日本語定義語彙による語釈文の記述に向けて

野呂 智哉　　徳田 雄洋

東京工業大学 大学院情報理工学研究科

概要

あらかじめ決められた少数の語彙ですべての語を定義することは，言語学習や言語処理研究において有用である．英語では，Longman や Oxford などの一部の辞書がそのような語彙 (定義語彙) にもとづいて作成されているが，日本語ではまだ存在しない．我々は，既存の国語辞書の語釈文に使用されている語の分布にもとづき，定義語彙の候補となる語を選定した．本稿では，我々が選定した候補語と一般に「基本語彙」と呼ばれる語彙に含まれる語を比較し，その違いを考察する．さらに，定義語彙を利用した語釈文の記述方法について考える．

1　はじめに

すべての語の意味を，あらかじめ決められた少数の語 (定義語彙) で記述することは，言語教育や自然言語処理研究にとって有用である．英語では，Longman Dictionary of Conemporary English (LDOCE) [10] や Oxford Advanced Learner's Dictionary (OALD) [4] などの辞書において，2000 語から 3000 語の定義語彙ですべての語釈文を記述している．実際，これらの辞書は英語学習者用辞書として広く利用され，さらに自然言語処理研究においても利用されている．一方，日本語では，定義語彙を明確に定めた上で作成された国語辞書は存在しない．「基本語彙」に関する研究はこれまでに数多くされてきている [6]．しかし，それらは主に児童や外国人日本語学習者が知っている (べき) 語に関する研究であり，語釈文を記述するための語というよりむしろ辞書に見出し語として登録すべき語を選定するための研究である．

我々は，限られた語彙で語釈文を記述することを目的とし，既存の国語辞書の語釈文に使用されている語の分布にもとづき，日本語定義語彙の候補となる語を選定した [8]．本稿では，その候補語と一般に「基本語彙」とと呼ばれる語彙に含まれる語を比較し，その違いを考察する．さらに，定義語彙を利用した語釈文の記述方法について考える．

2　選定方法

2.1　概略

本節では，定義語彙の候補となる語の選定手法 [8] の概略を説明する．本手法は，既存の国語辞書の語釈文に使用されている語の分布にもとづく手法である．手順は以下のとおりである．

1. 国語辞書中の各見出し語について，その語釈文中に出現する語との関係を表現したグラフ (単語参照グラフ) を作成する．
2. 作成した単語参照グラフをもとに，各語にスコア付けを行う．
3. スコアの高い語を定義語彙の候補語とする．
4. 各候補語について，定義語彙に採用すべきか否かを人手で判断する．

単語参照グラフは，国語辞書中の見出し語からその語釈文中の各語に対して有向枝をひくことによってできる有向グラフである．多品詞語や同じ表記で読みが異なる語を区別するため，各ノードは漢字表記 (正書法)，読み，品詞で区別する．ただし，助詞，助動詞，数字，固有名詞，記号はグラフから除外する．

各語へのスコア付けは，単語参照グラフに対応する隣接行列の，固有値が 1 のときの固有ベクトルを計算することにより行う．これは，PageRank [9] や LexRank [1] の原理と同じである．

2.2　評価実験

データとして岩波国語辞典コーパス [3] を使用した．岩波国語辞典コーパスは，岩波国語辞典第 5 版の全データに GDA タグ [2] などの XML タグを付与したものである．GDA タグのほかに，見出し語，漢字表記 (正書法)，品詞，例文など辞書特有のタグがある．本研究では，見出し語 (hd)，漢字表記 (orth)，品詞 (pos)，語釈文 (su)，語釈文中の各形態素 (n，v，ad など) の情報を利用し，例文 (eg)，文法的解説 (gram)，対義語 (ant)，語源 (etym)，見出し参照 (sr) などの補足的な情報は利用しない．また，語釈文中のかぎ括弧 “「」”，“『』” で囲まれた文字列は引用的な表現が多く，定義語彙を考え

る上では適切ではないものが多いので，先に述べたような補助的情報を表すタグが付与されていなくても対象から除外する．

先に述べた手法で単語参照グラフを作成すると，以下のような辞書や日本語特有の問題に直面する．

同一項目中に複数の漢字表記が存在する場合: 岩波国語辞典では，動詞「ひく」の項目には「引く」,「弾く」など 8 種類の漢字表記が登録されている．このように同一項目中に複数の漢字表記がまとめられている場合，単語参照グラフにおいても同一ノードとする．ただし，品詞が異なる場合は区別する．

品詞変換: 語釈文中の各形態素には，コーパスアノテーション方針の品詞分類にもとづく品詞が付与されているが，見出し語には岩波国語辞典の編者が定める品詞分類にもとづく品詞が付与されている．この 2 種類の品詞分類の違いを吸収するため，名詞，動詞など非常に粗い品詞分類を用意し，それぞれの品詞を変換する．

形態素区切り: 語釈文中の形態素の区切り方と，見出し語として登録されている語の単位が異なる場合がある．そこで，語釈文中の連続する 2 つの名詞または動詞が 1 語で見出し語として登録されている場合，その 2 語を 1 語に結合する．

表記の異なり: 「〜すること」の「こと」は平仮名で表記することが一般的であり，語釈文中でも平仮名で表記されるが，「こと」の項目では漢字表記として「事」が登録されている．「こと」という読みの語はほかにも「琴」,「古都」などがあるため，語釈文中に「こと」が出現した場合，それがどれに該当するか判断できない．そこで，このような場合は人手でノードの統合を行った．

以上の処理を行った結果，ノード数 69,013 個のの単語参照グラフが作成できた．このグラフに対し，Erkan らと同様，べき乗法により固有ベクトルの計算を行った．random walk のための減衰係数は 0.15，許容誤差は 10^{-4} とした．

スコア付けによる上位 50 語を表 1 に示す．ただし，スコアは 1 位を 1.000 としている．結果より,「ある (動詞，連体詞)」や「する」,「もの」などのごく一般的な語だけでなく,「意」,「物事」,「転」などの語釈文特有の (表現に使われる) 語も上位に現れることがわかる．逆に「そういう」や「A」,「B」は定義語彙として不適切であるように思われる．これらは岩波国語辞典には見出し語として登録されていない，いわゆる未登録語である．語釈文中のみに出現し，見出し語として登録されていない語は，単語参照グラフにおいて自身から他の語へ向かう

表 1　上位 50 語

順位	スコア	読み	漢字表記	品詞
1	1.000	ある	有る, 在る	動詞
2	0.7023	い	意	名詞
3	0.6274	ある	或る	連体詞
4	0.5927	こと	事	名詞
5	0.5315	する	為る	動詞
6	0.3305	もの	物, 者	名詞
7	0.2400	その	其の	連体詞
8	0.2118	ほう	方	名詞
9	0.1754	たつ	立つ, 建つ	動詞
10	0.1719	また	又, 復, 亦	接続詞
11	0.1713	いる	居る, 処る	動詞
12	0.1668	ひと	人	名詞
13	0.1664	つかう	使う, 遣う	動詞
14	0.1337	いく	行く, 往く	動詞
15	0.1333	なる	成る, 為る, 生る	動詞
16	0.1324	いう	言う, 云う, 謂う	動詞
17	0.1244	ものごと	物事	名詞
18	0.1191	どう	同	連体詞
19	0.1116	それ	其れ	代名詞
20	0.1079	とき	時, 刻	名詞
21	0.1074	てき	的	接尾辞
22	0.1020	そういう	そういう	連体詞
23	0.09682	じょうたい	状態	名詞
24	0.09165	あらわす	表す, 現す, 顕す, 著す	動詞
25	0.08968	いえる	言える	動詞
26	0.08780	えー	A	名詞
27	0.08585	てん	点	名詞
28	0.08526	とくに	特に	副詞
29	0.08491	ご	語	名詞
30	0.08449	いいあらわす	言い表す	動詞
31	0.08255	または	又は	接続詞
32	0.07285	えらびとる	選び取る	動詞
33	0.07053	ばあい	場合	名詞
34	0.06975	ところ	所, 処	名詞
35	0.06920	かたち	形	名詞
36	0.06873	ない	無い	形容詞
37	0.06855	ことがら	事柄	名詞
38	0.06709	びー	B	名詞
39	0.06507	やくにたつ	役に立つ	動詞
40	0.06227	われわれ	我我	代名詞
41	0.06109	じょし	助詞	名詞
42	0.06089	いいつける	言いつける	動詞
43	0.06079	てん	転	名詞
44	0.05989	えいご	英語	名詞
45	0.05972	じぶん	自分	名詞
46	0.05888	かた	方	接尾辞
47	0.05879	ため	為	名詞
48	0.05858	かく	書く, 描く	動詞
49	0.05794	かんがえる	考える, 勘える	動詞
50	0.05530	ふくし	副詞	名詞

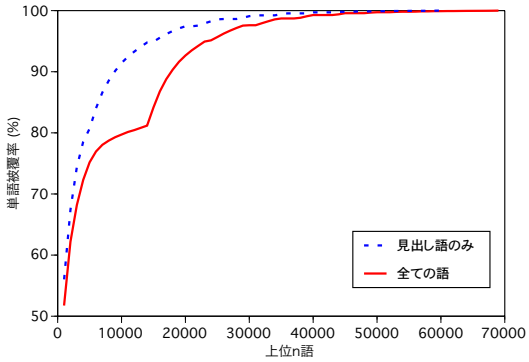

図 1　単語被覆率

枝が存在しないため，スコアが高くなる傾向にある．

今回の単語参照グラフの作成において，9,327 語が未登録語と判断された．その要因として，以下の 4 つがある．

品詞の不一致 (要因 1):　語釈文中の各形態素に付与されている品詞と見出し語に付与されている品詞の違いを吸収するため，粗い品詞分類を用意して変換しているが，それでも一致しない場合がある．

形態素区切りの不一致 (要因 2):　例えば，語釈文中では「そういう」が 1 形態素とされている一方，見出し語としては「そう」と「いう」があるだけで「そういう」が登録されていない．このような場合，未登録語と判断される．

データ不具合による除外 (要因 3):　コーパスデータの一部にフォーマットやアノテーションの不具合があり，それらは単語参照グラフを作成する前に除外した．ここで除外した見出し語はすべて未登録語と判断される．

真の未登録語 (要因 4):　例えば，語釈文中に「英語」という語が出現するが，見出し語としては登録されていない．また，「書ける」などの可能動詞も岩波国語辞典には登録されていない．これらは (我々の手法の問題ではなく) 真の未登録語である．

図 1 に単語被覆率を示す．単語被覆率とは，我々の手法による上位 n 個の単語の，全語釈文中に出現する単語に占める割合を表す．結果 (実線) より，$n = 10,000$ 前後で一度上昇が緩くなり，$n = 15,000$ あたりで再び上昇しはじめることが分かる．これは，未登録語と判断された語のすべてが上位 15,000 語の中に含まれていることが原因であると考えられる．先に述べたとおり，未登録語は語釈文中での出現頻度が低くてもスコアが高くなる傾向にあり，この結果はそれを反映している．未登録語と判断された語を除外し見出し語のみを対象とすると，上昇率は n が大きくなるにつれて次第に緩やかになっていくことが分かる．

表 2　見出し語被覆率 (%)

n	0	1	2	3	4	5+
見出し語被覆率	26.52	29.66	19.00	10.44	6.03	8.35

2.3　候補語の選定

スコア付けの結果をもとに，以下の 5,853 語を日本語定義語彙の候補語として選定した．

- 上位 15,000 語のうち，登録語と判断された語 (5,673 語)
- 上位 5,000 語のうち，要因 1 と 3 で未登録語と判断された語 (180 語)

未登録語と判断された語のすべてが上位 15,000 語の中に含まれていることを踏まえ，登録語は上位 15,000 語までを採用した．未登録語と判断された語は，実際よりスコアが高くなる傾向にあるため，上位 5,000 語までに限って本来登録語であるはずの語を人手でチェックし，180 語を候補語として採用した．

表 2 に見出し語被覆率を示す．見出し語被覆率とは，日本語定義語彙候補語に含まれない語が語釈文中に n 個出現する見出し語の，全見出し語に占める割合を表す．結果より，全見出し語の 26.52% は，その語釈文が定義語彙の候補語のみで記述されていることが分かる．つまり，この候補語をそのまま定義語彙とするならば，これらの語釈文を書き換える必要はない[*1]．

3　基本語彙との比較

提案手法によるスコア付けの結果と従来の「基本語彙」の違いを見るため，中央教育研究所と国立国語研究所がそれぞれまとめた教育用基本語彙 (それぞれ中央教育基本語彙，国語研基本語彙と呼ぶ) [7] と比較した．中央教育基本語彙は小学生を対象とし，4,323 語を選定している．国語研基本語彙は外国人のための日本語教育用として 6,060 語を選定している．

図 2 に再現率を示す．再現率とは，提案手法による上位 n 語と比較対象の基本語彙に共通して含まれる語の，比較対象の基本語彙に占める割合を表す．単語被覆率の場合と同様，$n = 10,000$ 前後の上昇が緩やかになっている．未登録語を除外し見出し語のみで比較した場合，比較対象の基本語彙のサイズと同じだけ提案手法による

[*1] たとえ定義語彙内の語のみで記述されていても，その語の使われ方によっては適切ではない場合もあるため，厳密には，すべてチェックする必要がある．詳細は第 4 節で述べる．

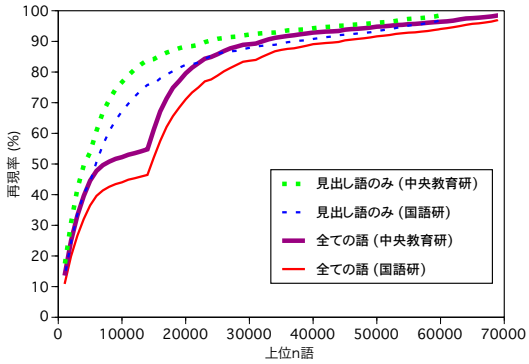

図 2　2 種類の基本語彙との比較 (再現率)

表 3　基本語彙に含まれないがスコアの高い語

順位	読み	漢字表記	品詞
51	てんじる	転じる	動詞
102	ようほう	用法	名詞
113	りゃく	略	名詞
372	ふるまい	振舞い	名詞
480	さししめす	指し示す	動詞

表 4　基本語彙に含まれるがスコアの低い語

順位	読み	漢字表記	品詞
20062	たいふう	台風	名詞
20095	おばあさん	お婆さん	名詞
31097	てつだう	手伝う	動詞
37796	かむ	噛む	動詞
47579	もちろん	勿論	副詞
65413	ところが	ところが	接続詞

上位の語を取り出して比較すると，共通して含まれる語は 50% 程度であることが分かる．

提案手法で上位に入るが基本語彙に含まれない語と，基本語彙に含まれるが提案手法によるスコアは低い語の一部を表 3，4 に示す．表 3 に示す語は，基本語彙としては適切ではないかもしれないが，語釈文を記述する上では必要と考えられる．一方，表 4 に示す語は日常的によく使う語であり，基本語彙に含めることは妥当であるが，口語的であったり，類義語で言い換えられるなどの理由により，語釈文の記述には不要である可能性が高い．

4　定義語彙による記述に向けて

本節では，定義語彙による語釈文の記述に関する問題点を述べる．我々は，笠原ら [5] と同様，既存の国語辞書の語釈文を初期状態とし，それを書き換えることにより定義語彙による記述を実現することを考えている．

1. 語釈文が定義語彙内の語のみで記述されている場合は，そのまま．
2. 定義語彙にない語が含まれる場合は，定義語彙内の語で置き換える．または，その部分を削除する．

語釈文中に含まれる定義語彙外の語が少ない場合には，置換や削除で書き換えられることが多い．以下に例を示す．括弧 (【】) で囲まれた語が定義語彙 (の候補語) にない語である．

痩せ山:　【土質】が悪く，木などが育たない山
夜業:　夜間に仕事をすること．その仕事．【よなべ】
並び無い:　他に並ぶものがないほどすぐれている．【比類】がない

「痩せ山」の場合，「土質」または「土質が悪い」を何か別の表現に換えることで解決できると考える．一方，「夜業」や「並び無い」では，「よなべ」や「比類がない」は見出し語と同義の表現を示しているだけで語義そのものとは異なるため，削除しても影響は少ないと考える．

定義語彙外の語が多い場合でも，基本的には語の置換や削除で解決可能と思われるが，場合によっては，語釈文全体を見直した方が良い場合がある．

ブザー:　【電磁石】を利用し，【振動板】の振動によって，【うなる】ような低い音を出させる装置．連絡・【通報】に使用する
毒蛾:　【黄色い】【中形】の，【蛾】の一種．夏，灯火に飛んで来る．雌の【毒針毛】が人の皮膚にさわると【強烈】な【かゆみ】を感じ，【高熱】を出す

語釈文が長くなると，定義語彙外の語が多くなる．それらは，語義そのものの説明だけでなく，文法的解説や歴史的背景に関する記述があったり，専門用語を用いて百科事典的な解説をしたりしていることが多い．例えば，動植物の場合，岩波国語辞典では外見的特徴だけでなく「〜科」，「〜目」のような生物学的分類や用途などあらゆる情報を載せている．一方，LDOCE では，外見的特徴と生息地域の説明だけであることが多く，記述が単純である．

雀:　われわれの生活に最も親しい野鳥で，背は茶色に黒い斑点があり，腹は白い，はたおりどり科の小鳥．稲の穂などをついばむので，農民にはきらわれるが，害虫をよく食べる．竹との取り合わせが，梅に鶯同様に周知 (岩波国語辞典)
sparrow:　a small brown bird, very common in many parts of the world (LDOCE)

より詳細な記述をするためには，より多くの (専門的な，特殊な，高度な) 語を必要とする．限られた少数の語で語釈文を記述する際には，語義そのものとそれ以外の説明は分けて考え[*2]，できる限り単純にする必要がある．

また，候補語の選定において，語の表記のみに注目し，意味を考慮していない．そのため，たとえ定義語彙内の語で語釈文を記述していたとしても，そのままで良いかどうか疑問に思うこともある．例えば，

寓目:　目をつけること．目をとめること

「つける」や「とめる」は定義語彙に含まれているが，「目をつける」や「目をとめる」のような表現が良いかどうかは考慮していない．多義語については，どの意味を語釈文の記述に利用するかを決める必要がある．

5　おわりに

本稿では，提案手法によるスコア付けの上位に入る語と従来の基本語彙を比較し，さらに定義語彙による語釈文の記述方法について検討した．我々が作成した定義語彙と従来の基本語彙では 50% 前後の語が両者に共通して含まれるが，一般的とは言い難いが語釈文の記述には必要となる語や，逆に日常的に使われるが語釈文の記述には不要となる語の存在も確認できた．語釈文の記述方法については，語釈文の内容を細分化し，定義語彙による記述を検討すべきものとそうでないものを区別すべきである．また，候補語の選定では表記のみに注目し，意味を考慮していないため，仮に候補語に挙がっている語だけで語釈文が記述されているとしても，その使われ方が妥当であるか否かを判断しなければならない．今後は，実際に語釈文の書き換えを試しながら，我々が作成した定義語彙を改善していく必要がある．

参考文献

[1] Güneş Erkan and Dragomir R. Radev. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, Vol. 22, pp. 457–479, 2004.

[2] 橋田浩一. GDA 日本語アノテーションマニュアル, 2005. http://i-content.org/gda/tagman.html.

[3] 橋田浩一.　岩波国語辞典のアノテーション — 照応・共参照・項構造 —, 2006.　http://www.i-content.org/rwcDB/iwanami/doc/tag.html.

[4] A. S. Hornby and Michael Ashby, editors. *Oxford Advanced Learner's Dictionary of Current English*. Oxford University Press, 2005.

[5] 笠原要, 佐藤浩史, フランシスボンド, 田中貴秋, 藤田早苗, 金杉友子, 天野成昭.「基本語意味データベース: Lexeed」の構築. 情報処理学会第 159 回自然言語処理研究会, pp. 75–82, 2004.

[6] 国立国語研究所. 日本語基本語彙 — 文献解題と研究 —. 国立国語研究所報告 116. 明治書院, 2000.

[7] 国立国語研究所. 教育基本語彙の基本的研究 — 教育基本語彙データベースの作成 —. 国立国語研究所報告 117. 明治書院, 2001.

[8] 野呂智哉, 徳田雄洋. 語釈文記述のための日本語定義語彙の構築に関する一考察. 言語処理学会第 13 回年次大会, pp. 626–629, 2007.

[9] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford University, 1998.

[10] Paul Proctor, editor. *Longman Dictionary of Contemporary English*. Longman, 2005.

[*2] 岩波国語辞典では，一部の語義以外に関する説明 (補足的説明) について，その先頭に記号「▽」をつけることによって語義そのものと区別している．本研究でもその情報を利用することで語義とそれ以外を分離することが考えられたが，岩波国語辞典コーパスではアノテーションの段階で記号「▽」が削除され，補足説明であることを表すタグも付与されていなかったため，分離できなかった．