コーパスを用いたビジネスカタカナ語語彙リストの構築

黄 海洪† 金丸 敏幸‡
†京都大学大学院人間・環境学研究科 ‡京都大学国際高等教育院
E-mail: †huang.haihong.46x@st.kyoto-u.ac.jp, ‡kanamaru@hi.h.kyoto-u.ac.jp

1 はじめに

グローバル化により多くの日本企業が海外に進出し、日系企業に就職することを目的とする日本語学習者は少なくない。職業に必要な日本語を教育する「職業目的の日本語（Japanese for Occupational Purposes: JOP）」の領域ではビジネス日本語（Japanese for Business Purposes: JBP）が注目され、そのニーズが高まっている。しかし、ビジネス分野の日本語の学習ニーズが高まっているにもかかわらず、ビジネス日本語教育は必ずしも現実に適応しているとは言えない現状がある。

本研究の目的は、コミュニケーション能力の向上に不可欠とされる語彙力に着目し、上級・超上級日本語学習者の学習効率の向上に役立てられるよう、ビジネス分野のカタカナ語（ビジネスカタカナ語）の選定を行うものである。日本のビジネスの世界では、カタカナ語が多用されているが、日本語学習者にとってカタカナ語は理解が困難なものの一つである。本研究では、ビジネスコーパスを用いて、ビジネス日本語教育用カタカナ語語彙リストを作成するとともに、ビジネス分野で用いられるカタカナ語の特性を明らかにする。本研究によるビジネスカタカナ語のリストは、日本語学習者、日本語母語話者双方にとって、円滑なビジネスコミュニケーションを行うために有用なものとなる。

2 リスト構築手順

2.1 ビジネスコーパスの構築

特定目的の言語教育において、語彙選定に関する研究はコーパス言語学と密接な関係を持つ。特定分野の専門語彙を教える教師にとって、どのような語を優先して教えるべきかといった問題に対してコーパスは有効かつ強力な解決策の一つになり得る。

本研究では、ビジネス分野におけるカタカナ専門用語を「ビジネスカタカナ語」と定義する。ビジネスカタカナ語を抽出するためには、ビジネス分野に特化したコーパスを構築する必要がある。コーパスは構築の目的により、大きく一般目的コーパスと特殊目的コーパスに分類される。特殊目的コーパスは、比較的規模が小さいため、研究や教育に利用しやすいという長所がある。以降、ビジネス分野に特化した自作コーパスを「ビジネスコーパス」と呼ぶ。

ビジネスコーパスを効率よく構築するには、言語資料の選定がきわめて重要である。しかし、ビジネス日本語の定義が明確にされていないため、ビジネスコーパスを構築するための言語資料の選定は容易ではない。

田野村[1]はインターネット上の日本語文書は膨大かつ多様で、言語研究資料として大きな価値と魅力があると述べている。カタカナ語は新陳代謝が激しい語であるため、言語資料としても更新頻度の高いものを参照することが望ましい。ウェブページは出版物に比べ、常に最新の用語が提供されるという点において、新規性の高いビジネスカタカナ語を抽出するのに理想的な言語リソースである。そこで、本研究ではインターネット上の日本語ウェブページをビジネスコーパスのデータとして使用することにした。

ビジネスコーパスの構築には、コーパス自動作成サービスである WebBootCat [2]を利用した。WebBootCat によるコーパス作成の手順は以下のとおりである。

(1) コーパスの名前をつける
(2) コーパスの言語を選択する
(3) コーパスのデータの規定方法を「シードワード」、「URLS」、「ウェブサイト」の 3 つから選択する

「シードワード」オプションとは、必要なデータを抽出するために用いる内容を規定する単語リストのことである。「URLS」オプションを使用すると、指定したウェブアドレスからテキストをダウンロードできる。「ウェブサイト」オプションは、ユーザーが入力したウェブサイト全体のコンテンツをダウンロードできる。本研究では、「シードワード」オプションを利用することにした。

「シードワード」には、ビジネス用語研究会[3]が新聞、雑誌などの資料から厳選した「常識として知っておきたいビジネス用語」から 33 語を選択した。表 1 に本研究で用いた WebBootCat 用の「シードワード」を示す。

表 1 シードワード 33 語

フロー	ソリューション	タスク
アーカイブ	レスポンス	イノベーション
チュートリアル	ユーティリティ	バーター
フェーズ	マター	コミット
インセンティブ	エグゼクティブ	レジュメ
シナジー	コンシューマー	エビデンス
コンテンポラリー	ストラテジー	サマリー
インタラクティブ	ライフハック	トレード・オフ
アジェンダ	ディスクロージャー	ロイヤルティ
リレーションシップ	エクスキューズ	パースペクティブ
スノッブ	コモンセンス	アカウンタビリティー

WebBootCat による URL の収集は、Bing 検索 API を利用している。コーパス用の URL の収集に際しては、1 タプル（tuple）に 3 語のシードワードが使用される。また、1 タプルで収集する URL は 10 個であり、シードワードの 3 語が同時に出現するウェブページの URL を一度に取得する。このプロセスを 100 回繰り返すことによって、シー

ドワードを含んだウェブページを収集する。続いて、収集したページの確認、コーパスのコンパイル、データクリーンなどの作業を実施し、同様の作業を 3 回繰り返す。最終的に、延べ語数 2,614,428 語、ファイル数 1,180 個のビジネスコーパスを作成した。

構築したビジネスコーパスの専門性については、KH Coder [4]を用いて確認した。KH Coder は、計量テキスト分析やテキストマイニングのためのフリーソフトウェアである。KH Coder の多次元尺度構成法を利用し、コーパスの内容を確認した。多次元尺度構成法は分析対象のコーパスに、どのような言葉が多く出現し、どの言葉と一緒に使用されているかを確認することができる。本研究で得られたビジネスコーパスの多次元尺度構成法の結果を図 1 に示す。

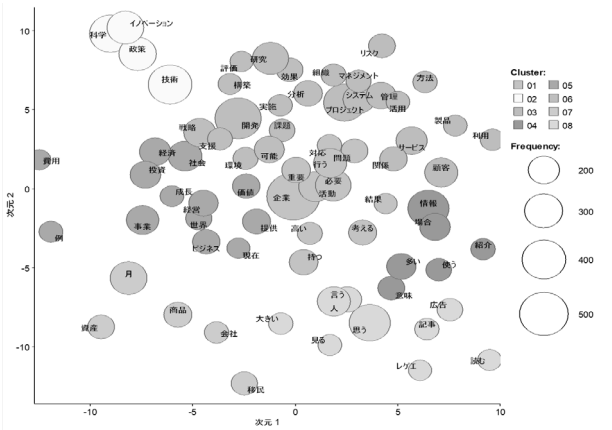

図 1　多次元尺度構成法による抽出結果

図 1 を見ると、中心に「企業」という語が出現している。その周囲には、「経営」、「ビジネス」、「商品」、「会社」、「経済」、「サービス」、「顧客」など、ビジネスに関連する単語が多く出現している。以上の結果により、今回得られたビジネスコーパスが専門性を有していると判断した。

2.2 ビジネスカタカナ語の抽出

寺嶋[5]は JSP 語彙の選定において、特定目的のもとに構築されたコーパスと大規模均衡コーパスである『現代日本語書き言葉均衡コーパス（以下：BCCWJ）』を用いれば、統計的に偏りがある特徴的な語を抽出できると述べている。中條・内山[6]は、TOEIC 模擬試験における単語の出現頻度と British National Corpus における単語の出現頻度を統計的に比較することによって、TOEIC の試験に特徴的に出現する語彙を抽出することに成功した。本研究も中條・内山に従い、BCCWJ における単語の出現頻度とビジネスコーパスにおける単語の出現頻度を統計的に比較し、ビジネスカタカナ語を抽出するという手法を用いる。具体的な抽出方法のフローチャートを図 2 に示す。

最初に、2.1 で述べたようにウェブコーパス構築ツールの WebBootCat を利用して、ビジネスウェブコーパスを構築する。次に、BCCWJ を基準コーパスとして、統計指標を計算し、リスト候補の語を抽出する。ここで抽出した語を対象として、日本語母語話者のビジネスパーソン 6 名による判定を行う。本研究では、ビジネスパーソン 6 名のうち、2 名以上がビジネス専門用語と判断した語をビジネスカタカナ語として認定する。最後に、認定された語をビジネスカタカナ語リストとしてまとめる。

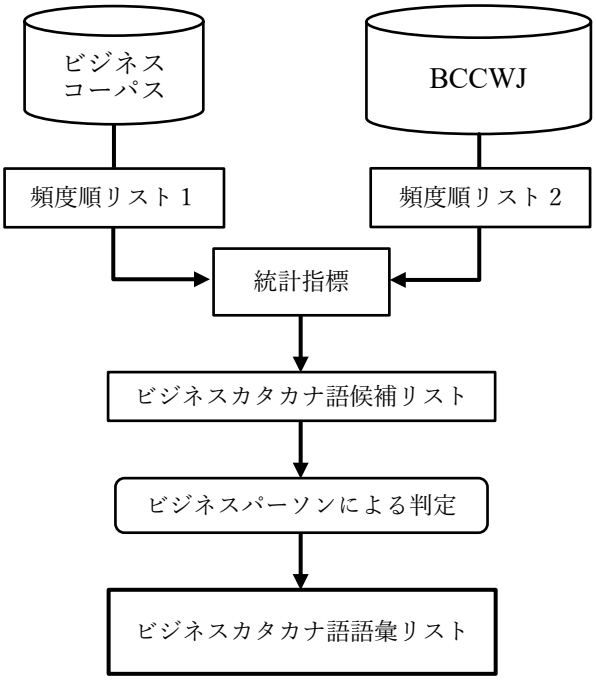

図 2　ビジネスカタカナ語の抽出フロー

ビジネスカタカナ語の抽出に用いる統計指標には、対数尤度比[7]（Log Likelihood Ratio: LLR）を用いた。LLR は、コーパス言語学などで特徴語を取り出すために用いられることの多い指標であり、下記の式(1)で求められる。

$$LLR = a\ log\ \frac{aN}{(a+b)(a+c)} + b\ log\ \frac{bN}{(a+b)(b+d)} + c\ log\ \frac{cN}{(a+c)(c+d)} + d\ log\ \frac{dN}{(b+d)(c+d)} \quad (1)$$

式(1)における各変数は以下のとおりである。

a : ビジネスコーパスの単語の出現頻度
b : BCCWJでの単語の出現頻度
c : ビジネスコーパスの延べ語数－a
d : BCCWJ の延べ語数－b
$N = a+b+c+d$

LLR を計算した後、ビジネスカタカナ語候補リストからビジネスパーソンの判定を経て、最終的なビジネスカタカナ語の語彙を選定する。選定の基本方針として、以下の 3 点を用いた。

(1) LLR の値の大きい順から採否を決める。
(2) 最終語数は、日本人成人の理解語彙数を参考にする。
(3) ビジネスパーソンが知っている語を採用する。

リスト作成にあたり、LLR の値が 5 より小さい語については最初に除外した。次に、「日本語を読むための語彙データベース(研究用)」[1]に含まれる固有名詞、フィラー、想定既知語彙、記号その他を除外した。最後に、以下のリストに含まれるカタカナ語を「一般カタカナ語」と定義し、これらの語についても除外した。

(1) 文化庁『外国人のための基本語用例辞典』[8]
(2) 国立国語研究所『日本語教育基本語彙七種比較対照表』[9]
(3) 国立国語研究所『日本語教育のための基本語彙調査』[10]
(4) 野本「簡約日本語」[11]
(5) 澤田「日本語教育のための基本外来語」[12]
(6) 国際交流基金『日本語能力試験出題基準』[13]
(7) 井上「カタカナ語（外来語）基本語彙 550 語」[14]

ビジネス日本語学習者を考えると、日本企業に就職する学習者の多くは、日本語能力試験 N1 相当の日本語運用能力を有していると想定される。N1 は旧日本語能力試験の 1 級に相当するレベルである。旧日本語能力試験の認定基準によると、1 級レベルの学習者の習得語彙数は 1 万語程度であるとされている。それに対して、日本人の理解語彙量の測定に関する調査である佐藤他[15]によると、日本人大学生の理解語彙量については 4 万語前後と推計されている。また、頻度順 5 万語までの語彙テストであっても高得点者が多いという結果が示されている。

したがって、上級ビジネス日本語学習者と日本人成人の語彙量の差は約 4 万語と推察される。日本語学習者にとって、この 4 万語の溝を埋めることが重要となる。以上のことを考慮すると、BCCWJ の頻度順リストの 1 万番台から 5 万番台に現れるカタカナ語は、上級ビジネス日本語学習者にとって、必要性が一番高い語彙であると考えられる。BCCWJ の頻度順ランクの 1 万〜5 万から抽出した結果、275 語のカタカナ語候補リストが抽出された。

ビジネスカタカナ語候補リストの中には、先行研究で選定されていない一般語(例:バスタブ、ロケーションなど)が依然として含まれている。これらのリストから、ビジネスカタカナ語として認定できる語を抽出するには、人手による判断が必要不可欠である。

投野他[16]は、「最新の大規模データに基づく頻度と、専門家による選定との両方が備わることで初めて十分な内容の語彙表が作られることになる」と主張している。本研究では、ビジネス経験を 10 年以上有する、職種の異なるビジネスパーソン 6 名の協力を得て、「ビジネス専門用語」と「一般語」の判別を行った。

以上の結果、最終的に 168 語をビジネスカタカナ語として判定した。この中で、ビジネスパーソン 6 名全員がビジネス専門用語と判断したのは、「エビデンス」、「インセンティブ」、「コンサルティング」、「デベロッパー」のわずか 4 語のみであった。4 語の中で LLR の値が一番高かったのは「エビデンス」であった。

3 考察

3.1 ビジネスカタカナ語の言語的特徴

一般的に、単語は長さが長ければ長いほど、学習しにくくなる。本研究で選定したビジネスカタカナ語 168 語の平均文字数は 5.77 であった。一方、日本語能力試験の出題基準におけるカタカナ語の平均文字数は 4.08 である。このように、ビジネスカタカナ語は全体的に文字数が多いため、語の後半部分を省略して用いる場合は少なくない。たとえば、「デフォルト」を「デフォ」としたり、「プレゼンテーション」を「プレゼン」と省略して用いたりする傾向が見られる。

次に、ビジネスカタカナ語の品詞的特徴について述べる。品詞別に見ると、「名詞-普通名詞-一般」、「名詞-普通名詞-サ変可能」、「形状詞-一般」、「名詞-普通名詞-形状詞可能」の 4 種類であった。それぞれの出現数は、137（81.55%）、23（13.69%）、6（3.57%）、2（1.19%）である。ビジネスカタカナ語の品詞構成割合を図 3 に示す。

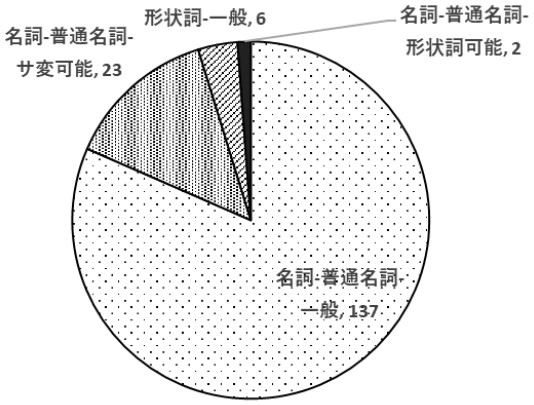

図 3　ビジネスカタカナ語の品詞構成

図 3 にあるように、圧倒的に多い品詞は、「名詞-普通名詞-一般」であり、全体の約 8 割を占めている。また、ビジネスカタカナ語には、「エビデンス」、「イノベーション」、「インセンティブ」といった抽象概念をあらわす名詞が多く含まれる。

図 4 は日本語の難易度測定システムである「帯 3」[2]を利用し、ビジネスカタカナ語に含まれる語の難易度スケールを測定したものである。「帯 3」で出力される難易度は 1 から 9 まで 9 段階である。難易度 1 が最も易しく、難易度 9 が最も難しい。図 4 から、難易度 5 以上のビジネスカタカナ語が全体の約 8 割を占めている。その中でも、難易度の最頻値は 9 であり、難易度 8 と 9 の単語の総数が半分以上に達することから、ビジネスカタカナ語は日本語母語話者にとっても難易度の高い語であることがわかる。当然のことながら、日本語学習者にとっては、さらに学習や理解が難しい語であると考えられる。

[1] 日本語を読むための語彙データベース（研究用）(http://www17408ui.sakura.ne.jp/tatsum/database.html)

[2] 帯 3 (http://kotoba.nuee.nagoya-u.ac.jp/sc/obi3/)

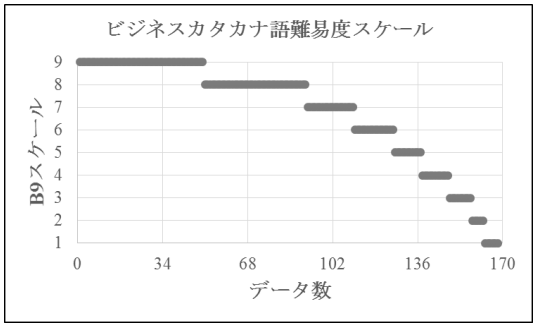

図4　帯3によるビジネスカタカナ語の難易度スケール

3.2 既存の語彙リストとの比較

本研究で得られたビジネスカタカナ語と、既存の2つのビジネス日本語語彙リストに含まれるカタカナ語とを比較し、語の重複について調査した。

(1) JALアカデミー『ビジネス日本語用例辞典』[17]（カタカナ語253語）
(2) KIT教材開発グループ『ビジネス用語集改訂版』[18]（カタカナ語162語）

リスト(1)に含まれていた語は以下の13語であった。

レジュメ、コンセンサス、レスポンス、フィードバック、プレゼンテーション、アウトソーシング、タイト、エキスパート、カスタマイズ、ポテンシャル、リニューアル、シビア、デベロッパー

リスト(2)に含まれていた語は以下の11語である。

イノベーション、シナジー、コンプライアンス、ステークホルダー、デフォルト、ブローカー、ポートフォリオ、ディスクロージャー、アウトソーシング、アパレル、アナリスト

以上の結果より、既存のビジネス用語集を使用したとしても15%程度のビジネスカタカナ語しか学習できないことが明らかとなった。

4 おわりに

本研究は、WebBootCatを利用し、約260万語からなるビジネスウェブコーパスを構築した上で、BCCWJを基準コーパスとして、統計指標LLRを計算し、ビジネスカタカナ語のリストを構築しようとしたものである。LLRの値とBCCWJの出現頻度を基準に、275語のリスト候補を抽出し、ビジネスパーソンによる判定を行って、最終的に168語からなるビジネスカタカナ語リストを構築した。

本研究により、ビジネス分野に多用されるカタカナ語の特徴が明らかとなった。形態的に見ると、ビジネスカタカナ語は文字数が多く、省略して使われるものも存在する。意味的に見ると、抽象的な概念を表す名詞の割合が高く、日本語として難解な語が多く含まれる。また、既存のビジネス日本語学習用語彙と重複する語は少なく、既存の語彙リストだけでは学習が難しいことも明らかとなった。

本研究で選定したビジネスカタカナ語リストは、今後、ビジネス日本語教育における語彙学習や教材開発などで活用することを目指す。

謝辞

本研究にあたり、公益財団法人日本漢字能力検定協会から、ビジネス日本語に関する貴重なデータをご提供いただきました。この場をお借りして厚く御礼申し上げます。

参考文献

[1] 田野村忠温 “BCCWJ に収められた新種の言語資料の特性について：データ重複の諸相とコーパス使用上の注意点，” 待兼山論叢.文化動態論篇， vol.46, pp.59-83, 2012.
[2] M. Baroni, A. Kilgarriff, J. Pomikalek, and P. Rychly, “WebBootCat: instant domain-specific corpora to support human translators,” Proceedings of EAMT， pp.247-252, 2006.
[3] ビジネス用語研究会, 知っているようで知らないビジネス用語辞典, 水王舎, 2018.
[4] 樋口耕一, 社会調査のための計量テキスト分析－内容分析の継承と発展を目指して－, ナカニシヤ出版, 2014.
[5] 寺嶋弘道, “日本語教育語彙を選定するための統計的指標－尤度比検定、カイ 2 乗検定、イエーツの補正公式の特徴－，” Polyglossia : the Asia-Pacific's voice in language and language teaching, vol.17， pp.71-83, 2009.
[6] 中條清美, 内山将夫, “統計的指標を利用した特徴語抽出に関する研究，” 関東甲信越英語教育学会研究紀要, vol.18， pp.99-108, 2004.
[7] T. Dunning, “Accurate methods for the statistics of surprise and coincidence,” Computational Linguistics, vol.19， pp.61-74, 1993.
[8] 文化庁, 外国人のための基本語用例辞典第二版, 大蔵省印刷局, 1975.
[9] 国立国語研究所, 日本語教育基本語彙七種比較対照表, 大蔵省印刷局, 1982.
[10] 国立国語研究所, 日本語教育のための基本語彙調査（国立国語研究所報告 78），秀英出版, 1984.
[11] 野本菊雄,“簡約日本語，”文林, vol.26， pp.1-36, 1992.
[12] 澤田田津子, “日本語教育のための基本外来語について，” 奈良教育大学紀要, vol.42， pp.225-239, 1993.
[13] 国際交流基金, 財団法人日本国際教育協会, 日本語能力試験出題基準, 凡人社, 1994.
[14] 井上道雄, “カタカナ語（外来語）基本語彙 550 語：その語彙特性と選定基準，” 神戸山手大学紀要, vol.6, pp.65-79, 2004.
[15] 佐藤尚子, 田中ますみ, 橋本美香, 松下達彦, 笹尾洋介, “使用頻度に基づく日本語語彙サイズテストの開発：50,000 語レベルまでの測定の試み，” 千葉大学国際教養学研究, vol.1， pp.15-25, 2017.
[16] 投野由紀夫, 加藤晴子, 小木曽智信, “学習語彙リスト作成の技法:日中英の視点から,” 応用言語学研究, vol.5， pp.43-59, 2003.
[17] JALアカデミー, ビジネス日本語用例辞典:英中韓対訳付き, アスク出版, 2008.
[18] KIT 教材開発グループ, ビジネス用語集改訂版, 凡人社, 2009.