コーパスに基づく日本語教育語彙リストの開発
-汎用性の高い日本語教育基本語彙6000語-

東京外国語大学大学院・Showa Boston Institute

本田ゆかり

要旨

本研究は、コーパスと統計に基づき「汎用性を重視した日本語教育読解語彙リスト」 (Version 2.00) (以下、「6000語リスト」) を開発した。これまで日本語教育では様々な語彙リストが作られている。従来、日本語教育語彙リストの開発は、小規模な語彙調査等を参考にして専門家の判断で語彙を選ぶのが一般的な方法であった。その代表的な例に旧日本語能力試験の『日本語能力試験出題基準』 (国際交流基金・日本国際教育支援協会 2002，以下，「出題基準」) がある。一方、近年はコーパスに基づく語彙リストの開発が進んだ。本田 (2019) ではコーパスと統計に基づく1万語の日本語教育リストを作成したが (以下、「1万語リスト」) 、本研究ではそれを日本語教育で利用しやすい形に編集、精緻化して語彙リストを作成し、6000語の基本語彙を選定した。また、この語彙リストに基づく語彙の汎用性レベルをチェックするツールをWeb上に公開した。

キーワード

語彙リスト　基本語彙　コーパス　日中同型語　日本語学習支援ツール

1. はじめに

本章では、「6000語リスト」開発の背景と目的について説明する。

1.1. 本語彙リストの開発の背景

語彙リストには「日本語能力試験出題のための語彙リスト」や「生活に必要な基本語彙」等それぞれ作成の目的があり、それによって選ばれる語彙も違う。日本語教育語彙リストはこれまでも数多く作られてきたが、従来このような個々の目的に沿って専門家が語彙項目を判断または判定し選ぶ方法が一般的だった。しかし、この方法は選定者の主観が語彙リストに影響するため、選定者が異なれば選ばれる語彙も異なる。

一方、コーパス準拠の語彙リストは語彙の頻度や分布に基づくので語彙の選定方法は客観的だが、コーパスサンプリングが選ばれる語彙にそのまま反映するため、どのようなコーパスに基づくかが重要である。

これを踏まえ、本田 (2019) では「現代日本語書き言葉均衡コーパス」(国立国語研究所 2011，以下、BCCWJ) に基づき 1 万語の日本語教育語彙リストを作成した (以下、「1万語リスト」) 。開発においてはBCCWJをそのまま利用するのではなく、BCCWJを

構成する媒体のバランスを検討し日本語教育向けに調整した上で語彙を選定した。そして、語彙リストを評価するためテキストカバー率調査を行ったところ、「１万語リスト」の語彙は、様々なジャンルのテキストで高いカバー率が示された。

しかし、これには課題も残った。BCCWJには日本語教科書コーパスが含まれていない。教科書は初級や中級レベルの日本語学習者が読む主な日本語テキストなので、日本語教育における語彙の重要度や汎用性を考える際には教科書も含めて分析したほうがいい。

語の単位の問題もある。「1万語リスト」の語彙項目は「茶まめ」の解析結果をもとにした短単位であるため、日本語教育における実用を考えると複合語に不足がある。

また「1万語リスト」では、頻度と分布からランク付けをすると別レベルに分かれやすい数字や対概念語（曜日や時間等）も除外していた。しかし、日本語教育ではこれらも語彙リストに入っていたほうが使いやすい。

さらに、漢字の知識がある学習者と、ない学習者の語彙の難易度の違いもある。漢字語彙の学びやすさは、漢字圏学習者と非漢字圏学習者で異なる。このことに配慮して、語彙リストに日中同型同義語に学びやすさの目安を示すタグが付いていると、日本語の教育や学習に役立つだろう。

本研究では上記の課題にも取り組み、実用的な語彙リストを作成したい。

1.2. 「6000語リスト」開発の目的

本研究では、1.1で述べた「1万語リスト」の問題点を修正し「6000語リスト」を開発する。1万語ではなく6000語を基本語彙としたのには、以下のような理由がある。

「1万語リスト」の開発過程で、コーパスの語彙頻度と分布には上位6～7000周辺に統計的な閾値が示された。それ以降の語彙項目のランクはコーパスサンプリングによって変動する可能性が高い。これについては松下 (2016) でも言及されているほか、英語教育でも投野 (2012) が明らかにしている。言語教育的な観点からは、この6000語付近の水準が汎用性の高い基本語彙と分野別に重要な専門語彙の峻別を示していると考えられる。そこで、本研究では上位6000語を日本語教育基本語彙として示す。なお、6000以下の語彙についても、参考情報として、基本語彙に準じる語彙または専門語彙としてリストに残す。

また、本語彙リストには漢字圏学習者にとっての漢字語彙の難易度を示す日中同形同義語タグも付与する。さらに、本語彙リストに基づき語彙の汎用性レベルをチェックするWebツールの開発も併せて行う。

2. 研究方法

ここでは、本語彙リストの作成方法について説明する。

2.1. 初級日本語教科書出現語彙調査

日本語教科書に出現する語彙を調査し、「1万語リスト」の順位を調整する。初級日本語教科書出現語彙調査は、本田 (2019) の結果に基づく。本田 (2019) で対象としたのは、以下の5種類の初級総合教科書である (表1) 。

表1　初級日本語教科書出現語彙調査で対象とした教科書

1	『みんなの日本語初級 1、2』 スリーエーネットワーク (2012-2013)
2	『初級日本語 げんき』 坂野永理他 (2020)
3	『できる日本語 初級』 嶋田和子監修 (2011)
4	『まるごと入門 初級 1、初級2』 国際交流基金 (2013-2014)
5	『大学生の日本語ともだち 1、2』　東京外国語大学留学生日本語教育センター (2017)

初級学習者にとって、母語話者向けの一般的な日本語の文章を読むことは難しい。そのため、日本語教科書が主な読み物になることが多く、初級学習者にとっては日本語教科書が読解テキストとして中級以上の学習者よりも重要であると考えられる。そこで、ここでは初級日本語教科書の語彙に焦点を当てる。

2.2. 「茶まめ」の語の単位と日本語教育語彙リスト

「1万語リスト」では、コーパスを「茶まめ」で形態素解析し、「語彙素」列を利用して語彙リストを作成したため、リストの語彙項目は短単位である。そのため、日本語教育では複合語として扱われる語彙が、リストでは2項目以上に分割されている場合がある。例えば、「積極的な」は「積極」と「的」に分かれる。

また、「語彙素」では同音の異表記がまとめられていることがあり、日本語教育では別の語として扱われる項目が1つにまとめられていることがある。例えば、「上がる」と「挙がる」はどちらも「上がる」にまとめられる。

「1万語リスト」では、このような複合語や異表記に対して特に補足説明はせず、「語彙素」列の表記をそのまま使っていた。しかし、日本語教育での実用性を考え、本語彙リストではこれらの項目を追記する。

2.3. 日中同型語

漢字圏学習者にとって、形も意味も同じである日中同型同義語は学習しやすい。語彙リ

ストの日中同型語にタグが付いていれば、日本語の学習や教育における利用に役立つ情報になる。

日中同型同義語には、形も意味も完全一致するものから、形は同じでも意味は部分的にしか一致しないものもある。部分一致の中でも一致の範囲に違いがあり、類推しやすいものも、しにくいものもある。部分一致項目は誤用が出やすく、同型異議語と比べてどちらが習得しやすいかは一概には言えない。また、日中同型語ではないが、漢字の知識があれば意味を類推しやすい日本語の漢字語彙もある (表2) 。

表2 日中同型語の類推しやすさと同型語でなくても類推できる漢字語彙

種別	意味の一致	類推しやすさ	中国語 ：「意味の和訳」
同型同義	完全一致	類推しやすい	結婚 ：「結婚」
同型同義	部分一致	類推しやすい～類推しにくい	正直 ：「まじめ」
同型異義	不一致	類推しにくい	同伴 ：「仲間」
該当なし		類推しやすい	医院 ：「病院」

日中同型同義語の分類については、松下達彦・陳夢夏・王雪竹・陳林柯 (2017) 「日中対照漢字語データベース」(Version 2.00) に詳しい。本研究ではこのデータを参照し、中国語を母語とする日本語教育専門家の協力を得て日中同型語タグを付与する。

2.4. 語彙リストの順位調整

「1万語リスト」は、コーパス頻度と分布統計 (*DP*: Gries 2008) 基づいて語彙を選定し、天野成昭他 (2000)『日本語の語彙特性　(第 1 期) 』の単語親密度を参照して日本語教育的な観点から順位調整を行った。この方法については、本田 (2019) で詳しく述べている。

「1万語リスト」ではリストのランク順に語彙のレベル分けも行った。旧日本語能力試験の「出題基準」においては、1級：10000語、2級：6000語、3級：1500語、4級：800語とされていたが、2級と3級の間隔が大きく、改定後はN1～N5の5レベルに分けられている。

英語教育では 、語彙知識のレベルを測定する“Vocabulary Levels Test” (Nation, 1983; Schmitt & Clapham, 2001; Beglar & Hunt, 1999) や “Vocabulary Size Test” (Nation & Beglar,2007) が開発されている。その語彙リストでは1000語 (word family) 区切りに語彙レベルが提示されている。そこで、本田 (2019) の「1万語リスト」では、英語教育の分け方を参考にしつつ、日本語教育における学習語彙数とレベル感を踏まえて、10000語を2000語 (初級程度) 、4000語 (初中級程度) 、6000語 (中級程度) 、8000語 (中級後半程度) 、10000語 (上級程度) のように2000語区切りで5レベルに分けて提示した。

「6000語リスト」では、「1万語リスト」に入っていなかった語彙項目 (複合語、異表記、対概念語、数詞、等) を追加したうえで、「初級教科書出現語彙調査」の結果に基づいて「1万語リスト」を順位調整して再配列し、上位6000語を基本語彙として選定する。た

だし、6000語以下の語彙も参考情報としてリストに残す。また、先行研究や日本語教育での実用性を考慮し、汎用性を軸とした語彙のレベル分けを行う。

3.結果

ここでは、第2章の方法に沿って作成した語彙リストについて説明する。

3.1. 語彙リストについて

「6000語リスト」には「1万語リスト」に入っていなかった語彙項目 (複合語、異表記、対概念語、数詞、等) を入れた。また、本田 (2019) の初級日本語教科書出現語彙調査の結果を踏まえ、出現の多かった語彙のランクを上げる等のレベル調整を行った。最後に、上位6000語を基本語彙として選定した。

3.1.1. 語彙リストに付いているタグと追加情報について

図1は「6000語リスト」の一部である。「見出し語」は、原則的に常用漢字表記に基づき、「読み」にその読み方を平仮名で示した。「見出し語」には、図1の「私【わたし／わたくし】」や「感ずる【感ずる／感じる】」のように情報を追記している項目がある。これは「茶まめ」の解析において「語彙素」にまとめられた項目のうち、日本語教育では別にしたほうがよい場合があるものを追記したものである。

見出し語	読み	初級教科書	日中同型語	出題基準・級	総頻度	分布統計（DP）	語彙素	読み	語種	品詞1
意味	いみ	◎		4	36046	0.35	意味	イミ	漢	名詞-普
続く	つづく	○		3	18900	0.330	続く	ツヅク	和	動詞-非
話	はなし	◎		4	45208	0.36	話	ハナシ	和	名詞-普
やる	やる	◎		4	82957	0.38	遣る	ヤル	和	動詞-非
私【わたし／わたくし】	わたくし・わたし	○		4	214686	0.42	私	ワタクシ	代名詞	和
違う	ちがう	◎		4	28965	0.35	違う	チガウ	和	動詞-一
生活	せいかつ	◎	◎	3	36791	0.36	生活	セイカツ	漢	名詞-普
程度	ていど		◎	2	18389	0.34	程度	テイド	漢	名詞-普
方	ほう	◎		4	85086	0.39	方	ホウ	漢	名詞-普
少し	すこし	◎		4	33425	0.36	少し	スコシ	和	副詞
もう	もう	◎		4	57550	0.38	もう	モウ	和	副詞
ある	ある	○		2	20357	0.34	或る	アル	和	連体詞
実際	じっさい			2	18531	0.34	実際	ジッサイ	漢	名詞-普
もらう	もらう	◎		3	38649	0.37	貰う	モラウ	和	動詞-非
使う	つかう	◎		4	64845	0.39	使う	ツカウ	和	動詞-一
残る	のこる	○		3	14943	0.33	残る	ノコル	和	動詞-一
与える	あたえる			2	19294	0.35	与える	アタエル	和	動詞-一
合わせる	あわせる			2	13516	0.34	合わせる	アワセル	和	動詞-一
既に	すでに			2	18527	0.35	既に	スデニ	和	副詞
感じる【感じる／感ずる】	かんずる			2	27045	0.36	感ずる	カンズル	混	動詞-一

図1　「6000語リスト」

「初級教科書」は「初級日本語教科書出現語彙調査」(本田 2019) の結果に基づき、以下のような基準でタグ付けをした。

◎：5～4種類の教科書に出現

○：3～2種類の教科書に出現

△：1～2種類の教科書に出現

「日中同型語」は、日本語と中国語の意味が完全一致する類推しやすいものだけに◎を記入した。2.3の通り、日中同型語のうち意味が不完全一致するケースには様々なパターンがあり、どの項目が学習しやすいのかは簡単に判断できない。これらは誤用を生みやすく、同型異義語より習得しにくい場合もある。この問題は日中同型語の習得研究が必要であり、本研究を超えたテーマである。そこで、本語彙リストでは、このような判断が難しいケースについては触れずに、利用者の使いやすさを優先的に考え、漢字圏学習者にとって習得が容易であることが明らかな項目に限り◎を付けることとした。これは、6000語中1327項目あった。

その他、本語彙リストには参考情報として「出題基準」の級（1級～4級）、コーパス出現頻度、分布統計　*(DP)* の値を入れたほか、茶まめ解析結果の「語彙素」「品詞」「語種」等の列を残した。

3.1.2.　語彙リストのレベル分け

ここでは、語彙リストのレベル分けについて述べる。「6000語リスト」に選定した6000語も、下位ランクの部分はコーパスサンプリングによってランク移動する可能性がある。コーパス頻度も低頻度になるにしたがって差が小さくなっていくことから、ランクが下になるにつれてリストのレベル分けも広い範囲を取って区分するのが妥当である。また、「出題基準」等を見ても4級が800語、3級が1500語、2級が6000語、1級が10000語のように語数が増えていくことや、レベルが上がるにつれて理解語彙としての語彙知識の範囲が広くなることから、日本語教育的観点からも上のレベルになるにしたがってレベル間の語数の幅を大きくしたほうが実用的であると考えられる。

そこで、本語彙リストでは6000語を基本語彙として選定しているが、Webツール版では実用性を重視し、1000語、2000語、4000語、7000語のように分けて、7000までを基本語彙にかかる可能性がある範囲とし、7001語以降12000語までを基本語彙に準じる語彙、さらに12000語以降を専門語彙として示した。

3.2. 語彙の汎用性チェックツール

「6000語リスト」を実装した「読解基本語彙チェッカー」(https://basic.chuta.jp/) を開発した。これは、学習する語彙が、本語彙リストから検索して汎用性の面からどの程度重要であるのかを調べるツールである。使い方は以下の通りである。

「読解基本語彙チェッカー」の入力画面 (図2) のボックスに調べたい語彙やテキストを入力し、右側の緑のペンマークを押すと、結果の画面 (図3) に移る。

結果の画面ではボックス内の語彙は色別表示される。また、画面右側の列にレベル分けされた語彙リスト、ボックスの下には各レベルの語数とその割合が示される。

なお、本ツールからWebツール版の語彙リストと関連する論文をダウンロードすることができる (準備中)。

図2 「読解基本語彙チェッカー」入力画面

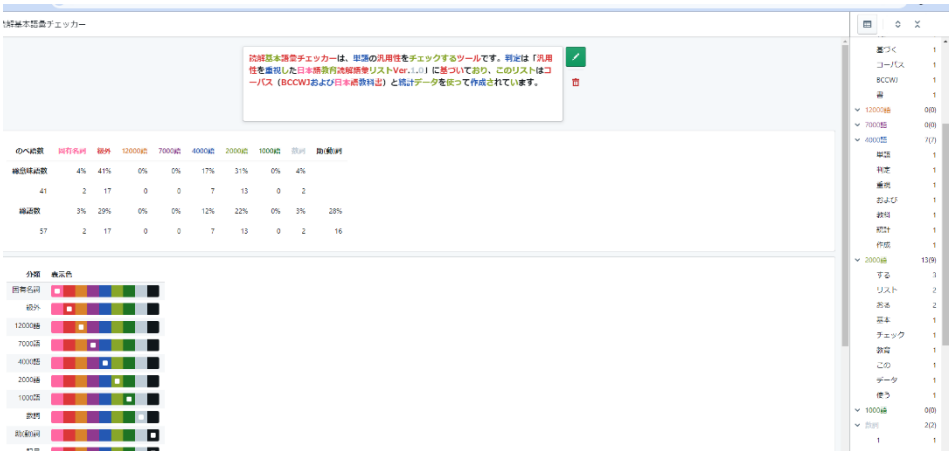

図3 「読解基本語彙チェッカー」結果表示画面

4. 今後の課題

今後は、対象とする日本語教科書を増やして研究を継続したい。今回対象とした5種類以外にも、初級レベルには『いろどり　生活の日本語』(国際交流基金 2020) や『初級日本語　とびら』(岡まゆみ他 2021) 等、利用者が多く新しい教科書が他にもある。また、初級だけでなく中級も対象にしてさらに教科書語彙調査を進めたい。

次に、本語彙リストをテキストカバー率調査によって評価する必要がある。テキストカバー率調査では、既存の語彙リストとの比較も行い、本語彙リストの特徴を明らかにしていきたい。

さらに、「出題基準」で示されている語彙の「級」が、汎用性の高さによってレベル分けを行った本語彙リストのレベルとどの程度一致しているのかを調べ、日本語教育の専門家が考える基本語彙と、汎用性の面から選ばれた基本語彙にどのような違いがあるのかを分析・考察したい。

謝辞

本研究を進めるうえで専門的なアドバイスをくださった東京外国語大学の投野由紀夫先と、Webツール開発において技術的な協力をしてくださった元東京国際大学の川村よし子先生に厚くお礼を申し上げます。

本研究はJSPS科研費 18K00708の助成を受けたものです。

引用文献

饗場淳子 (2011) 「日本語教育用語彙に共通する語についての一考察」『早稲田大学大学院教育学研究紀要』18号-2, pp.275-285.

天野成昭・近藤公久 (1999)『NTTデータベースシリーズ　日本語の語彙特性（第 1 期）』三省堂.

国際交流基金・(財) 日本語国際教育支援協会 (2006) 『日本語能力試験出題基準改定版』凡人社.

田中祐輔 (2016) 「初級総合教科書から見た語彙シラバス」山内博之監修・森篤嗣編『ニーズを踏まえた語彙シラバス』, pp. 3-31.

投野由紀夫・本田ゆかり (2016) 「第 2 章 教育語彙への応用」砂川有里子　(編) 『講座日本語コーパス 5 コーパスと日本語教育』朝倉書店, pp. 35-57.

本田ゆかり (2019) 「『初級日本語教科書共通語彙リスト』の開発」, 『日本語教育方法研究会誌』25巻2号, pp. 130-131.

本田ゆかり (2019) 「コーパスに基づく『読解基本語彙 1 万語』の選定」, 『日本語教育』172 号, pp. 118-132.

松下達彦　(2016) 「コーパス出現頻度から見た語彙シラバス」山内博之監修・森篤嗣編『ニーズを踏まえた語彙シラバス』, pp. 53-77.

Gries, S. T. (2008) Dispersions and adjusted frequencies in corpora. International Journal of Corpus Linguistics, 13 (4), 403–437

Nation, P. & Beglar, D. (2007). A vocabulary size test. The Language Teacher, 31(7), 9-13.

Tono, Y. (2013) Sampling biases and implications for better wordlist creation. Vocab @Vic conference, presentation slides. Victoria University of Wellington.