論　文

Twitterを利用した、特定の話題に特徴的な語彙の収集*

阿部一哉

0. はじめに

本稿の目的は、SNSの一つであるTwitterを利用して、特定の話題に特徴的なドイツ語の語彙を、収集する手法を提案することである。

例えばドイツ語授業の活動として、特定の話題についてドイツ語でことばのやり取りを行う際、その話題に特徴的なドイツ語語彙集があると便利である。

しかしながら「話題」は特定の形式を持った概念ではないため、話題に特徴的な語彙を収集するためには、何らかの工夫が必要である。

Twitterは、「ツイート」と称される140文字以内の短文の投稿を共有するウェブ上の情報サービスである(1)。ツイートには、個人ユーザーが「自分が休日に行った場所」を話題にするものもあれば、首相官邸が「総理の動き」を話題にしたものもある。ありとあらゆるユーザーが、ありとあらゆる話題についてのツイートを、日々投稿しているのである。

ところでTwitterには、「ハッシュタグ」と呼ばれる記法がある。これは、ツイートの話題をユーザーが意識的にラベル付けするために用いる記法で、ハッシュ記号（#）を特定の語の前につけることで表される。例えば、次のように用いられる。

Wie kann ich Asyl beantragen? Wo lerne ich Deutsch? **#Flüchtlinge** in Deutschland-DW-Spezial: http://www.dw.com/de/themen/erste-schritte-in-deutschland/s-32443…
（https://twitter.com/dw_deutsch/status/672072761313890304 太字は阿部）

これはドイチェ・ヴェレの2015年12月2日付けのツイートで、太字で示した**#Flüchtlinge**がハッシュタグである。ツイートを投稿するユーザーは、この記法を用いることで、当該ツイートがFlüchtlinge、すなわち「難民」を話題にしていることを、意図的に明示することができる。

そして、Twitterを閲覧しているユーザーは、任意のツイートにおける特定のハッシュタグをクリックすることで、そのハッシュタグを持つツイートを集約して表示することができる。

このようにハッシュタグは、特定の話題を意識的に明示するために用いられるが、そのことによって特定の話題についてのツイートを、集約することが可能になっている。

話題に特化した語彙を収集するためには、まずは特定の話題についての発話を収集する必要がある。そのために本稿では、このTwitterのハッシュタグに着目する。

2. 特定の話題に特化した語彙の収集

本稿では、特定の話題に特化した語彙を抽出するために、概略次の手法を採る。まずTwitterのハッシュタグに着目し、特定の話題についてのツイートを収集する。次に、そこから特徴的な語彙を抽出する。このような手法を確立するために、本章では次の二点の問題提起を取り扱う。

［1］ Twitterのハッシュタグに着目して、ツイートを収集することは可能か。可能であるとしたら、それはどのような方法によるのか。
［2］ 特定のハッシュタグに着目して得られたツイートの集合から、特徴的な語彙を抽出することは可能か。可能であるとしたら、それはどのような方法によるのか。

以下、これらの問題について分析を進めて行く。その際、ハッシュタグ#Flüchtlingeに着目して作業を進める。2.1節では問題提起［1］を、2.2節では問題提起［2］を扱う。

2.1 ハッシュタグに基づくツイートの収集

本節では、ハッシュタグ#Flüchtlingeを持つツイートの収集手法について見る。

Twitterは、任意のプログラム言語からサービスを利用するために、API（Application Processing Interface）と呼ばれる仕組を提供している。APIを利用すると、一定量のツイートを比較的短時間で収集することが可能である。よってここでは、ツイートを収集するために、APIを利用する[(2)]。

Twitterは様々なAPI[(3)]を提供しているが、今回はRESTスタイルのSearch APIを利用する。以下に挙げるツイートを収集する処理を、プログラム言語Pythonで記述し実行する。

(1) APIを利用して、インターネット経由でTwitterにアクセスする[(4)]。
(2) ハッシュタグ#Flüchtlingeを含み、ドイツ語で書かれたツイートを収集する。この処理により、最新の物から遡って、最大100ツイートを収集することができる。
(3) 処理（2）で得られるツイートのうち、最も古いツイートのidを利用して、

さらに古い100ツイートを取得する[5]。この処理を可能な限り繰り返す[6]。

(4) 得られるツイートはドキュメント・データベースのひとつ、mongoDBを利用して保存する（阿部2014，Kumar et al. 2014参照）。

この一連の処理を行った結果、収集することができたツイート・データは、概略以下の通りである。

ツイートが投稿された期間：2015年11月28日～12月4日
ツイート数：　　14903

現在では一般的に、電子化された大量の言語データを「コーパス」と呼び習わし、さまざまな分析手法が提案されている。本稿でも、得られたデータを「ツイート・コーパス」と名付け、コーパス分析手法を援用し、分析を行っていく。

以上本節では、Twitter APIを利用することにより、特定のハッシュタグを持つツイートを大量に収集することが可能であることを見た。

2.2　ツイート・コーパスに基づく語彙の抽出

本節では、得られたツイート・コーパスから、話題に特徴的な語彙を抽出する手法について見る。

ツイートは、本文だけではなく、ユーザー名や投稿日時など様々な要素からなる、複合データである。ここでは、本文のみを必要とする。ツイートデータは、Json形式で保存されている。Jsonとは、文字列、数値、リストといった様々なタイプのデータを、ラベル付けして整形したデータ形式である。ツイートのJsonデータの場合、本文は“text”と、ラベル付けされている。

そこで、再びPythonで処理を記述し、大量のツイートデータから“text”とラベル付けされた本文を抽出・集約したところ、ツイートコーパスは、概略次の規模のデータであることが分かった。

延べ語数：　　225,842語
重なり語数：　　25,798語

「延べ語数」とは、コーパス全体で使用されている語の数のことで、ドイツ語の場合空白で区切られた文字のかたまりを、数えて得られる。「重なり語数」とは形が異なる語の数のことで、例えばKatze「猫」がコーパス全体で百回用いられていた場合も、1と数える[7]。

このツイッター・コーパスを特徴づける語、すなわちハッシュタグ#Flüchtlingeで表される「難民」という話題について、特徴的な語とはどのようなものであろうか。ここでは、まずは語の使用頻度に着目したい。語の使用頻度とは、その語がコー

パスにおいて繰り返し用いられている回数のことを言う。コーパスの、例えばジャンルなどの特徴が、語の使用頻度分布に現れることが、すでに様々な先行研究において確かめられている（例えば石川2001）。ツイッター・コーパスでも、語の使用頻度に着目することで、特徴的な語を抽出することが可能になることが期待される。

そこで、ツイッター・コーパスで使用されている語の頻度を計算し、高いものから降順に並べ、上位30語を取り出すと、以下の通りになる。なお、語の後の（　）内の数字は頻度を表している。

RT（9703）, #Flüchtlinge（6658）, in（4110）, der（3637）, #Fluechtlinge（2839）, die（2758）, für（2593）, mit（2504）, –（2474）, und（2369）,（2006）, sie（1740）, vor（1670）, #flüchtlinge（1571）, einer（1460）, #syrien（1318）, Spende（1308）, Ort!）, https://t. co/sckoVALE67（1308）, @spendenhilfe:（1306）, von（1201）, auf（1166）, zu（1060）, nicht（1014）, ist（1002）, Die（978）, https://…（899）, im（820）, aus（818）, das（811）

このリストを見ると、かなり雑多なものが混じっている印象を受ける。

RT, #Flüchtlinge, @spendenhilfeなどは、ツイート特有の表記である。これらの「語」めいたものは、「話題に特徴的な語彙」としては、あまり相応しくない。また、http…で始まっているものは、他のWebサイトへのリンクであり、やはり「話題に特徴的な語彙」としては、相応しくないと言える。

また、語であっても、in, der, mit, undといった、機能語と呼ばれる部類のものは、具体的な意味内容が希薄で、「話題に特徴的な語彙」としては、相応しくないと言えるだろう。

更に、「語彙」という観点から見ると、istのような、動詞の一変化形を「一語」と数えるのか、seinという一つの辞書形に集約するのかというと、後者の方が妥当であると思われる。異なる変化形をひとつの辞書形に集約することを「レマ化」と呼ぶが、ここではツイートで使用されている語形をレマ化したほうが良い、ということになる。

このような観点から、上のリストを見直すと、ツイート・コーパスに特徴的な語の候補としては、Spendeかsein（←ist）ぐらいしか残らない。単純に使用頻度を見るだけでは、話題に特徴的な語を抽出することは難しい、と言える。

これらの問題点を解決するために、本稿では形態素解析と呼ばれる技術を援用する。形態素解析とは、テキストを文・語に分節し、品詞判定を行う処理のことである。ここでは、ドイツ語の形態素解析エンジンの一つ、TreeTaggerを使用する。TreeTaggerは、分節・品詞判定に加え、レマ化も行ってくれる。具体的には、Das ist ein Test.のような文が、次のような形に処理される。

Das	PDS	die
ist	VAFIN	sein
ein	ART	eine
Test	NN	Test
.	$.	.

一行が一語に対応している。左端のカラムが、文中での語形、真ん中のカラムが品詞（例えばNNは「普通名詞」）、右端のカラムが辞書形を示している。

このような処理をツイッター・コーパスの本文データに施すことで、上で挙げた問題の大部分が解決することが期待される。

そこで、TreeTaggerによる処理を行った上で、品詞を動詞、名詞、形容詞、副詞といったいわゆる「内容語」に限定しフィルタリングを行った。その上で、語の頻度を計算し、高いものから降順に並べ、上位30語を取り出すと、以下の通りになる。なお、ここでも語の後の（　）内の数字は頻度を表している。

<unknown>（91370）, RT（9707）, Ort（1334）, Spende（1309）, Syrien（798）, auch（731）, direkt（724）, Helfen（713）, helfen（693）, lieb（618）, Deutschland（618）, Flüchtling（511）, so（491）, kommen（455）, noch（447）, heute（440）, nur（439）, jetzt（379）, sagen（353）, nehmen（336）, mehr（332）, Polizei（330）, Frau（317）, geben（314）, wieder（311）, EU（305）, machen（302）, Land（294）, fordern（293）, hier（286）

ツイート特有の表記を「特徴的な語彙」から除外するべきことについてはすでに触れたとおりである。これらの語の辞書形は、このリストの場合、<unknown>と判断されているようである。従って、このリストから<unknown>を取り除くことにより、ツイート固有の表記をあらかた取り除くことができる[(8)]。

残った語はしかしながら依然、「話題に特徴的な語彙」というよりは、むしろあらゆる話題に関して共通して用いられる基本的なものが多い印象を受ける。例えばOrt「場所」, lieb「好ましい」, jetzt「今」, sagen「言う」, geben「与える」といった語は、極めて基本的な語彙に含まれる語である。なお日常的に基本的に用いられる語彙は、「基礎語彙」と呼ばれる。

従って、高頻度の内容語のリストに限定しても、なお話題に特徴的な語を抽出することは難しい。

高頻度の内容語リストから、「話題に特徴的な語彙」を得るためには、そこから基礎語彙に属する語を除去する必要があると言える。この作業を行うためには、基礎語彙に属する語の判定が必要になるが、ここでは便宜上既存の基礎語彙リストを使用したい。

ドイツ語の基礎語彙については、Jones&Tschirner（2006）を始めとする様々な

先行研究が認められる[9]が、ここでは、電子化されていて使いやすい、という理由で大薗（2014a）の基礎語彙集を利用する。

ツイッター・コーパスの内容語リストから、大薗の基礎語彙リストに記載のある語を除外した。その上で、語の頻度を計算し、高いものから降順に並べ、上位30語を取り出すと、以下の通りになる。なお、ここでも語の後の（　）内の数字は頻度を表している。

<unknown>（91370）, RT（9707）, Spende（1309）, Syrien（798）, Helfen（713）, %（247）, Http（224）, logo（223）, Provokation（214）, Weitersagen（164）, Untertitel|Untertiteln（163）, Massenschlägerei（162）, Asylunterkunft（150）, Merkel（142）, Flüchtlingsheim（129）, Syrer（125）, Polizeichef（99）, empören（92）, Einzelfallprüfung（89）, Bürgerbündnis（88）, Flüchtlingskrise（86）, D（86）, Slowakei（85）, Filiale（84）, Arbeitslosenstatistik（80）, de（73）, Schwächste（73）, vergewaltigen（72）, Flüchtlingsunterkunft（68）, zeihen|ziehen（67）

そして、上で述べたように、ここから<unknown>やRTといった、通常の語ではないものを取り除くと、次のリストになる。

Spende（1309）, Syrien（798）, Helfen（713）, Provokation（214）, Weitersagen（164）, Untertitel|Untertiteln（163）, Massenschlägerei（162）, Asylunterkunft（150）, Merkel（142）, Flüchtlingsheim（129）, Syrer（125）, Polizeichef（99）, empören（92）, Einzelfallprüfung（89）, Bürgerbündnis（88）, Flüchtlingskrise（86）, Slowakei（85）, Filiale（84）, Arbeitslosenstatistik（80）, Schwächste（73）, vergewaltigen（72）, Flüchtlingsunterkunft（68）, zeihen|ziehen（67）
※上のリストから7項目除外、計23語

Spende「寄付（金）」、Syrien「シリア」、Helfen「助ける（こと）」など、なるほど「難民（Flüchtlinge）」と関連のありそうな語が並んでいることが分かる。

以上、ツイッターのハッシュタグに着目することで、同じ話題についてのツイートを収集し、そこから話題に特徴的な語彙を抽出する手法についての分析を行った。その結果、本稿で提案する手法は一定の成果が得られることを示した。

3. 考察

Twitterのハッシュタグに着目することにより、話題に特徴的な語の抽出が、可能であることが確かめられた。以下、本章では、この分析結果について若干の考察を行う。

まず、手法についてであるが簡単なプログラミングの知識があれば、極めて容易に語彙リストを作成することができることが分かった。ハッシュタグを「話題」と呼ぶことについては更なる議論が必要であるが、技術的には可能、ということである。

次にツイート・データの活用法について。ツイートの中には、以下のように新聞サイトの記事へのリンクを持つものも多い。

45.000 Paar Schuhe sind in den Geschäften des GMS-Schuhfachverbunds als Spende für #Flüchtlinge gesammelt worden. https://www.schuhmarkt-news.de/business/organisationen/12-01-2016-gms-haendler-spenden-schuhe/…（太字は阿部）

ツイートデータを語彙収集のリソースとしてだけではなく、話題についての、長文テキストを収集するためのハブとして機能させることも可能であろう。ツイートデータには、これら以外にも様々な活用の可能性が潜んでいそうである。

次に、得られた語彙リストについてであるが、授業などで活用するには、更なる議論が必要であろう。ただし、どのような活用法を取るにせよ、リストにある語の使用事例は、いずれにせよ必要になるだろう。

使用事例については、学習者が簡単にPCやスマートフォンを使用して、確認することができる。例えば、Spende「寄付金」が、「難民（Flüchtlinge）」という話題において、どのよう事例で使用されているのか、具体的なツイートを確認したい場合、Twitterにおいて、以下の内容で検索を行えば良い。

Spende #Flüchtlinge lang：de

実際に検索を行うと、次のような一連のツイートが得られる。本文のみ抜粋して挙げる。

(1) "@Berlinale for #Refugees": Ihre Spende bei @betterplace_org unterstützt Projekte für #Flüchtlinge in #Berlin.
(2) #Merkel s Fluchtpunkt Chile? Warum die Aufregung? Wäre doch prima! (Zudem gut für Dtschld.) Bin zu Spende bereit! #annewill #Flüchtlinge
(3) Der Beginn von jedem Verständis, ist die gemeinsame Sprache: GLS-Bank IBAN DE18430609677039416400 #Flüchtlinge #Asylbewerber #Spende
(4) Wir unterstützen #Flüchtlinge in #Afghanistan–helfen Sie uns dabei mit einer Spende!

(5) 45.000 Paar Schuhe sind in den Geschäften des GMS-Schuhfachverbunds als Spende für #Flüchtlinge gesammelt worden.

このように、話題に特徴的な語が、「ホットな話題」において、具体的にどのような事例で使用されているのかを見ることにより、現在のドイツで、どのような言説がかわされているのかをうかがい知ることが可能になる。このようなアクティビティを授業に取り入れることは、ドイツの現在を知るために、有用であることは言うを待たない。

ただし、このような授業を行うためには、話題についての特徴的な語彙を、可能な限り素早く、何度も収集することが可能になる必要がある。

今回の分析では、特にプログラムの再利用を視野に入れずに作業を行ったため、素早い処理を繰り返し行うためには、処理をモジュール化する必要があるが、これについては今後の課題としたい。

4. 結語

本稿では、本来難しかった、「話題に特徴的な語彙」の収集が、Twitterのハッシュタグに着目することで可能になることを示した。

また、ハッシュタグの言語データとしての位置づけについては、慎重に議論を行う必要があること、また手法の再利用性を保証するために、処理のモジュール化が必要であること、の2点を今後の課題として確認した。

＊本項は平成27年度跡見学園女子大学特別研究「跡見学園女子大学ドイツ語e-learningシステムの構築と運用」の研究成果のひとつである。

参考文献

Jones, Randall & Tschirner, Erwin (2005): A Frequency Dictionary of German: Core Vocabulary for Learners (Routledge Frequency Dictionaries), Routledge.

Kumar, Shamanth, Morstatter, Fred & Liu, Huan (2014): Twitter Data Analytics. Springer.

阿部一哉 (2014):「新聞記事によるドイツ語コーパス構築」日本ドイツ語情報処理学会編ドイツ語情報処理研究 (24), 15-29, 2014

石川慎一郎 (2001):「テキストジャンルと構成語彙」『KELT』(神戸英語教育学会), 16 3-17 2001.

大薗正彦 (2014a):「基本語彙と頻度――実践と課題」恒川元行／大薗正彦 (編)『コーパス利用に基づくドイツ語研究――幅広いデータ収集と頻度から見直す』(日本独文学会研究叢書98), 49-64, 2014.

―― (2014a):「ドイツ語基本語彙リストの比較」『ドイツ文学論集』(日本独文学会中国四国支部) 47, 47-61. 2014.

注

(1) 'Twitter'『フリー百科事典 ウィキペディア日本語版』(http://ja.wikipedia.org/)。2016年1

月29日20時（日本時間）現在での最新版を取得。

(2) ブラウザ上で、同じハッシュタグを持つツイートを集約して閲覧することも可能だが、これは、人間が目で見て内容を理解していく、という一般的な用途に適していて、大量のデータ収集には不向きである。

(3) 開発者用のドキュメンテーション参照。短縮URL http://bit.ly/1LM58wx（2016年1月20日取得）

(4) その際、特殊な認証（OAuth認証）が取られるが、このために予めTwitterにアプリケーション登録を行っておく必要がある。

(5) ツイートのidは各ツイートに一意に振られた長整数形データ。新しいツイートほど値が大きくなる。

(6) Search APIでは、一つのアプリケーションが実行することができる検索の回数が、15分間で180回に限定されている。一回の検索はほぼ一瞬で終了するため、念のため検索を150回繰り返したら、処理をストップするよう指定する。本文（1）～（4）の処理で、最大15000ツイート収集することが可能であるが、検索条件に合致するツイートが過去一週間で15000に満たない場合は、その時点で処理は終了してしまう。また、Search APIでアクセスできるツイートは一週間前までのものである。

(7) コーパス言語学では、このようにコーパスの規模を概略的に示す場合、延べ語数、重なり語数を挙げることが多い。

(8) それでもRTのように、語と判定されたツイート固有の表記も若干残されているが、こういったものは目で見て取り除けば良い。

(9) 大薗（2014b）参照。