Vol.2016-DBS-163 No.8
Vol.2016-IFAT-123 No.8
2016/9/13

ツイート多言語分析に関する一検討

岡山愛[1,a)]　河合由起子[1,b)]　Muhammad Syafiq Mohd Pozi[1,c)]　Adam Jatowt[2,d)]

概要：本研究では，ジオタグ付ツイートデータを時間と場所と言語に基づき分析し，ユーザ行動に対する認知特性の解明を目指す．本稿では，認知特性の一要素として発信場所と言語形態の相違に着目し，特に多言語である欧州のツイートを対象とし，場所と言語の分析結果を可視化し，検証する．具体的には，まず，ツイッターユーザが登録時に設定した言語を母国語とし，ツイートで言及した言語を言及言語とし，ツイートユーザの多言語性を分析する．次に，ツイートにおける多言語話者の割合を明らかにする．さらに，母国語と言及言語の関係性となるヒートマップを構築し，母国語国と言及言語国間の物理的距離や言及言語の使用形態について考察する．

A Study of Microblogs Analysis based on Temporal -Spatial Language Divergence

Okayama Ai[1,a)]　Kawai Yukiko[1,b)]　Muhammad Syafiq Mohd Pozi[1,c)]　Adam Jatowt[2,d)]

1. はじめに

近年，ユーザの行動分析および可視化に関する研究において，ソーシャルネットワークサービス（SNS）データ，センサデータといった大量のストリーミングデータ分析技術が，国内外で広く注目されている．ジオタグ SNS を対象として，特定の店舗等で Check-in するユーザの移動軌跡を分析し，その店舗等のトレードエリアを抽出する手法 [1] や，タクシーに設置した GPS から取得した人々の移動パターンと地域に存在する施設のカテゴリ情報を用いて地域の機能性を発見する手法 [2] が実証されている．さらに，自然災害や疾病の流行を検出する手法 [3] や，一定領域の分析結果を地図の LOD に同期し可視化することで効果的な時空間解析が実証されている [4]．これまで著者らも，ユーザ行動分析として日本および米国の数ヶ月間のジオタグ付ツイートデータを分析し，データ発生位置とコンテンツ内容位置との差異，発生時間とコンテンツ内容時間との差異の分析，さらに位置と時間の関係性を考慮した時空間差異の分析および可視化に関する研究を行ってきた [5][6]．

本論文では，ジオタグ付ツイートデータからユーザ行動に対する認知特性の解明を目指し，ユーザ行動に対する認知特性抽出手法を提案する．本研究では，ユーザ行動に対する認知特性として，言及言語に着目する．具体的には，任意の発信位置と時刻における言語の相違（例えばパリにおけるフランス語とイタリア語，九州における福岡弁と標準語）に着目し，母国語と言及言語との差異，発信位置と各言語の発祥場所（母国語）との差異，さらに発信位置と言及言語との差異を抽出し，場所や時間における各出身地ごとの認知特性として検出する．これにより，多言語話者がツイートを発信する際に，いずれの言語を選択するかを解明でき，ユーザ行動との関連性や，各言語ごとの特性抽出につながる．本稿では，欧州におけるツイートユーザの多言語性を分析し，母国語と言及言語の言語形態の分析結果を可視化し，検証する．

2. システム概要

本研究は，ツイート発信位置，母国語，言及言語の違いによって生じる差異を可視化し，分析・検証することを目

[1] 京都産業大学
〒 603-8047 京都府京都市北区上賀茂本山
[2] 京都大学
〒 606-8501 京都市左京区吉田本町
[a)] g1344270@cc.kyoto-su.ac.jp
[b)] kawai@cc.kyoto-su.ac.jp
[c)] msyafiqpozi@gmail.com
[d)] adam@dl.kuis.kyoto-u.ac.jp

Vol.2016-DBS-163 No.8
Vol.2016-IFAT-123 No.8
2016/9/13

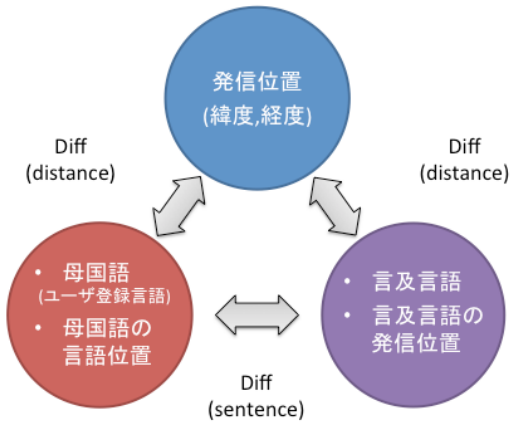

図 1 発信位置 (緯度経度), 母国語・言語位置, 言及言語・発信位置関係図

的とする（図 1）．既存研究では各言語のツイートの情報量分析としてツイートの言及言語と発信位置の関係性のみを対象としており [7] [8]，母国語と言及言語，母国語と発信位置の言語形態は分析されていない．本章では，母国語と言及言語，母国語と発信位置，言及言語と発信位置の 3 つの差異を可視化する手法について記述する．

2.1 ストリーミングツイートデータ取得とツイートの発信国名の付与

本論文では，多くの言語が使用されている欧州を対象としてストリーミングツイートデータを取得する．

まず，欧州の指定地域から重複を除いた緯度経度情報 を含むストリーミングツイートを The Streaming APIs[*1] を用いて取得する．指定地域は，1 度以上異なる南西および北東を指定することで，その 2 点に囲まれた矩形領域のストリーミングツイートを取得できる．

次に，ツイートの発信位置に基づき，発信国名を付与する．取得したストリーミングツイートの緯度経度情報から，Yahoo!ジオコーダ API[*2] を用いて住所を取得する．住所に含まれる国名を各ツイートの発信国として付与する．

以上より，ツイートユーザ ID，アイコン画像 URL，緯度，経度，ユーザ設定言語（母国語），発信国，ツイート内容，ツイート記述言語（言及言語），単語集合，取得時刻が取得付与でき，これらを一定時間管理する．

2.2 ツイートの言語分別

本研究では，図 1 に示すツイートについて発信国，言及言語，母国語の 3 点それぞれの差異を抽出する手段の 1 つとして，母国語と言及言語，母国語と発信位置，言及言語と発信位置の 3 つのヒートマップを作成する．本節では，ヒートマップに用いる発信国リスト・言及言語リスト・母国語リストの作成手順について述べる．

[*1] https://dev.twitter.com/streaming/overview
[*2] http://developer.yahoo.co.jp/webapi/map/openlocalplatform/v1/geocoder.html

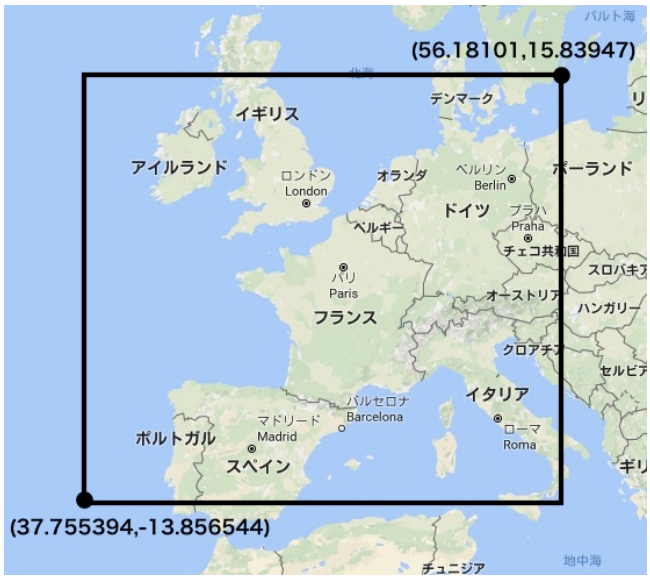

図 2 欧州ツイート取得範囲

発信国リストはツイートデータの発信国で出現する言語の集合，母国語リストはツイートデータのユーザ設定言語で出現する言語の集合，言及言語リストはツイートデータのツイート記述言語で出現する言語の集合である．

母国語と言及言語の差異

母国語リストと言及言語リストを入れ子に条件として問い合わせし，母国語と言及言語が任意の言語のときのツイート総数を取得する．

母国語（母国語発信位置）と発信位置の差異

母国語リストと発信位置リストを入れ子に条件として問い合わせし，母国語と発信位置が任意の言語（国名）のときのツイート総数を取得する．

言及言語（言及言語の発信位置）と発信位置の差異

言及言語リストと発信位置リストを入れ子に条件として問い合わせし，言及言語と発信位置が任意の言語（国名）のときのツイート総数を取得する．

以上の母国語と言及言語，母国語と発信位置，言及言語と発信位置の 3 種類の差異を用いて，ヒートマップを作成する．本論文では特に，母国語と言及言語間の差異を分析する．

3. 多言語分析に関する検証

本研究では，言語形態を可視化することを目的としており，多様な言語が使用されている欧州を対象とする．今回，2016 年 4 月 29 日から 7 月 30 日の約 3 ヶ月間，欧州の 19 ヵ国に対するツイートを収集できた（図 2）．本論文では，表 2 に示す欧州における 3 ヶ月間の 8,725,149 ツイートを検証対象とした．

表 1 母国語および言及言語表

en	英語	** tl	タガログ語	lv	ラトビア語	bg	ブルガリア語
* en-gb	英語（イギリス）	** ht	ハイチ語	* he	ヘブライ語	* ga	アイルランド語
* en-AU	英語（オーストラリア）	tr	トルコ語	ko	韓国語	sr	セルビア語
* en-IN	英語（インド）	ro	ルーマニア語	** zh	中国語	** iw	不明（未確認）
es	スペイン語	ru	ロシア語	* zh-cn	大陸中国語	** ne	ネパール語
es-MX	スペイン語（メキシコ）	no	ノルウェー語	* zh-Hans	簡体中国語	bn	ベンガル語
fr	フランス語	* nb	ノルウェー語（ブークモール）	* zh-Hant	繁体中国語	** ps	パシュトー語
* fr-CA	フランス語（カナダ）	* fil	フィリピン語	* zh-tw	台湾語	** ur	ウルドゥー語
* fr-BE	フランス語（ベルギー）	pl	ポーランド語	* gsw	アルザス語	** ka	グルジア語
de	ドイツ語	* id	インドネシア語	th	タイ語	* sq	アルバニア語
* de-CH	ドイツ語（スイス）	* gl	ガリシア語	** is	アイスランド語	** pa	パンジャーブ語
it	イタリア語	eu	バスク語	hi	ヒンディー語	** ta	タミル語
nl	オランダ語	* hr	クロアチア語	* ms	マレーシア語	** ckb	不明（未確認）
* nl-BE	オランダ語（ベルギー）	cy	ウェールズ語	* af	アフリカーンス語	** lo	ラーオ語
pt	ポルトガル語	ja	日本語	el	ギリシア語	** mr	マラーティー語
* pt-PT	ポルトガル語（ポルトガル）	ar	アラビア語	* xx-lc	不明（未確認）	** my	ビルマ語
da	デンマーク語	fi	フィンランド語	uk	ウクライナ語	** si	シンハラ語
* ca	カタルーニャ語	sl	スロベニア語	vi	ベトナム語	** sd	シンド語
cs	チェコ語	lt	リトアニア語	fa	ペルシア語		
** et	エストニア語	* bs	ボスニア語	* msa	マレー語		
sv	スウェーデン語	hu	ハンガリー語	* sk	スロバキア語		

*は母国語言語のみ出現.
**は言及言語のみ出現.

3.1 欧州におけるツイートの特徴

本節では，欧州における言語の多様性について検証する．表 1 に取得した母国語および言及言語の種類を示す．

まず，約 3 ヶ月間におけるツイートから取得できた母国語は 65 種類であった．これら母国語をツイート数の多い順にランキングすると，1 位が英語で半数を占めることがわかる（図 3）．次いで，スペイン語で約 17 %，3 位のフランス語以下では数%となり，10 位以下は 1 %以下であった．また日本語は約 0.35 %であった．なお，2 位のスペイン語の話者数は約 4 億 2000 万人 [*4]，3 位のフランス語の話者数は約 2 億 2000 万人 [*5] である．

次に，表 1 よりツイートで発言される言及言語は 53 種類であった．これら言及言語をツイート数の多い順にランキングすると，母国語同様に 1 位が英語で半数を占めていることがわかる（図 4）．次いで，スペイン語で約 13 %，3 位のフランス語以下では数%であった．

以上の結果より，母国語が 65 種類，言及言語が 53 種類で英語が半数以上になっていることから，母国語ではなく英語で情報を発信するユーザが多いことがわかる．実際に，例えば，日本語を母国語としているツイート数が 30,387（約 0.35 %）に対して，言及言語が日本語以外では英語が最も多いことが確認できた．

さらに，ノイズを削除したツイートデータに関しても検証を行った．今回，下記の 2 種類をノイズとした．

- 3 ヶ月間で発信したツイートが 2 ツイート以下のユーザが発信したツイート
  本研究では，図 1 で定義したツイート間の差異の可視化を目的としているため，同一ツイートユーザの 3 ツイート以上のツイートを対象とした．
- bot ツイート
  今回，3 ヶ月間で発信したツイート数でランキングを作成し，上位 300 アカウントを bot とみなしノイズとして除去した．閾値は，100 アカウントごとに上位 10 アカウントを 2 名による目視により判定し，6 割の bot アカウントを含む 300 とした．なお 400 アカウントで

表 2 ツイートストリーミングデータ

場所	開始日時	経過日時	ツイート数 [個]	量 [KB]
欧州	16-04-29	16-07-30	8,725,149	2,309

[*4] https://ja.wikipedia.org/wiki/スペイン語
[*5] https://ja.wikipedia.org/wiki/フランス語

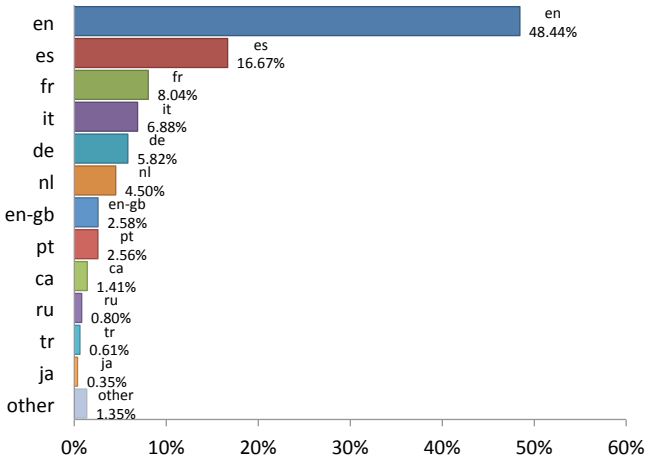

図 3 欧州ツイート母国語言語割合

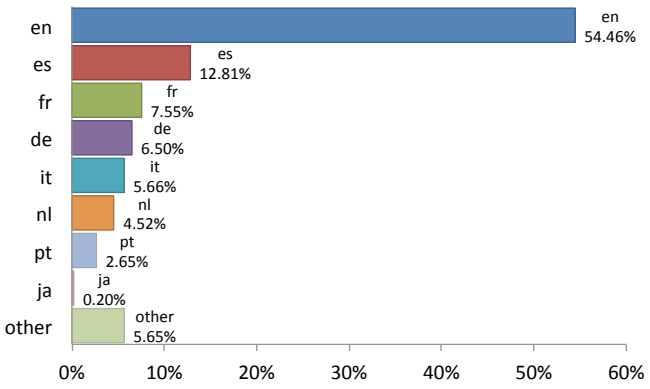

図 4 欧州ツイート言及言語の割合

は bot が 2 アカウント含まれていた．

ノイズを削除したツイートから取得できた母国語は 42 種類であった．ノイズ除去前から減少した 23 言語を，表 3 に示す．

表 3 ノイズ除去により削除された言語

ノイズ除去により削除された母国語
ベトナム語，アルザス語，スロベニア語，リトアニア語， フランス語（カナダ），スペイン語（メキシコ），ボスニア語， ノルウェー語（ブークモール），英語（オーストラリア）， 英語（インド），繁体中国語，ドイツ語（スイス），ヒンディー語， マレーシア語，アフリカーンス語，グルジア語，アルバニア語， フランス語（ベルギー），オランダ語（ベルギー），ウェールズ語
ノイズ除去により削除された言及言語
マラーティー語，タミル語，パンジャーブ語，シンド語

これらをノイズ除去前同様，ツイート数の多い順にランキングした結果を図 5 に示す．図 5 より，1 位は英語で半数を占め，次いで，スペイン語で約 14 %となり，ノイズ除去前と同程度であったが，3 位以下では順位がそれぞれフランス語からイタリア語，ドイツ語からオランダ語と順位が入れ替わり，7 位がポーランドと変化した．

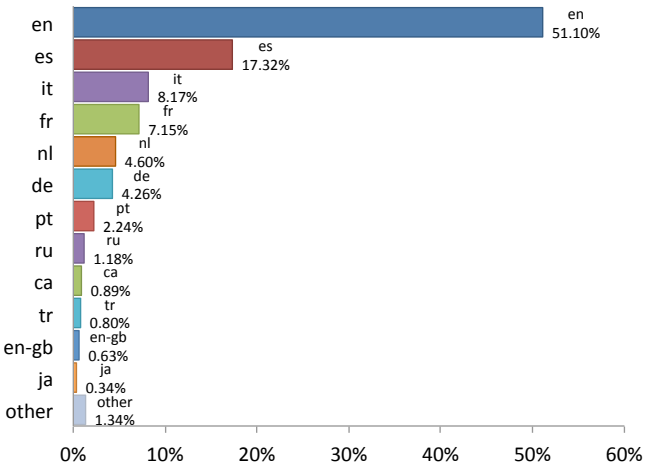

図 5 欧州ツイート母国語言語の割合 (ノイズ除去後)

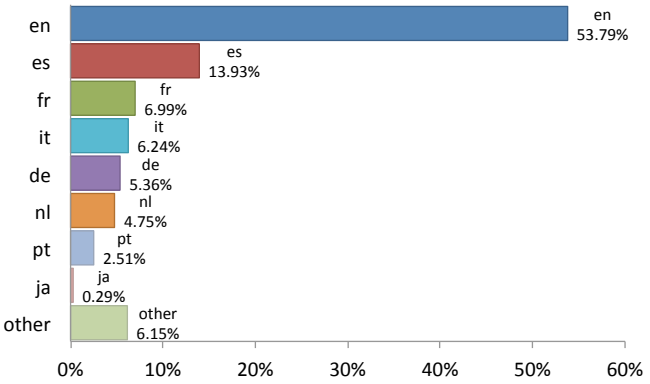

図 6 欧州ツイート言及言語の割合 (ノイズ除去後)

次に，ノイズを削除したツイートから取得できた言及言語は 49 種類となり，ノイズ除去前から減少した 4 言語は，マラーティー語，タミル語，パンジャーブ語，シンド語であった（表 3）．これらをノイズ除去前同様，ツイート数の多い順に言及言語ごとにランキングした結果を図 6 に示す．図 6 より，1 位は英語で半数を占め．次いで，スペイン語で約 14 %，3 位にフランス語で約 7 %となり，ノイズ除去前と同程度であったが，4 位以下では順位がドイツ語からイタリア語と順位が入れ替わった．オランダ語（6 位）とポルトガル語（7 位）はノイズ除去前同様の順位であった．なお，以上のことから英語以外を母国語とする場合，言及言語の英語を本研究では公用語とする．

3.2 欧州におけるユーザのツイートにおける多言語性

本節では，ユーザの多言語性を言及言語を用いて検証する．表 2 のツイートにおけるユーザ総数は 614,292 人であった．図 7 に Monolingual（単一言語話者）から Manylingual（五言語以上話者）のユーザの割合を各週ごとに示す．最初の 17 週目を除いて単一言語話者が平均 83.4 %を占めており，多くがツイートでは英語のみを用いていることがわかる．また，二言語話者以上は 16.6 %となるが，そのう

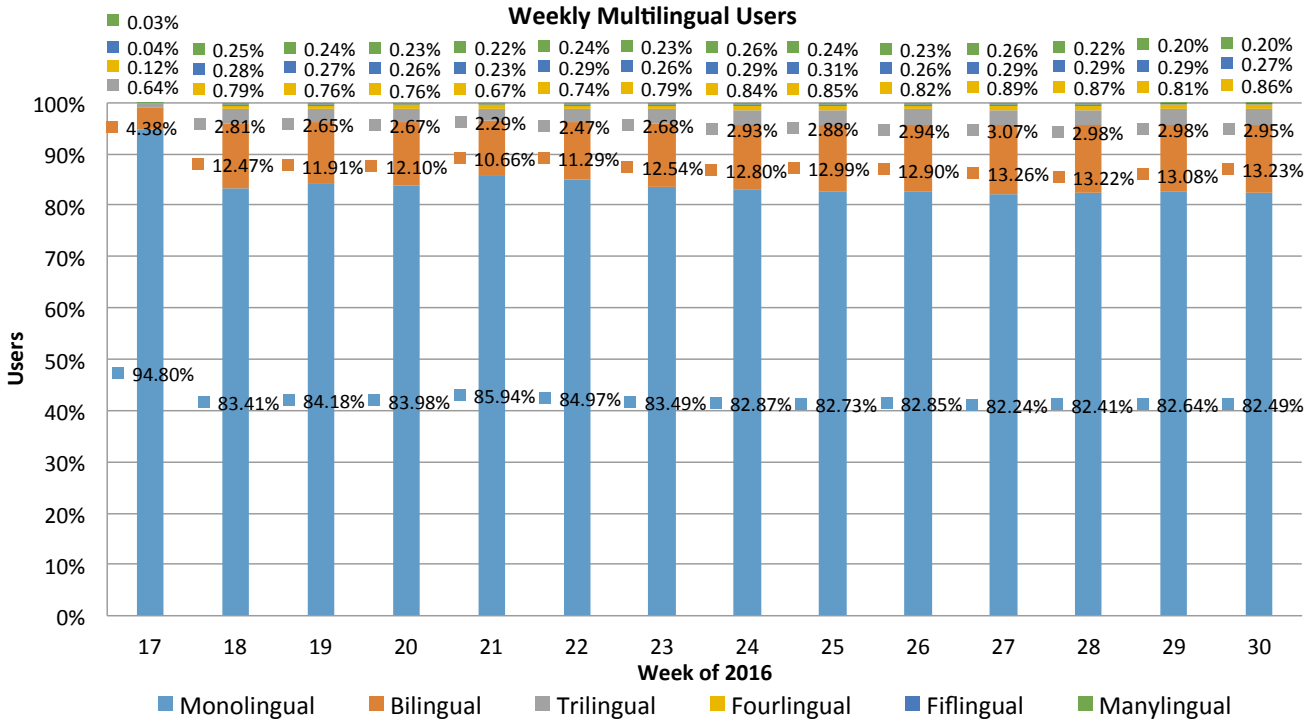

図 7 欧州多言語話者の割合

ち二言語を使用しているユーザが平均 12.5 %となることがわかる．前節の母国語と言及言語の差異を考慮すると，ツイートの内容や発信するときの状況によって母国語以外の言語を使いわけている可能性を示唆しているといえる．今後，これら多言語の話者がどこで言語を使い分けているかを検証予定である．

3.3 可視化ヒートマップの分析・検証

本節では，作成した母国語と言及言語のヒートマップ [*6] について検証する（図 8）．ヒートマップの作成には D3.js[*3] を使用した．今回，最大値と最小値の差が大きいため，ヒートマップのカラーは対数変換を行った．

ヒートマップの縦軸である言及言語はツイート数の多い順に上から 34 言語を配列し，横軸である母国語は左からツイート数の多い順に 56 言語を配列した．また，母国語と言及言語が同一のセルを正方形で，特徴的な結果のセルを丸で囲った．

少ない言及言語の発見

図 8 では，横軸の母国語に対して縦軸の言及言語の多くが用いられていることが分かるが，ロシア語が言及言語の行を見ると，母国語に対するツイート数が少ないことがわかる（図中「ru」の行）．これはロシア語が他の国ではあまりツイートの発話として利用されていないと言える．また，トルコ語，日本語，アラビア語にも同様のことが言える．これら言語は多様性が低く，母国語のユーザたちの間で多く使用する傾向にあるのではないかと考察される．

母国語より英語が多く使用されている言語の発見

図 8 より母国語が上位の言語を見ると，英語より母国語と同一の言語（正方形部分）を使用しているツイートが多いことがわかる．例えば，母国語をフランス語としている人たちは，英語よりフランス語でより多くツイートしている．しかし，下位にいくと母国語より英語の使用率が高くなる逆転現象が起こった．例えば，図中（∗1）のハンガリー語（hu），ギリシア語（el），ラトビア語（lv）が挙げられる．ハンガリー語の話者は約 1450 万人 [*7]，ギリシア語の話者は約 1200 万人 [*8]，ラトビア語の話者は約 190 万人 [*9] と少ないことからも，母国語ではなく公用語を意図的に使用している可能性が考えられる．

国の位置関係の影響

図 8 の（∗2）より，母国語がロシア語で言及言語がドイツ語のツイートが多いことがわかる．同様に（∗3）（∗4）の母国語がスペイン語に対して，言及言語が隣国のポルトガル語やイタリア語やドイツ語が多い．以上より，母国語国と言及言語国の物理的距離の近さの影響が一要因として考えられる．

移民（留学生）の発見

図 8 の（∗5）より，母国語が台湾語に対して欧州以外に位置する日本語のツイートが多いことがわかる．これらの

[*6] http://yklab.cse.kyoto-su.ac.jp/~okayama/Europe/EuropeLangdata_v2.html
[*3] http://ja.d3js.node.ws/
[*7] https://ja.wikipedia.org/wiki/ハンガリー語
[*8] https://ja.wikipedia.org/wiki/ギリシア語
[*9] https://ja.wikipedia.org/wiki/ラトビア語

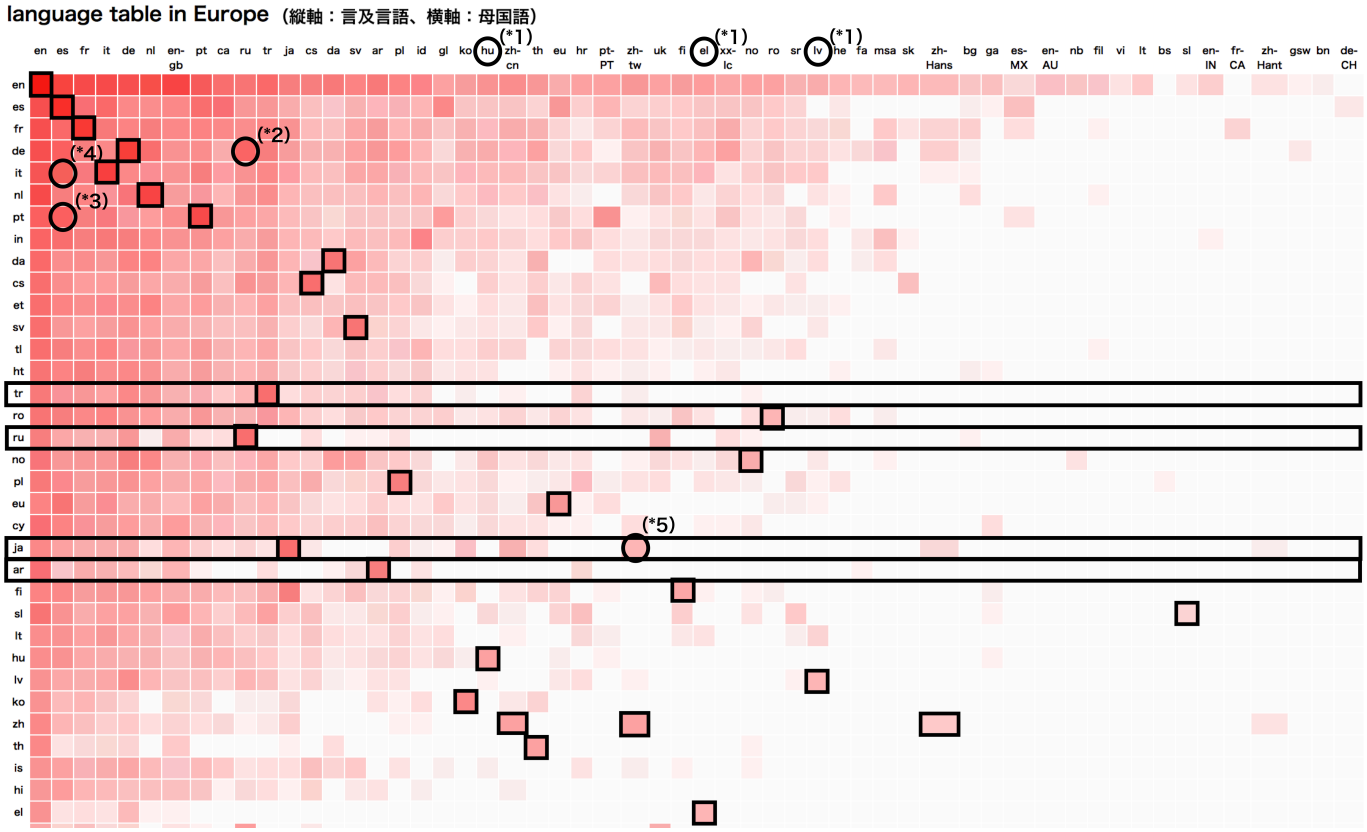

図 8 欧州言語ヒートマップ (縦軸：上位 34 言及言語，横軸：上位 56 母国語)*6

ツイートの中身に注目すると，台湾から日本への留学生のアカウントがあり，発信位置と母国語と言及言語の差異をとることで，移民（留学生）の発見ができる可能性を示唆している．

4. まとめ

本論文では，ユーザ行動に対する認知特性の解明を目指し，ユーザ行動に対する認知特性として，言語形態に着目し，任意の発信位置と時刻における言語の相違，母国語と言及言語との差異，発信位置と各言語の発祥場所（母国語）との差異，さらに発信位置と言及言語との差異を抽出し，場所や時間における各出身地ごとの言語形態の分析結果を可視化し，検証した．その結果，ツイートの内容や発信時の状況から，母国語以外の言語を使い分けている可能性が示された．また，母国語以外を使用する際に，各国の位置関係の影響や，移民するツイートユーザが多い国の発見ができた．今後，言及言語や母国語の発祥場所とは異なる国で，どこの場所でどの言及言語でを使用してツイートを発信しているかを分析し，ユーザ行動との関連性を明らかにし，各言語ごとの特性抽出を検討する．

謝辞

本研究の一部は，JSPS 科研費 16H01722 の助成を受けたものである．ここに記して謝意を表す．

参考文献

[1] Qu et al.: *Trade Area Analysis using User Generated Mobile Location Data*, WWW2013 (2013).
[2] Yuan et al.: *Discovering Regions of Different Functions in a City Using Human Mobility and POIs*, KDD2012 (2012).
[3] Sakaki et al.: *Earthquake shakes Twitter users: real-time event detection by social sensors*, WWW2010 (2010).
[4] Magdy, A., Alarabi, L., Al-Harthi, S., Musleh, M., Ghanem, T. M., Ghani, S., and Mokbel, M. F.: *Taghreed:A System for Querying, Analyzing, and Visualizing Geotagged Microblogs*, SIGSPATIAL 2014, pp. 163-172 (2014).
[5] Shoko Wakamiya, Adam Jatowt, Yukiko Kawai and Toyokazu Akiyama.: *Analyzing Global and Pairwise Collective Spatial Attention for Geo-social Event Detection in Microblogs*, WWW 2016, ACM Press, Montreal, Canada, demo paper pp. 263-266 (2016).
[6] Émilien Antoine, Adam Jatowt, Shoko Wakamiya, Yukiko Kawai, and Toyokazu Akiyama.: *Portraying Collective Spatial Attention in Twitter*, KDD 2015, pp. 39-48, Sydney, Australia, August (2015).
[7] Delia Mocanu, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang, AlessandroVespignani.: *The Twitter of Babel: Mapping World Languages through Microblogging Platforms*, PLoS ONE 8(4): e61981. doi: 10.1371/journal.pone.0061981.
[8] Graham Neubig, Kevin Duh.: ツイートの情報量について－情報理論に基づく多言語調査－, 言語処理学会 第 20 回年次大会発表論文集 (2014).