大規模言語モデルにおけるスペイン語の方言バイアス －地理的語彙変異の分析－

川崎義史
東京大学
ykawasaki@g.ecc.u-tokyo.ac.jp

概要

本稿では、大規模言語モデルにおけるスペイン語の方言バイアスを調査する。モデルの有する方言認識能力の解明は、方言的多様性を反映したモデルの開発のためにも重要である。モデルを仮想インフォーマントとみなし、実際の方言調査を模した多肢選択問題形式の実験を行った。スペイン語圏 21 カ国における、934 の語彙項目の地理的変異を扱った。実験の結果、モデルの方言認識能力は一様ではないことが判明した。モデルの方言認識能力は、モデルの訓練に利用されたと推定される方言ごとの学習データ量とは有意な相関を示さなかった。このことは、モデルの方言認識能力には、学習データ量以外の要因が影響している可能性を示唆している。

1 はじめに

スペイン語は、環大西洋地域を中心に 5.2 億人以上の母語話者を擁する大言語である [1]。スペイン語圏 21 カ国には多種多様な地理的語彙変異が存在する [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]。例えば、「自動車」を表すのに、スペインでは *coche* が、中南米では *carro* が主に使用される [8]。

実際のスペイン語の地理的語彙変異を大規模言語モデル (Large Language Model; LLM) はどの程度認識できるのだろうか？方言間の平等性の観点から、LLM においても特定の方言のみが優遇されることは望ましくない [13, 14, 15]。また、モデルの有する方言認識能力の解明は、方言的多様性を反映したモデルの開発のためにも重要である [16]。しかしながら、このような調査は数少なく規模も小さい [17]。

そこで、本稿では、LLM の有するスペイン語方言認識能力について大規模な調査を行う (図 1)。LLM を**仮想インフォーマント**とみなし、実際の方言調査を模した多肢選択問題形式の実験を行う。934 の語彙項目の地理的変異を調査対象とする。LLM の方言認識能力は、国別・方言域別に検証する。簡単のために、方言の最小単位は国とする。方言域の分類は、付録 A を参照されたい。

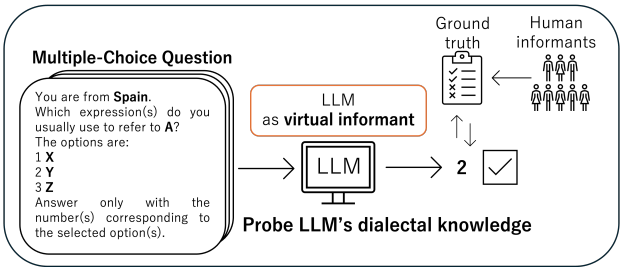

図 1 本研究の概略図：分かりやすさのために、スペイン語のプロンプトは英訳したものを示している。

実験の結果、LLM のスペイン語方言認識能力は一様ではないことが判明した。スペイン、赤道ギニア、メキシコ・中米、ラプラタ川流域の方言に対するモデルの認識能力は高い反面、チリ方言に対する認識能力は顕著に低かった。興味深いことに、モデルの方言認識能力は、モデルの訓練に利用されたと推定される方言ごとの学習データ量と有意な相関を示さなかった。このことは、単に当該方言の学習データを増やしても、その方言に対するモデルの認識能力が向上するとは限らないこと、また、モデルの方言認識能力には学習データ量以外の要因が影響している可能性を示唆している。

2 関連研究

LLM の有する方言認識能力を調査した研究は、主に英語を扱っている。これらの研究では、方言によりモデルの認識能力に差異が見られることが報告されている [13, 15]。そのため、モデル内の方言間の公平性 [18, 19] や言語（方言）バイアス [14] に関する懸念が提起されている。

スペイン語に関しては、Mayor-Rocher ら [17] が、

表 1　語彙項目 A141

項目番号	A141
語釈	*vehículo destinado al transporte de personas* “*vehicle intended for the transport of people*”
英語の対応語	CAR
変異形	*auto*, *automóvil*, *carro*, *coche*, *concho*, *máquina*

形態統語論的項目 20 と語彙項目 10 の計 30 項目を対象とした調査を行った。多肢選択問題形式により、方言域ごとにモデルの方言認識能力を検証した。実験の結果、スペインとラプラタ川流域の方言に対する認識能力が高い一方で、アンデスやメキシコ・中米の方言に対する認識能力は低いことが分かった。モデルの方言認識能力の違いは、各方言域の学習データ量の多寡によるものだと結論付けている。

3　手法

VARILEX　正解データは、VARILEX[8] から収集した。これは、スペイン語の地理的語彙変異に関する大規模なデータベースである。934 の語彙項目について、スペイン語圏 21 カ国における使用状況が記載されている。本稿では、VARILEX の改訂版を利用した[1)]。改訂版では、語彙項目の変異形（同義語）が当該国で優勢に使用されるか否かが二値で表現されている。

各語彙項目は、項目番号、スペイン語の語釈、英語の対応語、変異形の情報を含む。例として、表 1 に、語彙項目 A141 の情報を示す。項目ごとの変異形の数は大きく異なり、中央値は 8 である。また、国 – 項目のペアごとに、優勢な変異形の数も異なる。例えば、項目 A141 に関しては、スペインでは *coche* のみが優勢なのに対し、アルゼンチンでは複数の変異形（*auto*, *automóvil*, *coche*）が優勢である。

プロンプト　LLM を**仮想インフォーマント**とみなし、実際の方言調査を模した実験を行う。問題形式は、多肢選択問題とする。プロンプトにより、当該国で優勢に使用される変異形を選択肢の中から全て選ぶようにモデルに指示する。変異形は数字で番号付けされている。モデルには、変異形に対応する番号のみを返すように指示する。プロンプトはスペイン語で与える。プロンプトのテンプレートは付録 B を参照されたい。

評価指標　正解ラベルとモデルの予測をそれぞれ集合とみなし、両者の一致度を定量化する。一致度は、調整 Jaccard 係数（J_{adj}）により算出する。導出は、付録 C を参照されたい。この指標は、偶然の一致を補正した上で、二つの集合間の積集合の大きさを表すものである。この補正により、変異形の数や正答変異形の数（正答ラベル数）が異なる項目間の比較を可能にする。モデルの方言認識能力は、当該方言に関する全質問に対する指標の平均値により測定する。

1) https://h-ueda.sakura.ne.jp/varilex-r/

4　実験

4.1　実験設定

プロンプト　国 – 語彙項目の全組み合わせのうち変異形が 2 つ以上存在する 17,911 問をプロンプトに使用する。変異形は、質問ごとに無作為に並び替えた後に番号付けする。これにより、モデルによる特定の番号の選好の影響を緩和する [20]。質問の順序も無作為に並び替える。プロンプトで指定した回答形式に沿った回答のみを評価対象とする。

ベースライン　変異形が 3 つ以上の質問に対しては、最初の 3 つの変異形の回答（1/2/3）による結果をベースラインとする。「3 つ」というのは、質問ごとの正答ラベル数の平均 2.66 を整数に丸めた値である。変異形が 2 つの質問に対しては、両者を回答した結果をベースラインとする。

モデル　`gpt-4o`、`gpt-5.1`、`gpt-5.2` の 3 モデルを API 経由で利用した。GPT モデルに限定したのは、先行研究 [17] で他モデルを凌駕する性能を示したためである。本稿では、モデル間の比較よりも、語彙変異の包括的分析を主眼とする。

4.2　結果

全体結果　表 2 は全体結果を示している。N_Q と N_A は、それぞれ、質問数と有効回答数を表している。`gpt-5.1` が最高性能（0.338）を示した。`gpt-5.2` (0.336) と `gpt-4o`（0.314）は、僅かに劣る性能を示した。新しいモデルの方が高い性能を示し、またプロンプトの回答指示への追従能力が高かった。いずれのモデルもベースライン（0.110）を約 3 倍上回り、一定程度の方言認識能力を有していると言える。

国別結果　以下では、最高性能を示した `gpt-5.1` の結果を報告する。表 3 は、`gpt-5.1` の国別のモデル性能を示している。モデル性能は、スペイン（0.508）が最も高く、次いで、アルゼンチン（0.448）、

表 2　全体のモデル性能

Model	N_Q	N_A	J_{adj}
baseline	17,911	17,911	0.110
gpt-4o	17,911	17,853	0.314
gpt-5.1	17,911	17,911	**0.338**
gpt-5.2	17,911	17,910	0.336

表 3　gpt-5.1 の国別のモデル性能：ΔJ_{adj} は、ベースラインとの差分を表している。

Country	N_A	J_{adj}	baseline	ΔJ_{adj}
Spain	907	**0.508**	0.110	**0.398**
Equatorial Guinea	826	0.412	0.087	0.325
Cuba	920	0.317	0.120	0.197
Dominican Republic	892	0.339	0.102	0.237
Puerto Rico	917	0.306	0.138	0.168
Mexico	928	0.381	**0.140**	0.241
Guatemala	918	0.329	0.101	0.228
Honduras	921	0.332	0.114	0.218
El Salvador	717	0.310	0.102	0.208
Nicaragua	917	0.306	0.121	0.185
Costa Rica	563	0.308	0.103	0.205
Panama	832	0.279	0.092	0.187
Colombia	510	0.288	0.090	0.198
Venezuela	908	0.352	0.132	0.220
Ecuador	910	0.330	0.106	0.224
Peru	853	0.318	0.100	0.218
Bolivia	908	0.335	0.107	0.228
Chile	818	0.141	0.101	0.040
Paraguay	918	0.366	0.100	0.266
Uruguay	907	0.344	0.103	0.241
Argentina	921	0.448	0.125	0.323

赤道ギニア（0.412）、メキシコ（0.381）、パラグアイ（0.366）が高かった。一方、モデル性能が低いのは、チリ（0.141）、パナマ（0.279）、コロンビア（0.288）、ニカラグア（0.306）、プエルトリコ（0.306）となった。

5　考察

5.1　問題形式の影響

本節では、問題形式がモデル性能に与えうる影響を分析する。

全体　質問ごとの正答ラベル数とモデルが回答する変異形の数（モデル回答数）には、中程度の正の相関が見られた（$\rho = 0.437$（$p < 0.01$））[2)]。正の相関は、回答するべき変異形の数をモデルがある程度近似できていることを示唆している。ただし、正答かどうかは無関係である。モデル性能は、正答ラベル数（$\rho = -0.114$（$p < 0.01$））やモデル回答数（$\rho = -0.224$（$p < 0.01$））とは非常に弱い負の相関を示した。負の相関は、J_{adj} が偶然の一致を補正した結果生じるものであると考えられる。正答ラベル数やモデル回答数が大きくなるにつれ、偶然でも積集合が大きくなる可能性があるが、J_{adj} は偶然の一致を割り引いた値を返す。重要なことは、J_{adj} が語彙項目ごとの難易度やモデル回答数の影響を考慮できていることである。変異形の数が多い語彙項目ほど、モデルが正しく回答するのは困難になる。実際に、J_{adj} と質問ごとの変異形の数には、中程度の負の相関が見られる（$\rho = -0.591$（$p < 0.01$））。

国別　国別の J_{adj} は、平均正答ラベル数と有意な相関を示さない（$\rho = 0.183$（$p = 0.427$））（付録 D の図 3）。通常の Jaccard 係数を用いた場合は中程度の相関が見られるため（$\rho = 0.655$（$p < 0.01$））、真のモデル性能の評価が困難になる。国別の J_{adj} は、モデル回答数とも有意な相関を示さない（$\rho = 0.057$（$p = 0.806$））。これらの相関の欠如は、モデルの方言認識能力が、問題形式の設計に影響されておらず、真のモデル性能を反映していることを示唆する。さらに、国別の J_{adj} は、ベースラインとも有意な相関を示さない（$\rho = 0.246$（$p = 0.283$））。これらの性質は、J_{adj} がモデル性能を測る適切な指標であることを表している。

5.2　推定学習データ量の影響

学習データが豊富な方言ほど、LLM の方言認識能力が高くなるという知見が報告されている [17]。以下、この知見の再検証を行う。ただし、実際の学習データの構成は不明であることに注意を要する。先行研究 [17] に従い、国別の推定学習データ量を CEREAL[21] から算出した[3)]。データ量はトークン数で測った。分析の結果、先行研究 [17] とは異なり、推定学習データ量とモデルの方言認識能力には正の相関は見られず、逆に、有意ではないものの負の相関が見られた（$\rho = -0.305$（$p = 0.178$））。両変数の散布図を図 2 に示す。国は方言域ごとに色分けされている。

この結果は、方言ごとの推定学習データ量の違いのみでは、モデルの方言認識能力の差異を説明できないことを意味している。モデルの方言認識能力の差異には、学習データ量以外の要因 – 地域間の使用語彙の類似性、方言的特徴の消失による言語の標準化、学習データの構成的特性など – が影響している

2)　本稿ではスピアマンの順位相関係数を用いる。有意水準は $p = 0.05$ とする。

3)　https://zenodo.org/records/14771240

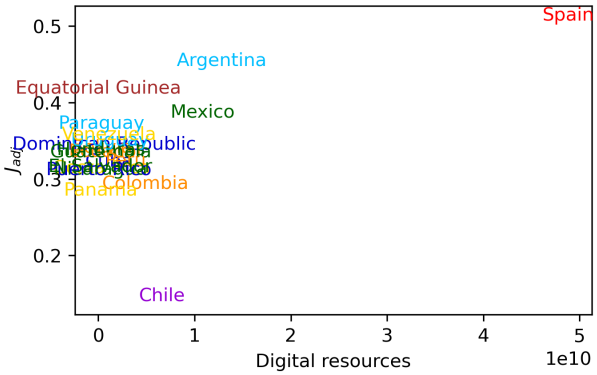

図 2　推定学習データ量と J_{adj} の散布図（$\rho = -0.305$ ($p = 0.178$)）：国は方言域ごとに色分けされている。

可能性がある。そのため、単に学習データ量を増やすだけでは、モデル性能が向上しない可能性がある。有効な改善策としては、方言データでのファインチューニングなどが考えられる。

推定学習データ量が最大の 3 カ国スペイン、アルゼンチン、メキシコについては、モデルの方言認識能力も最も高い部類に入る。しかし、推定学習データ量とモデル性能の正の相関は全体的には成り立たない。特に、下記二カ国の振る舞いが興味深い。

赤道ギニア　赤道ギニアは、推定学習データ量が最も少ないにも関わらず、モデルの方言認識能力は 3 番目に高かった。赤道ギニアとスペインの語彙使用パターンは類似している [22, 23]。そのため、推定学習データ量が最大のスペインの語彙使用パターンについて獲得された知識が転移した可能性がある。

チリ　対照的に、チリは推定学習データ量が 4 番目に大きいにも関わらず、モデルの方言認識能力は著しく低い結果となった。チリ方言の語彙使用パターンは特徴的である [24] ため、モデルによる認識は容易だと想定される。そのため、この結果は直観に反するものである。象徴的な例として、項目 A157（英語の *van*）に関して、モデルは、チリに固有の変異形（*furgón*/*furgonita*）の認識に失敗している。チリ方言に対するモデルの特異な振る舞いについては、今後、更なる調査が必要とされる。

表 4 に、方言域別の平均モデル性能を報告する。方言域別の結果は、主に国別の結果を反映している。スペイン方言に対する認識能力が最も高い。中南米の中では、ラプラタ川流域とメキシコ・中米の方言に対する認識能力が僅かに高い。次いで、アンティル諸島、大陸カリブ、アンデスが同等の結果となり、大きく離れてチリが続く。国別の場合と同様、推定学習データ量とモデルの方言認識能力には有意な相関は見られなかった（$\rho = 0.119$ ($p = 0.779$)）。

表 4　方言域別の平均モデル性能

方言域	J_{adj}
スペイン	**0.508**
赤道ギニア	0.412
アンティル諸島	0.321
メキシコ・中米	0.332
大陸カリブ	0.313
アンデス	0.318
チリ	0.141
ラプラタ川流域	0.386

6　おわりに

本稿では、LLM におけるスペイン語の方言バイアスを調査した。モデルを仮想インフォーマントとみなし、実際の方言調査を模した多肢選択問題形式の実験を行った。スペイン語圏 21 カ国における、934 の語彙項目の地理的変異を扱った。実験の結果、モデルの方言認識能力は一様ではないことが判明した。モデルの方言認識能力は、モデルの訓練に利用されたと推定される方言ごとの学習データ量とは有意な相関を示さなかった。このことは、単に当該方言の学習データ量を増やしても、その方言に対するモデルの認識能力が向上するとは限らないこと、また、モデルの方言認識能力には学習データ量以外の要因が影響している可能性を示唆している。

今後の課題は以下の四点である：(1) モデルの方言認識能力が、性別、年齢、教育レベルなどの社会言語学的な要因 [25] に影響されるか検証する必要がある。本稿では、これらの属性を欠く平均的なインフォーマントを想定した；(2) 本稿では、変異形の使用の有無が二値で表現されているが、実際には頻度の問題である。商用 GPT モデル以外のオープンモデルを使用することで、確率に基づく評価を行い、モデルの振る舞いをより詳細に分析することが求められる；(3) 方言の条件付け（例：メキシコ方言）によるテキスト生成タスクの評価を行うことで、モデル内で方言がどのように表現されているかより具体的な分析が必要である。方言の条件付けによるテキスト生成が実現すれば、方言別の合成データの構築 [14] も可能になる；(4) 本稿と同様の枠組みで、形態統語的変異 [26] などの言語変異についても調査を行い、モデルの有する方言バイアスを包括的に検証する必要がある。

謝辞

本研究の一部は、JSPS 科研費 JP23K12152 の助成を受けたものです。

参考文献

[1] Instituto Cervantes. **El español en el mundo 2025: Anuario del Instituto Cervantes**. Instituto Cervantes, 2025.

[2] Alonso Zamora Vicente. **Dialectología española (segunda edición muy aumentada)**. Gredos, 1968.

[3] John Lipski. **El español de América**. Cátedra, 1994.

[4] Manuel Alvar, editor. **Manual de dialectología hispánica. El Español de América**. Ariel, 1996.

[5] Manuel Alvar, editor. **Manual de dialectología hispánica: El español de España**. Ariel, 1996.

[6] Inés Fernández-Ordóñez. La lengua de Castilla y la formación del español, 2011.

[7] Bruno Gonçalves and David Sánchez. Crowdsourcing dialect characterization through Twitter. **PLoS ONE**, Vol. 9, p. e112074, 11 2014.

[8] Hiroto Ueda and Francisco Moreno-Fernández. VARILEX-R: Variación léxica en español del mundo / Datos revisados, 2016.

[9] Francisco Moreno-Fernández. **La lengua española en su geografía: Manual de dialectología hispánica**. Arco/Libros, 5 edition, 2020.

[10] Francisco Moreno-Fernández. **Variedades de la lengua española**. Routledge, 2020.

[11] Francisco Moreno-Fernández and Rocío Caravedo, editors. **Dialectología hispánica**. Routledge, 10 2023.

[12] Hiroto Ueda. **Dialectología del español y dialectometría**, pp. 87–104. Routledge, 2023.

[13] Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. Quantifying the dialect gap and its correlates across languages. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 7226–7245. Association for Computational Linguistics, 12 2023.

[14] Javier Muñoz-Basols, María del Mar Palomares Marín, and Francisco Moreno-Fernández. El Sesgo Lingüístico Digital (SLD) en la inteligencia artificial: Implicaciones para los modelos de lenguaje masivos en español. **Lengua y Sociedad**, Vol. 23, pp. 623–647, 12 2024.

[15] Fangru Lin, Shaoguang Mao, Emanuele La Malfa, Valentin Hofmann, Adrian de Wynter, Xun Wang, Si-Qing Chen, Michael J. Wooldridge, Janet B. Pierrehumbert, and Furu Wei. Assessing dialect fairness and robustness of large language models in reasoning tasks. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 6317–6342. Association for Computational Linguistics, 7 2025.

[16] María Grandury, Javier Aula-Blasco, Júlia Falcão, Clémentine Fourrier, Miguel González Saiz, Gonzalo Martínez, Gonzalo Santamaria Gomez, Rodrigo Agerri, Nuria Aldama García, Luis Chiruzzo, Javier Conde, Helena Gomez Adorno, Marta Guerrero Nieto, Guido Ivetta, Natàlia López Fuertes, Flor Miriam Plaza del Arco, María-Teresa Martín-Valdivia, Helena Montoro Zamorano, Carmen Muñoz Sanz, Pedro Reviriego, Leire Rosado Plaza, Alejandro Vaca Serrano, Estrella Vallecillo-Rodríguez, Jorge Vallego, and Irune Zubiaga. La leaderboard: A large language model leaderboard for Spanish varieties and languages of Spain and Latin America. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 32482–32524. Association for Computational Linguistics, 2025.

[17] Marina Mayor-Rocher, Cris del Pozo Huerta, Nina Melero, Gonzalo Martínez, María Grandury, and Pedro Reviriego. Es igual pero no es lo mismo: ¿Distinguen los LLMs las variedades del español? **Procesamiento del Lenguaje Natural**, Vol. 75, pp. 137–146, 9 2025.

[18] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. In **Proceedings of the Third Workshop on Abusive Language Online**, pp. 25–35. Association for Computational Linguistics, 2019.

[19] Isabel Orlanes Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey. **Computational Linguistics**, Vol. 50, pp. 1097–1179, 9 2024.

[20] Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In **Findings of the Association for Computational Linguistics: NAACL 2024**, pp. 2006–2017. Association for Computational Linguistics, 6 2024.

[21] Cristina España-Bonet and Alberto Barrón-Cedeño. Elote, choclo and mazorca: On the varieties of Spanish. In **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 3689–3711. Association for Computational Linguistics, 6 2024.

[22] Hiroto Ueda. Zonificación del español del mundo: Palabras y cosas de la vida urbana. In **Lingüística (ALFAL)**, Vol. 7, pp. 43–86, 1995.

[23] Hiroto Ueda. Zonificación múltiple de las ciudades hispanohablantes según el léxico urbano moderno: Análisis clúster y análisis de componentes principales. In Antonio Ruiz Tinoco, editor, **Jornadas sobre métodos informáticos en el tratamiento de las lenguas ibéricas**, pp. 121–140. Centro de Estudios Hispánicos de la Universidad Sofía, 2007.

[24] Abelardo San Martín Núñez. **El español en Chile**, pp. 216–226. Routledge, 10 2023.

[25] Janet Holmes. **An Introduction to Sociolinguistics**. Longman, 3 edition, 2008.

[26] Toshihiro Takagaki, Noritaka Fukushima, Masami Miyamoto, Antonio Ruiz Tinoco, Hiroto Ueda, Kimiyo Nishimura, and Kahori Umezaki. Variación gramatical del español en el mundo (VARIGRAMA), 2018.

[27] Raimundo Real and Juan M. Vargas. The probabilistic basis of Jaccard's index of similarity. **Systematic Biology**, Vol. 45, pp. 380–385, 9 1996.

A　スペイン語の方言域

スペイン語は、以下の 8 つの方言域に分類できる：

- **スペイン**
- **赤道ギニア**
- **アンティル諸島**：キューバ、ドミニカ共和国、プエルトリコ[4)]
- **メキシコ・中米**：メキシコ、グアテマラ、ホンジュラス、エルサルバドル、ニカラグア
- **大陸カリブ**：コスタリカ、パナマ、ベネズエラ
- **アンデス**：コロンビア、エクアドル、ペルー、ボリビア
- **チリ**
- **ラプラタ川流域**：アルゼンチン、ウルグアイ、パラグアイ

中南米諸国の分類は、Moreno-Fernández[9] に依拠した。スペインと赤道ギニアは別個に扱った。

B　プロンプトのテンプレート

以下、プロンプトのテンプレートを示す。質問ごとに、country、description、variant を指定する：

Responda a la siguiente pregunta. No tenga en cuenta las preguntas anteriores.
*Usted es de **[country]**.*
*¿Qué expresión(es) suele usar para referirse a «**[description]**»?*
Las opciones son:
*1 **[variant 1]***
*2 **[variant 2]***
*3 **[variant 3]***
...
Conteste solo con el número correspondiente a la opción. Puede elegir más de una opción; en ese caso, los números deberán ir separados por el signo «/ »en orden ascendente.

"Answer the following question. Do not take the previous questions into account.
*You are from **[country]**.*
*Which expression(s) do you usually use to refer to '**[description]**'?*
The options are:
*1 **[variant 1]***
*2 **[variant 2]***
*3 **[variant 3]***
...
Answer only with the number(s) corresponding to the selected option(s). You may choose more than one option; in that case, the numbers should be separated by the '/' sign in ascending order."

C　調整 Jaccard 係数

ある語彙項目に関して、当該国で優勢な変異形（正解ラベル）の集合を A、モデルが優勢だと予測した変異形の集合を B とする。通常の Jaccard 係数 [27] は、下記のように定義される：

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{X}{s+t-X}, \tag{1}$$

$|A| = s$、$|B| = t$、$X = |A \cap B|$ である。値が大きいほど、二つの集合の積集合の割合が大きいことを表す。しかし、集合が大きい場合には、偶然の一致によっても積集合は大きくなりうる。そこで、下記のように、偶然の一致を考慮した補正を行う：

$$J_{\text{adj}} = \frac{J - \mathbb{E}[J]}{1 - \mathbb{E}[J]}, \tag{2}$$

J は通常の Jaccard 係数、$\mathbb{E}[J]$ はその期待値を表す。

ここで、N 個の変異形から、無作為に s 個を選ぶことで A を、t 個を選ぶことで B を構成することを考える。この時、積集合 X の大きさは超幾何分布に従う：

$$X \sim \text{Hypergeometric}(N, t, s).$$

したがって、積集合 X の期待値は以下となる：

$$\mathbb{E}[X] = \frac{st}{N}.$$

一方、和集合 $|A \cup B| = s + t - X$ の期待値は以下となる：

$$\mathbb{E}[|A \cup B|] = s + t - \mathbb{E}[X] = s + t - \frac{st}{N}.$$

よって、J の期待値は以下のように近似できる：

$$\mathbb{E}[J] \approx \frac{\mathbb{E}[X]}{\mathbb{E}[|A \cup B|]} = \frac{\frac{st}{N}}{s+t-\frac{st}{N}} = \frac{st}{N(s+t)-st}. \tag{3}$$

この変換の結果、J_{adj} は以下の性質を持つ：

- $A = B$ の場合に限り、$J_{\text{adj}} = 1$
- $J = \mathbb{E}[J]$ の場合、$J_{\text{adj}} = 0$
- $J < \mathbb{E}[J]$ の場合、$J_{\text{adj}} < 0$

ただし、本稿では、J_{adj} が負値になる場合は、ゼロに切り上げた。

D　正答ラベル数と J_{adj} の散布図

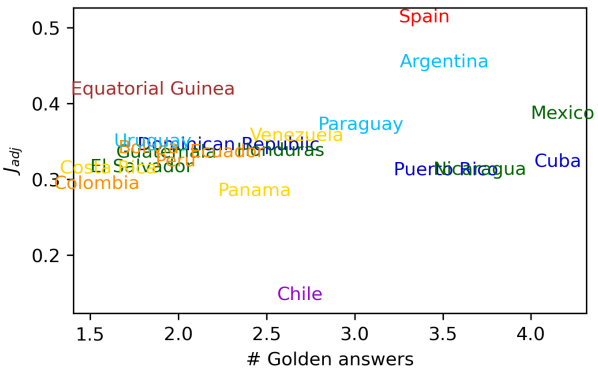

図 3　正答ラベル数と J_{adj} の散布図：$\rho = 0.183$ ($p = 0.427$)。国は方言域ごとに色分けされている。

4)　厳密には、プエルトリコは国ではなく、アメリカ合衆国の自治領である。