日本感性工学会論文誌 Vol.12 No.4 pp.511-518 (2013)

原著論文

内容推測に適したキーワード抽出のための日本語ストップワード

國府 久嗣*，山崎 治子**，野坂 政司*

* 北海道大学，** 岩手県立宮古高校

Japanese Stopword List Making for Keyword Extraction Suitable for Semantic Interpretation

Hisatsugu KOKUBU*, Haruko YAMAZAKI** and Masashi NOSAKA*

** Hokkaido University, 17-8 Kita-ku, Sapporo, Hokkaido 060-0817, Japan*
*** Iwate Prefectural Miyako High School, 2-1-1 Miyamachi, Miyako-shi, Iwate 027-0052, Japan*

Abstract : Extracting keywords from a target text data is essential for an analysis to describe substance characteristics of message content. We picked a use of a stopword filter from among alternatives because the method has the advantage that it is simple yet effective way. The filter we present was made up of non-content words and low-content words. Non-content-bearing words consisted mainly of function words and were gotten rid of by using part-of-speech (POS) tag information. High occurrence rate words in remaining had prospects of being keywords, however usually there were some low-content words like delexical verbs and so on. This article presents a stopword list obtained to come up with low-content words by sensuous manual procedures carried out using 40 text files from the CASTEL/J database and establishes it in the view of general versatility.
Keywords : *Stopword, Content analysis, Low-content word*

1. は じ め に

本稿では対象テクストの内容推測を目的としたキーワード抽出に有効なストップワード（処理対象から除外する語）とはどういうものであるかを検討し，その考察にもとづいてストップワードリストを作成・提示する．

1.1 利点と難点

使用目的に合致したキーワードを自然言語処理によって得る手法については様々なものがあるが「出現頻度が高い語の中から何らかの規準によって選別する」という考え方は通底している．このうちストップワードを用いる選別法の利点は「リストにある語かどうか」を判定するという単純処理だけで済む簡便さにある．また主たる分析過程で取り扱うデータ量を予め減ずることになるので計算機処理の負担軽減効果も期待できる．この点からストップワードの利用は情報検索分野で著しい［6］.

しかし適切なリストが既に存在している前提からの処理は容易と考えられるものの，このリスト自体をどう作成するかについては技術的問題とは別に考察すべき点がある．「どういう語を除外すべきか」という規準は用途によって大きく異なるため最適リストが一意に決まることは原理的に有り得ず，使用目的の明確化が必要である．

1.2 内容推測のためのキーワード

本稿でのストップワード使用目的は「対象テクストの内容面での特徴を強く反映した語をキーワードとして選別するため」である．これは内容分析に関するBerelson［3］の5つの分類のうち「to describe substance characteristics of message content」に沿っている［7］.

目的に合致したキーワードを抽出［注1］する方法について論じるには想定される分析手続き全体がどのように設計されているのかを示す必要がある．詳細について述べる紙幅はないが，簡略に概説すれば次の三段階で構成される．まずキーワード相互の関係性を距離的数値に換算する処理を行なう．次に，得られたクロス集計表を統計手法によって視覚化・クラスタリングする．最後に各クラスタに適したラベルを付与しながら対象テクストの内容を人間が解釈していく．この一連の分析作業を通して対象テクストを読解することなく内容が推測できる．テキストマイニングなどの探索的研究手法の一種であり，計算機処理による内容要約分野とも関連する．

現在ではKH Coder［注2］などのアプリケーションにも実装され，こうした分析は広く試みられている［2, 12］．なお距離的数値にはoverlap係数の改良版，視覚化には多次元尺度構成法（MDS），クラスタリングにはWard法の使用というのが本稿で想定する具体的な分析設計である［14-16］.

予定される一連の分析で端緒に位置されるキーワード抽出では個数に関する設定如何で抽出法の設計も変わってくるため，この点についてまず方針を明確にしなければならない．内容推測はキーワード数が増減してもそれに応じた精度の結果が得られ，また視覚化処理やクラスタリングでの視認性や解釈作業の煩雑さなどの問題点が考慮されねばならないことから，テクストの規模にかかわらず30程度に制限した方が効果的であるという仮説を採用する［16］．これは多次元尺度構成法を中核とするファセット分析を使って効果をあげた

Received 2013.02.18
Accepted 2013.07.22

リヴァプール式プロファイリング［4］での行動項目数［注3］が，理論的根拠について言及はないものの，概ねこの値前後に収束していることに拠る．

1.3 除去する対象の分類

キーワードから除去する対象は大まかに「非語」「非内容語」「低内容語」の三種類に分類できる．「非語」は句読点などの 区切り記号類が該当する．これはそもそも「語」ではないので当然除去されねばならない．「非内容語」については主たる構成要素が一般に「機能語」と呼ばれて日本語では助詞や助動詞などの品詞に分類されている語である．

Zipfの法則として知られる経験則［10］にあるように，限られた数のword typeが実際に使用されるword tokenの大部分を占めるという性質が自然言語には見られ，「機能語」は最上位の出現頻度を示すword typeであることが多い．内容推測には直接役立たないにもかかわらず候補にあがりやすく除去する必要がある．

「非語」「非内容語」は本稿の目的にはそぐわないため除去対象となるが，内容分析において常に不要とされるわけではない．これらの使用傾向から書き手の特徴を捉えること自体は十分可能である．宇治十帖の作者問題を扱った研究［17］やシェークスピアの正体に関する研究［5］などが古くから知られている．

「機能語」でないものは「内容語」とされ，詳細な分類については諸説あるが品詞では名詞や動詞，形容詞などが該当するといわれる．しかしこうした品詞の語であっても，対象テクストが「何をいっているのか」を推定するには極めて限定的かつ間接的な役割しか果たさないものがある．たとえば「代名詞」は名詞の一種であるが固有の意味をほとんど持たず，「食べている」の動詞「いる」などは非自立語である助動詞に近い働きをする．こうしたものも「非内容語」に分類した．

一旦どの品詞が該当するのか決めてしまえば「非内容語」は形態素解析器の品詞情報に基づいて除去が可能となる．しかし品詞情報によるフィルタリングでは排除されない語であるが内容推測に貢献しそうにないものも存在する．これを「低内容語」とした．この「低内容語」はリストに登録しておき，それをもとに除去するほかない．つまり本稿で作成するストップワードリストとは「キーワードとしての抽出が予想される程度に出現頻度が高い低内容語」のリストである．上記に示した考え方に基づき具体的には以下の手順でストップワードを選定した．

2. ストップワードの選定作業

2.1 使用したテクスト群

テクスト群としてはCASTEL／J［注4］に含まれる講談社現代新書33冊分のデータと講談社ブルーバックス7冊分の計40冊分を使う．前者には広い意味で「日本文化」を題材としているという緩やかなテーマの共通性がある．後者は科学読み物であってテーマの一貫性は特にない．出版年代は1967年から1992年までの範囲である．

各テクスト本文データには6桁の番号が振られ一冊につき6種類のファイルとして格納されている．全ファイル共通して先頭がJで始まり続く4桁分の英数字がテクスト識別に使われる．末尾の1桁は「分かち書き」の有無などファイルの加工状態を表す．このうち末尾が2の無加工テキストファイルだけを形態素解析器MeCab［注5］で品詞タグ付けして分析に用いた．

2.2 品詞タグ

MeCabの品詞タグはIPADIC［注6］にもとづいて付与される．この品詞情報で判別が可能な「非内容語」を取り除く．

IPADICによる品詞分類は4段階あり，MeCabでは「品詞」「品詞細分類1」「品詞細分類2」「品詞細分類3」となっている．「品詞」レベルでは「連体詞」「接続詞」「助詞」「助動詞」「連語」「副詞」「接頭詞」を除去した．「連語」は具体的には「について」「とかいう」などのことで形態素解析の間違いを減らすために便宜的に設けられた品詞分類である．

「品詞細分類1」では「形容動詞語幹」「副詞可能」「代名詞」「ナイ形容詞語幹」「特殊」「数」「接尾」および「非自立」を指定した．「品詞細分類1」で選定した項目は主に名詞の中から「非内容語」を除去する役割を担っている．また対象テクストが論説・評論文であることも鑑みて主観や修飾表現にかかわる度合いが高いものはなるべく除去する方針をとった．「品詞細分類2」「品詞細分類3」レベルでのフィルタリングは行なっていない．

2.3 ストップワードリスト

用意したテクスト群から順に個別テクストを取り出し，上位30の高頻度語中に「低内容語」があればリストに登録してキーワードから除去した．このとき「低内容語」であるかどうかの判断は感覚的なものであって明確な規準を先験的に用意していたわけではない．

作業を経て上位30語が「低内容語」以外で構成されたとみなせれば次のテクストに移った．このときストップワードリストは継承され，新たな「低内容語」があれば追加していった．以後40冊分のデータすべてで同様の感覚的な取捨選択作業を行なった結果，ストップワードリストが作成できた．

このリストに登録された語がそれぞれいかなる理由によって「低内容語」と判断されたのかは結果から遡って分類・再考察できる．これによって恣意的でなく妥当な選別基準を帰納的に求める．

3. リストの構成と複合語の取り扱い

40冊分のデータを使って作成したストップワードリストは100語程度の規模となった．その構成内容は大まかに「形容詞に関するもの」「動詞に関するもの」「その他」の3種類に分類できる．各分類項目について以下に考察を述べる．また複合語の問題についても触れる．

3.1 形容詞に関するもの

形容詞に該当するものは表1のようにまとめられる．カッコで括っているものには表記の揺れがある．対象データ全部の分析を通してキーワードに残った形容詞は「新しい」「深い（ふかい）」「若い」「美しい」「楽しい」「おもしろい」「すばらしい」であった．

対義語と対で出現した形容詞は「大きい／小さい」と「多い／少ない」「良い／悪い」の3例であり，いずれもストップワードとした．「高い／低い」「強い／弱い」「長い／短い」はペアの一方しか登場していない．はっきりと断定はしかねるが肯定か否定かでいえば肯定的意味合いで使えるものが出現しやすい傾向がある．

残ったものとストップワードの差については，後者の方が固有の意味が薄く多用されるものである点が指摘できる．事後的に集計表から実際に確認できたことだが，キーワードは特定少数または一つのテクストで高頻度出現し，ストップワードは多数のテクストで（少なくとも対義語の一方は）高頻度出現する傾向がみられた．

3.2 動詞に関するもの

英語でいえばdelexical verbに該当するものを中心に表2の語をストップワードとした．delexical verbは具体的には「be, come, go, get, give, put, take, make, do, have」等のことでbasic verbや light verbとも呼ばれる［11］．日本語では「基本動詞」「軽動詞」「機能動詞」などの用語がつかわれる．

論者によって定義は異なるが「多用されることで多義性を帯び，それによって意味の特定性が希薄化してしまった動詞」であるとはいえよう．表2では「非内容語」に分類された非自立の動詞らと同じ表記の語群（ラベル「併用」）にその傾向が強い．別ラベル「する」とした語群も同様に機能語的性質を持つ．

ラベル「叙述」の語群は論説・評論文に頻出する言い回しに関する動詞である．この語群も動詞そのものが持つ意味は軽くなっていると考えられる．

表1 形容詞のストップワード

否定	ない
程度	高い 多い 少ない 強い 大きい 小さい （長い ながい）
好悪	（良い よい いい） 悪い

表2 動詞のストップワード

併用	ある いる なる（行く いく）来る とる （見る みる）（言う いう）得る（過ぎる すぎる）
する	する やる（行う 行なう おこなう）（出来る できる）
叙述	（思う おもう）（考える かんがえる）わかる 見える 知る しれる いえる 示す 述べる（書く かく）よる
相違	異なる（違う ちがう）くらべる
入出	入れる（出る でる）（入る はいる）
雑	使う（用いる もちいる）（持つ もつ）（作る つくる） なす（起こる おこる）つく つける 聞く よぶ

このうち「しれる」についてはキーワード候補中に頻出したことについてやや納得し難いかもしれない．「しれる」に関しては「かもしれない」における下線部の用例がほとんどである．この表層形であれば出現事例が多いことに得心がいくであろう．リストには「レンマ化（見出し語化）」したものを載せていることによる弊害とはいえる．

ラベル「相違」「入出」は関係性についての動詞である．これを除去したのはキーワード抽出の最終的な目的が先述した通り関係性に基づいて語群を分割カテゴライズして意味構造を解釈することにあるからである．つまり語相互の関係性そのものを表す語は他の語に比べて内容推測への貢献度は低いと考えられる．

「雑」ラベルに分類される語もそれぞれ上述した3種類の理由（基本動詞，ジャンル文体，分析の都合）のいずれかに準拠してストップワードとした．

3.3 その他

形容詞と動詞以外でストップワードに該当するものは表3の通りである．

「誤認」としたのは形態素解析における誤認でフィルタリングをくぐり抜けてしまった語群である．「かれる」は平仮名表記の代名詞「彼」の誤認識によってキーワード候補中に出現してしまっている．「つまり」は言い換えの副詞などとしてではなく「パイプのつまり」の下線部的な名詞として誤認識してしまった事例である．「お」は敬語をテーマとするテクストにおいて「接頭詞」として本来品詞タグで除去されるべきものが，用例が多すぎたことなどにより「感動詞」に多数取り違えられてキーワードに残った．この種の誤認識は特定のテクストにだけ発生する．

ラベル「関係」は動詞に関して言及した3番目の理由「分析の都合」つまり関係性に該当する語群である．

「冗長」としたのは今回分析対象としたデータに関して，自明であったり文脈上同義語とされる語が他に頻出していたりするなどの理由で「低内容語」と判定した語群である．「わが国」と「日本」は今回のデータ群では同義であるが，たとえば外国語の原書を翻訳した日本語テクストを分析する際には同義とはならない．「わが国」の方が多義的であり，その分意味が軽く希薄であるとみなせる．重複した場合はこちらを除去する判断となる．

「雑」ラベルの語群はそれぞれ雑多な理由でストップワードとしている．「そのもの」「一つ」「あと」は「名詞 一般」としてIPADICに登録されているので品詞タグで除去されなかったが内容推定への貢献は期待し難い．

表3 その他のストップワード

誤認	かれる つまり お
関係	上 下（次 つぎ）
冗長	わが国 自分 人（人々 人びと）
一字	別 他 間 話 例 形 日 家 手 名 身
雑	そのもの 一つ あと a

「a」はひとつのテクストにおいてのみ「Xにおけるリーダーa」のように代名詞的に多数用いられたことでキーワード候補として出現した．書き手およびテクストの特徴をよく表した語であることは間違いないが意味の希薄性の観点から除去した．

残りの語群はラベル「一字」に分類したものである．これは他の分類に含まれない一文字語からなる．ストップワードであると判断した理由はそれぞれ先述してきた規準のいずれかには該当する．しかしもっとも重視した点はこれらが一文字であることである．一文字語は複数文字からなる語よりも固有の意味が希薄であって内容推測に貢献し難いと考えた．

3.4 複合語

形態素解析器 MeCab による語の解析は厳密に形態素レベルへの分解を指向しているわけではないが，日本語話者が日常生活で「これは語である」と認識するものよりは細かく単位を扱っている．このため解析結果には自立の程度が低く多義的で意味が希薄な一文字語が頻出しやすい．

文でも文節でも句でも語でも形態素でもそうであるが，自然言語では複数の単位が連なることで意味の特定がなされる仕組みが存在している．特徴語抽出には複合語・合成語といった長単位が適しているという意見があり［13］，こうした需要に応えるTermExtract［注7］のようなアプリケーションも存在する．

本稿でのキーワード抽出は語同士の関係性を分析することによる内容推測に資することにあるので「社会集団」「日本社会」「社会構造」といったレベル（長単位）での複合語を含むと重複する部分が多く冗長となってかえって解釈が難しくなる可能性が高い．「社会，集団，日本，構造」という語群（短単位）によるクラスタが見出されれば十分であり，これらから自然に思い付ける程度の複合語であればわざわざ明示する必要性は見出し難く，二種類の単位が混ざることも好ましいとは言い難い．

このことは一文字語においても同様と考えられる．しかし，たとえば「地図」がテーマのJ03692において一文字語「図」をそのままキーワードとするか「低内容語」とみなして除去してしまうよりも表4にある「国絵図」「図屏風」のような想起が難しい特殊な複合語として明示できた方が分析には好都合である．このとき一文字語「図」そのものは複合語の一部分としては残ったのであるからキーワードから消去してしまってかまわないと考えられる．

表4　複合語の事例

J03692	日本図　世界図　行基図　地形図　伊能図　図屏風　国絵図
J06552	先進国
J06642	神々　神まつり
J06752	蛇神　竜蛇　蛇信仰　世界蛇　祖霊
J06982	神々
J08682	謙譲語　尊敬語
J08872	断眠
J09192	現代語
J10042	虫プロ
JB7052	脳生理学
JB8462	断酒
JB8522	耐性菌
JB8572	CH　H2O
JB8722	脳機能

以上のことから一文字語についてのみ，キーワード候補の語群と組み合わせて作成した複合語が高頻度で出現しているかを確認する処理過程を設けた．このとき複合語形成にかかわった一文字語は一律にではなく基準を設定して除去の有無を判定する．これはたとえば一文字語「蛇」が2位の語に6倍以上の差をつけて最頻出であるJ06752において「蛇神」「竜蛇」「蛇信仰」など辞書にない複合語を抽出できたからといって「蛇」そのものを除去してしまっては内容推測に重大な障害をもたらすからである．

こうした複合語は特殊で専門性の高い内容を扱ったテクストでキーワードに加わる傾向がみられた．

4. リストの評価

出来上がったリストをもとにストップワードを分類し，キーワードとの差異や区分基準に関して考察を行なった．「固有意味の希薄性」を基幹としつつ「品詞分類の間違い」「ジャンル文体」「分析の都合」等々の様々な個別事由も確認できた．しかしストップワード選定に関して完全な恣意性の排除は困難であり議論の余地は残される．この章ではリストの効果および汎用可能性について検討していく．

4.1 Tag Cloudによる効果の比較

ストップワードによる効果を簡易Tag Cloud［注8］を使って確認する．40冊分すべての場合を提示するのが望ましいが，紙幅の都合上1例のみ取り上げる．

図1では出現頻度のキーワード内最大最小値差に対する比率に応じて文字サイズを算出し，最頻出語と各語間の共起頻度をもとに対数尤度比を計算して4段階の濃淡をつけた．語の並びは中心から左回りで外にいくほど出現頻度の低い語となっている．

同じテクストについてストップワードリストを空にした場合は図2の結果となる．「する」が中心にあるのはサ変動詞が多く使われていることを示している．文体の特徴についてはある程度推察できる材料とはなるが，内容推測については間接的な情報しかもたらさない．その情報についても「する」は40冊全部で最頻出であるので「日本語」か「論説・評論文」

図1　J01052でのキーワード抽出結果

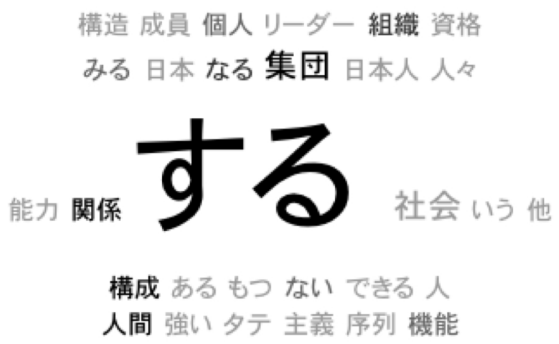


図2　J01052での低内容語を含む抽出結果

の特徴としかいえない．濃淡から「構成する」「関係する」が多用されていることは推察できる．

品詞タグによる除去さえも無化すれば図3になる．助詞が最頻出グループを形成しており,「社会」「集団」「日本」「関係」の4語程度しか内容推測に使えそうなものは見当たらない．それでも単純な内容推測は不可能ではないがノイズが多すぎる．

ここで提示したのは1例についてのみであるが，同様のことは他のテクストにおいても概ね成り立つであろうことは想像に容易い．ストップワードによるキーワード抽出への貢献は十分だと判断できる．

4.2　登録語の遍在傾向

このリストが効果的で汎用性に富むのならば40冊分のテクストいずれにおいても登録語が多数出現しているはずである．一方，各テクストに偏在する不要語の集成にすぎないのであれば逆の結果となるだろう．前者の傾向が強ければ必要登録語数の飽和が期待可能である．つまりリストの大きさは一定数で頭打ちとなり完成され，その後はどんな未知のテクストにも対応できることになる．後者であればこの種のリスト作成にはあまり意味がないことになる．

自然言語データでは語の出現傾向はその分量によって大きな影響を受けるため規模の違うテクスト同士における比較は難しいとされる［1］．ここでは簡単のため登録語について「ほとんどすべてのテクストで高頻度」「すべてのテクストで出現するが低頻度の場合もある」「複数のテクストで高頻度だが出現しないこともある」「ひとつのテクストでだけ高頻度」という4段階で大まかに評価［注9］を下した．「遍在／偏在(非遍在)」という極性を設定して「遍在度」の高低を判定している．

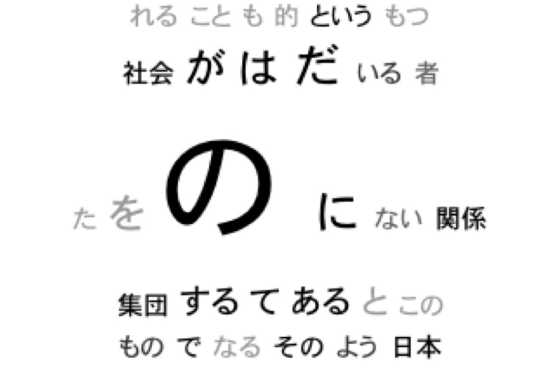


図3　J01052での非内容語までを含む抽出結果

まず「形容詞」は表5の通りである．表記の揺れがあるものについては漢字表記だけを示したが評価は全表記込みとした．揺れについていうとJ08872のみで「ながい」60回に対して「長い」1回［注10］となったが，他では「ながい」がほぼ0だった．「良い」は漢字混じりが他のふたつに対して少なく「よい」と「いい」は全体を通して拮抗していた．これらの各表記は排他的に用いられる傾向があり，書き手の個人文体に関する特徴と考えられる．

「良い」はほとんどのテクストで高出現頻度だったがJ03692でだけ全く出現しなかった．クロス集計表からは表記だけでなく同義表現の排他的使用という傾向も見受けられることから，現時点では発見できていないもののJ03692では「良い」を使わないで同様の内容を表す何らかの代替手段が用いられている可能性が指摘できる．

遍在度については「低」に該当する語の存在事由について特に考察が必要である．「小さい」は宇宙について書かれたJB8732でだけ「相互作用のきわめて小さいこの粒子」のように多用される．「確率」「質量」「エントロピー」などの用語についてそれらが極小であると述べる文脈で使われており，このテクストに固有の用法である．「悪い」もJ09692が形容詞・副詞をテーマとするため「悪い」という語そのものが言及対象となる特徴的な使い方がされている．両者ではそれぞれの対義語「大きい」「良い」も連動して出現頻度が増加したことが確認できた．

これらのことから「大きい／小さい」や「良い／悪い」をストップワードとすることで対象テクストから取得すべき情報に欠損が生じる場合があることがわかる．しかし30語制限のもとでの内容推測に資するのであればここで毀損されるような情報は必須とまではいえない．「良い／悪い」は形容詞について説明する際の一例としてあげられたものに過ぎず，「相互作用」についても「小さい／大きい」ではなく「相互」「作用」が「粒子」に関係するということがわかれば十分ではないだろうか．いずれにせよ検討の余地は残ることになる．

一方,「動詞」は表6のように分類できた．遍在度「やや高」と「やや低」との差は非出現テクストがあるかどうかであり，頻度分布の総体としては後者が前者よりも高い事例もある．

表5　形容詞遍在度

高	ない
やや高	多い 強い 大きい 長い
やや低	良い 高い 少ない
低	小さい 悪い

表6　動詞遍在度

高	ある　なる　見る　言う　思う　考える　する　出来る　持つ
やや高	いる　来る　得る　わかる　知る　書く　よる　行う　出る　入る　使う　作る　つく　つける
やや低	行く　とる　見える　しれる　いえる　示す　述べる　やる　異なる　違う　用いる　聞く　起こる
低	なす　くらべる　入れる　よぶ

「低」に分類されたものでは「なす」が地図の作成についてのJ03692で「四国が一塊をなす楕円状の一島として表わされ」のように多用される．書き手の文体的特徴や癖を表してはいるが「つくる」の言い換えと考えてキーワードからは外した．

「入れる」は化学反応について書かれているJB8572に「金属の酸化物を少量入れると」のような形で頻出する．特徴的な使用ではあるが「出る／入る」と同じく関係性について直接述べる語であることからストップワードに残した．「くらべる」（J06552）についても同様である．

「よぶ」は「これを特異点とよぶことにする」などの用例に見られる通り「いう」の言い換えとしてJB8732で使われていた．このテクストでは用語解説が多いことに起因している．

「その他」は表7である．前ふたつと比べて遍在度が低い語群であることがわかる．また「低」のうち「お／上／下」の3語が敬語について書かれたJ08682に，「名／あと」がエッセイの書き方についてのJ04102に高頻度出現という具合に特定のテクストに偏る傾向が見られる．

以上のことから，遍在度が低く個別テクストに固有の不要語がこのリストにも一定程度含まれていることは確認できた．しかしリスト全体に占める割合は大きくなく，それらの存在は少数の限られたテクストに起因することもわかった．

4.3 別データでの使用例

評価の最後として汎用性を確認するため，作成に使った40冊以外でリストを使用する．日本に関する論説・評論文ということで緩やかなテーマの共通性はあるが執筆時期が古く表記や使用語彙が今回の例と異質なものとして戦前に書かれた『最終戦争論』［注11］を選んだ．結果は図4の通りである．赤い丸で囲った3語に問題があるものの概ね良好な結果が得られた．

表7　その他遍在度

高	自分　人
やや高	次　間
やや低	かれる　わが国　人々　他　話　例　形　日　家　身　そのもの　一つ
低	つまり　お　上　下　別　手　名　あと　a

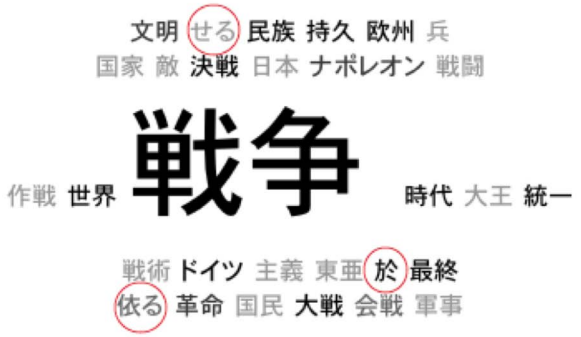


図4　最終戦争論でのキーワード抽出結果

「せる」は書き手の文体上の特徴を強く反映して残ったのだと考えられる．「ドイツを屈伏させることが怪しくなって来た」のように使役で使うのに加えて「戦争術の徹底せる進歩」のように現代語では「した」にあたる文語複合語［注12］として多用されている．後者についてMeCabは誤認識して「せる」にレンマ化してしまった．

「於」は「に於て」等が「において」と認識されず漢字部分が残ったのだと考えられる．「依る」も表記が登録語のものと違っていたため除去できていない．

表記の揺れ問題は思い付く限りのパターンをあらかじめ登録語に加えておくという方法で対処できそうである．「せる」については書き手の個人文体上の特徴と思われるので事前に予期することは難しいかもしれない．

次に『最終戦争論』と同時期に書かれた別テクストとして九鬼周造の評論･エッセイ［注13］から図5のようにキーワードを抽出した．「渋」「対」といった一文字語が目につく．前者は「甘味」の対概念であるので「低内容語」とは言い難く削除の必要はないだろう．もし「味」がキーワードに入っていれば複合語「渋味」の形にできたかもしれない．「対」は「対する」が「対」と「する」に分離されて後者が消去されてしまった形である．技術的な改善策は必要だろうが現状でも許容範囲とは思われる．

最後に現代語による翻訳の例としてとりあげたのが図6である．ヴィトゲンシュタインの『青色本』［注14］からキーワード抽出を行なった．言語論をテーマにした難解な哲学書であるため特異な語の組み合わせになっているが妥当な結果が得られた．

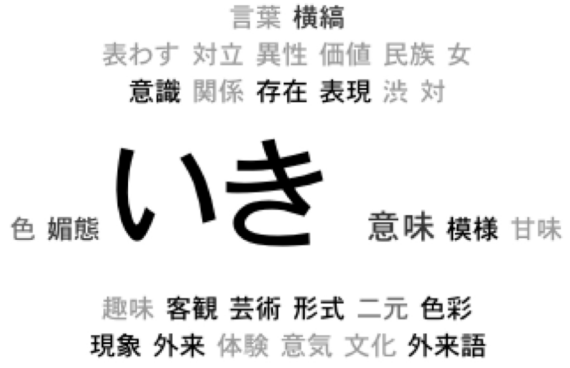


図5　九鬼周造の評論でのキーワード抽出結果

図6　青色本でのキーワード抽出結果

以上3例での使用では0から3程度の「低内容語」がキーワードに含まれていた．このことから現時点でのリストが完全な汎用性を持つとはいえないことがわかる．しかし同時にかなりの程度このリストが有効に機能することも判明した．

5. おわりに

本稿で考察したストップワードリストは内容推測に利用可能な30語のキーワードを獲得するためのものであり，現代語で書かれた新書40冊分のデータから作成した．

登録語彙数は98で表記の揺れのあるものを同一語とすれば75語である．比較するとたとえば英語におけるSMARTシステムのリスト［8］は571語でこれよりずっと多い．表層語をそのまま登録しているため規模が大きくなっているのだが，表層形を集約するタイプのものでは英語のみならず仏語などでも200から300語程度の規模に収まる［9］ようである．ストップワードを「非内容語」と「低内容語」に区分して後者のみでリストを作成したことが本稿の特色であり，これによって小さく見通しのよいリストが得られた．

リストは効果的であると同時に汎用性を持たねばならないが，汎用性は「登録語数はいずれ飽和する」という仮説に依拠している．本稿のリスト作成で登録語数の変移をJ01052からJB8732まで順に40冊分記録したところ新出現ペースの減少というかたちでこれに沿った傾向が確認できた．

図7では表記違いを一語にまとめた場合とそうでない場合との累積度数を示している．いずれの場合も新たな登録語数は減少していく傾向が見られ，特に前者で顕著であった．

「低内容語」の内訳については，文法的な事由によるものやジャンルに関するものは数も限られており比較的少量のテクスト群によっても登録語数の飽和がみられた．その反対に書き手の癖に関するものやテーマの特異性に由来するものは豊富なバリエーションを持ち，新出現登録語の排除は容易ではないと予測される．こちらをいかに効率的に処理してリストの汎用性を高めていくか，また，キーワード語数制限値を大きく変化させたときにストップワード抽出工程が被る影響についてなどは今後の課題としたい．

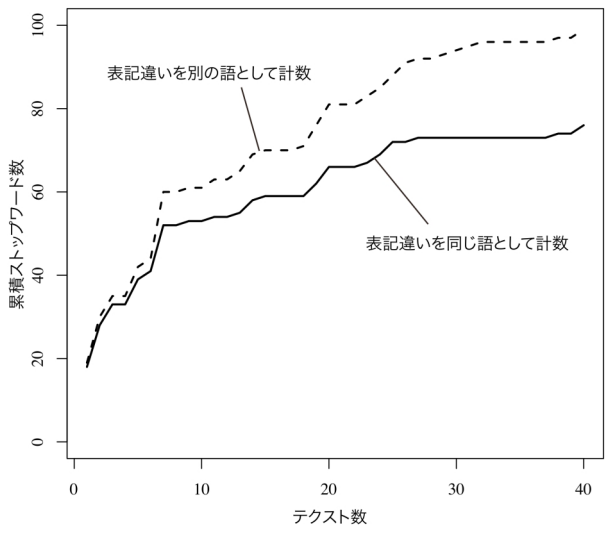


図7　低内容語の累積度数

注

［注1］ 主に社会学ではここでいうキーワードに該当するものをcodeと呼び，抽出作業をcodingという．

［注2］ http ://khc.sourceforge.net/

［注3］ Canterらの分析ではComputer CodingではなくHuman Codingが用いられている．

［注4］ http ://www.gsk.or.jp/　國府が使用許諾を得ている．

［注5］ http ://mechab.googlecode.com/svn/trunk/mecab/doc/index.html

［注6］ http ://mecab.sourceforge.net/src

［注7］ http ://gensen.dl.itc.u-tokyo.ac.jp/termextract.html

［注8］ Word Cloud，Collocation Cloudなど用途に応じて別の呼び方を使う場合もあるが，ここでは単にTag Cloudと呼ぶ．

［注9］ 高頻度の規準はキーワードに入る程度かどうかである．Okapi BM25やTF-IDF（Term Frequency-Inverse Document Frequency）など数値化する手法もあるが使用しない．

［注10］ MeCabが「短眠者も長眠者も」の「長」を「長い」にレンマタイズしたため．

［注11］ http://www.aozora.gr.jp/cards/000230/files/1154_23278.html　石原莞爾（1940）『最終戦争論』青空文庫

［注12］ 「す」の未然形と「り」の連体形からなる．

［注13］ http ://www.aozora.gr.jp/index_pages/person65.html 九鬼周造『いきの構造』『外来語の所感』『かれいの贈物』『祇園の枝垂桜』『偶然の産んだ駄洒落』『小唄のレコード』『伝統と進取』青空文庫

［注14］ http://www.geocities.jp/mickindex/wittgenstein/witt_blue_jp.html ルートヴィヒ・ウィットゲンシュタイン，ミック訳『青色本』プロジェクト杉田玄白

参考文献

［1］ Baayen, R. H. : Word Frequency Distributions, Kluwer Academic Publishers, 2001.

［2］ Baayen, R. H. : Analyzing Linguistic Data : A Practical Introduction to Statistics using R, Cambridge University Press, 2008.

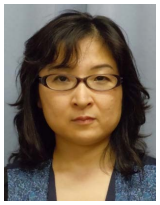
［3］ Berelson, B. : Content analysis in communication research, Hafner, 1952.

［4］ Canter, D. and Heritage, R. : A Multivariate model of sexual offence behaviour : developments in “offender profiling”, The Journal of Forensic Psychiatry, 1 (2), pp.185-212, 1990.

［5］ Efron, B. and Thisted, R. : Estimating the number of unseen

species : How many words did Shakespeare know?, Biometrika, 63, pp.435-447, 1976.
［6］ Lo, R. Tsz-Wai, He, B. and Ounis, I. : Automatically Building a Stopword List for an Information Retrieval System, Proceedings of the Fifth Dutch-Belgian Information Retrieval Workshop, 5, pp.17-24, 2005.
［7］ Neuendorf, K. A. : The Content Analysis Guidebook, Sage, 2002.
［8］ Salton, G. and McGill, M. J. : Introduction to Modern Information Retrieval, McGraw-Hill, 1983.
［9］ Savoy, J. : A Stemming Procedure and Stopword List for General French Corpora, Journal of the American Society for Information Science, 50 (10), pp.944-952, 1999.
［10］ Zipf, G. K. : The Psycho-Biology of Language, Houghton Mifflin, 1935.
［11］ 相沢佳子：英語基本動詞の豊かな世界－名詞との結合にみる意味の拡大，開拓社，1999.
［12］ 石川慎一郎，前田忠彦，山崎誠：言語研究のための統計入門，くろしお出版，2010.
［13］ 小椋秀樹，山口昌也，西川賢哉，石塚京子，木村睦子：『日本語話し言葉コーパス』の形態論情報の概要 ver.1.0，独立行政法人国立国語研究所，2004.
［14］ 國府久嗣：計算機を用いた日本語文学テクストの内容分析，国際広報メディアジャーナル，3，pp.109-132，2005.
［15］ 國府久嗣，園田勝英：コロケーションに着目した日本語テクストのメッセージ分析，自然言語処理研究会報告，178(3)，pp.15-20，2007.
［16］ 國府久嗣，山崎治子：意味推測に用いる語彙抽出数の非干渉性，自然言語処理研究会報告，207(10)，pp.1-7，2012.
［17］ 安本美典：文体統計による筆者推定－源氏物語，宇治十帖の作者について－，心理学評論，2(1)，pp.147-156，1958.

國府　久嗣（非会員）
1997年 北海道大学大学院文学研究科修士課程修了．修士（国文学）．2004年 北海道大学大学院国際広報メディア研究科博士課程前期修了．修士（国際広報メディア）．2013年 北海道大学大学院国際広報メディア・観光学院博士課程後期単位取得退学．文学理論および統計，自然言語処理等を複合した研究領域「計量テクスト論」に取り組む．



山崎　治子（非会員）
1999年 北海道大学大学院文学研究科修士課程修了．修士（国文学）．現在，岩手県立宮古高等学校教諭．放送部顧問としてドキュメンタリーやドラマ制作，朗読，アナウンスの指導を担当する．2012年より岩手県高等学校文化連盟放送専門部理事．

野坂　政司（正会員）
1977年 北海道大学大学院文学研究科単位取得退学．現在，北海道大学情報基盤センター特任教授．CALLに関連する教材・授業法開発などを中心に，デジタルコンテンツの生成，保存，提供について総合的な研究に従事．所属学会にeラーニング教育学会，情報文化学会，日本アメリカ文学会がある．