言語学研究の支援を目的とした大規模時系列ウェブアーカイブからの新造語のマイニング

鍛治伸裕† 宇野良子†† 喜連川優†

† 東京大学生産技術研究所
†† 東京大学大学院総合文化研究科
E-mail: †{kaji,kitsure}@tkl.iis.u-tokyo.ac.jp, ††ryoko@sacral.c.u-tokyo.ac.jp

あらまし 言語学においては，これまでに新造語は様々な観点から注目されてきた．しかしながら，新造語のデータを網羅的に取得することは困難であり，このことは，従来の言語学において円滑な研究活動を阻害する大きな要因となっていた．この問題を解消するために，我々は大規模なアーカイブの活用を試みている．ウェブ上には新造語を含んだテキストが豊富に存在するため，ウェブアーカイブを言語学者が自由に研究利用することができるようになれば，新造語の分析を行ううえで大きな助けになることが期待される．本論文では，大規模時系列ウェブアーカイブを用いた新造語の言語学研究を支援するための我々の取り組みについて報告する．また，ウェブアーカイブを活用することによって，言語学的に興味深い知見をいくつか得ることができたので，その一部についても述べる．

キーワード 新造語，言語学，自然言語処理，時系列ウェブアーカイブ

Mining Neologisms from a Large Diachronic Web Archive for Supporting Linguistic Research

Nobuhiro KAJI†, Ryoko UNO††, and Masaru KITSUREGAWA†

† Institute of Industrial Science, University of Tokyo
†† Graduate School of Arts and Sciences, University of Tokyo
E-mail: †{kaji,kitsure}@tkl.iis.u-tokyo.ac.jp, ††ryoko@sacral.c.u-tokyo.ac.jp

Abstract Neologisms have been an object of study in linguistics. However, because of the lack of linguistic resources, it has been difficult for linguists to investigate neologisms. To resolve this situation, we are investigating to make use of a Web archive (e.g. the Internet Archive) for such linguistic research. Because we can find plenty of neologisms on the Web, a Web archive is an unquestionably valuable resource for the linguistics community. This paper reports our attempt to support linguistic analysis of neologisms based on a large Web archive. We also show our initial results on linguistic analysis of neologisms using a time-series Web archive.

Key words neologism, linguistics, natural language processing, diachronic Web archive

1. はじめに

我々人間の話す言葉は変化するものであり，そこでは新しい単語 (新造語) が日々造り出される．例えば「サボる」という動詞は，今でこそ標準的な日本語として認識されているが，元来は怠業を意味するフランス語の sabotage に由来する造語である．同様の新造語として，グーグルを使ってウェブ検索を行う，という意味の動詞「ググる」などを挙げることができる．

言語学では，新造語の分析に関する研究が数多く存在する．例えば日本語の新動詞に関する研究として，Tsujimura と Davis は日本語新動詞の形態的特徴の分析を行っている [1]．特に，新動詞の語幹形成に関する規則はこれまでにも多くの研究者によって分析が行われてきている [2], [3]．一方，新造語は社会言語学の立場からも盛んに研究が行われてきている [4]．Yonekawa [5] は新造語の使用を，規範からの逸脱として位置付けて分析を行っている [4], [6].

しかしながらこれまで言語学では，新造語を分析するための有効な方法論が確立されておらず，このことが円滑な研究活動を阻害する大きな要因となっている．新造語を分析する方法としては，まず考えられるのは内省である．すなわち，母語話者が自分の言語直感のみにもとづいて言語理論を構築するという方法であり，いわば思考実験である．このやり方では，当然の

ことながら分析者が知らない新造語は内省にあがらないため，少数の恣意的な新造語を分析対象とせざるを得ないことが問題となる．世の中に十分浸透していない新造語や，閉じたコミュニティでしか使われない新造語を網羅的に調べあげることは極めて困難である．内省によらない分析方法としては，出版物などのテキストを実際に調査することによって，新造語の分析を行うことが考えられる．しかし，我々の知る限り，新造語の分析に利用可能な言語資源は十分に整備されていない．

一方，近年のウェブの急速な発達とデータベース技術の進展に伴い，大規模なウェブアーカイブが利用可能になりつつある．(注1) ウェブ上には新造語を含んだテキストが豊富に存在するため，ウェブアーカイブを言語学者が自由に研究利用することができるようになれば，新造語の分析を行う上で大きな助けとなることが期待できる．特に時系列ウェブアーカイブが利用できる場合には，新造語の通時的な変化過程が調査可能になることが大きな利点である．つまり，新造語が発生して伝搬していく過程，そして場合によっては，衰退していく様子も観察することができるようになると考えられる．

しかしながら，大規模なウェブアーカイブを新造語データベースとして活用するためには，そこに含まれる新造語を検出して索引付けを行う必要があるが，これが技術的課題となる．単純には，形態素解析器を適用することになるが，そのような方法には 2 つの問題が存在する．まず 1 つ目の問題として，大部分の新造語は解析辞書に登録されていないため，既存の解析器では新造語を含むテキストを正しく解析できない．人手で辞書を拡張することも考えられるが，作業者にとって未知の新造語には対応できない．さらに，辞書の整備にかかる膨大な作業コストも無視することができない．そして 2 つ目の問題として，仮に形態素解析器がウェブテキストを完全に解析できたとしても，どの単語が新造語であって，どれが新造語でないのかを判定する方法は自明ではない．

我々はこの問題の解決を図るために，ウェブアーカイブからの新造語のマイニングに取り組んでいる．より厳密には，新造語マイニングを，テキストから未知語 (辞書未登録語) を自動獲得する問題と，獲得した未知語の中から新造語を発見する問題の 2 つに切り分けて，個別の技術開発を進めている．前者には識別モデルの適用，後者には Dynamic Time Warping (DTW) 距離 [7] にもとづく時系列分析の適用を検討している．

これまでに開発した技術と大規模時系列ウェブアーカイブ [8] を用いて，新造語の認知言語学的分析についても平行して研究を進めている [9]．大規模な実データを活用することによって，言語学的に興味深い知見をいくつか得ることに成功している．

2. 未知語獲得

未知語の獲得には識別モデルにもとづく手法を用いる [10]．手法の詳細は文献を参照されたい．

(注1)：http://www.archive.org/index.php
http://dbpubs.stanford.edu:8091/~testbed/doc2/WebBase
http://www.webinfomall.cn

表 1　未知語に割り当てる品詞

普通名詞，サ変名詞，母音動詞，子音動詞カ行
子音動詞サ行，子音動詞タ行，子音動詞ハ行
子音動詞マ行，子音動詞ラ行，子音動詞ワ行
子音動詞ザ変，イ形容詞，ナ形容詞

2.1 問題設定

まずは未知語獲得という問題を整理にする．本論文では，コーパスと辞書が与えられたとき，そのコーパスから未知語を抽出して，それに適切な品詞を割り当てる問題を未知語獲得と呼ぶ．ただし，動詞などの活用語は，活用形によって表層形が変化するため語幹を考える．例えば，動詞「ググる」の場合は語幹「ググ」を抽出する．未知語に割り当てる品詞は，先行研究 [11]〜[13] を参考にして，表 1 の 13 種類とした．すなわち，ここに記載されていない品詞に分類される未知語は獲得対象としない．品詞の定義は，基本的には JUMAN 辞書の品詞細分類に従っているが，活用語については活用型を品詞とみなしている．

2.2 候補抽出

大規模テキストから未知語を獲得するためには，その候補を抽出することが必要となる．単純には，テキスト中の全ての部分文字列が未知語候補となりうるが，そのような方法は非効率である．

そこで，文脈情報を用いて効率的に候補の絞り込みを行う．以下の具体例を考える．

(1) a. ノートを コピったのだ
　　b. の文字を コピれない.
　　c. 図書館で コピりましょ

「ノートを」「の文字を」「ったのだ」「れない.」などの文字列は，子音動詞ラ行 (の語幹) の前後に出現しやすい．このような文字列のことを，その品詞の弁別的先行文字列，弁別的後続文字列と呼ぶこととする．もし表 1 中の各品詞に対して，その品詞の弁別的先行文字列と後続文字列を大量に取得することができれば，それらに囲まれて出現している文字列 (上の例における「コピ」) を，その品詞の未知語候補として抽出することが出来る．

ある品詞 t の弁別的先行文字列は，既存の辞書を用いてコーパスから自動抽出する．まず文字列 p に対して次のスコアを定義する．

$$coverage(p) = |\{w \in \mathcal{W}_t | 0 < f(pw)\}|$$

式中の $f(pw)$ は文字列 pw の頻度であり，$\mathcal{W}_t$ は品詞 t が割り当てられている辞書登録語 (活用語の場合は語幹) の集合である．$coverage(p)$ は，p が何種類の辞書登録語の直前に出現したかを表しており，この値が大きいほど p は品詞 t に特徴的と言える．そのため $coverage(p)$ が閾値を越える p を品詞 t の弁別的先行文字列とする．実験では，閾値は $\sqrt{|\mathcal{W}_t|}$ に設定し，辞書には JUMAN 辞書(注2)を用いた．なお，弁別的後続文字列も

(注2)：http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html

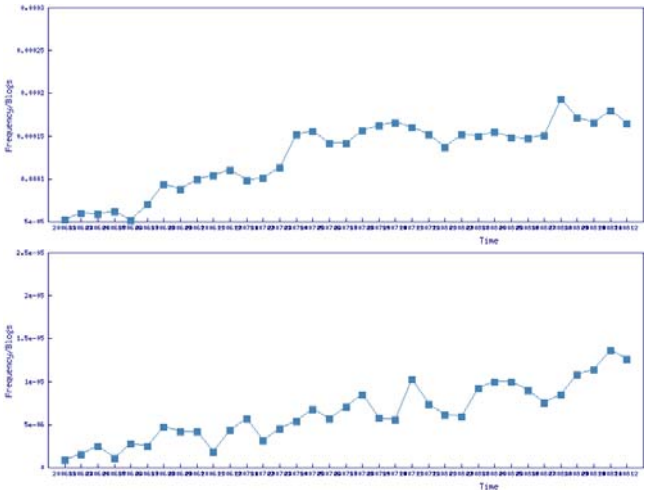

図 1　新造語の頻度の通時的変遷 (上: 写メる，下: モフる)

表 2　獲得した未知語

品詞	具体例
普通名詞	戦メリ，ようつべ，ダメンズ，駄ゲー，ぬこ，ウサ公
母音動詞	ぶちギレる，ウケる，イぢける，チャラける
子音動詞ラ行	ググる，写メる，キョドる，タゲる，ポシャる
イ形容詞	うざい，チャラい，あっつい，もっさい

同様の方法で取得する．

2.3　識別モデルの適用

抽出された候補が未知語であるか否かを識別モデルを用いて判定する．学習器は品詞ごとに別々に構築するが，そのときの計算効率を考慮して，高速なオンライン最大マージン学習器の 1 つである Passive Aggressive Algorithms [14] を用いる．

素性は，その候補をコーパスから抽出するときに使われた弁別的文字列を 2 値素性として使う．しかし，弁別的文字列の長さが大きい場合，素性ベクトルが疎になって事例間で素性が共有されにくくなるので，弁別的先行文字列の全ての接尾文字列も素性とする．例えば，弁別的先行文字列「ノートを」からは接尾文字列「ートを」「トを」「を」も素性として利用する．同様に，弁別的後続文字列の全ての接頭文字列も素性とする．

2.4　獲得した未知語の例

これまでに，327GB の日本語テキスト (UTF8 エンコーディング) に対して提案手法を適用して，46,500 の未知語の獲得をしている．既存の辞書をベンチマークにして適合率と再現率を調査したところ，それぞれ 86.4%，82.8%であり，良好な精度であることが確認された．表 2 に獲得した未知語の例を示す．「ようつべ」「ぬこ」「ググる」「写メる」など，多くの新造語が獲得されていることが分かる．

3.　時系列分析

獲得された未知語の中には，新造語ではないものも含まれている．そのため，未知語の中から新造語を発見するための支援を行う．新造語には，通時的な頻度が時間経過とともに単調に増加していくという特徴がある．例えば，図 1 は，2006 年から 2008 年の間に収集したブログテキストを用いて，新造語「写メる」と「モフる」の頻度を月単位で表示したものである．頻度の変遷の様子が酷似していることが分かる．このような通時頻度に見られる類似性にもとづいて新造語の発見支援を行うために，DTW 距離を用いた方法を検討中である [7].

表 3　「ファブリーズする」と「ファブる」に対する「を格」と「に格」の出現数.

	を格	に格
ファブリーズする	27	93
ファブる	53	36

4.　認知言語学的分析

上記の手法と大規模時系列ウェブアーカイブ [8] を用いて，新造語の認知言語学的分析についても研究を進めている [9]．その結果，言語学的に興味深い知見をいくつか得ることができた．その一例として，新動詞「ファブる」における格交替の分析結果について簡単に紹介する．

新動詞「ファブる」は動詞「ファブリーズする」を縮退させた単語である．この 2 つの動詞の用法の違いを分析するために，2006 年に収集されたウェブテキストを対象として調査を行った．その結果，2 つの動詞の間に格交替現象が起こっていることが分かった．元の動詞「ファブリーズする」では「ソファにファブリーズする」のように，ファブリーズをかける対象 (例では「ソファ」) が格助詞「に」をとることが多かった．一方，新動詞「ファブる」の場合は，対象が格助詞「を」がとる例が多く観察された．実際に，2 つの動詞を含むテキストを 1000 文ずつ抽出し，ファブリーズをかける対象が格助詞「を」と「に」をとった回数を調べると表 3 のような結果となった．χ^2 検定を行ったところ，2 つの動詞の用法には統計的に有意な差 ($p < 0.01$) があることが分かった．以下に，典型的な用例を示す．

(2) a.　帰ってから服にファブリーズしなければ．
　　b.　汗臭いスーツをファブらなきゃ．
　　c.　かろうじて衣装にファブったあとは即寝でした．

格交替に伴う意味変化は，言語学で広く議論されている研究

課題である [15]．我々も格交替がこの 2 つの動詞の意味にどのような影響を与えているのかについて，さらに分析を進めている．詳細は文献 [9] に譲るが,「ファブる」が「ファブリーズをする」という文字通りの意味とは，やや異なる意味合いで使用されている例がいくつか観察されている．

(3) a. いらないもの，臭いものを指してそれファブっておいてーという言い方もあるってさ．
b. お前も田中のきもさ知ったらファブりたくなるよ．

さらに，同様の現象が，他の新動詞についても観察されるのかどうかについても調査を進める予定である．

5. 関 連 研 究

新造語のデータベースとしては，Rice University Neologisms Database [16] が存在する．このデータベースには，大学生が収集した英語の新造語が約 5500 登録されている．しかし，これは用例データベースではなく，基本的には新造語のリストである．これに対して我々の取り組みは，ウェブを新造語の用例データベースとみなして，その通時的振る舞いを明らかにしようとする試みである．

言語学，特にコーパス言語学の分野においては，ウェブテキストを研究目的に活用するための方法がこれまでにも議論されている. [17]〜[19]．しかしながら，我々の知る限りにおいて，大規模な時系列ウェブアーカイブを活用したという報告は存在しない．一方，工学的な立場から比較した場合，これらの先行研究では既存の検索エンジンのラッパー作成が主に議論されており，新造語のマイニングに関連する技術は提案されていない．

自然言語処理の分野では，形態素解析器の精度向上を目的として，コーパスから未知語獲得を行うための方法が議論されている [11]〜[13], [20]．我々の提案手法は，未知語獲得を核としており，これらの研究と関連が深い．しかし，言語学研究の支援を目的としている点が大きく異なる．

6. お わ り に

大規模な時系列ウェブアーカイブは，言語学者にとって貴重な言語資源であるにもかかわらず，これまで積極的に研究利用されることは少なかった．その理由の一つとして，既存のテキスト解析技術では，言語学者のニーズに必ずしも答えることができなかったことが挙げられる．そのため，時系列ウェブアーカイブを言語研究に活用するためには工学的な支援技術の開発が必要不可欠である．

本論文では，言語学研究の支援を目的として，大規模な時系列ウェブアーカイブから新造語を発見するための我々の取り組みについて述べた．研究は未だ萌芽的な段階にあるが，これまでのところ大規模ウェブテキストから大量の未知語を獲得することに成功している．そして，その結果得られた新造語と大規模時系列ウェブアーカイブを活用することによって，新造語の言語学的分析を行い，一定の成果をあげつつある．

近年では，ウェブアーカイブに代表されるように，我々が利用可能なテキストデータは急速に増大している．そうしたテキストデータの言語学研究を支援するためにもテキスト解析技術は今後さらに重要性を増してくると考えている．

文　　献

[1] N. Tsujimura and S. Davis: “A construction approach to innovative verbs in Japanese”, Handbook of the International Conference on Construction Grammar (2008).

[2] P. T. Sato: “Denominal verbs with -r: A response to de Chene”, Papers in Japanese Linguistics, **10**, pp. 149–169 (1985).

[3] B. de Chene: “*r*-epenthesis and the Japanese verbs”, Papers in Japanese Linguistics, **10**, pp. 172–207 (1985).

[4] L. Anderson and P. Trudgill: “Bad Language”, Blackwell (1990).

[5] A. Yonekawa: “Scientific Studies on the Language of Adolescents (wakamonogo o kagaku suru)”, Meiji Shoin (1998).

[6] W. Labov: “Sociolinguistic Patterns”, University of Pennsylvania Press (1972).

[7] F. Peng, F. Feng and A. McCallum: “Dynamic programming algorithm optimization for spoken word recognition”, IEEE Transactions on Acoustics, Speech, and Signal Processing, pp. 43–49 (1978).

[8] M. Kitsuregawa, T. Tamura, M. Toyoda and N. Kaji: “Socio-Sense: A system for analyzing the societal behavior from long term Web archive”, Proceedings of APWeb, pp. 1–8 (2008).

[9] 宇野良子, 鍛治伸裕, 喜連川優：“新動詞の認知言語学的分析：大規模時系列ウェブコーパスと言語処理技術が可能にする言語のダイナミズム研究”, 言語処理学会第１５回年次大会 (2009).

[10] 鍛治伸裕, 喜連川優：“文脈にもとづく未知語獲得における識別モデルの適用”, 言語処理学会第１５回年次大会 (2009).

[11] 福島健一, 鍛治伸裕, 喜連川優：“機械学習を用いたカタカナ用言の獲得”, 言語処理学会第１３回年次大会, pp. 815–818 (2006).

[12] 桑江常則, 佐藤理史, 藤田篤：“後続ひらがな列に基づく語の活用型推定”, 情報処理学会研究報告 NL-186-2, pp. 7–12 (2008).

[13] Y. Murawaki and S. Kurohashi: “Online acquisition of Japanese unknown morphemes using morphological constraints”, Proceedings of EMNLP, pp. 429–437 (2008).

[14] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shawartz and Y. Singer: “Online passive-aggressive algorithms”, Journal of Machine Learning Research, **7**, pp. 551–583 (2006).

[15] S. Andersson: “On the role of deep structure in semantic interpretation”, Foundations of Language, **7**, pp. 387–396 (1971).

[16] S. Kemmer: “The Rice university neologisms database”, http://esa4.rice.edu/ ling215 (2007).

[17] A. Kehoe and A. Renouf: “WebCorp: Applying the Web to linguistics and linguistics to the Web”, Proceedings of WWW, pp. 7–11 (2002).

[18] B. Morley, A. Renouf and A. Kehoe: “Linguistic research with XML/RDF-aware WebCorp tool”, Proceedings of WWW (2003).

[19] W. H. Fletcher: “Making the Web more useful as a source for linguistic corpora”, North American Symposium on Corpus Linguistics (2002).

[20] S. Mori and M. Nagao: “Word extraction from corpora and its part-of-speech estimation using distributional analysis”, Proceedings of COLING, pp. 1119–1122 (1996).