大規模言語モデルによるコーパスアノテーション精度の検証
—英語・スペイン語の前置詞句を対象として—

稲生秀俊[1] 堀江舞柚[1,2,3]
[1]東京外国語大学大学院 [2]マドリード・カルロス3世大学大学院 [3]日本学術振興会特別研究員
{inoo.hidetoshi.w0, horie.mayu.s0}@tufs.ac.jp

概要

本研究は、behavioral profiles [1] に基づくコーパス分析において必要となる形式的・意味的アノテーションを対象に、大規模言語モデル（LLM）の実用性を検証した。英語とスペイン語のコーパスから抽出した前置詞句の用例に対し、手作業と LLM によるアノテーションを比較し、正解率および一致度指標に基づく統計分析を用いて評価を行った。さらに、zero-shot/few-shot プロンプト設計の違いが精度に与える影響、言語間差についても検討した。結果、LLM のモデル、プロンプト設計、タグ項目や言語によってタグ付け精度には大きな差が見られ、LLM を研究利用する際はプロンプト設計と人手検証を組み合わせたハイブリッド運用が不可欠であることを示した。

1 はじめに

認知言語学における用法基盤モデルの理念を体現する behavioral profiles [1] の手法を用いたコーパス分析では、用例ごとに形式的・意味的な複数の変数を付与する必要がある。しかし、これらのアノテーションは手作業に依存することが多く、処理可能なデータ量や一貫性・客観性の確保に限界がある。その結果、大規模データを用いた分析はサンプル調査にとどまらざるを得ない場合が多い。

近年、LLM の発展によりアノテーション作業を自動化・省力化できる可能性が指摘されている。一方で、LLM の出力を研究データとして利用するためには、その精度や安定性を検証し、どのような条件下で信頼可能であるかを明らかにする必要がある。

本研究では、英語およびスペイン語のコーパスデータを対象に、手作業によるアノテーションを基準として、LLM によるタグ付け精度を体系的に検証する。特に、タグごとの差異、プロンプト設計（zero-shot/few-shot）の影響、および言語間での傾向の違いに焦点を当て、LLM を言語学研究に導入する際の方法論的含意を明らかにすることを目的とする。

本稿の構成は以下の通りである。第 2 節では先行研究を整理し、第 3 節では分析対象のデータおよびコーパスを概観する。第 4 節ではタグ付けおよび分析手法を示し、第 5 節で LLM によるタグ付け結果を示した上で考察を行う。最後に第 6 節でまとめと今後の課題を述べる。

2 先行研究

Feuerriegel et al. [2] は、研究目的に応じて辞書ベースから LLM まで適切な NLP モデルを選択する必要性を指摘するとともに、研究者にとっての NLP の判断過程の「解釈可能性」と「正確性（人間の判断との一致度）」のトレードオフを考慮した研究設計の重要性を強調している。

Yu et al. [3] は、謝罪発話行為を対象に、GPT-3.5/GPT-4 と手作業のアノテーションを比較し、LLM による語用・談話アノテーションの可能性を検証した。高い正確性（92.7%）を報告する一方で、文脈依存的な誤判定やプロンプト設計の影響といった課題を指摘している。LLM を一次的なアノテーション手段として用い、人手で検証・修正するハイブリッド運用の有効性を示唆する。

Mi et al. [4] は、LLM がイディオムの文脈に応じた意味を正確に解釈する能力を評価した。結果、モデルによってはプロンプトのわずかな違いが結果の大きな違いをもたらすことや one-shot プロンプトがかえって性能を低下させること、現行の LLM は文脈理解に不十分な点があることなどを指摘している。

これらの先行研究を踏まえ、本研究では LLM の導入範囲をアノテーション作業に限定し、手作業によるタグ付けを基準として、モデルおよびプロンプト設計の違いがアノテーション精度に与える影響を検証する。特に、複数言語を対象とすることで、LLM の性能が言語特性に依存する可能性についても検討する。

3 データ

英語の分析では in the soup という前置詞句を対象とする。in the soup には、料理の「スープの中に」という字義通りの意味と、「苦境に陥って」というイディオムの意味とがある。さらに、文脈によってはこのいずれでもない修辞的な意味に使用されることもある。

分析対象とする in the soup の用例データは、the Corpus of Historical American English（COHA）[5] の全文データから自作 Python コードで 231 例を抽出した。in the soup を含む文およびそれに先行する 2 文も併せて抽出し、文脈から意味を判別しやすくした。

スペイン語の分析対象は、前置詞句 en el aire （英訳 in the air）である。en el aire には、物理的に「空気中にある」という意味のほか、「不安定・危うい状況にある」という慣用的意味がある（RAE, n.d.: s.v. *aire*[1]）[6]。本研究では前者を字義通りの用法、後者をターゲットとするイディオム的用法とし、それ以外の比喩的用法は修辞的用法として区別した。

スペイン語の用例は Corpus del Español del Siglo XXI（CORPES）1.3 版 [7] から抽出した。2001 年から 2025 年までのスペインの書き言葉データに限定し、コンコーダンス検索で得られた en el aire の用例から 3,697 件をダウンロードした。

4 分析方法

4.1 検証手法

分析の準備として、英語およびスペイン語コーパスからダウンロードした用例の中から、それぞれ100例をランダムサンプリングした。次に、言語学的分析を想定した変数に基づき、各 100 例に対して手作業によるアノテーションを行った（具体的なアノテーション項目は 4.2. 節参照）。英語データは稲生が、スペイン語データは堀江が担当した。

その後、LLM によるタグ付けに用いる各言語のプロンプトを作成した。zero-shot/few-shot の両条件に共通の説明を与え、few-shot 条件にはタグ付け例を 3 例追加した。（付録 A を参照。）

次に、作成したプロンプトを用いてタグ付けを実行した。各言語 100 例からなる入力データは csv 形式で作成し、プロンプトは直接入力した。使用した LLM は ChatGPT（5.2 Thinking）[8] および Gemini（3 Flash Web 版）[9] である。最後に、手作業によるアノテーション結果を正解とみなし、LLM による結果と比較することで精度を算出した。精度の評価指標としては、正解率（%）および、偶然一致を補正した 2 者アノテーター間一致度指標である Cohen's κ [10] を用いた。Cohen's κ は、正解率では捉えにくいラベル分布の偏りを補正した一致度指標であり、本研究ではアノテーション精度の補助的評価指標として用いる。κ 値の解釈は Landis and Koch [11] および小町 [12] に基づき、以下の区分を参照した（詳細な解釈は 5 章で議論する）。

Cohen's κ の解釈の目安（κ 値：一致強度）
<0.00： 不一致（Poor）
0.00-0.20： 少し一致（Slight）
0.21-0.40： 普通の一致（Fair）
0.40-0.60： かなり一致（Moderate）
0.60-0.80： すごく一致（Substantial）
0.80-1.00： ほぼ一致（Almost Perfect）

4.2 アノテーション

behavioral profiles [1]を参考に、以下の意味的・形式的変数を設定した。

- 意味｛字義通り/イディオム/修辞｝
- 関連動詞｛不定詞で抽出｝
- 動詞の法｛不定詞/叙述（直説） /接続/命令…｝
- 動詞の時制｛現在/現在完了…｝
- 文の態｛能動/受動｝
- 文の意味極性｛肯定/否定｝
- 修飾用法｛制限/叙述｝
- 節の種類｛主節/従属節｝
- 関連名詞｛名詞（句）を抽出｝
- 人称｛1/2/3｝
- 名詞の数｛単数/複数｝
- 名詞の有生性｛人/その他有生物/無生物｝
- 修飾の有無｛有/無｝

英語およびスペイン語の各言語的特性や研究上の関心に応じて、一部タグは取捨選択して適用した。判断が困難な用例については「−」を記入することとした。（具体的なタグ付け例については付録 A の # Examples 参照。）

5 結果と考察

以下では、手作業によるタグを正解とみなし、LLM

によるタグ付け結果を表にまとめる。

なお、モデルによっては、与えた 100 例すべてに対してタグ付けが行われない場合があった。その際は、実際の出力数を明示した上で、100 例に換算して正解率を算出する（正解数/出力数×100）。

5.1 アノテーション精度の比較

5.1.1 英語データ

英語データの分析結果を表 1 に示す。

表 1 分析結果の比較（英語）

LLM	ChatGPT				Gemini			
shot	zero		few		zero		few	
出力	100		100		100		100	
指標	%	κ	%	κ	%	κ	%	κ
意味	40	.12	72	.56	87	.77	81	.68
動詞	25	.10	85	.82	83	.81	81	.79
法	87	.00	94	.73	92	.57	93	.46
時制	63	.28	83	.72	87	.79	80	.70
態	87	.00	99	.95	99	.96	93	.74
極性	22	.00	51	-.01	97	.71	91	.50
用法	76	.00	53	.20	49	.16	68	.18
名詞	4	.04	69	.68	76	.75	73	.72
人称数	47	.05	85	.79	87	.89	87	.82
有生	63	.13	90	.80	91	.82	85	.83
平均	51	.07	78	.62	85	.72	83	.64
中央	51	.04	84	.72	89	.78	83	.71
SD	29	.09	16	.30	14	.22	8	.20

英語では、入力した全 100 件に対して出力が得られた。

LLM 別の精度を見ると、全体の正解率（%）は Gemini zero-shot ＞ Gemini few-shot ＞ ChatGPT few-shot ＞ ChatGPT zero-shot であり、全般的には Gemini の zero-shot が最も高い正解率を示した。両 LLM を平均するとタグ付け精度が高い項目は態（94.5）、法（90）、有生性（82.25）、時制（78.25）であり、精度が低い項目は名詞（55.5）、用法（61.5）、極性 （65.25）であった。

Zero-shot と few-shot の比較では、ChatGPT は zero-shot から few-shot によって正解率が向上した一方、Gemini は few-shot ではむしろ正解率が低下する結果となった。ChatGPT では few-shot で精度が低下したのは用法のみであった。Gemini では法と人称・数の 2 項目以外のすべての項目で正解率が低下した。

Cohen's κ [10] による一致度評価では、ChatGPT の zero-shot が.07 で「少し一致」であったが、few-shot によって.62 の「かなり一致」まで向上した。Gemini は κ においても.72 から.64 に低下したが、いずれも解釈の目安としては「すごく一致」の範囲に収まっている。個別の項目については、ChatGPT zero-shot の法、用法、有生性や、Gemini few-shot の法、極性、用法のように正解率がある程度高いにもかかわらず κ が低い項目があった。これはタグが特定の値に偏っていた場合に、偶然の正解とみなされることが影響していると考えられる。

5.1.2 スペイン語データ

スペイン語データの分析結果を表 2 に示す。

表 2 分析結果の比較 （スペイン語）

LLM	ChatGPT				Gemini			
shot	zero		few		zero		few	
出力	100		100		53		39	
指標	%	κ	%	κ	%	κ	%	κ
意味	74	.28	41	.22	81	.61	82	.66
動詞	34	.33	42	.41	96	.96	85	.84
法	65	.43	75	.57	94	.90	92	.82
時制	58	.44	68	.55	94	.92	90	.86
極性	93	.50	99	.55	100	1.	97	.79
節	57	.25	54	.25	62	.38	64	.25
名詞	24	.24	23	.23	83	.83	80	.79
人称	96	-.01	96	.49	96	.49	100	NaN
数	78	.50	76	.46	91	.79	92	.81
有生	86	-.02	83	.11	92	.68	92	.67
修飾	57	.13	82	.24	96	.73	95	.48
平均	66	.28	67	.40	90	.75	88	.70
中央	65	.28	75	.42	94	.77	92	.70
SD	23	.19	24	.23	11	.21	10	.19

全体の正解率は Gemini zero-shot ＞ Gemini few-shot ＞ ChatGPT few-shot ＞ ChatGPT zero-shot であった。タグ項目別では、極性（97.4）、人称（97.1）、有生性（88.4）、数（84.2）といった比較的形式的な特徴は高精度でタグ付けされた一方、名詞（52.38）、節の種類（59.34）、動詞（64.21）、意味（69.55）では精度が低かった。

ChatGPT は出力数が安定しているものの、項目間の精度差が大きい傾向が見られた。これに対し Gemini は、多くの項目で高い正解率を示したが、

zero-shot では 53 例、few-shot ではさらに少ない 39 例しか出力されないという問題が確認された。また、few-shot 条件では、存在しないデータを出力するという例も確認された。このことから、出力されなかった例を評価に含めるか否かによって、Gemini の精度は大きく変動し、評価方法によっては 30%前後と算定される可能性がある。

zero-shot と few-shot の比較では、精度が向上する項目と低下する項目の両方が存在し、特に ChatGPT においてその傾向が顕著であった。

Cohen's κ [10] による一致度評価の結果、ChatGPT の平均 κ 値は zero-shot 条件で .28、few-shot 条件で .40 となり、few-shot による改善が確認されたものの、その一致度は中程度にとどまった。一方、Gemini の平均 κ 値は zero-shot 条件で .75、few-shot 条件で .70 と、両条件ともに非常に高い一致度を示し、few-shot による明確な改善は見られなかった。これらの結果から、few-shot プロンプティングの効果はモデル間で一様ではないことが示唆される。

また、ChatGPT の zero-shot 条件では、正解率が高いにもかかわらず κ 値が極めて低い、あるいは負の値を示すカテゴリが観察された（人称 96%/κ = -0.01、有生 86%/κ = -0.02）。これは、特定のラベルへの偏りが強い予測が行われている可能性を示す。

5.2 考察

以上の結果から、LLM によるアノテーション精度は、使用するモデル、プロンプト設計、タグ項目、対象言語の組み合わせによって大きく左右される[i]ことが明らかとなった。

第一に、モデル間の差異として、英語・スペイン語のいずれにおいても Gemini の zero-shot が最も高い精度を示した点が注目される。この結果は、LLM が事前学習段階で獲得している知識や分野適性が、言語学的アノテーションにも影響を与える可能性を示唆している。

第二に、プロンプト設計の効果はモデル間で一様ではなかった。ChatGPT では few-shot による平均的な精度向上が確認された一方で、Gemini では精度の低下や出力数の減少が観察された。これは、本研究で用いた 3 例という限られた例示数の影響である可能性を否定できない。しかし、例示数を増やした英語データの追加実験[ii]において、やはり ChatGPT にのみに平均正解率の向上が見られたことから、few-shot の有効性はモデルに依存的であり、必ずしも汎用的に有効な手法とはいえないことが示唆される。

第三に、タグ項目による精度差も顕著であった。両言語に共通して精度が低かった名詞のタグは、長距離の統語依存関係や、複数の文にまたがる意味・談話文脈の統合を必要とする項目であり、LLM の分析能力に一定の制約があることを示している。一方、有生性などの項目は、局所的かつ形式的な手がかりに基づいて判断可能であるため、高い精度が得られたと考える。

第四に、言語間の差異として、特にスペイン語データにおける Gemini の出力数の不安定性は、LLM を研究利用する際の再現性・信頼性の観点から重要な課題である。特に多言語データを扱う場合に、分析結果の比較可能性を損なう要因となり得る。

以上を踏まえると、現時点における LLM のアノテーション利用は、完全な自動化よりも、一次的なタグ付けを担い、人手による検証・修正を前提とした補助的ツールとして位置づけるのが妥当である。この見解は、LLM の研究利用における慎重な位置づけを提唱する Yu et al. [3] の指摘とも整合する。

6 おわりに

本研究は、LLM を用いたコーパスデータのアノテーション精度を検証するパイロット調査を行った。その結果、LLM モデル、プロンプト設計、タグ項目、言語の違いによって精度に大きな差が生じることが明らかとなった。特に、Gemini の zero-shot は高精度を示した一方、few-shot ではモデルによって精度が低下するなど、プロンプト設計やモデル特性の影響が大きいことが示唆された。

これらの知見は、今後の大規模コーパス分析において LLM を導入する際に、単独での自動化ではなく、人手による検証と組み合わせたハイブリッド運用が必要であることを示す。今後は、対象言語や対象となる言語現象の拡張、出力の再現性・安定性の評価、他分野への応用可能性の検討が課題となる。

[i] 英語の ChatGPT 条件間の差は大きい（Cohen's *d* [13] = 1.14）のに対し、スペイン語では小さかった（*d* = .07）。ChatGPT と Gemini の zero-shot 間では両言語ともに大きな効果量（*d* = 1.50, 1.23）が観察された。

[ii] few-shot の例を 21 件に増やして試行したところ、ChatGPT では平均正解率 91%、平均 κ =.67、Gemini では同じく 85%、κ =.71 であった。

謝辞

本研究の遂行にあたり、有益なコメントを賜りました大谷直輝先生（東京外国語大学准教授）、川上茂信先生（東京外国語大学名誉教授）、佐近優太先生（神田外語大学講師）、ならびに福田航平氏（東京外国語大学大学院生）に深く感謝申し上げます。

参考文献

1. Gries, Stefan Th. (2010) Behavioral profiles: A fine-grained and quantitative approach in corpus-based lexical semantics. *The Mental Lexicon* 5(3): 323-346. DOI: https://doi.org/10.1075/ml.5.3.04gri

2. Feuerriegel, Stefan, Abdurahman Maarouf, Dominik Bär, Dominique Geissler, Jonas Schweisthal, Nicolas Pröllochs, Claire E. Robertson, Steve Rathje, Jochen Hartmann, Saif M. Mohammad, Oded Netzer, Alexandra A. Siegel, Barbara Plank and Jay J. Van Bavel (2025) Using natural language processing to analyse text data in behavioural science. *Nature Reviews Psychology* 4: 96–111. https://doi.org/10.1038/s44159-024-00392-z

3. Yu, Danni, Luyang Li, Hang Su and Matteo Fuoli (2024) Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology. *International Journal of Corpus Linguistics* 29:4: 534–561. https://doi.org/10.1075/ijcl.23087.yu

4. **Mi, Maggie, Aline Villavicencio and Nafise Sadat Moosavi** (2025) Rolling the DICE on Idiomaticity: How LLMs Fail to Grasp Context. https://arxiv.org/abs/2410.16069v2

5. Davies, Mark (2010) *The Corpus of Historical American English* (*COHA*) full-text data.

6. REAL ACADEMIA ESPAÑOLA: *Diccionario de la lengua española*, 23.ª ed., [versión 23.8 en línea]. <https://dle.rae.es> [Fecha de la consulta: 2025/12/28].

7. REAL ACADEMIA ESPAÑOLA: Banco de datos (CORPES XXI) [en línea]. *Corpus del Español del Siglo XXI* (*CORPES*). <http://www.rae.es> [Fecha de la consulta: 2025/10/15]

8. OpenAI (2025-26). ChatGPT (Version 5.2) [Large language model]. https://chatgpt.com/

9. Google (2025-26) Gemini (Version 3 Flash Web 版) [Large language model]. https://gemini.google.com/

10. Cohen, Jacob (1960) A Coefficient of Agreement for Nominal Scales. *Education and Psychological Measurement* 20(1): 37-46.

11. Landis, J. Richard and Gary G. Koch (1977) The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33(1): 159-174.

12. 小町守 (2024) 『自然言語処理の教科書』東京：技術評論社.

13. Cohen, Jacob (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

A 付録：few-shot プロンプトの例（英語、in the soup）

The attached file contains English language data for linguistic analysis. The first row contains column titles, and data begins from the second row onwards. Below, "Column A through Column O" refer to the column names displayed when this file is loaded in Excel. Columns are separated by commas. At the current stage, data is entered in Columns A through E.

The target phrase for analysis is "in the soup".
Content of existing data:
Column A （No.）: Unique row number. Since this data is extracted from a larger dataset, the numbers are not consecutive. This data will not be used in the current task.
Column B （genre）: The genre of the example in that row.
Column C （second preceding sentence）: The sentence that appears two sentences before the sentence containing the target phrase （Column E）. （Depending on the source data, a preceding sentence may not exist.）
Column D （preceding sentence）: The sentence that appears one sentence before the sentence containing the target phrase （Column E）. （Depending on the source data, a preceding sentence may not exist.）
Column E （sentence）: The sentence containing the target phrase.
Content of columns to be filled:
Based on the data already entered in Columns B through E, please enter appropriate data （e.g., idiomatic, he, 3SG, prove, positive, indicative, present, attributive, active, inanimate, etc.） in Columns F through O. For all columns except Columns G and I, select the most appropriate option from the given choices. Once all data has been entered, output the complete data from Columns A through O in CSV format.
Column F （sense）: Whether the target phrase in the sentence in Column E is used literally, idiomatically, or in a figurative （non-idiomatic） sense. Idiomatic meaning refers to conventional, non-compositional meanings such as 'in trouble' or 'in a bad situation'. Literal meaning refers to being physically in soup as food. Figurative meaning refers to meanings that are neither literal nor idiomatic as defined above. Options: literal, idiomatic, figurative
Column G （noun）: Extract in its surface form the subject of the clause containing the target phrase, or the noun, pronoun, interrogative word, or noun phrase that is most syntactically and semantically related to the target phrase. If unknown or not applicable: '-'
Column H （person）: The person and number （1, 2, 3; SG, PL） of the noun or noun phrase in Column G. Options: 1SG, 2SG, 3SG, 1PL, 2PL, 3PL
Column G. Options: 1SG, 2SG, 3SG, 1PL, 2PL, 3PL
Column I (verb): Extract in lemma form the predicate verb of the clause in which the target phrase is used in the sentence in Column E, or if there is no predicate verb, the verb most semantically related to the target phrase in the example sentence. If there is no verb in the example sentence but it can be considered an ellipsis, enter the elided verb. If the example sentence is a verbless sentence, enter 'none'.
Column J (polarity): Whether the meaning of the sentence in Column E is positive or negative. Options: positive, negative
Column K (mood): The mood of the sentence in Column E. Options: indicative, imperative, subjunctive
Column L (tense): The tense (present, past, future) and aspect (progressive, perfect) of the sentence in Column E. Aspect may not be present in some cases. Options: present, past, future, present progressive, past progressive, future progressive, present perfect, past perfect, future perfect
Column M (modification): Whether the target phrase in the sentence in Column E functions as a predicate (predicative) or as a modifier (attributive). Options: predicative, attributive
Column N (voice): Whether the sentence in Column E is in active or passive voice. Options: active, passive
Column O (animacy): The animacy of the subject or semantic subject of the sentence in Column E (human should be distinguished from other animate entities. Animate includes non-human animals, and inanimate includes both inanimate objects and body parts). Options: human, animate, inanimate

Examples are given below:
Example 1:
Second preceding sentence: I altered the recipe slightly because I did n't include the red Thai curry paste and I substituted carrots for the red pepper and quinoa for faro because I had those items in the house .
Preceding sentence: Overall , I really enjoyed this recipe and I was really glad I got to @ @ @ @ @ @ @ @ @ @ .
Sentence: Unlike the salad , where I was a little put off by the slightly bitter taste of the kale , in the soup the taste really complemented the sweetness of the potatoes .
Data to be filled in Columns F through O： literal,taste,3SG,complement,positive,indicative,past,predicative,active,inanimate
（以下略）