日本語版CoLAの構築の舞台裏

染谷大河　大関洋平

東京大学

{taiga98-0809, oseki}@g.ecc.u-tokyo.ac.jp

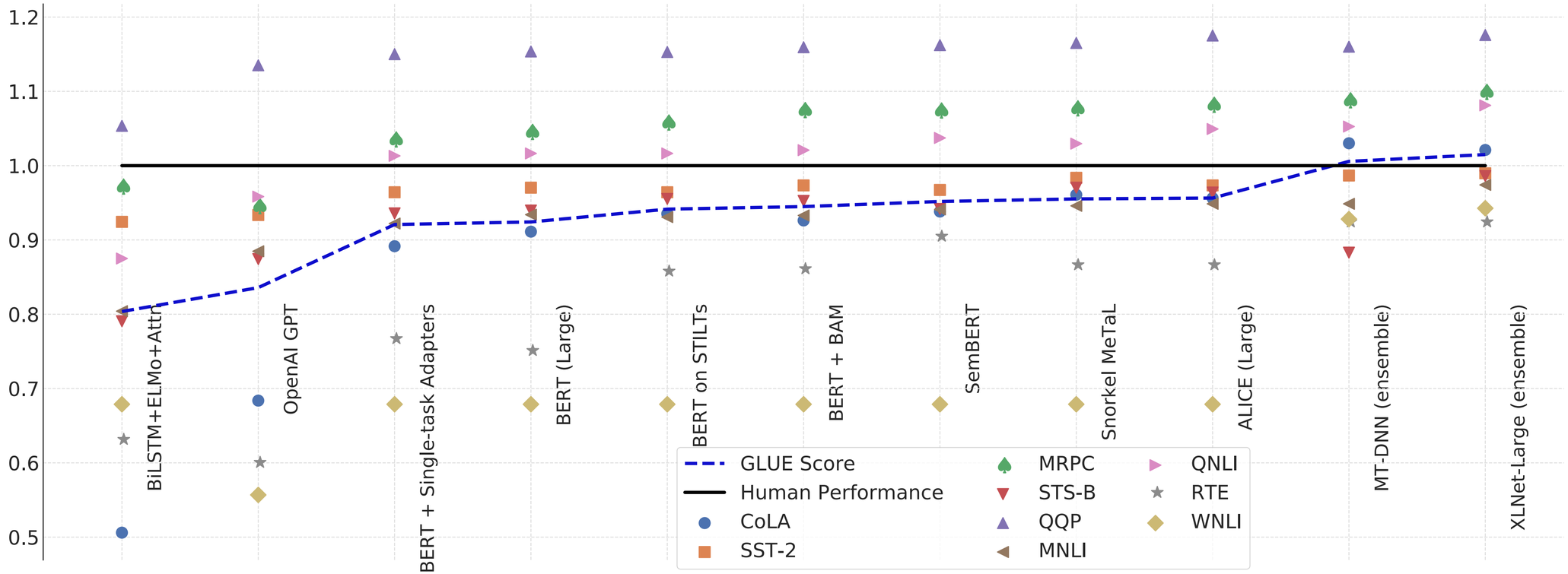
@NLP2022
Workshop on Japanese Evaluation Dataset

Outline

- 自己紹介
- 先行研究・JCoLA概要
- JCoLAの作成過程と苦労
- JCoLAのこれから

言語モデルの大規模データセットでの評価

- 大規模データセットを用いた評価（ベンチマーキング）
  - GLUE (General Language Understanding Evaluation) (Wang+, 2019)



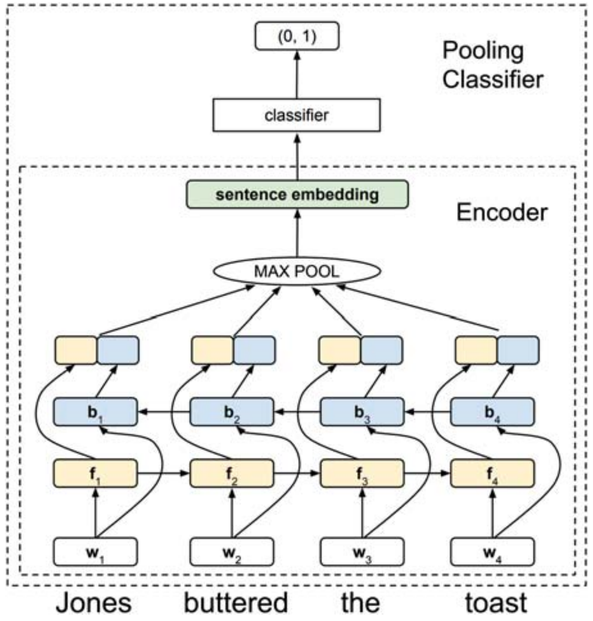
「統語知識」に重点を置いた評価

- Targeted Syntactic Evaluation (Linzen+, 2016)
  - CoLA (Corpus of Linguistic Acceptability) (Warstadt+, 2019)
  - BLiMP (Benchmark of Linguistic Minimal Pairs) (Warstadt+, 2020)

Phenomenon	N	Acceptable Example	Unacceptable Example
Anaphor agr.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*
Arg. structure	9	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
Binding	7	*Carlos said that Lori helped him.*	*Carlos said that Lori helped himself.*
Control/raising	5	*There was bound to be a fish escaping.*	*There was unable to be a fish escaping.*
Det.-noun agr.	8	*Rachelle had bought that chair.*	*Rachelle had bought that chairs.*
Ellipsis	2	*Anne's doctor cleans one important book and Stacey cleans a few.*	*Anne's doctor cleans one book and Stacey cleans a few important.*
Filler-gap	7	*Brett knew what many waiters find.*	*Brett knew that many waiters find.*
Irregular forms	2	*Aaron broke the unicycle.*	*Aaron broken the unicycle.*
Island effects	8	*Whose hat should Tonya wear?*	*Whose should Tonya wear hat?*
NPI licensing	7	*The truck has clearly tipped over.*	*The truck has ever tipped over.*
Quantifiers	4	*No boy knew fewer than six guys.*	*No boy knew at most six guys.*
Subject-verb agr.	6	*These casseroles disgust Kayla.*	*These casseroles disgusts Kayla.*

関連研究：Warstadt et al. (2019)

- Corpus of Linguistic Acceptability (CoLA)
  - 統語論の論文や教科書から例文を収集（１万文程度）
  - モデルの出力から容認度を算出するのに工夫が必要



Label	Sentence
*	The more books I ask to whom he will give, the more he reads.
✓	I said that my father, he was tight as a hoot-owl.
✓	The jeweller inscribed the ring with the name.
*	many evidence was provided.
✓	They can sing.
✓	The men would have been all working.
*	Who do you think that will question Seamus first?
*	Usually, any lion is majestic.
✓	The gardener planted roses in the garden.
✓	I wrote Blair a letter, but I tore it up before I sent it.

$$\text{Word LP Min-N} = \min_{N}\left\{-\frac{\log p_m(w)}{\log p_u(w)}, w \in \xi\right\}$$

$$\text{Word LP Mean} = \frac{\sum_{w\in\xi} -(\log p_m(w)/\log p_u(w))}{|\xi|}$$

$$\text{Word LP Mean-Q1} = \frac{\sum_{w\in \mathrm{WL}_{Q1}} -(\log p_m(w)/\log p_u(w))}{|\mathrm{WL}_{Q1}|}$$

$$\text{Word LP Mean-Q2} = \frac{\sum_{w\in \mathrm{WL}_{Q2}} -(\log p_m(w)/\log p_u(w))}{|\mathrm{WL}_{Q2}|}$$

関連研究：Warstadt et al. (2020)

- Benchmark of Linguistic Minimal Pairs (BLiMP)
  - ミニマルペアを自動生成し、統語現象ごとに分類
  - 言語モデルの出力を直接使用できる（正文により高い尤度を付与できるか）

Phenomenon	N	Acceptable Example	Unacceptable Example
Anaphor agr.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*
Arg. structure	9	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
Binding	7	*Carlos said that Lori helped him.*	*Carlos said that Lori helped himself.*
Control/raising	5	*There was bound to be a fish escaping.*	*There was unable to be a fish escaping.*
Det.-noun agr.	8	*Rachelle had bought that chair.*	*Rachelle had bought that chairs.*
Ellipsis	2	*Anne's doctor cleans one important book and Stacey cleans a few.*	*Anne's doctor cleans one book and Stacey cleans a few important.*
Filler-gap	7	*Brett knew what many waiters find.*	*Brett knew that many waiters find.*
Irregular forms	2	*Aaron broke the unicycle.*	*Aaron broken the unicycle.*
Island effects	8	*Whose hat should Tonya wear?*	*Whose should Tonya wear hat?*
NPI licensing	7	*The truck has clearly tipped over.*	*The truck has ever tipped over.*
Quantifiers	4	*No boy knew fewer than six guys.*	*No boy knew at most six guys.*
Subject-verb agr.	6	*These casseroles disgust Kayla.*	*These casseroles disgusts Kayla.*

問題点①：生起確率と容認性判断の橋渡し

- CoLAのようにデータセットがペアになっていない場合、言語モデルの出力する文の生起確率から直接容認性の予測を行うことはできない。
  - 言語モデルとは別に２値分類を行う分類器を学習する必要がある。
  - 予測の良し悪しが何に起因するものなのかわからない

    → ミニマルペアによる比較が適当

問題点②：自動生成データの質

- BLiMPのように自動生成したデータには、単純なパターンも含まれる。
- 理論言語学の論文で問題となるような、より複雑な判断を必要とする例文についての検証はできていない（Class III judgments, see Marantz, 2005; Linzen and Oseki, 2018)。

  → 言語学の論文等から例文を構成するべき？

Phenomenon	N	Acceptable Example	Unacceptable Example
Anaphor agr.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*
Arg. structure	9	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
Binding	7	*Carlos said that Lori helped him.*	*Carlos said that Lori helped himself.*
Control/raising	5	*There was bound to be a fish escaping.*	*There was unable to be a fish escaping.*
Det.-noun agr.	8	*Rachelle had bought that chair.*	*Rachelle had bought that chairs.*
Ellipsis	2	*Anne's doctor cleans one important book and Stacey cleans a few.*	*Anne's doctor cleans one book and Stacey cleans a few important.*
Filler-gap	7	*Brett knew what many waiters find.*	*Brett knew that many waiters find.*
Irregular forms	2	*Aaron broke the unicycle.*	*Aaron broken the unicycle.*
Island effects	8	*Whose hat should Tonya wear?*	*Whose should Tonya wear hat?*
NPI licensing	7	*The truck has clearly tipped over.*	*The truck has ever tipped over.*
Quantifiers	4	*No boy knew fewer than six guys.*	*No boy knew at most six guys.*
Subject-verb agr.	6	*These casseroles disgust Kayla.*	*These casseroles disgusts Kayla.*

問題点③：対象となっている言語・統語現象の偏り

	English	Russian	Hebrew	French	Basque	Italian	Chinese	Japanese
Subject-verb Agreement	Linzen et al. (2016), Gulordava et al. (2018), Marvin & Linzen (2018), An et al. (2019), Warstadt et al. (2019), Futrell et al. (2018), Bernardy & Lappin (2017), Mueller et al. (2020)	Gulordava et al. (2018), Mueller et al. (2020)	Gulordava et al. (2018), Mueller et al. (2020)	Gulordava et al. (2018), Mueller et al. (2020), An et al. (2019)	Ravfogel et al. (2018)	Gulordava et al. (2018), Mueller et al. (2020), Trotta et al. (2021)	Xiang et al. (2021)	
Filler-gap/ Island effects	Wilcox et al. (2018, 2019), Chaves (2020), Da Costa & Chaves (2020), Chowdhury & Zamparelli (2018, 2019), Warstadt et al. (2019)					Trotta et al. (2021)	Xiang et al. (2021)	
Anaphora/ Binding	Marvin & Linzen (2018), Warstadt et al. (2019), Futrell et al. (2018)					Trotta et al. (2021)	Xiang et al. (2021)	
Negative Polarity Items	Wilcox et al. (2019), Marvin & Linzen (2018), Futrell et al. (2018), Jumelet & Hupkes (2018)							
Argument structure	Warstadt et al. (2019), Kann et al.(2019), Chowdhury & Zamparelli (2019)						Xiang et al. (2021)	

JCoLA (Japanese Corpus of Linguistic Acceptability)

- 理論言語学のジャーナル論文から例文を抽出

  -> Class III の容認性判断についての評価が可能

- 例文をもとにミニマルペアを作成

  -> 言語モデルの出力を直接用いて評価が可能

- これまで大規模かつ、多様な統語現象を扱うデータセットが構築されて来なかった「日本語」を対象にしたデータセット

  -> これまで検証が進んでいなかった言語での検証が可能

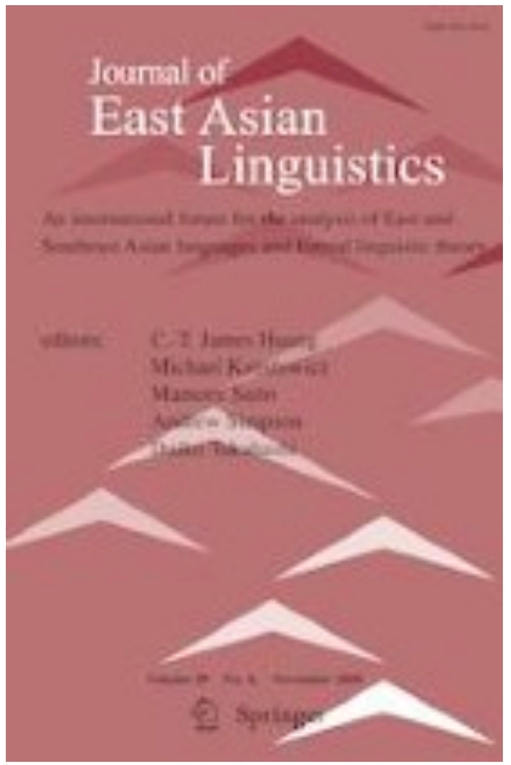
Outline

- 自己紹介
- 先行研究・JCoLA概要
- JCoLAの作成過程と苦労
- JCoLAのこれから

データ収集

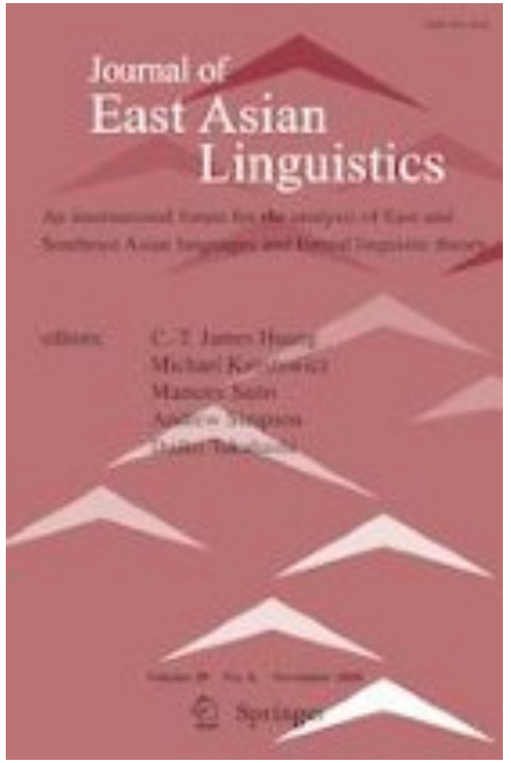
- Journal of East Asian Linguistics (2006 - 2015)
  - 日本語統語論に関する論文 （28本）から、
    合計で2323文を抽出。

- 脚注・付録を含むが、構造分析のために提示された
  例文等は除いて抽出した。



データ収集の裏側

- Journal of East Asian Linguistics (2006 - 2015)
  - 収録されている全133本の論文の内容を確認しつつ選んだ
    日本語統語論に関する論文 （28本）から、
    合計で2323文を人手で一つずつ抽出。
    （グロス・日本語訳を含む）



- 脚注・付録を含むが、構造分析のために提示された
  例文等は、論文本文を丁寧に読み、構造としての
  提示であると確認した上で除いて抽出した。

タイプ分類

- 個別の統語現象ごとにモデルを評価するために、
  データポイントを３つの粒度で分類

- 大分類（type）
  - 容認性判断の性質や本文中の提示のされ方に基づく分類
- 中分類（phenomenon）
  - 扱っている統語現象ごとの分類
    - 二つ以上に分類される場合は中分類2（phenomenon-2）を使う。
- 小分類（paradigm）
  - 中分類よりもさらに細かい分類

大分類（type）

- Acceptability -> 純粋な容認性判断
- Interpretation -> 特定の文脈・解釈の元での容認性判断
- Coreference -> 指示詞等の同一指示解釈を問題としている容認性判断
- Lexical -> 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 -> 既出例文の繰り返しである例文
- Variation -> 理論構築に無関係な要素のみにしか違いがない例文

中分類（phenomenon, phenomenon-2）

- 統語現象ごとに11の中分類に分類
  - BLiMPの12分類+Others*をもとに、日本語のデータに合わせて分類を追加・削除して構成

統語現象	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

*Others: 比較表現等が含まれる。後に説明するようにミニマルペア作成時には削除してある。

中分類（phenomenon, phenomenon-2）の裏側

- 統語現象ごとに11の中分類に分類
  - BLiMPの12分類+Others*をもとに、日本語のデータに合わせて分類を追加・削除して構成
  - phenomenonごとの偏りがなるべく小さくなるように、粒度や項目をアジャイルに変更しながら作成した
    - 結果としてやはり何周も見る必要がある

統語現象	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

*Others: 比較表現等が含まれる。後に説明するようにミニマルペア作成時には削除してある。

小分類（paradigm）

- 中分類の下位分類（39種類）
- 例：
  - island effects
    - complex NP island
    - adjunct island
    - specificity island
    - negative island
    - factive island

小分類（paradigm）の裏側

- 中分類の下位分類（39種類）
  - こちらも粒度等を都度調整しながら分類
- 例：
  - island effects
    - complex NP island
    - adjunct island
    - specificity island
    - negative island
    - factive island

ミニマルペアの作成

- まず、以下の条件を満たす例文（負例）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、負例に対しては対応する正例が存在する（cf. Sprouse et al., 2013）
  という前提のもと、以上で抽出した負例のそれぞれに対応する正例を作成

ミニマルペアの作成の裏側

- それぞれの負例に対して、
  1. 論文中提示されている正例
  2. 本文を読みつつ対応する正例を作例

  のいずれかを行うことにより、正例を作成した。
  - 論文の意図を汲み取りつつも、系列長がなるべく変化しないように作例
    - とはいえ、扱う統語現象の性質上系列長がずれてしまうものもある
      - e.g. 省略現象など

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貶すのは誰をですか？	初めて会う人が貶すのは誰をですか？
	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？	初めて会う人が貶すのは誰をですか？
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芋は大きかった。	トムが食べたことが明らかな芋は大きかった。
NPI licensing	NPI	今回は誰が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。

Outline

- 自己紹介
- 先行研究・JCoLA概要
- JCoLAの作成過程と苦労
- JCoLAのこれから

JCoLAの問題点

- ベンチマークとしてはデータセットがやや少ない
  - CoLA: 10,657（文）
  - ItaCoLA: 9,722（文）
  - BLiMP: 67,000（ミニマルペア）
  - CLiMP: 16,000（ミニマルペア）
  - JCoLA: 369（ペア）・2,323（文）

- 各統語現象ごとに平均40ペアほどを確保できている

今後の展望

- より大規模なデータセットの構築
  - 既に化学論文からの図表の抽出等に使われている、
    情報抽出の技術の言語学論文への応用（NLP 2022）
- 容認度のアノテーションの妥当性？
  - 一部、論文内のアノテーションに違和感
    - → 人間の容認性判断を取り検証する予定
- データの見直し、整理
  - → 近日中に公開予定

ご清聴ありがとうございました。
ご質問・コメント等よろしくお願いします。