コーパスの異なりと単語親密度を活用した日本語共通基礎語彙の抽出

松田真希子，児玉茂昭，竹元勇太，石坂達也（長岡技術科学大学）
森篤嗣（国立国語研究所），川村よし子（東京国際大学），山本和英（長岡技術科学大学）
E-mail :{matsuda@vos, kodamas@kjs }+nagaokaut.ac.jp,
amori@ninjal.ac.jp, kawamura@tiu.ac.jp, {takemoto, ishisaka, yamamoto}@jnlp.org

1　はじめに

日本語共通基礎語彙の抽出は，言語処理や初中等国語教育など様々な分野で意義があるが，特に地域在住外国人に対する日本語支援においては緊急の課題である。現在日本には十分な日本語力を持たない外国人が数十万人存在しているが，あらゆる公的情報を多言語で提供するのはコスト的にも負担がかかり，支援策として最善ではない。そのため伝達に使用される日本語をわかりやすくすることが非常に重要となるが，語彙や文型等の制限に関して明確な規格化を行い，その規格に従って自動的にやさしい日本語に変換する技術はまだ開発されていない[1]。そこで，本研究ではやさしい日本語の規格化の一つとして日本語共通基礎語彙の抽出を行う。本発表ではその経過報告として，基礎語彙抽出に関する全体構想と複数の Web コーパスの出現頻度や単語親密度から抽出した基礎語彙候補語について報告する。

2　基本語彙と基礎語彙

日本語で情報の伝達を行うときに最低限必要である語彙集合を定める試みはこれまでに多く行われている。そうした語彙集合は，大きく基本語彙と基礎語彙に分けることができる。この二つは類似した概念であるが，その目的は異なる。

基本語彙は特定の領域や対象者を想定した「まず学ぶべき語」であり，ある特定の目的に特化した語彙集合であって，多くの場合，教育的観点からの選定が行われている。日本語については，初中等国語教育に基づく阪本の教育基本語彙[2]や，外国人のための日本語能力試験出題基準(JLPT)[3]，などがそれにあたる。

これに対して基礎語彙はその言語の中核をなす最低限必要不可欠な語彙集合を抽出することが重視される。これを目指したものでは，英語では Longman の”English dictionary”が有名である。

一方日本語で基礎語彙を志向した語彙表としては Ogden の”Basic English”から影響を受けて定められた土居[3]の語彙表，国立国語研究所の日本語教育基本語彙[5]などがある。また，醍醐プロジェクト[6]では既存の語彙表の情報を付与した日本語基本語彙表を公開している。しかし 4 節に示す理由により，十分に成功した基礎語彙は存在していない。

3　本研究の立場

本研究においては，第 1 節で述べたように日本語の基礎語彙抽出を目的としている。基本語彙の選定を目的としないのは以下のような理由による。基本語彙は「この語を覚えれば目標としている言語行動に役立つ」ということが前提となっているため，語彙選定にあたっては，目標言語行動との関係性が強いものを選ぶ必要が生じる。外国人に対して情報を提供する際に最低限必要な日本語の選定においても，情報提供を目的とした基本語彙を選定する必要がある。しかし，日本語の構成上最低限必要な基礎語彙が曖昧なまま基本語彙を選定すると，その意味的な網羅性と使用目的に対する妥当性を満たすことができない可能性がある。このため，まず日本語共通基礎語彙を確定させ，その上で目標言語行動に応じた語彙を追加していくことによって，網羅性と妥当性の双方を確保することにする。

4　基礎語彙抽出方法

基礎語彙の抽出にあたっては，次の 3 点を満たす語彙を抽出することが重要である。それらは，

1)　出現頻度が高いこと
2)　意味領域を十分にカバーしていること
3)　意味的な排他性が高いこと

とまとめることができる。

まず，コーパスにおける出現頻度が高い語彙は，使用されることの多い基本的な語彙であると考えることができる。次に，抽出された語彙あるいは語彙の組み合わせが日本語の語彙集合全体の意味領域をできる限りカバーしていなければ，言い換えを行う際に適切な語彙を見出すことができないという不都合が生じる。さらに，抽出される基礎

語彙の集合は，可能な限り小さいことが望ましいので，抽出された語彙集合内で言い換えが可能な語彙は可能な限り排除し，言い換えを行うことが不可能な意味的な排他性の高い語彙集合を用意する必要がある。

既存の基礎語彙リストは，意味領域の網羅性が重視されており，上に述べた被覆率などは考慮されていない。一方，基本語彙については，学習者にとっての親密度が重視されており，「万年筆」（JLPT4 級）のような文房具名など出現頻度を考慮したものとは考えられないモノ名詞が多く含まれている。

本研究では，上記の 3 点のうち，出現頻度の高さに着目し，異なるコーパスに共通して出現頻度の高い語彙を抽出する。また，基礎語彙の抽出にあたっては，地名・人名などの固有名詞や未知語は排除することとする。

5　コーパスの異なりを活用した基礎語彙選定

本研究では，まず基礎語彙の候補となる語彙を収集するために，分野の異なる 4 種のコーパスを収集し，分析に用いた。それらのコーパスの基本情報を表 1 に示す。

表 1 コーパスの基本情報

	総単語数	異なり語数#1	#2	コーパスのタイプ
Wikipedia	171,772,307	792,685	7,278	学術文文語的
Yahoo! 知恵袋	57,679,832	162,097	3,870	会話文口語的
日本経済新聞	692,754,923	315,267	3,599	報道文文語的
livedoor Blog	1,424,278,564	86,296	4,210	日記文口語的

#1 未知語，記号，固有名詞を除く

#2 上位からの累積頻度 90%越え時の単語順位

コーパスとして使用したのは，Yahoo!知恵袋，livedoor Blog，日本経済新聞, Wikipedia のそれぞれから収集した日本語文である。これら 4 種のコーパスは表 1 に示したように，日本語が使用される様々な場面をカバーしている。このため，まず，これらのコーパスに高い頻度で出現する語彙を抽出し，更にそれらの各コーパスから抽出された高頻度語彙の集合に共通する語を抽出することで，日本語が使用される様々な場面で使用可能な共通基礎語彙の集合が抽出できる。

ここで抽出する基礎語彙の集合の大きさについては，各コーパス内の出現語彙 90%以上の被覆率を持たせることを目標とした。その結果，Wikipedia 以外は 5,000 語以下で 90%をカバーしたため，各コーパスの上位 5000 語を分析対象とした。また，前述の通り，固有名詞，記号及び未知語は語彙集合から除き，「ている」，「てある」のような語連続についても，「て＋いる」，「て＋ある」のように分割した。

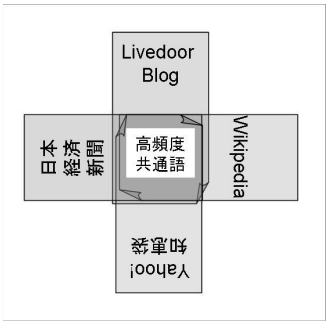

図 1　高頻度共通語抽出のイメージ

次に，このようにして抽出した共通語彙の集合の各語彙に対して，親密度判定を行った。親密度判定については，5 名の評価者が『日本語の語彙特性』[7]の方式に従い，最も親密度の高いものを 7，最も低いものを 1 として 7 段階での評価を行いリスト化した。このデータを基礎語彙構築に用いるための基礎データとして用いることとした。

6　実験結果と考察

表 1 で示したコーパスに対して形態素解析器 ChaSen を用いて分かち書き単語分割を行い，各コーパスにおいて，出現頻度の高かった語の上位 5,000 語をリスト化した。また，4 種のコーパスに現れた語彙について，各コーパスでの頻度を加算し，その頻度の総和の順に並べた場合の上位 5,000 語をリスト化した。[1]

その結果上位 100 語にあがった語を，以下の表 2 に示す。一番左の項が抽出された語彙，真中の項が品詞分類，一番右側が品詞分類の細目である。「の」が複数個所にあらわれていることからわかるように，このリストにおいては，表記が同一であっても文法的な機能などに違いがある場合には別の語彙であるとみなしてリスト化している。[2]

[1] Livedoor blog のコーパス規模が大きいため，頻度の総和ではなく，各コーパスにおける各語彙の出現頻度の割合を求め，その総和から上位 5,000 語を選出したが，結果はほぼ同じだった。

[2] このように表記においては同一であるが文法的機能

表 2　上位 100 語のリスト

の_助詞_連体化	ある_助動詞_助動詞
は_助詞_係助詞	まで_助詞_副助詞
た_助動詞_助動詞	と_助詞_接続助詞
に_助詞_格助詞	円_名詞_接尾
を_助詞_格助詞	よ_助詞_終助詞
て_助詞_接続助詞	的_名詞_接尾
が_助詞_格助詞	人_名詞_接尾
する_動詞_自立	その_連体詞_連体詞
だ_助動詞_助動詞	さん_名詞_接尾
で_助詞_格助詞	くる_動詞_非自立
と_助詞_格助詞	ない_形容詞_自立
も_助詞_係助詞	ん_助動詞_助動詞
1_名詞_数	って_助詞_格助詞
ます_助動詞_助動詞	けど_助詞_接続助詞
いる_動詞_非自立	人_名詞_一般
2_名詞_数	たい_助動詞_助動詞
ない_助動詞_助動詞	できる_動詞_自立
です_助動詞_助動詞	言う_動詞_自立
0_名詞_数	から_助詞_接続助詞
の_名詞_非自立	へ_助詞_格助詞
か_助詞_副助詞/並立助詞/終助詞	これ_名詞_代名詞
	見る_動詞_自立
3_名詞_数	という_助詞_格助詞
なる_動詞_自立	者_名詞_接尾
こと_名詞_非自立	百_名詞_数
5_名詞_数	で_助詞_接続助詞
が_助詞_接続助詞	として_助詞_格助詞
れる_動詞_接尾	それ_名詞_代名詞
から_助詞_格助詞	もの_名詞_非自立
ある_動詞_自立	だけ_助詞_副助詞
に_助詞_副詞化	られる_動詞_接尾
4_名詞_数	な_助詞_終助詞
9_名詞_数	万_名詞_数
てる_動詞_非自立	お_接頭詞_名詞接続
6_名詞_数	の_助詞_格助詞
ん_名詞_非自立	いる_動詞_自立
と_助詞_並立助詞	自分_名詞_一般
8_名詞_数	今日_名詞_副詞可能
十_名詞_数	やる_動詞_自立
7_名詞_数	行く_動詞_自立
う_助動詞_助動詞	いい_形容詞_自立
など_助詞_副助詞	しまう_動詞_非自立
日_名詞_接尾	千_名詞_数
思う_動詞_自立	し_助詞_接続助詞
年_名詞_接尾	前_名詞_副詞可能
や_助詞_並立助詞	出る_動詞_自立
この_連体詞_連体詞	今_名詞_副詞可能
ね_助詞_終助詞	せる_動詞_接尾
いう_動詞_自立	何_名詞_代名詞
よう_名詞_非自立	私_名詞_代名詞

や意味について異なる語彙の取り扱いについては，今後検討する必要があるものと思われる。

一見してわかるように，上位 100 位までに含まれている語彙には，助詞や助動詞などの機能語や，「いる，ある」などの基本的な動詞，「日，年，人」などの基本的な名詞が多く含まれている。

さらに，この上位 5,000 語に対して，第 4 節で述べた親密度の付与を行った。表 3 は，リスト化した 5,000 語の中に，表中に示した親密度以上の語がいくつあったかを示している。またそれらの語彙が，コーパス全体のどの程度の割合をカバーしているかの比率も示した。

表 3　親密度別被覆率

親密度	語数	Yahoo 知恵袋	Wikipedia	日本経済新聞	Livedoor
7	2,454	75.78%	65.74%	66.04%	75.45%
6.5	4,149	85.03%	78.41%	82.51%	84.64%
6	4,659	86.05%	80.45%	87.86%	85.90%
6 未満	5,000	87.33%	82.26%	89.68%	87.98%

表 3 から被覆率のみを取り出してグラフ化したものが図 2 である。

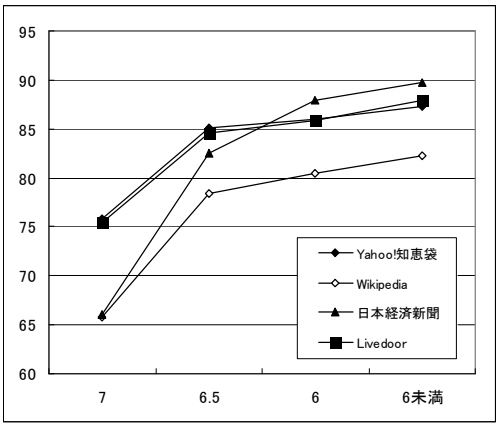

図 2　親密度別被覆率

図 2 に見られるように，Wikipedia を除くと，親密度 6.5 以上を与えられた 4,149 語は，コーパス全体の約 8 割をカバーしている。[3]

最後に，日本語能力試験の出題基準との比較を行った結果を示す。日本語能力試験の各級において出題の基準となっている語彙の総数は，それぞれ，表 4 に示すとおりである。

[3] リスト中の表記の重複する語彙を一つとみなした場合，語数はさらに減少するものと考えられる。

表 4　日本語能力試験出題基準級別語彙数

級	総語彙数	級	総語彙数
1 級	2,977	3 級	701
2 級	3,648	4 級	755

選定した 5,000 語の語彙リストと日本語能力試験出題基準の各級に含まれる語彙との重なり語彙の総数と，表 4 に示した各級の語彙総数に対する重なり語彙総数の割合を表 5 に示す。

表 5　級別重なり語彙数

	7+	割合	6+	割合	総数	割合
1 級	180	6.0%	650	21.8%	680	22.8%
2 級	814	22.3%	1,532	42.0%	1,572	43.1%
3 級	363	51.7%	474	67.6%	481	68.6%
4 級	529	70.1%	624	82.6%	646	85.6%
級外	453		1,158		1,359	

表 5 からは以下のことがわかる。

1. 日本語能力試験の級が下がるほど，重なり語彙の比率は上昇する。このことは，選定した 5,000 語の語彙リストが，より基本的で初歩的な語彙を，発展的な語彙よりもよくカバーしていることを意味している。
2. 親密度 7 以上と判定された親密度の高い語彙は，級が下がるに従って総数に対する比率が高くなる。このことは，親密度の高い語彙はより基礎的な語彙の中に多いことを意味している。
3. 級外と判定される語彙が約 2 割程度存在している。このことは日本語能力試験の出題基準語彙の集合と，本研究で選出した 5,000 語の集合との間にある程度の異なりが存在していることを意味している。

以上から判断すると，今回選定した 5,000 語の語彙リストは，日本語の基礎語彙を考察する上での出発点として使用可能であると考えられる。

7　今後の課題

今後の課題としては，まず，注 2 でも述べたように，同一の表記を用いる機能的に異なる語彙をどのように取り扱うかという問題を解決しなければならない。また逆に，表記は異なるが同じ意味をあらわす語彙については統合を行う必要がある。ただし，この場合には「同じ意味」とは何かについての考察をきちんと行う必要がある。たとえば，自立動詞の「行く」と補助動詞の「いく」は同一の意味を持っていて，統合することが可能かという問題がある。同様の問題は，他の語彙についても生じる可能性がある。また，連合して相や法をあらわす機能表現をどのように取り扱うかという問題も存在する。たとえば，「てある」，「ている」，「かもしれない」などについては，連合した形式を基礎語彙とみなす必要があるかもしれない。

さらに，モノ名詞を基礎語彙から除外すべきかどうかという点に関しても十分な議論が必要である。モノ名詞は動作性名詞や動詞や形容詞等のコトに関する語彙に比べ時代の影響を受けやすく，可変的で増加傾向にあるため，基礎語彙からできる限り除外したほうがよいかもしれないが，その一方でモノ名詞なしでは十分な情報が伝達できないという問題がある。基本レベルカテゴリーのモノ名詞を残す等，実際の言い換えを行う中で絞込みを行いたい。

今回の研究においては，第 3 節に挙げた三つの条件のうち高頻度であるかどうかのみに注目をして語彙の選定を行った。今後は言い換えコーパスを作成し，どのような語が言い換えられなかったかに関するリスト化を行い，本研究の成果をもとに，残りの高被覆率，高排他性を満たすような語彙リストを作成し，報告する予定である。

参考文献

[1]庵功雄，岩田一成，森篤嗣，「やさしい日本語」を用いた公文書の書き換え－多文化共生と日本語教育文法の接点を求めて, 2009 年度日本語教育学会秋季大会予稿集，日本語教育学会，135-140, 2009

[2] 阪本一郎，教育基本語彙，牧書店，1958

[3] 国際交流基金，日本語能力試験出題基準，凡人社, 1994.

[4] 土居光知，基礎日本語，六星館, 1933.

[5] 国立国語研究所，日本語教育のための基本語彙調査, 1984.

[6] 佐藤理史，醍醐プロジェクト，参照 URI: http://sslab.nuee.nagoya-u.ac.jp/~sato/research/daigo.html.

[7] 天野成昭・笠原 要・近藤公久，日本語の語彙特性第 4 期, NTT 出版, 2008