現代日本語コーパス比較分析のための中間語彙層の抽出と応用

ホドシチェク ボル †　　山 元 啓 史 ‡

† 国立国語研究所言語資源研究系
‡ 東京工業大学大学院社会理工学研究科

要 旨

本研究は、現代日本語コーパスにおける中間語彙層の役割に焦点を当てその応用について示すものである。ここでいう中間語彙層とは、個々の単語の情報量を計算し、その情報量の序列において中間に位置する語彙である。一般的には、序列として使用頻度が用いられているが、本研究では各単語が持つ情報量の分布を用いる。頻度の高い語は機能語が多く、文の構造や語の係り受けを明示するが、内容としての情報量は少ない。一方、頻度の低い語は、トピック・内容をよく表した語か、珍しい語、固有名詞などである。頻度の低い語を利用すると領域毎に特化された情報が顕著になるあまり、共有語彙が少なくなり、文書間の比較や時系列の変動が分析しにくくなる。中間語彙は上記の 2 つの問題点について相互比較、時系列比較を可能にする語彙集合である。本稿ではその抽出の手法と応用例について述べる。

Analysis and Application of Mid-Rank Lexicons of Modern Japanese

Bor Hodošček †　　Hilofumi Yamamoto ‡

† Department of Corpus Studies, National Institute of Japanese Language and Linguistics
‡ Graduate School of Decision Science and Technology, Tokyo Institute of Technology

Abstract

The present study focuses on the role of mid-rank words within modern Japanese. Taking all the words and ranking them according to their information content, mid-rank words are defined as words positioned at the center of the ranked list. While words are commonly ranked according to their frequency, the present study instead utilizes the distribution comprising the information content of all words. Frequently occurring words are often function words which express the structure of the sentence and the grammatical relations between words, but are otherwise low in information content. On the other hand, infrequent words often either express the content or the topic of the sentence or are proper nouns or other rare words. Utilizing infrequent words for comparisons between documents and for the analysis of change across time is fraught with problems because words become too domain-specific which leads to insufficient shared vocabulary between documents for meaningful comparisons and analyses. This study shows how the set of mid-rank words satisfies the aforementioned problems and makes it possible to compare documents and time-series data.

1　はじめに

本研究は、現代日本語コーパスにおける中間語彙層の役割に焦点を当てたものである。ここでいう中間語彙層とは、個々の単語の情報量を計算し、その情報量の序列において中間に位置する語彙である。語彙の序列 (rank-order) の分布としてはジップの第２法則 (Zipf, 1949) が有名であり、日本語ではＬ字型分布 (水谷, 1975; 中野, 1976) となるといわれている。これらの研究では、序列として使用頻度が用いられているが、本研究では各単語が持つ情報量の分布を用いる。頻度の高い語は機能語が多く、文の構造や語の係り受けを明示するが、内容としての情報量は少ない。一方、頻度の低い語は、トピック・内容をよく表した語か、珍しい語、固有名詞などである。情報量の少ない語彙は、機能語として扱われ、英語の情報処理では stop word としてあらかじめ処理から取り除かれることが多い。ところが、情報量の中ぐらいに位置する語彙（ここでいう中間語彙層）はトピックを決定するほどの内容性もなく、かといって機能語としての役割だけでもなく、実に捕らえどころのない語であり、未だその本質はわからない。では、なぜそのような語彙が重要なのか。

2　なぜ中間語彙層が重要なのか

コーパスが入手可能になり (前川, 2008)、データサイエンスとしての言語学が行われつつある。従来、目に見えなかった微妙な違い、内省の及ばない時代をさかのぼる資料の分析、議論だけでは決着がつかなかった数々の問題について、大量データ・機械処理によって、きわめて機械的で主観性を排除した分析が可能になる。

中でもコーパス利用の言語学に有効な分析の１つは比較である。単なる比較なら、従来より多くの言語学者が目を皿のようにして幾万もの資料を比較してきている。しかし、言語学者が目を使ったとたん、言語学者の先入観も含まれてしまう。機械処理はこの主観性の問題を解決してくれる。

さて、２つの時代の言語（あるいはコーパス）を比較する際、それらの間で共有する語がまったくない場合（あるいはその反対にどの時代にも同じ語が大量に出てくる場合)、言語の変化は分からない。トピックに依存する単語（あるいはすべての時期に均等に大量に出現する機能語）よりも、継続的に変化（変動）が追跡できる語彙がデータとして望ましい。

言語全体を分析利用するのではなく、目的に適切な語は何であるのかを考えることによって、比較の方法が確立できれば、あとは数理的な問題として解決していけばよい。言語学以外にも、考古学、民俗学、歴史学、文献学など数理的な比較をデータベースによって行う領域においては中間的な語彙の抽出方法の検討は有効なはずである。

これまでに中間的な語彙の取扱いについては、Luhn (1968, 120) や長尾 (1983, 28) など少なからずある[1]。Luhn (1968) は論文概要の自動生成システムを開発するにあたり、図 1 にあるように、頻度上位と頻度下位の語彙をカットし、中間の語彙を生成に利用した。破線はそのランクに位置する語の貢献度を示しており、正規分布を仮定している。長尾 (1983, 28) も自動抄録の生成において、中間の語彙を有効としている[2]。しかしながら、両者とも upper-cutoff、lower-cutoff の位置は経験的なものとしており、どこで切り分けるべきかについては議論されていない。

さて、頻度上位、中位、下位に分割したとき、それぞれの語彙層にはどのような特徴があるのだろうか。頻度下位の語は特定の文章にしか見られず、人名、地名など固有名詞や専門用語の文書検索に利用される。頻度上位の語は統語関係をつかさどる機能語である。また、上位に位置する「、」や助詞の使い方は文章の内容やジャンルの影響が小さくかつ書き手のクセが出やすいことから著者の推定に利用できることが実証されている (金, 1994, 1997)。

英語学においても、頻度上位の用語の出現パターンを利用して、著者の推定を行う研究が行われている。Burrows (2002) は “Delta” という尺度を導入し、全体頻度上位 150 語の機能語を抽出し、分析対象の各文書について、それぞれ 150 語の出現頻度と各文書における Z スコアを算出した。その上で著者不詳文書の Z スコアのパターンがどの著者のものと一致するかを計算し、著者の推定を行った。

このように語彙の序列を上位、下位に分割すると、そこには言語の特徴を分析する上で便利な役割が見える。おそらく中位においても何らかの分析において便利な特徴があるものと考える。では、実際に日本語コーパスにおける中間語彙層に焦点を当て、その抽出と応用を検討する。

[1] 中間に位置する語彙を層として意識的にとらえた研究は見当たらない。

[2] 長尾 (1983) は Luhn (1968) のベルカーブ辺りに語の長さの分布を割り当てている点が少々異なる

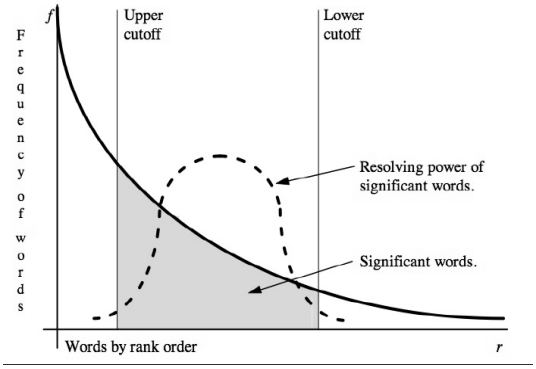

図 1: Hyperbolic curve relating occurrence frequency with rank order; adapted from Luhn (1968, 120)

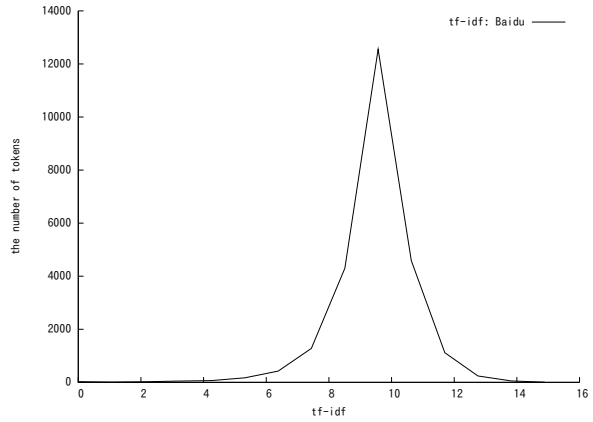

図 2: Baidu 時系列コーパスにおける *tf-idf* の分布: *tf-idf* の平均は 9.970、SD は 1.229 であった。1 σ 以上の語はトピックを明示する語、-1 σ 以下の語は助詞、助動詞、基本動詞「ある」「いる」のような文の構造を明示する語が見られる。平均を挟んで上下 1 σ の範囲にある語、つまりトピックにも、文構造にもいまいち貢献しない語、かつ、広範囲に渡って量的に見られる語を基準に談話の特徴語を客観的に選ぶ。N-gram データで文章数はわからないので、Baidu コーパスでは *idf* は計算できない。そこで、「現代日本語書き言葉均衡コーパス BCCWJ」(前川, 2008) によって計算した。

3 中間語彙層の抽出

実際に中間語彙層の抽出を検討する。次節との関連もあり、「Baidu ブログ・掲示板時間軸コーパス」(バイドゥ株式会社, 2010, 以下、Baidu コーパス) を材料として用いる。各単語の情報量の重み付けとしては *tf-idf* (Spärck Jones, 1972) を用いる。図 2 は *tf-idf* の分布で、表 1 はその値において上位、中位、下位あたり用語の一覧である。下位は「に」「の」「は」「が」「で」「を」など頻度の常に多く、キーワードになりにくい語である。上位は固有名詞が多く、内容が何となくわかりそうであるが、頻度は低い。上位の語は、どの文章にも見られるわけではないだけでなく、言語処理に用いられる辞書では未登録扱いされる可能性が高い。中位は従来、分析において、あまり特徴がないために注目されてこなかったが、これらの語はどの時代、どの文献にもある程度見られることが予想されるため、時系列コーパスを用いれば、相対的な変動（増減）が観察できる。

図 2 を見ると実際のコーパスでは、尖りすぎではあるが、ほぼ平均値を中心として両裾野を持つ分布である。そこで、分布の拡がり（標準偏差）を計算し、中位を平均値から上下 1 σ をとり、右を上位、左を下位として自動的に 3 分割することができる[3]。

[3] 1 σ を上下のカットに使う積極的な根拠は検討中であるが、合理的な理由としてはつぎのように考える。上下 1 σ はガウス分布における変曲点である。分布にそって、ある点を同時に左右裾野に辿っていく時、2 つの点は語数を減らす方向（下へ向かう）のほうが、減らさない方向（左右に広がる方向）より大きいが、変曲点を境に減らさない方向（左右に広がる）のほうが大きくなる。つまり、変曲点は中間語彙層の区切りと考える。

表 1: *tf-idf* の上位、中位、下位の用語: 上位は固有名詞などトピックが特定できるようなキーワード性の高い語。下位はどの文章にも必ず出てくるような助詞、助動詞など文型を示す語。中位は内容語ではあるが、トピックが特定できるほど、語に特徴はない。数値は *tf-idf* 値。いずれも順位は降順で、上位は上から 1、中位は平均値から、下位は下から、それぞれ 10 語をリストアップした。集計は基本形ではなく、出現形（実際にコーパスに出てきたまま）で集計されている。

	上位		中位		下位	
1	15.9491	ルコ	9.9702	旧家	0.8213	か
2	15.8162	レイ子	9.9702	衣替え	0.7878	を
3	15.4672	浜北	9.9702	起爆	0.7046	で
4	15.2879	カイラ	9.9702	つと	0.6009	、
5	15.0323	セリーナ	9.9702	噛ま	0.4852	に
6	14.8390	Bo	9.9702	ハイハイ	0.4537	が
7	14.6914	育江	9.9701	祝福	0.4383	は
8	14.6903	昌浩	9.9701	描ける	0.4341	て
9	14.6244	保呂	9.9700	煮詰め	0.3497	。
10	14.5512	リンゼイ	9.9700	レンタル	0.2305	の

4 中間語彙層の応用

中間語彙層の応用として、時系列コーパスにおける「うわさ」の検出を試みる。「うわさ」は「話の盛り上がりの現象の 1 つ」と考え、中間語彙層にあたる用語によって増減している箇所の特定を試みる。

Twitter や Facebook など時系列に整理したリソースの発達により、時系列で人間の行動を分析する研究やシステムが出てきた。また、Google トレンドのように、サーチエンジンで検索された検索語を時系列に整理し、流行りのことばを可視化するシステムも出てきた[4]。Komori et al. (2012) は、Twitter の即時性、周期性の特徴に目をつけ、肯定的態度、否定的態度について調査し、フーリエによるモデル化を行い、昼間の態度はほぼ肯定的、夜間の態度は否定的であることを発見した。Nikolov (2012) は、過去の Twitter に投稿された時系列テキスト（ツイート）を用いて、トレンド[5]を予測する研究を行った。この研究では、あるトレンドワードの発生、増加、衰退の過程を数理的に調査し、トレンド増加のきっかけとなった時期を予測した。発生の推移と予測は今後の対策として重要である。しかし、実際に未来を予測する際、我々あらかじめトピックの内容を知らないはずである。どのキーワードでそれは検索できるのか、わからないはずである。はじめからトピックの名称やキーワードがわかった「うわさ」があるのではなく、はじめはぼんやりとした話が語り継がれて、いつしか用語らしきものが使われだし、それがだんだん一般的になって、市民権を得るのであろう。こういうプロセスは一般的なことばの発生と同じであるが、「うわさ」が他の用語の発生や衰退と異なる点とはいったい何であろうか。

4.1 「うわさ」の生態

「うわさ」は言語の営みによって生じる現象のひとつである。目にも見えず、手でも触われず、興味、損失、危険性が高いと感ずれば感ずるほど、不安は募り、伝達スピードも加速する。伝達のチャンネルは無数に存在するため、一度成長したものを鎮圧するのはきわめて困難となる (林, 2007, 55–60)。オルポート・ポストマン (2008) は次の式によって「うわさ」を形式化した。

$$R \sim I \times A$$

「うわさ」の流布量 (R) が内容の重大さ (I) と表現の曖昧さ (A) の積に比例することを経験則から示したものである。重要でなければ、どうでもよいことは伝わらない。気になるが、いまいち内容が不確かなので、不安が募り、曖昧なまま次から次へと内容が伝えられていく。常識や周知の事実のように内容が明確なものは話題にも上らない。だが、デマのようなあやしい情報は、受けつがれていくうちに、次第に短くなり、要約され、平易になっていくという傾向をもっている (オルポート・ポストマン, 2008, 90)。「うわさ」には、平準化（情報が短く要約されて平易化すること）、強調（情報の中からある要素が選ばれ、誇張されること）、同化（知的あるいは感情的な状況で情報が歪められること）などの現象

[4] `http://www.google.co.jp/trends/`。これらにより現時点の急上昇ワードを見せるだけでなく、過去にさかのぼって話題の盛衰が一瞥できる便利さや、検索語がいつどの程度の量使われたか、またそれに伴う代表的な記事は何であるのかを表示できるようなった。

[5] 時代の趨勢、潮流、流行のこと。広辞苑第六版 (新村, 2008)

が見られる (林, 2007, 55–6)。「うわさ」の初期状態には、1) 内容、表現共にはっきりしない、2) 歪められている可能性がある、3) 聞き手にとって重要で、関心の高い内容を持つ、などの諸特性があると言えよう。

4.2 「うわさ」の検出と中間語彙層

では、実際に「うわさ」の箇所を時系列データより検出するにはどのようにすればよいのだろうか。本研究では、「話の盛り上がり現象の特定」に的を絞って議論を進める。

「うわさ」は「話の盛り上がりの現象の 1 つ」と考え、キーワードではなく、（どこにでも見られるごく）一般的な用語が時系列の中で急激に増減している箇所を検出する方法を用いる。

「うわさ」の初期的段階は、それを指し示す記号が統一されていない可能性が高く、それは「うわさ」文の検知において問題となる。たとえば、文を単語に分割する際、新語は解析辞書に当然含まれていないことも考えられるし、固有名詞が辞書になければ、いくつかの語に分割され、何の特徴もない語になってしまう[6]。

経済の変動や Twitter のような秒刻みの変動の場合には、増減の反転する場合も多く、頂点の特定はむずかしいが、月ごとの離散数（単語の頻度集計）の場合には、あまり微動は考えられにくい。そこで、単語の頻度の状態を恒常と異常の 2 つの状態と考え、シューハート（W.A.Shewhart）の管理図の考え方を用い、異常な状態を特定する[7]。上記により、Baidu コーパスの全用語を計算し、頻度は 3 σ 以上、*tf-idf* は 1 σ 以下に絞り込むと、39,989 の用語が得られた。

4.3 考察

図 3 は、Baidu コーパス 127 カ月間において、100 カ月以上出現した用語で、その用語の各月の頻度平均 200 以上、標準偏差 3 σ 以上を検索し、抽出された用語のうち、顕著な盛り上がりを見せた用語「アメリカ」の例である。「アメリカ」は、126 カ月に出現し、ほぼ毎月かなりの頻度で見られる。*tf-idf* は 7.59 で、情報量は中位である。用語「アメリカ」は、2008 年 10 月急激な盛り上がりを見せる。当時、アメリカではニューヨーク証券取引市場のダウ平均株価は史上最大の 777 ドルの暴落を記録している[8]。

以上、用語「アメリカ」を通して顕著な盛り上がり現象を特定したことを示した。これを可能にしたのは、ある特定の時だけにその用語が使われるのではなく、時を隔てて使われ、相対的な頻度の変化が観察できたからである。「話の盛り上がり現象」の特定では時系列という場合の比較を行ったのであるが、この考え方は時系列でなくても、文書間の比較においても応用できるはずである。

5 おわりに

中間語彙層は文書もしくは文をはじめとする言語情報の比較、文書内容の比較や分析を通して、有効な語彙であることを示した。*tf-idf* による情報量変換において見られることは示したが、他の重み付け計算については実施していないので、中間語彙層による分析および分類の精度に与える影響についてはまだ不明である。

参考文献

オルポート, G. W.・L. J. ポストマン (2008)『デマの心理学』，岩波モダンクラシックス，岩波書店.

バイドゥ株式会社 (2010)「Baidu ブログ·掲示板時間軸コーパス」. http://www.baidu.jp/corpus/.

Burrows, John F. (2002) "'Delta': A measure of stylistic difference and a guide to likely authorship", *Literary and Linguistic Computing*, Vol. 17, No. 3, pp. 267–287.

林幸雄 (2007)『噂の拡がり方—ネットワーク科学で世界を読み解く (DOJIN 選書 9)』，化学同人.

金明哲 (1994)「読点の打ち方と文章の分類」,『計量国語学』，第 19 巻，第 7 号，p317–330 頁，dec 月.

[6] たとえば、「ワールドトレードセンター」は「ワールド／トレード／センター」に分割されるなど。

[7] 管理図は、古典的かつ固定的な選別基準である。佐藤 (1968, 63–4) がわかりやすく説明している。簡単に説明すれば、時系列データは規則的な間隔で得られたデータであるから、そのデータ群から、平均値などの特性値を算出し、それをグラフ化する。平均値を中心として、上方管理限界 (UCL: Upper Control Limit)、下方管理限界 (LCL: Lower Control Limit) の 2 本の線を書き、限界を越えた値を異常値と見做す。シューハート管理図による管理方法は、3 シグマ法とも呼ばれ、標準偏差の 3 倍以上の値を異常値としている。

[8] Wikipedia「世界金融危機 (2007 年–)」によると、2008 年 9 月 29 日にアメリカ合衆国下院が緊急経済安定化法案を一旦否決したのを機に、ニューヨーク証券取引市場のダウ平均株価は史上最大の 777 ドルの暴落を記録したとのこと。

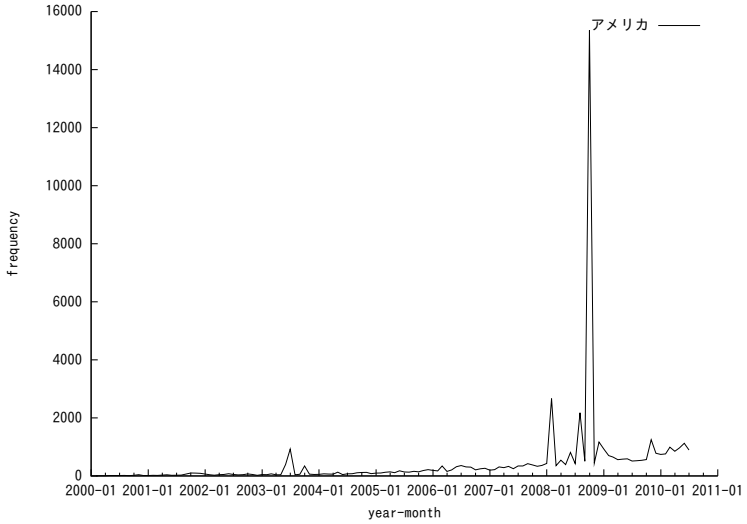

図 3: 盛り上がり箇所の検出「アメリカ」の例: 127 カ月間において、100 カ月以上出現した用語で、その用語の各月の頻度平均 200 以上、標準偏差 3 σ 以上を検索。すると、「アメリカ」は 1 カ所のみ顕著な盛り上がりが見られた。2008 年 10 月 10.72 シグマの隔たりで、15358 回出現している。このころ、アメリカ合衆国下院が緊急経済安定化法案を一旦否決したのを機に、ニューヨーク証券取引市場のダウ平均株価は史上最大の 777 ドルの暴落を記録した。この図には 2001 年 9 月 11 日の同時多発テロの盛り上がりは見られないが、Baidu コーパスの初期のデータ量が少ないのが原因していると見ている。

——— (1997)「助詞分布に基づいた日記の書き手の認識」,『計量国語学』, 第 20 巻, 第 8 号, 357–367 頁, dec 月.

Komori, Masashi, Naohiro Matsumura, Asako Miura, and Chika Nagaoka (2012) "Relationships between Periodic Behaviors in Microblogging and the Users' Baseline Mood", in Teruhisa Hochin and Roger Y. Lee eds. *SNPD*, pp. 405–410: IEEE Computer Society.

Luhn, Hans Peter (1968) *HP Luhn: Pioneer of Information Science: Selected Works*: Spartan Books.

前川喜久雄 (2008)「KOTONOHA『現代日本語書き言葉均衡コーパス』の開発:〈特集〉資料研究の現在」,『日本語の研究』, 第 4 巻, 第 1 号, 82–95 頁, 1 月.

水谷静夫 (1975)「短い作品の語彙の量的構造昭和初期流行歌の調査から 1」,『計量国語学』, 第 72 号, 1–12 頁.

長尾真 (1983)『言語工学』, 人工知能シリーズ 2, 昭晃堂.

中野洋 (1976)「いわゆる L 字型分布からはずれる語彙量の分布について」,『計量国語学』, 第 76 号, 25–31 頁.

新村出 (2008)『広辞苑 第六版 (普通版)』, 岩波書店, 第 6 版.

Nikolov, Stanislav (2012) "Trend or No Trend: A Novel Nonparametric Method for Classifying Time Series", Ph.D. dissertation, Massachusetts Institute of Technology.

佐藤信 (1968)『推計学のすすめ - 決定と計画の科学 (ブルーバックス)』, 講談社.

Spärck Jones, Karen (1972) "A Statistical Interpretation of Term Specificity and Its Application in Retrieval", *Journal of Documentation*, Vol. 28, pp. 11–21.

Zipf, George Kingsley (1949) *Human Behavior and The Principle of Least Effort, An Introduction to Human Ecology*: Addison-Wesley Press Inc.