Vol.2015-CH-106 No.6
2015/5/16

ロシア語による SNS からの情報の収集・分析に向けて

世利彰規[†1]

本発表ではロシア語で書かれたソーシャルメディアを用いた言説分析の経過を報告する．今回は，ツイッターを資料として用いる．まず先行する電子化されたロシア語の言語資料について紹介する．次にツイッターからデータを取得する上で使用した API やツールについて説明する．さらにロシア語を分析する上での，ロシア語の文法に関わる固有の問題について述べる．最終的に「日本」というキーワードで検索して取得したツイートを数量や共起頻度などの角度から分析し，ツイッターを使用するロシア人がもっている日本についてのイメージを取り出すことを目指す.

Toward Text Mining from SNS in Russian

SERI AKINORI[†1]

This presentation is a progress report of text mining in Russian SNS. In this presentation, we use Twitter as data. First, we introduce digitalized Russian text data. Second, we explain how to get data from Twitter using API and libraries. Third, we mention peculiar problems of Russian language on mining texts. Finally we try analyzing Russian Tweets with keywords ‘Japan’ focusing on amount of Tweets and co-occurrence information. The aim of this research is to consider way to extract public opinion and impression on Japan that Russian Twitter-User have.

1. はじめに

ロシア語のテクスト処理に関わる状況を概観した上で，ロシア語で発信された SNS のうち，Twitter を処理・分析することで，Twitter を使用しているロシア人が日本に対して抱いているイメージを探ることを試みる．

2. 研究背景

筆者は，これまで書かれたロシア文学の言語に注目してきており，前回の人文科学とコンピュータ研究会での発表はロシア文学テクストの分析に基づく発表を行なったが，現代ロシア社会の言説を読み取ることにも関心が出てきた．ここからはリアルタイムに発信される生きた言葉に焦点を当てる．

3. 先行研究

ここではロシア・日本におけるロシア語処理の流れについて概観する．ロシア語をコンピュータで処理しようとする試みは，ロシアだけではなく，ヨーロッパでも行なわれてきた．例えば，主にドイツのテュービンゲン大学のロシア語コーパス[1]やスウェーデンのウプサラ大学のコーパスはロシア語研究者によく用いられてきた．

10 年前に使用することのできたテュービンゲン大学コーパスやウプサラ大学のロシア語コーパスは，現在は使うことができなくなっている．非公開となった理由については現在調査中である。

3.1 ロシアにおけるロシア語処理

もっとも大規模なロシア語のコーパスとして，ロシアナショナルコーパス[2]が挙げられる．このコーパスは British National Corpus をモデルとしており，およそ 3’000 万語が収録されている．ありとあらゆるジャンル，文体，地域のロシア語をカバーしており，ロシア語研究やロシア語教育が想定される用途として挙げられている．最終的には 2 億語を収録することを目的としている．

現存するロシア語の言語資料としては,「ロシア語テクスト自動処理（Aot）[3]」と呼ばれるサイトが挙げられる．日本の青空文庫に相当する「マクシム・マシュコフの図書館」4内のテクスト内の文章を形態素解析して検索する．形態素解析はルールベースであり，辞書とプログラムはこのサイト[5]から入手できる．

3.2 日本におけるコンピュータを用いたロシア語研究の概観

北海道大学の浦井康男は，日本においてはじめてロシア語研究にコンピュータを応用しようと試みた．形態素解析器[6][7][8]やドストエフスキーや 18 世紀ロシア文学者の文学作品について，データベースや KWIC を作成した[9][10][11]．開発言語には主に prolog を使用している．

東京外国語大学の成定久美子は，形態素解析器や自動タグ付け器の作成[12][13]を言語処理の立場からはじめて行なった．語学教育用ツールの開発も行なっている．実装言語には Perl が使用されている．

名古屋大学の水野晶子[14]のページ．ウェブ上のデータを用いたロシア語研究[15]のほか，語学教育用のツールの作成も行なっている．使用言語は Perl である．

言語処理の立場とは別に，東京外国語大学のロシア語学科のメンバーを中心として，ロシア語コーパス研究会が発足し,活動をおこなっている.東京外国語大学のチームは，ロシア語の話しことばコーパス[16]やウェブ上の辞書[17]

†1 東京大学大学院人文社会系研究科
University of Tokyo, Graduate School of Humanities and Sociology

ⓒ2015 Information Processing Society of Japan

の開発を行なっている．

4. 分析に使用する資料について

4.1 ロシアにおける SNS

4.1.1 ロシア固有の SNS

ロシアの SNS 事情について概観する[18]．ロシアにおいてよく用いられている SNS として「オドノクラスニキ」と「フコンタクチェ」の二つが挙げられる．

2006 年から運用されている「オドノクラスニキ（Odnoklassniki）[19]」（ロシア語で「クラスメート」の意味）は 130’000’000 人が使用している[20]．（参考となる数値として，ロシアの総人口は現在 143’500’000 人であるとされる[21]）「オドノクラスニキ」（ロシア語で「クラスメート」の意味）のネットワークを用いると，かつての同窓生が一同に集うことができる．そのため，ボーイスカウトで知り合った友達や近所の友達と再会するために使われている．

2006 年末から始まった「フコンタクチェ（Vkontakte）[22]」（「接触している」の意味）は 125’000’000 人[23]が使用している．これも承認した知り合いのみについて交流をおこなう閉じた SNS であるといえる．また，facebook を意識した作りとなっており，外観も facebook に近い．

これらは閉じた SNS であり，無関係な第三者が面識のない利用者の書き込みを閲覧したり，抜き出したりすることはできない．統計からロシア人の多くがこれらのメディアを用いていることがわかる．

4.1.2 ツイッターとの比較

それに対して，ロシアではおよそ 10’000’000 人[24]がツイッターを用いている．上に挙げた 2 つのソーシャルメディア（オドノクラスニキとフコンタクチェ）と比較しても，12 分の 1 の人口しかツイッターを使用していないことになる．ここから，ツイッターを用いているロシア人には何らかのバイアスがかかっていることが予想される．つまり，ツイッターによって発信される社会的な意見は全体の一部にすぎないことになる．このことを考慮に入れて，ツイッターに関して分析を行なっていく．

また，Twitter 社は公式的にはツイッターを SNS と見なしていないが[25]，今回の発表ではツイッターも SNS の一つと見なすことにする．

4.2 ツイッターについての説明

現時点でツイッターを処理するために用意されている API には主に以下の 3 つが存在する．

4.2.1 Rest API

もっともよく用いられる API で，ツイートの入出力に用いられる．アクセスは，15 分に 180 回までと制限されている．

4.2.2 Streaming API

リアルタイムで流れるツイートを取得する．取得するツイートの言語や地域は指定することができる．

4.2.3 Search API

検索するキーワードを指定してツイートの収集を行なう．Rest API を基礎としたものである．今回はクエリを指定するこの Search API を使用してツイートを取得した．

5. ロシア語処理における問題

5.1 ロシア語のコトバの問題とその解決方法

ロシア語は屈折語に分類され，英語などと比べても多くの語形変化形をもつ．ロシア語をコンピュータで処理する場合，語形変化をどのように取り扱うかについて述べる．

正規表現を用いてロシア語の活用・曲用を処理することができる．

ロシア語の名詞・形容詞には男性・女性・中性の 3 つの性があり，主格・属格（生格）・与格・具格（造格）・前置格の 6 つの格がある

動詞の時制形はおおまかに現在形と過去形の形があり，それぞれ 6 つの形をもつ．

これらの活用・曲用の語尾は大まかにタイプ分けすることができる（あらかじめ有限の文法ルールを定義しておくことで，ほとんどの語の品詞を判別することができる）ので，正規表現による処理に向いている．

今回，「日本」を意味する名詞および「日本の」を意味する形容詞について検索を行なった．名詞は 3 通り，形容詞は 7 通りの曲用形を設定した．

5.2 ロシア語の形態素解析

上に述べたように，ロシア語の品詞は活用・曲用語尾によってほとんど規則的に特定することができる．そのため，ロシア語の形態素解析は予め定義されたルールに従って行なわれることが多い．

先行する形態素解析器には多くのものがあるが，筆者の知る限り，おおむねザリズニャークの文法辞書[26]を用いたルールベースのものがほとんどを占める．この文法辞書は，ロシア語に含まれる語彙の形態変化について網羅的に記述したものである．浦井や成定らの研究でもこの辞書のルールを参考にして形態素解析システムを構築している．

5.2.1 NLTK.snowball のステマー

python の言語処理用のライブラリ NLTK の中にもロシア語を処理するためのステマーが存在する．

NLTK のステマーもまた，Porter のアルゴリズム[27]を用いたルールベースのものである．

5.2.2 OpenCorpora の形態素解析

ロシアにおいて，独自に品詞情報をタグ付けしたコーパスを作成しようとするプロジェクトが進行している．それが OpenCorpora である．OpenCorpora[28] は，2009 年に始まった人手でタグ付きコーパスを構築しようというプロジェクトである．現段階では辞書のみが使用できる．システム概要をみると，ルールベースのようである．アノテーシ

ⓒ2015 Information Processing Society of Japan

ョン用のプログラムは公開されており[29], perl や python, php など様々な言語で使うことができる.

6. 「日本」というキーワードを含むツイートの収集・分析の試み

ここからは，実際に「日本」を意味する語を含んだツイートを Search API を用いて収集し，どのようなことが分かるか考えてみる.

今回は，1）ツイートそのものとなるテクスト，2）ツイートが発信された時間，3）ツイートが発信された場所の 3 つの情報を分析に用いる．データ収集に使用する Search API では，1）テクスト情報は text タグ，2）時間情報は created_at タグ，場所情報は user.location タグによって取得することができる.

6.1 時差の問題

ツイッターがいつ発信されたかを表す created_at タグからは現地時間の時間情報を取得することができる．「日本」に関するツイートを対象とし，ツイート数の増減を調べていく．ツイート数の増加は日本で起きた何らかの出来事に起因する可能性がある．この場合，ツイートが発信されたロシア時間と日本時間との差についても考慮する必要がある.

日本標準時（JST）は協定世界時より 9 時間進んでいる．ロシアは 11 の標準時をもち, モスクワは協定世界時よりも 4 時間進んでいる（UTC+4）.

ツイッターの created_at タグに記載される時間は UTC time（協定世界時）に統一されている[30]．そのため，協定世界時を基準にツイートが発信されたロシア時間とツイートに取り上げられた出来事の起こった日本時間の差を考えるとよい.

6.2 収集に使用した Twitter の API

今回の発表では,「日本」というキーワードを含んだロシア語のツイートを収集・分析し，そこから分かることを考えてみる．今回使用する Search API は LL 言語 Python において TwitterSearch という名のラッパーを通して使うことができる.

今回の調査で抜き出したツイートに関する情報は，ツイートそのものとなるテクストと時間，場所の 3 つである.

今回の分析では，API は Search API を用いている．クエリを指定して 1 週間分のツイートを取得することができる．使用した理由は，ある程度の量（1 週間の間につぶやかれたツイート）必要なツイートを的確に（ロシア語でつぶやかれたツイート）取得することができるためである．Streaming API でも試行してみたが，現在リアルタイムでつぶやかれているツイートしか取得できなかった．また，言語をロシア語で指定しても，別の言語のツイートが取得されることもあったため，今回は採用しなかった.

次に，Search API の特徴について説明する．まず，指定するクエリの大文字，小文字は区別されない，という特徴がある．実際，クエリを「JAPAN」としても「japan」として同じ期間中につぶやかれたツイート数を数えてみたところ, 同じ結果が返ってきていた. つまり,「JAPAN」も「Japan」も「japan」も大文字小文字を区別せずに，一つのクエリで取得することができることになる.

6.3 ツイート数からのデータ：比率・割合

まず所定の期間内にロシア語で発信された「日本」というキーワードを含むツイートの数を数え上げ，ツイート数の増減から何が見えてくるのかを調べてみる．今回の調査では,「日本」を意味する名詞,「日本の」を意味する形容詞をクエリとしてツイートを収集した．分析に用いるツイートには内容的には同じリツイートも含んでいる.

ロシア語で発信された「日本」を含むツイートの変動は現実の世界の出来事と関連があるのだろうか．分かりやすい例として, 2014 年 12 月 8 日から 15 日までの一週間にロシア語で発信された「日本」を含む 77257 ツイートの時間的な増減を見てみよう.

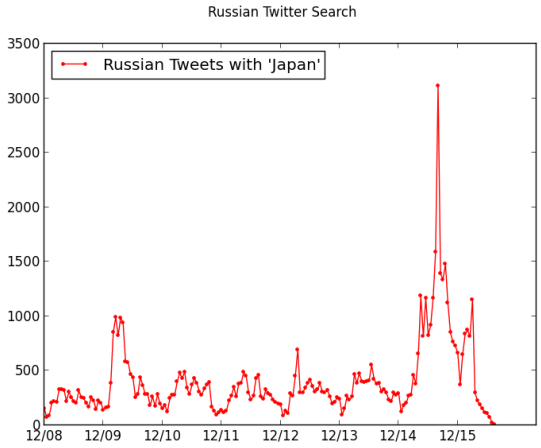

図 1 年末における「日本」を含んだロシア語ツイート数の変遷（2014/12/8-2014/12/15）

上のグラフから，14 日から 15 日にかけてツイート数が増えていることが分かる．2014 年 12 月 15 日は，日本において衆議院選挙が行なわれた日であり，このことがニュースなどのメディアに流れたため，ツイート数の増加が起こったと予測される．日本における政治・経済の何らかの動きがツイート数の増減の一因となっていると考えられる.

「日本」という語だけを含むツイート数だけでは分かることは限られる．他のキーワードを含むツイート数と比較して明らかになることを考えてみる．例えば，ロシアにおいて「アメリカ」というキーワードを含むツイート数に比べて「日本」を含むツイート数はどれほどの差があるのだろうか. 以下では 2015 年 3 月 20 日から 26 日にかけてロシ

ⓒ2015 Information Processing Society of Japan

ア語で発信された「日本」および「アメリカ」を含むツイートの変動を見てみる．日本に関するデータ数は上記と同じ方法で収集した 97234 ツイートである．アメリカに関するツイートは，「アメリカ」「USA」を意味する名詞および「アメリカの」を意味する形容詞を検索語に指定して収集し，92353 ツイートを得ることができた．これらを描画すると次のようになる．参考として，リツイートを点線で示した．

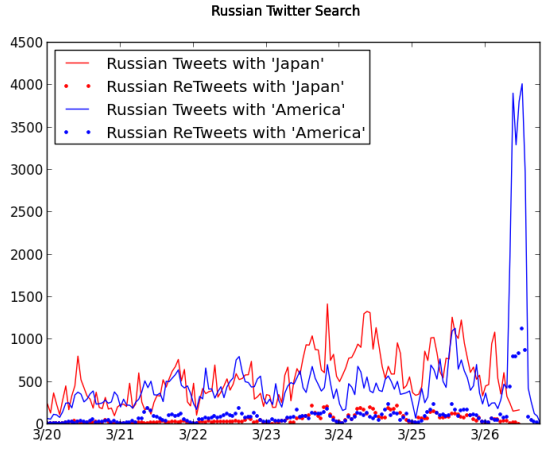

図 2 「日本」および「アメリカ」を含むロシア語ツイートの変遷（2015/03/20-/2015/3/26）

上に挙げたグラフから分かるように，平常時には日本を意味する語を含むツイート数の変動もアメリカを意味する語を含むツイート数の変動も重なっている．期間内のほとんどの時間において，ツイート数の推移はふつう横ばいであるが，突然なんらかの理由でツイート数の急激な上昇が起こっている．

6.4 共起情報から読み取れること

ロシア語で発信された「日本」を意味する語を含むツイートに関する今回の調査・分析は，最終的にロシア人が日本についてどのような意見やイメージを持っているのか，その一端を明らかにすることを目的としている．ツイート数だけではなく，ツイートに書かれていることも対象にする必要がある．そこで，ツイート内のテクスト情報の語の共起情報を調査した．

今回の調査・分析では，単語 N グラムを共起情報とみなした．2015 年 3 月 20 日から 26 日にかけて「日本」を意味する語が含まれた 97234 ツイート内を調査した．その結果得られた単語 3 グラム，4 グラム，5 グラムの降順リストの上位 5 位までを以下に挙げる．

япония осуществила запуск	631
запустила новый спутник	550
япония запустила новый	550
запуск спутника шпиона	533
осуществила запуск спутника	531

表 1 単語 3 グラム

япония запустила новый спутник	550
япония осуществила запуск спутника	531
осуществила запуск спутника шпиона	530
для любителей японской кухни	395
любителей японской кухни рецепт	395

表 2 単語 4 グラム

япония осуществила запуск спутника шпиона	530
для любителей японской кухни рецепт	395
япония запустила новый спутник шпион	221
осуществила запуск спутника шпиона когаку	199
запуск японского спутника шпиона когаку	159

表 3 単語 5 グラム

集計されたこれらの N グラム単語列でまず読み取れるのは，2015 年 3 月 26 日に日本が情報衛星を打ち上げたことについて取り上げている．このことについてツイートは否定的であり，「日本がスパイ衛生を打ち上げた」旨の内容であることが分かる．

次に読み取れることが，日本料理への関心である．上の N グラム単語列には「日本料理愛好家へのレシピ」という内容のものがある．ロシアでは寿司や焼き鳥など日本料理への関心が強まっており，そのような関心がツイートにも現れた可能性がある．

普通にツイートをしている場合，ここまで多くの単語列が重複することはありえない．そのため，これらの N グラム単語列はリツイートによるものも多いことが推測される．

今回ここで述べたことは，あくまでも印象的・直感的なものである．これらの手法からより有効なことが分かるような方法や仕組みを考える必要がある．

7. まとめとこれからの課題

今回の発表では，ロシアにおける電子化されたテキストデータについて概観した．そして，ロシア語によって発信された日本を題材としているツイッターを収集・分析し，ツイッターを使用しているロシア人が日本に対してどのような意見や印象を抱いているのか，探るという試みをおこなった．

7.1 これまでの分析から分かることのまとめ

今回おこなったことは，1)「日本」というキーワードを含むロシア語ツイート数の推移，2)「日本」というキーワードを含んだロシア語ツイート内における共起情報の 2 つの調査である．ここから，「日本」を含むロシア語ツイートの増減は日本における出来事をリアルタイムに反映していることが明らかになった．

ⓒ2015 Information Processing Society of Japan

7.2 今回の調査・分析の問題点

第 6 節では,「日本」というキーワードを含むロシア語ツイートを収集・分析することによって，ロシア人が日本に対して抱いているイメージの一部を明らかにする試みをおこなった．この試みには，当然のことながら，いくつかの問題点が存在する．

まず, ツイート数やツイート内の共起情報の結果を見て，ツイート数の増減の原因を推し量った．これは，原因を後付けするにすぎず，何か新しい情報を明らかにするまでには至っていない．

また，何度か述べているように，ツイッターを用いているのはロシア人全体の一部であり，発信される意見や情報にも偏りが存在することが予想される。そのため，ツイッター以外にデータを入手する方法を考える必要がある．

7.3 今後の課題・方針

上述したように, 今回おこなった分析の試みは, 印象的・直感的なものである．より有効・有益な結果を得るためには統計的な性質に着目して分析方法を考える必要がある．

今回は，データ収集に試行錯誤したため，分析して何かが言えるほどのデータを集めることができなかった．一週間以上のスパンでツイートを収集し，より多くのデータを用いた本格的な分析を行なうことも今後の課題となる．

今回の調査・分析ではツイッターを SNS と見なしてデータの収集・分析をおこなったが，ツイッターは社会全体の意見のほんの一部しか反映しておらず，分析によって得られた結果から何かを主張することは難しい．したがって，違うメディアを用いて分析を行なうことも今後は検討したい．とりわけ，ツイッター以外のメディア（例えばロシア語で書かれた YouTube のコメントやブログなど）からデータを収集，分析することが考えられる．

また，単語についての共起情報だけでなく，不読符号列[31]と呼ばれる記号列に着目して分析を進めるという方針も考えられる．

参考文献

1) 小林潔「ドイツ・テュービンゲン大学のロシア語コーパス,『ロシア語研究』, 第 16 号, pp.64-85, 2003
2) http://ruscorpora.ru/[2015/4/1 閲覧]
3) http://www.aot.ru/search1.html[2015/4/1 閲覧]
4) http://www.lib.ru/[2015/4/1 閲覧]
5) http://www.aot.ru/download.php[2015/4/1 閲覧]
6) 浦井康男「ロシア語の形態生成と Prolog 系」『Japanese Association of Russian Scholars(27)』pp.114-115, 1995
7) 浦井康男「ロシア語の形態生成と Prolog 系-3　入力と出力」『Memoirs of the Faculty of Education, Fukui University Series 1 Humanities (50)』pp.1-26, 1995
8) 岡本哲也・浦井康男・山本佳代子「ロシア語の形態生成と Prolog 系(II) : 動詞」『Bulletin of the University of Electro-Communications 6(2)』pp.157-169, 1993
9) 浦井康男「近代ロシア文章語(散文)形成期の諸作家における造語体系について　- ラヂシチェフ，カラムジン，プーシキンの場合 -」 『北海道大学文学研究科紀要. The annual report on cultural science 121』pp.21-87, 2007
10) 浦井康男「近代ロシア文章語の語彙発達について : 現代語頻度辞書の観点から」『Japanese Association of Russian Scholars(35)』p.119, 2003
11) 浦井康男「カラムジンにおける造語法について : ラジシチェフと比較して」『Japanese Association of Russian Scholars(33)』p.135, 2001
12) 成定久美子・佐野洋「ロシア語タグ付きコーパス検索機能を備えた語学教材作成支援ツール」『IPSJ SIG Notes 2005(104)』pp.49-56, 2005
13) 成定久美子・佐野洋「ロシア語タグ付きコーパスの作成」『情報科学技術レターズ 3』pp.339-342, 2004
14) http://www.remus.dti.ne.jp/~a012548m/[2015/4/1 閲覧]
15) 水野晶子・藤村逸子「身体部位名詞を伴う再帰構文における格の問題―ロシア語とフランス語の比較」『国際開発研究フォーラム(26)』（名古屋大学大学院国際開発研究科）, pp. 65-87, 2004
16) http://cblle.tufs.ac.jp/multilingual_corpus/ru/[2015/4/1 閲覧]
17) http://cblle.tufs.ac.jp/dic/ru/[2015/4/1 閲覧]
18) http://jp.rbth.com/articles/2012/10/22/39563.html[2015/4/1 閲覧]
19) http://ok.ru[2015/4/1 閲覧]
20) http://www.slideshare.net/MikePonomarenko/eng-smm-rusbrief[2015/4/1 閲覧]
21) http://data.worldbank.org/indicator/SP.POP.TOTL[2015/4/1 閲覧]
22) https://vk.com[2015/4/1 閲覧]
23) http://www.slideshare.net/MikePonomarenko/eng-smm-rusbrief[2015/4/1 閲覧]
24) http://www.slideshare.net/MikePonomarenko/eng-smm-rusbrief[2015/4/1 閲覧]
25) http://www.cnet.com/news/twitters-not-a-social-network/[2015/4/1 閲覧]
26) *Зализняк А.А.*, Грамматический словарь русского языка. Словоизменение, M. 1980
27) Porter, M. *An algorithm for suffix stripping*. “Program 14.3”, pp.130-137, 1980
28) http://opencorpora.org[2015/4/1 閲覧]
29) https://github.com/OpenCorpora/opencorpora[2015/4/1 閲覧]
30) https://dev.twitter.com/overview/api/tweets[2015/4/1 閲覧]
31) 中嶋邦裕・渡辺靖彦・松本浩輝・西村涼・岡田至弘「日本語と英語のツイート末尾で多用される不読符号列の比較」『研究報告自然言語処理. 2013-NL-214. (15)』pp.1-7, 2013

ⓒ2015 Information Processing Society of Japan