# 既存学習語彙表の再評価の試み
## —学習サービスの視点から—

小野真嗣[*1]・曽我聡起[*2]・菊地真人[*3]・田邉鉄[*4]
Email: onomasa@mmm.muroran-it.ac.jp

*1：室蘭工業大学大学院工学研究科ひと文化系領域
*2：公立千歳科学技術大学理工学部情報システム工学科
*3：名古屋工業大学大学院工学研究科つくり領域
*4：北海道大学情報基盤センターデジタルコンテンツ研究部門

◎Key Words　　語彙頻度表，学習語彙表，WordNet，上位語と下位語

## 1. はじめに

語彙表とは、短時間で素早く効率よく学習するための基礎資料として存在価値があり、市販の英単語帳などに応用されるものである。言語情報処理技術が大衆化し、機械処理による言語分析が進むにつれ、多種多様な語彙表が公開されてきた。一方、語彙表が多く生成されると、今度は語彙表自体をどのように選択すればよいのかという問題も見えてきた。語彙表は生成される数に比べて、その評価数は少なく、統合的な利活用も少ない。

本発表では、既存の学習語彙表を収集・集約して、データの再活用を図り、新たな語彙学習法や提供サービスを見出すために語彙表分析を試み、その処理結果を報告する。具体的には、①英語の実態調査としての語彙頻度表や、②英語の学習指標としての学習語彙表の双方の特徴を踏まえ、③概念辞書などの語彙関係性データベースの情報と組合せることで、各語彙が持つ新たな特性の解明を試みる。1 語 1 語覚える単調な作業となる語彙学習について、難易度と頻出度と相関度の３点からなるグループ化を通じて再分析することにより、これまでにない有効性、効果、満足度が得られる語彙学習提供サービスの開発を目指す。本発表は、その処理過程で得られた分析結果の中間報告とする。

## 2. 先行研究

1990 年代以降、汎用英語コーパスの普及とコンピュータを用いた機械処理の語彙研究が隆盛したことにより、様々な語彙表が各研究者により編纂されている。本節では、本研究で利用するデータセットとしての既存英語語彙表について、それぞれの特性についてまとめる。

### 2.1 語彙頻度表データ

語彙の使用状況を調査する目的で、コーパスに基づく語彙頻度調査は、Kucera & Francis (1967)を皮切りに始められた。アメリカ英語における総語数約 100 万語により編纂された Brown Corpus に基づく語彙頻度表作成により、初めてコンピュータ解析に基づいた語彙使用状況が明らかとなった。

その後、1990 年代に入り、英国においてコーパスに基づく語彙研究は進み、Kilgarriff (1995)はイギリス英語における総語数約 1 億語により編纂された British National Corpus（以下、BNC）に基づく語彙頻度表(以下、BNC 語彙表)を作成し、大規模データによる頻度分析がさらに進展した。

また、Windows95 の爆発的な普及以降は、言語学者も容易に自分で機械的に語彙分析を行える時代となり、自ら集めた言語資料データに基づき、様々な分析が進められ、今日に至っている。英語教育の分野では、杉浦(2002)が一例であり、高校英語の検定教科書に掲載されている英語語彙を収集し、語彙表を編纂して使用語彙の検討を行っている。本研究では、英語の使用実態に基づいた調査を行うため、BNC 語彙表を分析尺度として用いた。

### 2.2 学習語彙表データ

一方、学校教育における教材作成や語彙学習の難易度の観点から、学習英語語彙表もこれまでに編纂されてきた。園田(1996)は語彙の効率的な学習を促し、学生の英語学習のための手引として、7454 語を学習語彙に選定し、各語にレベル 1 から 5 の難易度を付与する形でリスト化して、一般的には北大語彙表として認知されている。

学会組織としても学習語彙表の試みがあり、大学英語教育学会(JACET)の英語語彙研究会が中心となり、JACET8000 が編纂され、8000 語が学習語彙として選定されている。これは前節で述べた BNC 語彙表と学会独自のサブコーパスに基づいて、中学校および高等学校の教育現場状況に配慮した学習語彙表として編纂されたものであり、「日本人英語学習者のための科学的教育語彙表」と表現されている。この他、全国高等専門学校英語教育学会(COCET)が編纂した理工系の科学技術英語語彙を指向した COCET3300 の他、出版社ではアルク教育社が「中学生から一般社会人まで全ての英語学習者が段階を追って効率的に学べるよう、有用性と重要性を考慮して選定」したとされる標準語彙水準 12000(SVL12000)などが編纂されている。

本研究では、英語の学習向け難易度の指標として、園田(1996)による北大語彙表を用いた。

### 2.3 機械可読の概念辞書データ

前節までの語彙表の他、コンピュータプログラム化された語彙データセットとして、WordNet と呼ばれる機械可読の概念辞書が存在する。WordNet は、ある語が synset と呼ばれる同義語のグループに分類されており、簡単な語義のほか、他の同義語グループとの関係性が標準出力

で示すことができるデータ集である。各語は上位語や下位語と呼ばれる語群との関係性を示す情報を有しており、適切な検索をかけることによって、語の意味や関係性が出力され、データとして抽出することができる。本研究では、主に上位語と下位語の関係性に絞り、既存の語彙表情報と組合せる形で、学習者が覚えるべき総語彙の範囲内の言語世界観における小さな語彙関係性の可視化を試み、いわゆる Learner's Minimized WordNet の構築を試みた。

### 2.4 問題の所在と研究の仮説

これまで英語語彙表の活用は、学習者にとっては、主に高校受験や大学受験のための効率的な英語語彙習得の学習活動のために用いられてきた。一方、昨今では教育現場のデジタル化が進みつつある中で、英語教育環境も大きな変化を迎えている。小学校においては外国語教育の波が押し寄せ、従来の 5.6 年生に限られていた外国語活動から、3 年生以上へ外国語活動が引き下げられ、5 年生以上は教科化して英語教育が本格的に始まった。大学英語教育や一般社会人向けには e-Learning コンテンツは多様に揃っているが、小学校英語教育においては、アクティビティ向けの語彙選定やデジタル機材を通じた学習法、および学習提供のサービス手法については、まだ発展途上段階の分野と言える。

これまでの研究の過程においては、言語学分野による語彙表編纂と情報学分野における概念辞書構築にそれぞれ分かれ、その分野内に限って行われてきた経緯がある。双方のデータの関連性について教育工学の観点、また小学校英語教育の観点からの学際的研究は始まったばかりであり、この点についての一層の研究が求められる。そこで、本研究ではその学際的研究の試みとして、使用頻度に基づく学習語彙間の相互関係性の調査と関係性の可視化について実験・調査を行った。

## 3. 学習語彙の相互関係性の調査分析

### 3.1 調査分析対象

本研究では学習語彙間の相互関係性を調査するためでも、従前のような新たに独自の語彙表を生成することはせず、既存の英語語彙表を再活用することとした。語彙使用の頻度と学習すべき語彙の重要性や難易度の捉え方については、Lonsdale(2013)が参考となる。Lonsdale によると、図 1 に示すように、1000 語の習得により日常会話の 85%の語彙をカバーすることに繋がり、3000 語の習得では 98%の語彙をカバーすることになると述べている。そのため、核となる語の習得が必要であり、使用頻度の高い語彙から習得した方が効率的である点に言及している。

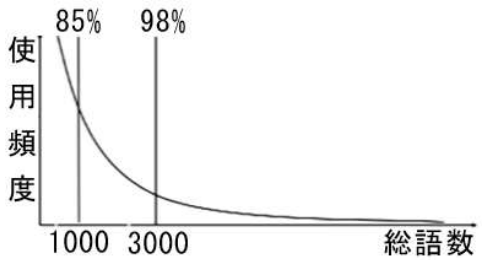


図 1 総語数と使用頻度の関係(Lonsdale(2013)より引用)

具体的な語彙データに基づく調査を行う上で、語彙頻度の情報としては、BNC 語彙表を用いることとし、一方の学習語彙の情報としては、北大語彙表を用いることとした。それぞれの語彙表の掲載語彙の統計情報については、次節の表 1 と 2 に示す。調査分析の基準として Lonsdale の尺度に倣い、語彙頻度表については 1000 語ごとに区切って語彙数を計算し、上位 1000 位以内と 3000 位以内のレベルにおける語彙を調査対象とした。一方、学習語彙表については北大語彙表で示されている語彙難易度の指標をそのまま用い、レベル 1 から 3 までを調査対象とした。

### 3.2 語彙の生起状況に関する分析と考察

本研究は萌芽研究である。分析を容易に進めるために、まずは出来る限り少なく、核となる語彙に焦点を絞り、分析を始めることとした。前節までに述べた通り、小学校英語を念頭に置き、取扱う語彙頻度情報は、英語の基本五文型の主要素となる名詞と動詞に絞り、頻度順位の上位 3000 語以内にある 2105 語を対象とした。

表 1 語彙頻度表(BNC 語彙表)の収録語彙数

| 頻度順位 | 名詞語彙数 | 動詞語彙数 | 小計 |
|---|---|---|---|
| 上位 1000 位以内 | 419 | 218 | 637 |
| 上位 2000 位以内 | 954 | 419 | 1373 |
| 上位 3000 位以内 | 1477 | 628 | 2105 |
| 3001 位以降 | 1785 | 653 | 2438 |
| 左記品詞内小計 | 3262 | 1281 | 4543 |
| 掲載総語数 | | 6318 | |

一方、学習語彙の絞込みにおいても、目安となる Lonsdale の尺度である上位 1000 語と 3000 語に注目し、北大語彙表のレベル 1 から 3 までを調査対象とし、4659 語を分析とした。

表 2 学習語彙表(北大語彙)の収録語彙数

| 難易レベル | 語彙数 | 語彙指標 |
|---|---|---|
| レベル 1 のみ | 785 | 中学必修 |
| レベル 2 以下 | 2563 | 高校必修 |
| レベル 3 以下 | 4659 | 大学受験 |
| レベル 4 以下 | 6179 | 大学基本 |
| レベル 5 含む合計 | 7453 | 大学上級 |

表 3 語彙頻度と学習語彙の生起分布比較

| BNC＼北大語彙 | 中学必修 | 高校必修 | 大学受験 |
|---|---|---|---|
| 動詞上位 1000 位 | 129 | 197 | 197 |
| 動詞上位 2000 位 | 148 | 308 | 330 |
| 動詞上位 3000 位 | 154 | 389 | 439 |
| 名詞上位 1000 位 | 184 | 366 | 379 |
| 名詞上位 2000 位 | 256 | 693 | 772 |
| 名詞上位 3000 位 | 274 | 901 | 1058 |
| 合計上位 1000 位 | 313 | 563 | 576 |
| 合計上位 2000 位 | 404 | 1001 | 1102 |
| 合計上位 3000 位 | 428 | 1290 | 1497 |

BNC 語彙表において、名詞と動詞に限って使用頻度が高く上位 3000 位以内に入っており、かつ、北大語彙表において学習語彙として重要性が高くレベル 3 以内に入っている語、つまり、これら双方の語彙表において共通となる語を対象に、概念辞書である WordNet の上位語および下位語の関連性を抽出する作業を行った。語彙頻度と学習語彙の生起分布比較については、表 3 に示した。

表 3 は語彙の生起数を表示したものであるが、実際の語彙の例示を次の(1)と(2)で示す。紙面の関係で全て示すことはできないため、頻度上位 1000 位以内に生起する中学必修名詞 184 語、および頻度上位 1000 位以内に生起する中学必修動詞 129 語のみ例示する。

(1) 頻度上位 1000 位以内に生起する中学必修名詞 184 語

act, age, air, animal, answer, area, arm, baby, bed, board, body, book, box, boy, brother, building, business, car, care, case, chance, change, child, church, city, class, club, company, condition, country, course, cup, date, daughter, day, doctor, dog, door, end, evening, example, eye, face, fact, family, father, feeling, few, field, figure, fire, floor, food, foot, force, form, friend, front, game, garden, girl, glass, government, ground, group, hair, hand, head, heart, home, horse, hotel, hour, house, husband, idea, interest, job, kind, king, land, language, leg, letter, library, life, light, line, list, lot, machine, man, market, matter, measure, member, mile, mind, minute, moment, money, month, morning, mother, music, name, news, night, note, number, office, order, page, paper, parent, part, party, pattern, people, percent, person, picture, piece, place, plan, plant, point, power, president, price, problem, product, question, reason, record, report, rest, result, road, room, rule, school, science, sea, season, service, shop, side, son, sort, sound, space, state, station, step, story, street, student, study, summer, system, table, teacher, test, thing, thought, time, top, town, tree, value, village, voice, wall, war, water, way, week, wife, window, woman, word, world, year

(2) 頻度上位 1000 位以内に生起する中学必修動詞 129 語

add, agree, allow, appear, arrive, ask, base, be, become, begin, believe, break, bring, build, buy, call, carry, catch, cause, choose, close, come, continue, cover, cut, decide, die, do, draw, drive, drop, eat, enjoy, enter, explain, fall, feel, fight, fill, find, finish, fly, follow, forget, get, give, go, grow, happen, have, hear, help, hit, hold, hope, include, introduce, keep, kill, know, lead, learn, leave, let, lie, listen, live, look, lose, love, make, mean, meet, move, need, open, pass, pay, pick, play, present, pull, put, raise, reach, read, receive, remain, remember, return, rise, run, say, see, seem, sell, send, serve, set, show, sit, smile, speak, spend, stand, start, stay, stop, suppose, take, talk, teach, tell, thank, think, throw, train, try, turn, understand, use, visit, wait, walk, want, watch, wear, work, write

表 3 や(1)、(2)による分析結果から考察する。初修学習における英語語彙において、頻度 1000 位以内に限ると、動詞よりも名詞の語彙数が多く、比に換算すると、およそ 1:1.43 となるが、頻度 3000 位以内まで拡大すると、動詞と名詞の語彙生起数の差はさらに開き、その比は 1:1.78 となる。さらに、高校必修レベルや大学受験レベルになると、その差は一層広がり、頻度 3000 位以内において高校必修レベルでは 1:2.31 に、大学受験レベルでは 1:2.41 となり、物事そのものの知識量に比例しているように感じられる。一方、動詞は基本五文型に共通して出現し、どの文においても英文の核となる存在になることから、動詞の語彙数の伸びは、中学必修レベルから大学受験レベルにおいて 310 語ほどの増加であるのに対し、名詞の語彙数の伸びは、874 語もの増加となっていることが判明した。学習過程や時間経過とともに 564 語分をより多く名詞を覚えなければならない点は、物事の事象などの知識量の増加に反映している可能性があり、中学必修レベルにおける動詞と名詞の学習するべき語彙数の差が 55 語である点からも想像できる。これらの点についてグラフにより可視化したものが、図 2 である。

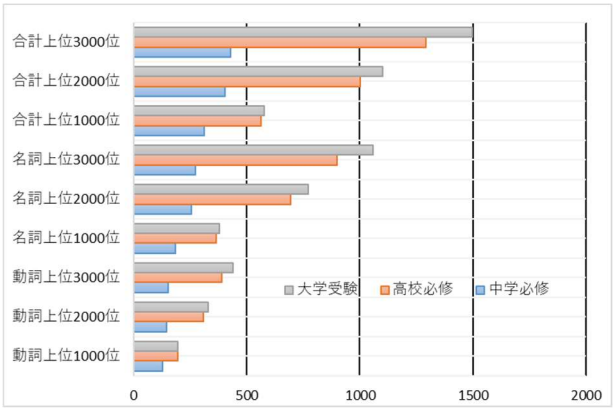


図 2 語彙頻度と学習語彙の品詞別生起分布グラフ

### 3.3 概念辞書情報の付与と可視化

表 3 に示す通り、小学校英語を念頭に、核となる英語語彙を見出し、合計上位 1000 語以内の 313 語を対象にして、その各語に対する概念辞書情報を関連付ける作業を行った。しかしながら、300 語程度では他の語との関連性はまだまだ薄く、図 3 のような期待していた語の階層構

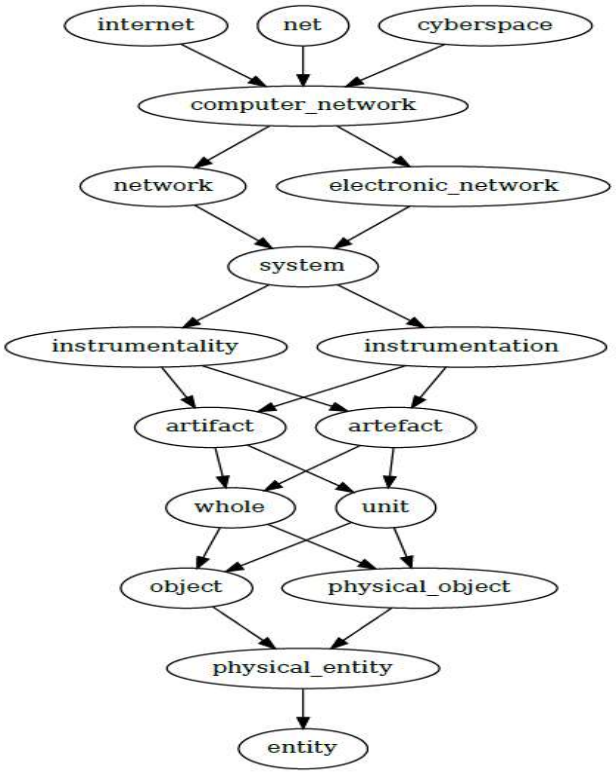


図 3 制限内の語彙世界における語彙関連性の表示例

造は得られなかった。図 3 は総語彙を約 15000 語に拡張した際に、ようやく語の関係性が階層構造となって現れたものである。頻度上位 3000 語の語彙世界となる範囲においては、名詞と動詞からなる 15000 語の 10%である約 1500 語の語彙世界の中では、まだ関係性を示すに至る十分な語彙数が揃っていないものと現時点では考えられる。今後はこの点のさらなる継続的な研究が必要となる。

### 3.4 今後の課題：概念辞書内に存在しない語の存在

最後に、前節において、語の関係性を階層構造として示すに至らなかった点は述べたものの、その副次的な結果として、上位語・下位語といった他の語との関係性を持たない語の存在も見えてきた。これは、いわゆる「代用の利かない基本語」と考えられるのかもしれないが、頻度上位の 1000 語、2000 語、3000 語の各範囲内において、それぞれ 31 語、55 語、80 語が他の語との関係性を持たないことがわかった。こちらも、一層の研究が必要となる。

(3) 頻度上位 3000 語以内の語彙世界において、WordNet 上において、上位語・下位語を有さない 80 語

| |
|---|
| acid, advice, airport, apple, artist, beer, brick, camera, century, childhood, cigarette, citizen, city, clock, corridor, critic, daughter, decade, democracy, diary, diet, director, disaster, doubt, drama, economics, equivalent, farmer, festival, few, god, grammar, guitar, hospital, insist, king, knee, lake, laugh, lawyer, marry, merchant, mile, mine, mountain, museum, night, parliament, passenger, percent, phone, plastic, poem, poetry, pollution, pot, potato, poverty, president, prince, prisoner, pupil, refugee, republic, rid, sea, sheep, shrug, son, studio, television, tennis, thank, tourist, town, valley, video, violence, week, year |

## 4. おわりに

本稿は、中間発表としての位置付けで、既存英語語彙表の再評価として、従来の目的別に編纂された単体での語彙表利用から、各語彙表情報をコンピュータ処理を加えることで融合し、教育改善に応用することを目的に調査を行った。また、概念辞書を用いて各語が有する他語との関係性を辞書記述レベルから図示する形で可視化を試み、語彙の学習定着を図る手法を試みた。

本研究では、語彙頻度表と学習語彙表の統合を通じて、概念辞書との掛け合わせを試みたが、語彙の世界観の創出には他の語彙表を用いることで、違った様相が示されることも期待され、SVL12000 などにおける上位 3000 語レベルでは、「英語の基礎をなす必須単語」、「日常生活で活躍する英単語」、「楽しく会話がはずむ英単語」と定義付されているため、また異なる結果が得られる可能性がある。小学校英語語彙を検討する上では、SVL12000 も研究の価値があると考える。引続き既存語彙表の有効利用や再活用の検討を続け、教育現場に求められる効果、効率、満足度の高い語彙学習の提供サービスについて研究を進めていく所存である。


## 謝辞

本研究は JSPS 科研費 JP22K02825 の助成を受けたものです。

本研究の遂行にあたり、研究協力者として室蘭工業大学大学院修士課程の西山幹泰さん、細川大和さん、ならびに名古屋工業大学大学院修士課程の按田将吾さんには、データ処理プログラミングの実装において、細部にわたりご支援をいただきました。ここに感謝いたします。


## 参考文献


(1) 杉浦千早：“高校英語教科書語彙リストの作成と使用語彙の検討”，Language Education & Technology，39 号，外国語教育メディア学会，pp.117-136.（2002）.

(2) 園田勝英：“大学生用英語語彙表のための基礎的研究”，言語文化部研究報告叢書, 7 巻, 北海道大学言語文化部, p.200.（1996）.

(3) Kilgarriff, A.: BNC Database and Word Frequency Lists, (1995) Retrieved from https://www.kilgarriff.co.uk/bnc-readme.html (June 30, 2022).

(4) Lonsdale, C.: How to learn any language in six months, (2013) Retrieved from https://youtu.be/d0yGdNEWdn0 (June 30, 2022).

(5) Kucera, H. & Francis, W.N.: Computational Analysis of Present Day American English, Brown University Press, p.424. (1967)

(6) Miller, G.A.: WordNet: An Electronic Lexical Database, MIT Press, p.423. (1998)