日本語話者の項省略判断に関するアノテーションとモデリング

石月由紀子 [1,4]　栗林樹生 [1,2]　松林優一郎 [1,4]　笹野遼平 [3,4]　乾健太郎 [1,4]
[1] 東北大学　[2]Langsmith 株式会社　[3] 名古屋大学　[4] 理化学研究所
yukiko.ishizuki.p7@dc.tohoku.ac.jp
{kuribayashi,y.m,kentaro.inui}@tohoku.ac.jp sasano@i.nagoya-u.ac.jp

概要

日本語ではしばしば主格や目的格などの項が省略される．項の省略の可否は構文の妥当性といった制約や母語話者の選好によって判断される場合があり，母語話者は省略の容認度を潜在的に規定していることが想定される．しかしながら，そのような母語話者の判断のモデリングは，既存の省略解析の枠組みには含まれてこなかった．本研究では，読み手の省略判断に関するデータの収集と，そのデータを用いた省略判断モデルの構築を行い，母語話者と自然言語処理モデルの省略判断についてその傾向を調査した． 収集したデータは BCCWJ に対する差分データとして公開予定[1]である．

1　はじめに

日本語は項省略がしばしば発生する言語である．例えば（1a）では，「マックが」を省略した方が自然であり，一方（1b）では「小塚原の刑場に」を省略しないほうが自然である[2]．

(1a) マックは椅子をつかみ、前後逆に置いた。そこに**マックが**またがり、無言で画像に見入った。

(1b) 一六六〇年頃（万治年間）、幕府は、牢死者や刑死者を弔うため、本所に回向院を立てさせた。さらに一六六七年（寛文七）、その別院として、**小塚原の刑場に**建てられたのが、この回向院だそうな。

本研究ではこのような項の省略判断について，(i) ある項が表出・省略されているべきかという人間の読み手が下す判断のデータ収集・分析と，(ii) 自然言語処理モデルを用いた省略判断予測モデルの構築を行う．項の省略の可否に関しては統語的な制約などについて少数の事例のもと議論されてきた [2, 3, 4, 5] が，本研究ではそれらの研究と相補的に，実際に人間の読み手が文章を読んだ際に判断する省略の容認度について比較的大規模なコーパスにアノテーションする．計算モデルによる省略判断の予測については，日本語の読みやすさの自動判定・文章執筆支援（項を省略・表出すべきかの自動推薦）といった応用的出口を見据えるとともに，既存の言語モデルが談話的な現象である省略判断を行えるのかという学術的な分析として位置づける．

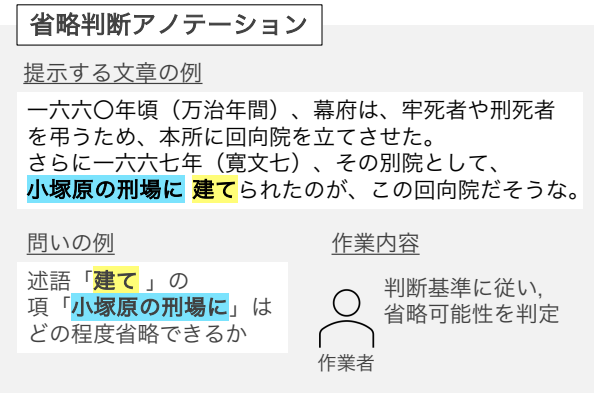


図 1　アノテーション作業の外観

日本人大学生 5 名を対象にしたデータ収集の結果，人間の作業者間での省略判断は概ね一致し，個人差が生じる事例は限定的であることが分かった．また，省略の是非の判断根拠もあわせて追跡する方式で作業を行い，分析の結果，統語的容認性や省略要素の復元可能性といった言語の制約的な部分では人間の省略判断が比較的一致し，それらでは説明できない選好の部分では判断が揺れる傾向があった．

言語モデルによる省略判断の予測に関する実験では，日本語 BERT を用いたモデルを作成し，その性能を調査した．言語の制約的な要因で決まる事例については，モデルは概ね正しい判断を下し，一方で文の自然さといった選好によって省略の可否が決定する事例については，言語モデルの予測性能が人間

1）https://github.com/cl-tohoku/JAOJ
2）(1a) は BCCWJ-DepParaPas[1] 内の 00003_A_PB59_00001 の文章，(1b) は 00004_A_PB22_00002 の文章より引用した．

表 1　省略の程度を説明するものとして想定される因子

カテゴリ	因子	内容	例（対象の項は太字で示す.）
制約	1. 解釈可能性	項が省略された場合でも補われるべき情報が定まる	コウモリは夜行性で、昼間は洞窟の中で**コウモリが**ぶら下がって休んでいます。
制約	2. 構文的妥当性	項が省略・表出されると文法的に不適切	果たして**誰に**聞けばよいものだろうか？
制約	3. 述語項構造以外の解釈変化	省略によって助詞等が失われ，文意が変化する	**播磨であれば、**畿内にも近く気候も温暖で作物も豊かです。
制約	4. 文脈上の重要さ	慣習的に省略される項であり，その項が定まらなくても読みに影響が生じない	このヒバ材には滅菌作用のほかに、（中略）血圧や心拍数を落ち着かせ、**人に**安らぎを与える効果もあると言われています。
選好	5. その他の自然さ	文章の流暢さなどから省略・表出させるのが自然	このことを、**私たちは**もう一度謙虚に考えてみる必要があるのではないでしょうか。

の作業者間に対して劣るという結果が得られた．

2　省略判断アノテーション

2.1　タスク設定とアノテーション基準

日本語母語話者の項省略の容認度とその一致の程度を調査することを目的とし，省略判断のデータを収集した．図 1 にアノテーションタスクの外観を示す．作業者には文章内のある述語と，それと同一文内にある述語の項が 1 つ示される．作業者は，指定された述語の項の表現が省略可能かを判断する．なおこの項は (i) 元の文章でも表出している場合と (ii) 元の文章では省略されているが特定の位置に復元されている場合の 2 通りがあり，作業者は判断時にどちらの事例かを知ることはできない．アノテーション対象の述語-項の収集方法は 2.3 節で説明する．

アノテーション基準は，省略に関する言語の制約的性質を予備的に調査した結果を踏まえ，実際に著者らで予備アノテーション作業を行いながら設計した．まず，判断の程度を定義するにあたり，省略判断を，言語の制約的な性質に基づく強い判断と，作業者のある種の選好に基づくものとに分け，これに「省略」「表出」の方向を示した 4 値を省略の程度として定めた．最終的な人間の判断が制約的な要因によって説明される場合は「省略（制約）」及び「表出（制約）」と定義し，選好に基づく要因によってのみ説明される場合を「省略（選好）」及び「表出（選好）」と定義した．さらに，制約的な要因としてどのような因子があるかについて，省略の潜在的な判断要因を著者内で議論し，少なくとも制約的な因子については著者間での判断に揺れがないことを基準として，表 1 のとおり因子を定めた．ただし，ここで例外的に因子 3 については，項が表出した文と省略された文のどちらを選ぶかが書き手の表現したい文意によってのみ定まる，読み手には判断不可能な例の存在が判明したため，新たに「判断不可」というラベルを追加し，最終的に 5 つの程度について作業者に判断させた．

実際の作業では，作業者は 5 つの段階を直接判断するのではなく，表 1 の各因子に対しての判断を下すことによって間接的に 5 つの段階を判断する．これによって，作業者がどの判断要因を用いて判断しているのかを分析できるようにした．例えば，1 節で挙げた例 (1a) については省略するのが自然であると判断されるが，判断の根拠は表 1 の 5 つの因子への判断として次のように記述できる: (因子 1) 項が省略された場合にも補われるべき情報が定まる，(因子 2) 項の有無は構文的正しさに影響しない，(因子 3) 省略に伴う助詞機能の喪失によって文意が変化しない，(因子 4) 慣習的に省略される項ではない，(因子 5) 文章の流暢さから省略するのが自然．最終的なラベルはこれらの因子への回答の組み合わせによって選択される．この場合は，制約に関する因子 (1-4) の組み合わせからは制約としての判断は下されず，選好に関する因子（5）で項を省略すべきと判断したため，最終的に付与される省略の容認度は「省略（選好）」となる．

実際の作業時には，その項がどのような潜在的要因を持ち，最終的にどの容認度のラベルが付与されるべきかを質問フローチャートに答えていく形で回答する．このフローチャートが因子に対する回答の組み合わせから 5 値の容認度ラベルへの対応を定めている．フローチャートの詳細は付録 A に示す．

2.2　作業手順とインターフェース

作業の具体的手順を説明する．回答には専用のツールを用いた（付録 B に記載）．作業画面上には，図 1 に概略したように，特定の述語・項が強調され

た文とその前方文脈が提示される．作業者は画面の上から順に文章を読解し，対象となる項がどの程度省略されるべきかを，その文を読んだ時点までの情報と事前に与えたアノテーション基準に沿って判断する．回答後に画面上の「次の問題」ボタンを押すと，次の対象述語-項までの文章が表示される．作業者は直前に読んだ文の続きから読み進め，対象述語まで読み進めたら問いに回答する作業を繰り返す．2.1 節で説明した通り，コーパス上での実際の文は省略判断時に提示されていないが，読み誤ったまま回答し続ける状況を防ぐため，判断終了後，次の事例に進む時点でコーパス上に本来記述されていた文が提示される．文内に問うべき項が複数存在する場合は，無作為に 1 つを作業対象とした．また，既に回答した事例を遡って回答し直すことは禁止した．

この作業を日本語母語話者 5 名により実施した．本研究では，作業者数が 5 名と少数のため，読解力の水準を揃える目的で同一大学の学生を選出した．事前にアノテーション対象外の文章を用いて訓練作業を実施し，アノテーション基準や作業内容に対する理解度を確認した後，本実施を行った．

2.3 アノテーション対象データ

述語項構造情報が付与されたコーパス，BCCWJ-DepParaPas [1] の書籍ドメイン 32 文章をアノテーション対象とした．対象となる述語-項ペアのサンプリング手法は付録 C に示す．同コーパスには文章中で省略されている項に照応関係及び格関係が付与されているため，省略された項が何かを特定可能である．このデータを用いて，項を文中に強制的に表出させた状態でその項の省略可能性を作業者に問う．この際，項が省略されていた場合は，文中のいずれかの位置に表出させて作業者に提示する必要がある．項をどこへ復元すべきかの情報はコーパスに付与されていないため，事前準備として省略された項をなるべく自然な位置に補った．作業手順は付録 D に記載する．最終的に 1,054 事例の述語-項ペアを収集し，これにコーパス上で既に項が表出しているものを加えた 2,392 事例を今回の対象とした．

3 アノテーション結果

作業者間の判断の一致度　作業者の省略判断の一致度について，Krippendorff の alpha[3] の値が 0.84 と高く，判断が概ね一致することが確認された [6]．

作業結果の集約とラベル分布の分析　得られた 5 名分の判断ラベルについて，省略判断の予測モデルを構築するため 1 つの代表ラベルに集約し，訓練・開発・評価データを作成した[4]．集約の際は，省略（制約）＜省略（選好）＜判断不可＜表出（選好）＜表出（制約）というラベル間の順序を仮定し，5 名の回答の中央値を判断の代表値とした．結果として，回答の集約時に中央値が「判断不可」に分類される事例は 1 事例のみであったため，今回の分析ではこの事例を除外し，以降は省略の程度を省略（制約），省略（選好），表出（選好），表出（制約）の 4 値として分析を行う．

表 2　データセット内の判断ラベル分布

		省略		表出	
分割	データ数	制約	選好	選好	制約
訓練	1,475	29.4%	15.7%	11.5%	43.8%
開発	457	31.4%	8.1%	14.2%	46.2%
評価	459	31.6%	11.8%	9.6%	47.1%
全体	2,391	30.2%	13.3%	11.6%	44.9%

集約後のデータセット内のラベル分布を表 2 に示す．ラベルの割合は表出 (制約) が最も多く（約 45%），制約と選好の間では制約の事例が全体の約 75% を占め，選好の事例は 25% 程度となった．加えて，この集約後のラベルと各作業者のラベルの間での混同行列を確認したところ（図 2 左），作業者間の判断の揺れは主に隣接するラベルで見られ，離れたラベルで揺れることは稀であることが分かった．また，表出と省略の間をまたぐ揺れは主に選好に基づく判断で起こっていることが分かった．

コーパスと作業者判断の比較　読み手と書き手の判断に差違があるかを分析するため，コーパス中での表出・省略を正解とした場合に，作業者の判断の中央値がこれと一致するかを確かめた．結果として，作業者の結果を表出・省略の 2 値予測とみなした場合のコーパスに対する正解率は 97.0%であり，例外的な事例はあるにせよ，読み手である作業者と書き手の省略選択は概ね一致することが示された．

4 実験：省略判断モデル

人間の省略判断を現行のニューラル言語モデルがどの程度予測可能か調査する．

3) 3 名以上かつ順序尺度に対応する一致度の尺度で，一般的には alpha が 0.677 以上であれば高い一致度であるとされる．

4) データセットは記事単位で分割し，同一記事の事例が横断的に含まれないようにした．また，データセット間でラベル分布が大きく異ならないことを確認した．

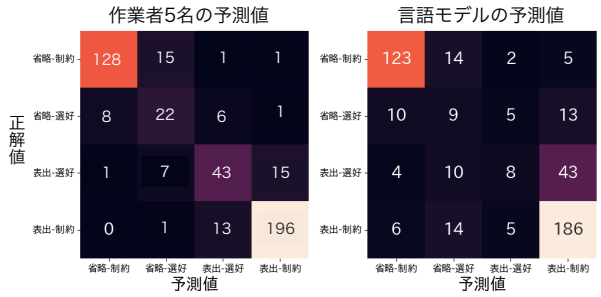


図 2 開発データにおける作業者と言語モデルの混同行列．左図は作業者 5 名の回答とその中央値の混同行列で，値は事例数の平均値．右図は BERT-large の混同行列．

表 3 作業者間と言語モデルの判断性能（F_1 値）．作業者間の値は，中央値を正解として各作業者の F_1 値を求め，5 人の値の平均を取ったもので，性能の上限値とみなせる．

	平均		省略		表出	
モデル	Macro	Micro	制約	選好	選好	制約
作業者間	76.0	83.7	88.9	61.2	61.4	92.7
BERT-base	52.1	69.6	74.2	33.7	19.4	81.0
BERT-large	50.2	70.8	76.0	29.6	10.9	84.2

4.1 実験設定

モデル 言語モデルとして日本語 Wikipedia で事前学習済みの Transformer 言語モデル（BERT-base-japanese,BERT-large-japanese[5)]）を用いた．対象の述語-項ペアを含む 1 文と，その前方文脈をモデルの入力とし，省略判断の 4 カテゴリ { 省略（制約），省略（選好），表出（選好），表出（制約）} の多値分類モデルを訓練した．モデルの入力と訓練時のハイパーパラメータと詳細は付録 E と F に示す．

評価尺度 上記 4 値を名義尺度として F_1 値を計算した．順序を考慮した評価は今後の課題とする．

4.2 結果

作業者と言語モデルの比較 作業者の判断を予測するモデルを学習した結果を表 3 に示す．制約により可否が定まる事例では相対的に性能が高く，一方で選好により可否が定まる事例では，人間の作業者間の F_1 値と比べると，モデルの予測性能は劣ることが示された．

また，言語モデルの予測に関する混同行列を観察したところ（図 2 右），言語モデルでは離れたラベル間で選択を誤る事例が人間より多くなる傾向が見られ，特に人間が「表出（選好）」を選んでいる事例に対して「表出（制約）」のラベルを予測するケースが多く見られた．

要因ごとの分析 人間の回答の判断根拠とモデルの正解事例との関連を分析し，言語モデルが予測可能な事例の性質を調査した．(i) 省略されると補われるべき情報が定まらず（因子 1），その場合に書き手の意図が伝わらない（因子 4）項と，(ii) 省略されても項が一意に定まる事例のうち（因子 1），表出または省略により構文的妥当性が損なわれる（因子 2）項で正答率が高かった．前者は新情報の項，後者は構文の容認性判断で判定される項であり，これらの項への判断は概ね達成できていると示唆される．

5 関連研究

人間と計算モデルの対照 人間と計算モデル間の判断の比較は，文の容認性判断のモデリング [7, 8, 9, 10, 11] などに始まり，自然言語処理分野において盛んに行われてきた．本研究では言語モデルの追学習で省略判断予測をしており，言語学的な仮説に直接答えるものではないが，作成したデータは計算モデルから得られる統計量 (情報量など) との対照といった言語学的検証にも活用できる．

省略解析 典型的な省略解析 [12, 13, 14] は，与えられた文の項省略を検知し，省略されている情報を復元するという問題設定である．本研究は，文章中のある項を省略・表出させるべきかという，実際の言語運用に焦点を当てた問題設定を導入している．

6 おわりに

本研究では，日本語母語話者の項省略判断について，コーパスの事例を元に 2,392 事例のデータを収集した．また，このデータを元に項省略判断モデルを構築し，現行のニューラル言語モデルが人間の判断を予測できるかを分析した．

今回収集したデータは作業者数が 5 名であり，作業者間の判断の差違をより詳細に分析するには至らなかったが，今後データ規模を拡大し，選好に基づく判断の揺れについてその因子を分析することや，個人の読解力の差違による判断の揺れの分析，書き手の文章記述力の差による書き手と読み手の判断のずれ等の分析も検討していきたい．また，モデルの予測性能を向上させることで，省略判断の予測に基づく文章の可読性評価や，日本語文章執筆支援といった技術の開発を目指したい．

5) https://huggingface.co/cl-tohoku/bert-base-japanese-v2, https://huggingface.co/cl-tohoku/bert-large-japanese

謝辞

本研究は科研費 JP19K12112 の助成を受けたものです.

参考文献

[1] 浅原正幸, 大村舞. Bccwj-depparapas:『現代日本語書き言葉均衡コーパス』係り受け・並列構造と述語項構造・共参照アノテーションの重ね合わせと可視化. 言語処理学会第 22 回年次大会発表論文集, pp. 489–492, 2016.

[2] Satoshi Oku. **A theory of selection and reconstruction in the minimalist perspective**. University of Connecticut, 1998.

[3] Mamoru Saito. Notes on east asian argument ellipsis. **LANGUAGE RESEARCH**, Vol. 43, pp. 203–227, 2007.

[4] Serkan Sener and Takahashi Daiko. Argument ellipsis in japanese and turkish. **MIT Working Papers in Linguistics 61 : Proceedings of the 6th Workshop on Altaic Formal Linguistics : Department of Linguistics and Philosophy. MIT**, pp. 325–339, 2010.

[5] Yuta Sakamoto. Phases and argument ellipsis in japanese. **Journal of East Asian Linguistics 2016 25:3**, Vol. 25, pp. 243–274, 7 2016.

[6] Klaus Krippendorff. Computing krippendorff's alpha-reliability. 2011.

[7] Sebastian Schuster and Tal Linzen. When a sentence does not introduce a discourse entity, transformer-based models still sometimes refer to it. In **North American Chapter of the Association for Computational Linguistics**, 2022.

[8] James A. Michaelov and Benjamin K. Bergen. Do language models make human-like predictions about the coreferents of italian anaphoric zero pronouns? In **International Conference on Computational Linguistics**, 2022.

[9] Allyson Ettinger. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 34–48, 1 2020.

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

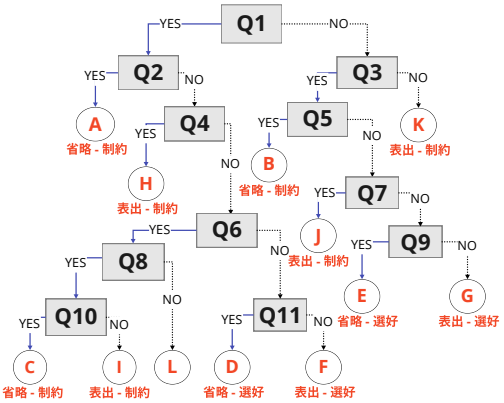
[11] Riki Fujihara, Tatsuki Kuribayashi, Kaori Abe, Ryoko Tokuhisa, and Kentaro Inui. Topicalization in language models: A case study on Japanese. In **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 851–862, Gyeongju, Republic of Korea, October 2022. International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

[12] Ryuto Konno, Yuichiroh Matsubayashi, Shun Kiyono, Hiroki Ouchi, Ryo Takahashi, and Kentaro Inui. An empirical study of contextual data augmentation for Japanese zero anaphora resolution. In **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 4956–4968, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[13] Nobuhiro Ueda, Daisuke Kawahara, and Sadao Kurohashi. BERT-based cohesion analysis of Japanese texts. In **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 1323–1333, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

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

A　フローチャート



因子	質問項目	内容
解釈可能性	Q1	提示された項が表出しない場合でも，そこに補われるべき項が一意に定まり，なおかつそれが今回指定された項と同一のものである．
構文的妥当性	Q2, Q5	項を表出すると明らかに冗長，または，構文としておかしくなる．
	Q4, Q7	表出させないと構文的に不自然になる．または，補う際に非常に困難を伴う．
述語項構造以外の解釈の変化	Q6	省略によって排他や対比，強調といった助詞による機能が失われ，前提が変化するなど前後の読みに影響を与えてしまう．
	Q8	元の文で項が省略されているかを予測できる．
	Q10	元の文では読み手の意図を伝えるために省略されていると思う．
文脈上の重要さ	Q3	省略した場合に項を一意に定めなくても，書き手が読み手に伝えようと意図する文意の大枠は変わらない．
その他の自然さ	Q9，Q11	省略する方が自然．

図 3　アノテーション時に使用したフローチャート

2.1 節で説明する因子に基づいた判断アノテーションを実現するにあたり，各因子に対応する質問項目への回答の組み合わせから最終的な省略の程度ラベルを決定する質問フローチャートを作成した (図 3)．フローチャート内のノードは各因子に対応した質問であり，その回答結果により次の質問項目が定まるという形式で因子間の依存関係が表現される．最終的に辿り着く終端記号は 5 つの省略の程度ラベルのいずれかに対応する．フローチャート上の終端記号は，各質問項目への回答の履歴を弁別出来るように便宜上それぞれ異なるアルファベットを割り当てている．例えば，「項が省略された場合に補われるべき情報が定まり（解釈可能性），かつ，項を表出すると構文的に不適切である（構文的妥当性）から，省略の程度は『省略（制約）』」という判断となる事例は，実際の作業時にはフローチャートの上から Q1 → YES → Q2 → YES → A というパスを辿り，結果として「省略（制約）」のラベルが割り当てられる．

B　作業ツール

図 4 に作業者が利用する作業ツールの実際の画面を示す．画面上部には記事 ID，事例 ID，対象となる述語-項ペアの情報が示され，画面下部には特定の述語・項が強調された文とその前方文脈が提示される．作業者は，現在の事例への判断を終えた後，画面右上にある「→」ボタンを押すことで次の対象述語-項に進む．事例を遡るためのボタンはタスクの制約上存在しない．

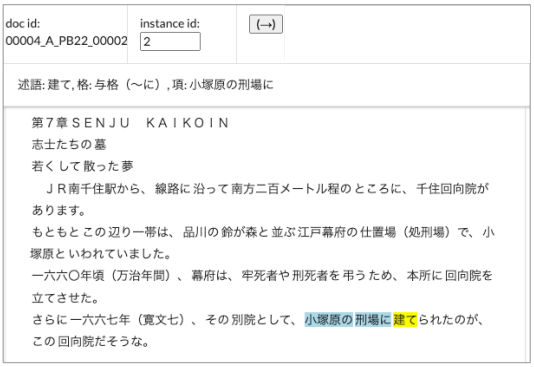


図 4　実際の作業画面の例

C　アノテーション対象

本研究ではコーパス上に出現する項のうち，ガ格，ニ格，ヲ格の項を対象とし，述語については和語の用言を対象とした（「サ変名詞+する」は対象に含まない）．また，機能性の強い述語（ある，なる，やる）を予め除外している．元コーパスでは 1 文の中に複数の述語が含まれる場合があるが，2.1 節や 2.2 節で記したように，作業者に元コーパスの文と，省略された項を補った文の双方を見せる可能性が存在する都合上，アノテーション対象の項が 1 文に対して 1 つとなるようにサンプル数を制限した．1 文に複数の述語が存在する場合には無作為に対象となる述語と項のペアを決定した．

D　省略されている項の補填作業

コーパス上で省略されている項を文中に補填する作業は，省略判断を行う作業者と異なる言語学専攻の大学生 2 名で行い，作業者間の合意のもと挿入位置・表出形を一意に定めた．さらに，別の作業者 1 名によって，この項の挿入によって不自然な文章となった事例を取り除いた．

E　言語モデルへの入力系列

使用した 2 つのモデルの入力最大系列長にあわせ，対象述語-項を含む 1 文を入力の最終文とし，この文の文末から遡って系列長がサブワード 512 トークンとなるように前方文脈の単語系列を定めた．入力系列中では，対象述語と項のトークン列の前後を，それぞれ特殊トークン <unused0>,<unused1> で囲むことで，判断対象となる述語と項の位置を表した．

F　ハイパーパラメータ

追学習には 4 つの GPU（NVIDIA RTX A6000）を用いた．BERT-base モデルのバッチサイズは 16, 学習率は 3e − 05，BERT-large モデルのバッチサイズは 8, 学習率は 5e − 05 とした．エポック数については，開発データに対する損失関数の値が 3 エポック連続で改善しない場合には早期終了するよう設定した．その他のハイパーパラメータは Hugging Face Transgformers の TrainingArguments クラス[6)] のデフォルト値に従う．

6)　https://huggingface.co/docs/transformers/main_classes/trainer