては、以下のような主語敬語 (subject honorification) の例がある。

(3)　a. **伊藤先生が**メアリーを**お褒めになった**。
　　b:***私が**メアリーを**お褒めになった**。

(3) を正解するためには、主語と述語の一致関係を捉え、「お褒めになった」という表現が、一人称主語では許されないということを正しく理解する必要があるが、このような日本語の一致現象を言語モデルは必ずしも捉えることができないということが明らかになった。

以上より、英語を中心とした欧米の言語を対象とした既存研究では、一定の統語知識を備えていることが確認され、かつ様々な下流タスクで高い性能を発揮する LSTM・Transformer ベースの言語モデルも、N-gram 言語モデルが学習する以上の統語知識を必ずしも得ることができていないことが明らかになった。特に、日本語特有の表現である敬語表現等を含め、比較的長距離の依存関係を伴う統語現象についての精度が低いことが確認された。

6　おわりに

本研究では、既存のデータセットの問題点を解決しつつ、日本語を対象とした統語的評価のための初めてのデータセットである JCoLA を構築した。JCoLA を用いた評価では、LSTM・Transformer ベースのモデルも長距離の依存関係を伴う統語現象を捉えきれないことが示唆された。JCoLA を用いたより詳細な言語モデルの統語的評価、及びデータセットの規模の拡大は今後の課題としたい。

謝辞

本論文は、筆者が東京大学の卒業研究として行った内容を記したものです。また、本研究は JST さきがけ JPMJPR21C2 の支援を受けたものです。

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł Ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30, pp. 5998–6008. Curran Associates, Inc., 2017.

[2] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In **International Conference on Learning Representations**, 2019.

[3] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. **CoRR**, Vol. abs/1905.00537, , 2019.

[4] Noam Chomsky. **Syntactic structures**. Mouton, 1957.

[5] Martin B H Everaert, Marinus A C Huybregts, Noam Chomsky, Robert C Berwick, and Johan J Bolhuis. Structures, not strings: Linguistics as part of the cognitive sciences. **Trends Cogn. Sci.**, Vol. 19, No. 12, pp. 729–743, December 2015.

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

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

[8] Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 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**, pp. 211–221, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[9] Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 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)**, pp. 1195–1205, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[10] Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 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)**, pp. 32–42, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[11] Rui P Chaves. What don't RNN language models learn about Filler-Gap dependencies? **Proceedings of the Society for Computation in Linguistics**, Vol. 3, No. 1, pp. 20–30, 2020.

[12] Shauli Ravfogel, Yoav Goldberg, and Francis Tyers. 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**, pp. 98–107, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[13] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. **Transactions of the Association for Computational Linguistics**, Vol. 7, pp. 625–641, November 2019.

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

[15] Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 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**, pp. 2784–2790, Online, April 2021. Association for Computational Linguistics.

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

[17] Aixiu An, Peng Qian, Ethan Wilcox, and Roger Levy. 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)**, pp. 2888–2899, Hong Kong, China, November 2019. Association for Computational Linguistics.

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

[19] Richard Futrell, Ethan Wilcox, Takashi Morita, and Roger Levy. RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency. September 2018.

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

[21] Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 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)**, pp. 5203–5217, Online, August 2021. Association for Computational Linguistics.

[22] Takumitsu Kudo. Mecab : Yet another part-of-speech and morphological analyzer. 2005.

[23] Taku Kudo and John Richardson. 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**, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[24] Jey Han Lau, Alexander Clark, and Shalom Lappin. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. **Cogn. Sci.**, Vol. 41, No. 5, pp. 1202–1241, July 2017.

A 中分類 (*phenomenon*) の基準

ARGUMENT STRUCTURE
動詞の項構造に関わる容認性判断。
例えば、項の順序や格標示に関わる容認性判断等が含まれる。
(1) a. 太郎が花子に会う。
b. *太郎が花子を会う。 (Takahashi, 2006)

BINDING
名詞句の束縛関係に関する容認性判断。
例えば、相互代名詞の同一指示解釈に関する容認性判断等が含まれる。
(2) a. 彼ら$_i$にお互い$_i$の母親からそのことを伝えた。
b. *お互い$_i$の母親から彼ら$_i$にそのことを伝えた。
(Kishimoto, 2012)

FILLER-GAP
移動した構成素と移動元の空所の依存関係に関する容認性判断。
例えば、wh 疑問文や分裂文に関わる容認性判断等が含まれる。
(3) a. 何を誰も読まなかったの？
b. ?*誰も何を読まなかったの？ (Tomioka, 2009)

ELLIPSIS
文中の要素の省略可能性に関する容認性判断。
例えば、動詞句や名詞句を省略できる環境に関する容認性判断等が含まれる。
(4) a. 晴れの日は良いが、雨の日は落ち込む。
b. *晴れの日は良いが、雨のは落ち込む。
(Saito et al., 2008)

ISLAND EFFECTS
島の制約に関する容認性判断。
例えば、複合名詞句制約 (complex NP constraint) や付加詞条件 (adjunct condition) に関する容認性判断等が含まれる。
(5) a. 太郎が昨日花子に会った人を探している。
b. *太郎が昨日会った人を探しているのは花子にだ。
(Takahashi, 2006, 正例は筆者が作例)

NPI LICENSING
否定極性項目 (negative polarity items, NPIs) の出現環境に関する容認性判断。
例えば、「誰」と「も」で構成される表現に関する容認性判断等が含まれる。
(6) a. 今回は誰から寄付を呼びかけもしなかった。
b. *今回は誰が寄付を呼びかけもしなかった。
(Kishimoto, 2012)

QUANTIFIERS
数量詞 (quantifiers) の分布に関する容認性判断。
例えば、遊離数量詞 (floating quantifiers) に関する容認性判断等が含まれる。
(7) a. 太郎が CD を友達に 2 人送った。
b. *太郎が友達に 2 人 CD を送った。
(Tsujioka, 2011)

CONTROL/RAISING
コントロール (control) や繰り上げ (raising) 構文に関する容認性判断。
例えば、コントロール動詞と V-V 複合動詞を形成できる動詞の種類に関する容認性判断等が含まれる。
(7) a. 座り損ねる。
b. *転び損ねる。 (Kishimoto, 2012)

VERBAL AGREEMENT
BLiMP では、主語と動詞の数の一致に関する例文を、SUBJECT-VERB AGREEMENT としてまとめている。JCoLA では、より一般に主語の性質が動詞の形態に反映される現象や、動詞が主語の性質に制約を与えるような現象に関する容認性判断を含む中分類として VERBAL AGREEMENT を採用した。
例えば、主語敬語 (subject honorification) に関する容認性判断等が含まれる。
(8) a. 伊藤先生がメアリーをお褒めになった。
b. *メアリーが伊藤先生をお褒めになった。
(Kishimoto, 2012)

NOMINAL STRUCTURE
BLiMP では、限定詞 (determiner) と名詞の一致に関する例文を、DETERMINER-NOUN AGREEMENT としてまとめている。JCoLA では、より一般に名詞句の内部構造に関わる容認性判断を含む中分類として、NOMINAL STRUCTURE を採用した。
例えば、「の」の分布に関する容認性判断等が含まれる。
(9) a. 私が昨日見た人。
b. *私が昨日見たの人。 (Saito et al., 2008)

MORPHOLOGY
BLiMP では、動詞の過去分詞の活用が正しく行われているかに関する例文を、IRREGULAR FORMS としてまとめている。JCoLA では、より幅広く形態論に関する容認性判断を含む中分類として MORPHOLOGY を採用した。
例えば、形容動詞の活用に関する容認性判断等が含まれる。
(10) a. その粒子は計測可能だとジョンは思っている。
b. *その粒子は計測可能なとジョンは思っている。
(Sudo, 2015)

OTHERS
BLiMP を元にして作成した以上の中分類のいずれにも当てはまらない例文。この場合は、小分類 (paradigm) にて該当する統語現象に基づいて分類している。例えば、比較表現に関する容認性判断等が含まれる。
(11) a. ジョンはベッドが長い以上に背が高い。
b. *ジョンはベッドが長いより背が高い。
(Sudo, 2015)

B JCoLA のミニマルペア例

phenomenon	paradigm	負例	正例
ARGUMENT STRUCTURE	case	太郎がその本に読んだ。	太郎がその本を読んだ。
	passive	家が大工に建てさせられた。	大工が家を建てさせられた。
	scrambling	最も太郎が面白かった人を取材した。	太郎が最も面白かった人を取材した。
	animacy	ジョンにはお金が居る。	ジョンには兄弟が居る。
	aspect	太郎がプールで 1 時間で泳いだ。	太郎がプールで 1 時間泳いだ。
	internal argument	ジョンが息子を自殺した。	ジョンが息子を自慢した。
VERBAL AGREEMENT	subject honorification	健が山田先生にお会いになった。	健に山田先生がお会いになった。
	person constraint	あなたは寒いです。	私は寒いです。
BINDING	weak crossover	初めてそいつに会う人が呪すのは誰をですか？	初めて会う人が呪すのは誰をですか？
	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？	花子が誰にそいつが書いた論文を修正させたのですか？
	anaphor	自分$_i$の先生$_i$には学生がわかる。	先生$_i$には自分$_i$の学生がわかる。
	reciprocal	お互い$_i$の母親から彼ら$_i$にそのことを伝えた。	彼ら$_i$にお互い$_i$の母親からそのことを伝えた。
ELLIPSIS	nominal ellipsis	晴れの日は良いが、雨のは落ち込む。	晴れの日は良いが、雨の日は落ち込む。
	adjunct ellipsis	太郎がその理由で解雇された後、花子も解雇された。	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。
	parasitic-gap	初めて会う人が誰を貶しますか？	初めて会う人が呪すのは誰をですか？
MORPHOLOGY	part of speech	子供そう。	美味しそう。
	idiom	太郎の忠告は花子には糠にも釘だった。	太郎の忠告は花子には糠に釘だった。
	reflexive	強い地震のため建物が自壊をした。	強い地震のため建物が自壊した。
	inflection	それは計測可能だ粒子だ。	それは計測可能な粒子だ。
	nominalization	原稿に手の入れ方は人それぞれだ。	原稿への手の入れ方は人それぞれだ。
	honorification	伊藤先生からそのことを話しておいでになる。	伊藤先生からそのことをお話になっている。
QUANTIFIERS	floating quantifiers	学生が家を 4 人買った。	学生が 4 人家を買った。
	universal quantifiers	みんながみんな大学へ行かない。	みんながみんな大学へ行く訳ではない。
	classifier	3 本ずつのその鉛筆。	その 3 本ずつの鉛筆。
	negation	ジョンはメアリーが賢い以上に賢くない。	ジョンはメアリーが賢い以上に賢い。
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。	太郎が昨日花子に会った人を探している。
	adjunct island	太郎が読んだから花子が怒ったのはその本をだ。	太郎がその本を読んだから花子が怒った。
	specificity island	ジョンはそのメアリーより高い指輪を買った。	ジョンはメアリーより高い指輪を買った。
	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。	ジョンはメアリーが雇ったより賢い人を見つけた。
	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。	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思っていたのの欠陥を指摘した。
FILLER-GAP	intervention effects	誰も何を読まなかったの？	何を誰も読まなかったの？
	relative clause	山田先生はこの本をなったことはお読みだ。	山田先生はこの本をお読みになった。
	cleft	山田先生がなったのはこの本のお読みだ。	山田先生がこの本をお読みになった。
	resumptive pronoun	トムがそれらを食べたことが明らかな竿は大きかった。	トムが食べたことが明らかな竿は大きかった。
NPI licensing	NPI	今回は誰から寄付を呼びかけもしなかった。	今回は誰から寄付を呼びかけもしなかった。
	NCI	ジョンがもし何も盗んだら、逮捕されるだろう。	ジョンがもし何か盗んだら、逮捕されるだろう。
Nominal structure	modifier	私が昨日見たの人は素敵だった。	私が昨日見た人は素敵だった。
	measure phrase	このビルは低さ 20 メートルある。	このビルは高さ 20 メートルある。
CONTROL/RAISING	subject control	転び損ねる。	座り損ねる。