様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　　８　日現在

機関番号：　３２６６５
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３７０５５１

研究課題名（和文）WWW検索による日本語研究の総合的展開

研究課題名（英文）Integrated Development of Japanese linguistics by WWW searching

研究代表者
荻野　綱男（OGINO, Tsunao）

日本大学・文理学部・教授

研究者番号：００１１１４４３

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：WWW検索を通じて日本語の使われ方を調査し、日本語使用のいくつかの側面を明らかにした。第1に、平成26年度は方言語形の分布調査を行い、どのような方言語形がどの地域に分布するかを明らかにした。第2に、平成27年度はことばのゆれの調査を行い、特に、漢字・ひらがな・カタカナのいずれでも表記できる語について、それぞれの表記の持つ特徴を明らかにした。第3に、平成28年度は新語の普及の問題を取り上げ、Twitter での発信日付を手がかりにして、新語がどのように普及していくのか、そのプロセスを明らかにした。

研究成果の概要（英文）：　I investigated the usage of Japanese language by WWW search engines and clarified some aspects of Japanese usage.　First, in 2014, I investigated distribution patterns of dialectal wordforms and I could grasped clearly what kind of words had spread to what kind of areas.　Second, in 2015, I investigated fluctuations of notations of Japanese words and I indicated some characteristics of kanji, hiragana, and katakana notations in the case of the words which can be written in three characters.　Third, in 2016, I picked up the problem of spreading patterns of neologisms using Twitter's transmission dates and I revealed the speed and tendency of newly created wordforms.

研究分野：　日本語学

キーワード：　WWW　方言語形　ゆれ　新語

１．研究開始当初の背景

WWW の検索エンジンが手軽に使えるようになり、それを利用して WWW の検索ができるようになってきた。

そこで､WWW 上の大量の日本語文書を日本語の使用例とみなして、日本語のデータとして使おうという試みが始まった。

ただし、検索エンジンは、しばしば改訂され、あるときにできた検索方法が時間をあけて再度行おうとするとうまくいかない例がしばしばある。また、検索件数が大きく変動することもあり、検索エンジンがどの程度信頼できるのか、疑問に思える面もある。

そこで､WWW を日本語のデータベースとして利用して、いくつかの課題を追究しながら、検索エンジンの特徴を探り、有効な使い方を考えることが求められた。

２．研究の目的

WWW および検索エンジンをどのように使えば日本語研究として有効なのかを明らかにすることが大きな目的である。

この目的を達成するためには、実際に具体的なテーマを設定して、それを明らかにするべく、検索エンジンを使ってみるということが重要になってくる。

ほぼ、毎年一つのテーマを決め、それを調べることで検索エンジンの特徴を明らかにし、また、同時に、日本語学的に興味深い現象を探ることを目的とした。

具体的には、以下の三つのテーマを設定した。

平成 26 年度は方言語形の分布調査である。WWW の資料を用いて各種方言語形がどの地域に分布しているかが明らかになれば、現地調査などのお金と手間と時間がかかる調査手法を用いなくとも、とりあえずこのあたりに分布するというような概略がわかるだけでも、意味は大きい。

平成 27 年度はことばのゆれを調査した。漢字・ひらがな・カタカナのいずれでも表記できる語をいくつか選び、それぞれの用例多数を検索し、それぞれの文字で書かれた場合にどのような特徴があるのかを探ることにした。

平成 28 年度は新語の普及の問題にアプローチした。Twitter は、送信の日付が明確にわかるという特徴を持つ。そこで､Twitter に蓄積されている膨大なデータを検索し、いつの時点でそれぞれの新語が使われているのかを明らかにしようと考えた。

これらの三つの研究テーマを通じて、WWW をどのように利用するのかを探ることができた。

研究目的は、あくまで日本語研究が中心であるが、同時に、WWW の使い方を開発するような面も無視できない。学生などにとっても、手軽に第１次資料にアクセスし自分で分析するようなことができるようになれば、自力でさまざまなテーマが追求できるようになる可能性が出てくる。このようにして、コーパス言語学が発展することを目指すことも（やや長期的な視点としては）目的の一つといえよう。

３．研究の方法

WWW の検索は、検索エンジンを使えばきわめて簡単にできると考えられている。しかし、実はそうではない。適当な単語を検索エンジンの検索窓に入れて、出てきた検索件数がそのまま WWW の用例数だなどと単純に考えてはまずい。

検索エンジンの検索結果は、とりあえず原文を機械的に形態素解析して、それぞれの単語を抽出し、インデックスを作成しておき、それを検索時に高速で検索して結果を利用者に返すようになっている。したがって、検索結果には間違いが含まれることになる。どれくらい間違いが含まれるかは何ともいえない。このことは将来にわたって問題として残るだろう。中には意図的に間違った言葉遣いをしていると思われる例さえある。

というわけで、検索は一瞬で終わるとしても、実際は、そこから用例の整理という長い作業が始まるのである。

WWW を検索して得られた用例を一つずつ確認し、本当に当該の用例として扱っていいのか、検討する必要がある。ここは、機械的に行うことがむずかしい。どうしても人手で（人の目で）行わなければならない。

そこで、３年間にわたって、研究補助者２名を雇用することにした。ある程度作業に慣れてもらわないと、効率的な検索・整理はできない。少数の人にある程度長期にわたって作業を行ってもらうことで信頼できる結果が取り出せるものである。そのようなことから、研究室内でパソコン２台を設置し、それぞれの人に１台ずつ使ってもらい、長期的に作業してもらう形にした。

４．研究成果

(1) 方言語形の分布調査

WWW を用いて方言語形の分布地域を探ってみた。

適当な方言語形を取り上げ、その分布地域を探るという課題である。

どんな結果になったか、例として「かっぱぐ」（意味は「はがす」ということ。「掻き剥ぐ」の変化したもの）を取り上げよう。

この語形を検索すると、用例によっては、使われる地域がわかることがある。

http://yaplog.jp/konpei_3/archive/378 には、次のような用例がある。

栃木の方言・・7・「かっぱぐ/かっつぁぐ」編

June 25 [Sat], 2011, 21:51

◆かっぱぐ・・・・剥がすというような意味ですね。(最後の「ぐ」は鼻濁音で)

例文１・・あんちゃんがまだ寝でっから、学校遅れるよーって布団をかっぱいで来な!
訳・・兄ちゃんがまだ寝てるから、学校に遅れないように布団を剥がしておいで！

例文２・・隅田川で死体が上がったの？うん、薙（むしろ）掛げてあっから、かっぱいで顔を見でみな・・。
訳・・・隅田川で死体が上がったの？うん、薙を掛けてあるから、剥がして顔を確認してみな・・。

この例では、栃木県の方言の例として「かっぱぐ」をあげている。これは、使用例（書き手自身が使っている例）ではなく、議論や説明のために「かっぱぐ」を取り上げているもので、こういうのを「参照例」と呼ぶことにする。辞書記述などは典型的な参照例である。
このように１例ずつ検討して、参照例と使用例をわけ、それぞれでどの地域の方言としているかを見ていくと、以下のような結果になる。

参照例
県別
福島県 2　群馬県 1　栃木県 2　茨城県 4　埼玉県 3
市町村別
埼玉県深谷市 1　埼玉県羽生市 1　茨城県大子町 1　茨城県古河市 1
使用例
県別
群馬県 1　東京都 2　大阪府 1
市町村別
栃木県三上町 1 →河内郡　東京都世田谷区用賀 1　東京都練馬区大泉学園 1　東京都墨田区 1　千葉県柏市 1　富山県高岡市 1

このようにして使用例と参照例を分けると、使用例では、その書き手がどこに住んでいるのか、どこの出身かがわかる場合は少なく、一方、参照例ではどの地域の方言かが書かれている場合が多い。上記のようなわずかの例から帰納するのは問題が多いが、それでも北関東あたりに分布するのではないかということがわかる。参照例が数十例～２００例くらいあると、かなり確実な分布地域がわかるものである。
このような調査を２５０語ほどの方言語形について試みた。全般に、分布地域がわかる場合が多く、WWW を調べるだけで、だいたいの分布地域を把握できることが明らかになった。

(2) ことばのゆれの調査
漢字・ひらがな・カタカナのいずれでも表記できる語を選び、それぞれの用例多数を検索し、それぞれの文字で書かれた場合にどのような特徴があるのかを探ることにした。
たとえば、朝顔・あさがお・アサガオを取り上げよう。
違いがもっともわかりやすいのは、複合語を調べた場合である。
「朝顔」の場合は、いずれの複合語も成立が古いものである。「朝顔七輪」は七輪の一種であるが、形状がアサガオの花に似ていることから命名されたものであろう。七輪の歴史と同じくらい古いようだ。「朝顔鉢」も鉢の一種でアサガオの形に似ているものである。各地の焼き物として作られているから、相当古い。「肥後朝顔」は、アサガオの品種名であるが、江戸後期に出現した州浜系統が九州に渡り、熊本で栽培されていたものに由来すると考えられる。「桔梗朝顔」「変化朝顔」なども同様である。
「アサガオ」の場合は、品種名の中でも比較的新しいものはカタカナで表記されることが多い。「ホシアサガオ」、「マメアサガオ」、「ツクバネアサガオ」、「マルバアサガオ」、「タイワンアサガオ」、「チョウセンアサガオ」、「宿根アサガオ」などがそれに該当する。「宇宙アサガオ」は、2012 年に宇宙で被曝させたアサガオである。
「あさがお」の場合は、「あさがお栽培」を除いてほぼすべて組織・会社・団体の名前が並んでいる。「あさがお連」は阿波踊りの組である。「あさがお薬局」「あさがお歯科」「あさがお整骨院」「あさがお法律事務所」「あさがおクリニック」など、いずれも専門職と結びついており、名前を柔らかいものにすることで親しみやすい効果がある。「あさがお公園」「あさがおクラブ」「あさがお組」「あさがお保育園」など、子供関連の組織名などにひらがなが多用される傾向にある。
このように、３種類の表記を持つ語でも、それぞれの表記は特徴的な使われ方がされるものである。
このような表記のゆれが見られる語を１０組ほど調べた。

(3) 新語の普及
Twitter は、送信の日付が明確にわかるという特徴を持つ。そこで、Twitter に蓄積されている膨大なデータを検索し、いつの時点でそれぞれの新語が使われているのかを明らかにしようと考えた。
https://twitter.com/ から全部のツイートに対して適当な語を検索することができる。
ただし、検索結果が新しいものから順に並んで出力されるだけで、「検索件数」は表示されない。そこで、検索結果の中から先頭５０件を数えて、５０件目の日時を確認し、検索件数を推定するという方法をとることにした。
たとえば、「エモい」を調べた場合を例にしよう。
まず、Twitter で、日付指定して、新語調

査をする。日付指定の仕方は、「エモい since:2016-01-01 until:2016-12-31 」で検索すると、2016-12-31 から 2016-01-01 の順で表示されるので、2016-12-31 から 50 件数えて、計算して、1 年分の使用件数を推定する。

例：2016-12-31 から 2016-12-23 で 50 件の場合、8 日で 50 件なので、365÷8×50＝2,281　2,281 件/年と計算する。

例：2016-12-31 だけで 50 件の場合、時間を確認して、最初の Twitter の時間 15:59 から 50 件めの Twitter の時間 13:53 を引いて、2 時間 06 分。これを分に直して、24 時間＝1440 分　2 時間 06 分＝126 分

1440÷126×50×365＝208,571 件/年

こういうことで、毎年の「エモい」の用例数は以下のように計算できる。

2016 年　208,571 件/年
2015 年　86,164 件/年
2014 年　75,301 件/年
2013 年　65,536 件/年
2012 年　58,450 件/年
2011 年　37,489 件/年
2010 年　89,693 件/年
2009 年　2,028 件/年
2008 年　261 件/年
2007 年　21 件/年
2006 年　0 件
（以下、これ以前はすべて０件）

この結果から、「エモい」は 2010 年に第１次流行の時期があって、急激に使われるようになり、2016 年にはさらに数倍に伸びるという結果になっている。

ただし、毎年のツイートの総数が増えている可能性があるので、上記の結果がそのまま新語の普及過程を表しているわけではない。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）

① 荻野綱男(2016.6.25)「表記のゆれと単語の意味・用法の違い――朝顔・アサガオ・あさがおを含む複合語を例に――」語文第 155 輯，pp.左 1-8，査読有

〔学会発表〕（計　　件）

〔図書〕（計　　件）

〔産業財産権〕

○出願状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
荻野綱男（OGINO, Tsunao）
日本大学・文理学部・教授
研究者番号：00111443