LGBT に関する G20 各国の Wikipedia テキストの分析

Analyzing G20 Countries' Wikipedia Texts on LGBT

大原圭人， 間中駿介， 赤木茅， 江草遼平， 橋本隆子

Keito Ohara, Shunsuke Manaka, Kaya Akagi, Ryohei Egusa, Takako Hashimoto

千葉商科大学

Chiba University of Commerce

Abstract: In order to promote LGBT participation and acceptance in society, we need to develop indicators to compare the current status of each society. This study clarifies the differences in LGBT awareness in the language space of each country by analyzing the differences in perceptions of LGBT among the G20 countries that appear in Wikipedia articles.

This paper uses a natural language processing model (BERT) to vectorize Wikipedia articles in each country's language and then uses dimensionality compression, clustering, and other processing to clarify the characteristics of LGBT awareness in each language space. We also find a correlation between LGBT awareness and the Gender Gap Index in each language space.

1. はじめに

LGBT とは,「レズビアン」,「ゲイ」,「バイセクシュアル」,「トランスジェンダー」の頭文字を組み合わせた用語である．性的少数者の総称として用いられる場合が多く,LGBT のうち,「L」「G」「B」の三者は性的指向に関わる類型であり,「T」は性自認に関する類型である．「レズビアン」は,女性同士のロマンティックな感情や性的な関係を持つ女性を指し,「ゲイ」は,通常,男性同士のロマンティックな感情や性的な関係を指すが,広義には同性愛者全般を指すこともあるとされている．「バイセクシュアル」は異性愛者や同性愛者との両方に対して感情的な結びつきを持つ人々を意味し,「トランスジェンダー」は生まれた性別と自己認識が異なると感じる人々とされている．性別適合手術を受けるかどうかに関わらず,多くの異なるアイデンティティが存在する．そのほかにもクィアやインターセックス等が挙げられるが,今回は用語として認知度が高く,一般的に用いられる LGBT に焦点を当てる．

昨今 SDGs の 17 の目標の「5．ジェンダー平等を実現しよう」や,「ダイバーシティ＆インクルージョン」として性の多様性が注目されている．しかしながら，ダイバーシティの推進の度合いは国毎に大きく異なり，また，自国内での認識や取り組みについては日常生活の中で実感することができるのに対して，他国の事情については身近でない現状がある．これはダイバーシティの推進においてモデルケースや戦略を策定するに当たり不都合である．グローバル化が進む現代社会において,各国・地域の LGBT 含むダイバーシティに関する認識の差異,世界基準での自国の現状を理解することは，重要である．

そこで本論文では，各国の LGBT に関する認識について, Wikipedia の記事の記述に基づいて明らかにすること目的とする．Wikipedia は，ボランティアの共同作業によって執筆及び作成されるフリーのインターネット百科事典である．Wikipedia には同一項目に関する記事が言語毎に作成されており，内容の詳細さや記述の厚みは，その言語を使用して記事を作成する作成者によって大きく異なる．そこで，Wikipedia から「LGBT」の記事をデータとして取得し,分析することで, 各国の LGBT に関する認識について一部を明らかにすることができると推察される．

したがって,本稿では, ベクトル化した Wikipedia の記事をテキストマイニング手法(文書類類似度,クラスタリング)によって分析し,国家ごとの LGBT に対する記述の性質の違い分析し,更にジェンダー・ギャップ指数との関連を論じる．

2. 先行研究

秋田大学教育文化学部のアマルギオアイレ・ディアナ・エレナ氏らは,「日本社会における LGBT に対する認識と 透明化のプロセス」において日本社会の LGBT に対して態度について考察している.[1]

独立行政法人の労働政策研究・研修機構は,「諸外

国の LGBT の就労をめぐる状況」において,欧米先進国を対象に就労をめぐる状況や雇用主の取り組みについて述べている．[2]

先行研究では,対象の一か国に対する分析や「就労」等の特定分野を対象とした研究が行われている．我々も同様に LGBT に焦点を当てているが,本論文では Wikipedia の書き込みを対象として,LGBT の認知に対して多角的な情報を取り込み,分析を行っている点で異なると言える．

3. 研究手法・分析結果

3.1. データ

本論文では,G20 の各国を研究対象として選定した．G20 は国際経済協力のための主要フォーラムであり,G20 メンバーの GDP 合計は世界の約 8 割以上を占めていることから,経済分野において大きな影響力を有していると言える．G20 各国が世界に与える影響力の大きさに注目し,選定の対象とした．

記述データは Wikipedia における「LGBT」というタイトルの記事から,言語毎に脚注と参考文献を省いた項目を抜粋した．この時，選定対象の国とその国の公用語あるいは最も使用されている言語での記事を紐づけた．言語と選定国の対応については，表 1 の通りである．

表 1 言語と選定国の対応

言語	国名
German	ドイツ
Spanish	メキシコ，アルゼンチン
France	フランス
English	アメリカ，イギリス， カナダ，オーストラリア,南アフリカ
Portuguese	ブラジル
Italian	イタリア
Indonesian	インドネシア
Korean	韓国
Chinese	中国
Japanese	日本
Hindi	インド
Turkish	トルコ
Arabic	サウジアラビア
Russian	ロシア

また世界経済フォーラムによるジェンダー・ギャップ指数について,「ジェンダー･ギャップ指数 2022」[3]における「総合」「経済」「教育」「政治」「健康」の 5 分野の各国のランクに基づいて順位データを採用した．ジェンダー・ギャップ指数は主に男女格差に関する指標であるが，多様性の実現の観点から，ジェンダー・ギャップの少ない国ほど社会における LGBT に関わるダイバーシティの推進を達成できていると仮定する．

表 2 ジェンダー・ギャップ指数ランキング

言語	総合	経済	教育	健康	政治
German	6	88	82	64	5
Spanish	33	110	62	49	15
France	36	95	1	41	26
English	40	51	1	76	39
Portuguese	57	86	73	1	56
Italian	79	104	60	95	64
Indonesian	87	87	106	73	81
Korean	105	114	104	46	88
Chinese	107	45	123	145	114
Japanese	125	123	47	59	138
Hindi	127	142	26	142	59
Turkish	129	113	99	100	118
Arabic	131	130	87	114	131
Russian					

Wikipedia データについては,各言語のテキストを一度英語に翻訳し,その後日本語に翻訳するという手法を採用した．これは,各言語の記事を直接日本語に翻訳した際に細かいニュアンスが直訳的に翻訳され,正確なデータ収集に問題が生じるリスクを考慮したためである．

3.2. 類似度比較

国ごとの違いを定量的に比較することを目的とし，文章類似度の比較を行う．まず，Wikipedia から取得した文章データから，自然言語処理モデル BERT(Bidirectional Encoder Representations from Transformers)を用いて言語ベクトルを計算した．各言語の記事間の類似度を言語ベクトルのコサイン距離によって求めた結果を図 1 に表す．文章が一致している場合（類似度=1.0）を最も明るい色,類似度が低くなるにつれ,オレンジ,赤,紫と暗い色に変化していくように可視化している．

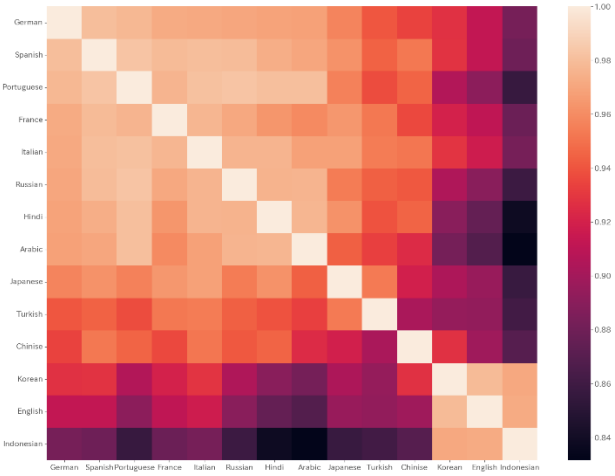

図 1 German を基準とした文章類似度のヒートマップ

図 1 より,German—Japanese,Turkish—Indonesian に二分していることが分かる．

表 3 は, Gender Gap 指数の各項目のランキングと，各項目において最もランクの高い言語に対するコサイン類似度との相関係数である．例として,総合ではドイツのランクが最も高いために,各言語のドイツとのコサイン類似度と,ドイツを除いた総合ランキングの相関係数を求めている．なお,English 及び,Spanish はそれぞれ複数国該当するため,国は,その言語圏の最も人口の多い国をランキングの値として用いた(English;米国, Spanish;メキシコ).

表 3 類似度との相関

Gender Gap 指数 項目	類似度との相関係数
総合	-0.312
経済	-0.498
健康	-0.024
教育	-0.604
政治	-0.426

表 3 より,Gender Gap 指数における教育,及び経済と言語類似度には相関関係が認められる.なお,類似度は大きい(1 に近い)ほど類似しているのに対して，ランキングは小さいほど評価が高いため負の相関となっている.

3.3. BERT を用いたクラスタリング

文章類似度比較で求められた文章ベクトルを,PCA を用いて次元削減した.その後，ward 法で階層クラスタリングを行い，5 つのクラスタに分類した(図 1).

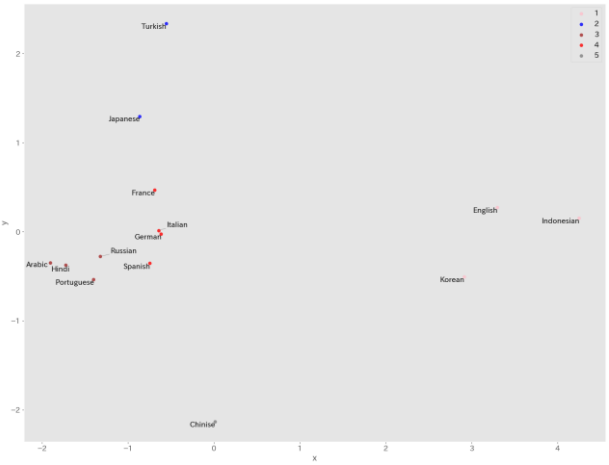

図 2 クラスタリング

表 4 クラスタリング番号と該当言語

クラスタ番号	言語名
1	English, Korean, Indonesian
2	Turkish, Japanese
3	Portuguese, Hindi, Russian, Arabic
4	German, Spanish, France, Italian
5	Chinese

図3-7は各クラスタに含まれる記事から作成した，ワードクラウドである．なお，グループごとの特徴を観察するために分析対象となるデータのうち 50%以上の言語で共通する形態素を stop word に設定し，ワードクラウド作成に用いるデータから除外した．該当する頻出単語には，LGBT,LGBTQ などが挙げられる．残ったデータから，名詞・形容詞・動詞を用いて作成した.

図 3 クラスタ 1 の言語の Wikipedia 記事における

記述をもとにしたワードクラウド

クラスタ 1 では「SGM,フラッグ,切り離す,アライ」が頻出用語として抽出された.「切り離す」は,主に English で言及されている.白人主導の LGBT コミュニティに異議を唱えた集団が,自分たちをそのコミュニティから切り離して考えるといった意味で使用されている.フラッグとは,レインボーフラッグを指し,English 圏ではこれまで性差別による歴史等から自殺への言及がされていることが分かる.「アライ」とは,LGBT では無いが LGBT の人たちの活動を支持し,支援している人たちのことを指す.現在「アライ」をイニシャルに追加するか否か論争が巻き起こり,抽出されたことが分かった.

図 4 クラスタ 2 の言語の Wikipedia 記事における記述をもとにしたワードクラウド

クラスタ 2 では「ヨーロッパ,国際,犯罪,キリスト,古代」が頻出用語として抽出された.Japanese は,他国よりも文章量が多く,特に歴史の項目が長尺で記述されていた.多くの項目が国際的な視点から記述されていた中で,具体事例では日本の反 LGBT 組織が関係し増悪犯罪を受けやすい等の記述があり,日本独自のものであることが分かった.その当時は LGBT への認知が低かったことが読み取れる.Turkish は,歴史についての記述のみであった.古代ギリシャから中世ヨーロッパまで,同性愛を禁じたキリスト教との関係を含めて記述されていた.歴史について深く言及している一方で,現在の認識は記述されていないことから,認知が進んでいないと考察した.

図 5 クラスタ 3 の言語の Wikipedia 記事における記述をもとにしたワードクラウド

クラスタ 3 では「公共,関心,エル」が頻出用語として抽出された.この集合は Hindi や Arabic 等,記述量自体が少ない傾向にある.そのため,女性が「公共」の場を得ること,加えて LGBT に関心を持つことについて記述されていることが分かった.共通して,現時点での認知度は高いとは言えないものの,前向きな姿勢であることが分かった.

図 6 クラスタ 4 の言語の Wikipedia 記事における記述をもとにしたワードクラウド

クラスタ 4 では「人口,指標,デー,タカイ,レベル」が頻出用語として抽出された.この集合は,ヨーロッパ諸国で形成されており,国際女性デーが採択された地域であることから,総じて LGBT への認知や理解が進んでいると言える.「タカイ」「レベル」で国や企業から受け入れられていることから,これらの言葉が頻出として出現している.この集合では全言語に国際女性デーの記述があるが,ドイツでは項目として「国際行動デー」がある.ヨーロッパ諸国は他の地域と比較して高い水準での認知が波及し,寛容に受け入れられていることが分かった.

図 7 クラスタ 5 の言語の Wikipedia 記事における記述をもとにしたワードクラウド

クラスタ 5 では「台湾,検索,フラッグ」等が頻出度の高い用語として抽出された.フラッグはクラスタ 1 同様,レインボーフラッグを指し,LGBT コミュニティを意味するものである.「検索」は,LGBT 資源中心（リソースセンター）で中国語の記事や情報を入手できる場として記述されていた.Chinese の記事では台湾との比較が多い.

4. 考察

今回の分析から,Wikipedia における「LGBT」というタイトルの記事と「ジェンダー・ギャップ指数」

特に教育には相関がみられ, 各国の教育の平等性が, 言語毎の記事の内容と関連があることが示された. 本稿の分析は因果に関する分析を行ってはないが, 教育によって性的マイノリティの言論空間への参加と, WEB 空間における平等意識を醸成するという関係性は, 想像に難くない.

またクラスタリングとワードクラウドの分析から, 各クラスタには地域特性や記述量に共通点があり, 定期的なクラスタリングによって今後の動向を容易に追跡できると考察する. 今後は, 手法をさらに検討し, 分析を深め, 各国における認知の違いを明確にしていきたいと考えている。

謝辞

本研究は,「千葉商科大学・数理データサイエンス教育プログラム」における「特別講義(データサイエンス)」の一環であり,千葉商科大学 基盤教育機構による助成を受けている.

参考文献

[1] 秋田大学教育文化学部,アマルギオアイレ・ディアナ・エレナ,「日本社会における LGBT に対する認識と 透明化のプロセス」,(2018.01) https://www.akita-u.ac.jp/honbu/global/ja/abroad/inbound/pdf/report_2018_01.pdf

[2] 独立行政法人, 労働政策研究・研修機構,「諸外国の LGBT の就労をめぐる状況」,(2016.05.31) https://www.jil.go.jp/foreign/report/2016/0531_01.html

[3] 世界経済フォーラム,「ジェンダー・ギャップ指数 2022」,(2022.08) https://www.gender.go.jp/public/kyodosankaku/2022/202208/202208_07.html#:~:text=2022%E5%B9%B4%E3%81%AE%E6%97%A5%E6%9C%AC%E3%81%AE,%E4%BD%8E%E3%81%84%E7%B5%90%E6%9E%9C%E3%81%A8%E3%81%AA%E3%82%8A%E3%81%BE%E3%81%97%E3%81%9F%E3%80%82