明海大学大学院応用言語学研究科紀要
応用言語学研究 No.5 抜刷
(2003年3月)

第5回明海大学大学院応用言語学研究科セミナー　講演

学習語彙リスト作成の技法：日中英の視点から

投野 由紀夫
加藤　晴子
小木曽 智信

Graduate School of Applied Linguistics
Meikai University

学習語彙リスト作成の技法：日中英の視点から

投野由紀夫・加藤晴子・小木曽智信

1. はじめに

本シンポジウムの目的は、学習語彙表の作成の現状と課題を日・中・英という多言語の視点から比較検討することである。具体的には、学習語彙表の歴史的変遷を各言語ごとに概観し、その作成の動機や背景を考察する。さらに、この 20 年ほどでコンピュータによる語彙表作成が急速に発達してきた経緯を踏まえ、言語教育に携わる我々でも比較的容易に語彙表作成が出来るようになってきた周辺事情および語彙表作成過程の実際を紹介する。最後に三言語それぞれの立場から、学習語彙表のあり方について考察を加える。

2. 学習語彙表作成の歴史とその背景

2.1. 英語の場合

英語の語彙表作成は大きく電子化以前と電子化以後に分けて考えることが出来る。特に電子化以前は語彙表作成が当時の教育に対する考え方に大きく影響を受けた時代であり、電子化以後はコーパス言語学の隆盛がその最大の特徴ということが出来る。

2.1.1. 電子化以前

英語の語彙表についてその起源をたどるのは学問的に興味あるところであるが、ここでは話を 19 世紀米国で盛んになったスペリング改革運動以降から始めたい[1]。20 世紀の初頭には主としてこの 19 世紀にずっと行われたスペリング統一の教育的効果を検証するために Elbridge が 3 万 5000 語の新聞データから約 6000 語の語彙調査を行ったり (Elbridge 1911)、Anderson が 36 万語の規模の手書きデータから約 9223 語の語彙リストを作成し、スペリングエラーを分析している (Anderson 1921)。

さらに 1920 年代から第二次大戦前にかけては行動主義心理学の影響で教育測定運動が盛んになり、言語に関しても精密な最小単位に分割していき、それを正確に測定することが言語能力の測定や教材作成に不可欠との立場からさまざまな言語統計が取られた。この頃の代表的な研究者は Edward L. Thorndike である。彼は動物実験による学習心理学の分野で大きな功績があるが、それを人間の学習理論にも応用し、最初の知能テストや幼児の読み書き能力の測定指標、また学習語彙表の面でも非常に貢献した人である。1921 年には当時としては大規模な 400 万語の言語資料に基づいた学習基礎語彙確定のための語彙表を作成している (Thorndike 1921)。さらにこの時すでにテキストの偏りによる語彙統計のぶれを予測し、分布 (range) の考え方を打ち出している。1944 年には Lorge と一緒に 1800 万語のコーパスから 3 万語の語彙リストを作成している (Thorndike & Lorge 1944)。この当時はすべて印刷物からの人手によるカウントだったことを考えると、その

[1] 英語語彙表の歴史に関しては Fries (1950)を参照。

資料の規模はきわめて大きいといえよう。その集大成ともいえる Thorndike & Lorge (1944) などは戦後の日本の英語学習辞典の1つの重要語指標の柱として、長く活用された。

このように最初の学習語彙表は主として米国におけるスペリング改革・教育のため、また初等教育を中心とした学習基礎語彙確定のためであり、そのような時代のニーズや社会背景があって生まれてきた。また行動主義の影響が強くあったため、学習語彙表の作成手順などに当初からかなり綿密な標本抽出や頻度分布統計が開発された。これらの考え方の多くは後の電子化以後の語彙表作りにも生かされていることを特筆しておく。

リスト名	語数	人手(M)/コンピュータ(C)	利用した資料の規模
Elbridge (1911)	6,000	M	35,000 語
Anderson (1921)	9,000	M	360,000 語
Thorndike (1921)	10,000	M	400 万語; 分布 (range)
Horn (1926)	36,000	M	16 種類の語彙表を合成
Lorge & Thorndike (1938)	1,200	M	225 万語, 意味別頻度
Thorndike & Lorge (1944)	30,000	M	1800 万語
Dale & Chall (1948, 1995[2])	3177	M	Readability 研究のため
West (1953)	2,000	M	500 万語, 意味別頻度

表 1: 電子化以前の主要な英語語彙表[2]

もう1つコンピュータ導入前の動きとして特筆すべきは、日本における Harold E Palmer, A.S. Hornby らの学習語彙表作成であった。彼らは日本における英語教育の実践を踏まえて、学習基本語彙の選定に尽力した。彼らの語彙表作成の最大の特徴はデータよりも経験を重視したこと、また単語単体よりも「語彙の使われ方」を強調して活用語彙表の原型を目指したことであった。Palmer が Carnegie Report をまとめるにあたり、データからの客観的な頻度表だけではよい語彙表にならないと言ったのは示唆的である。

その他、英語では早くから意味別の頻度表 (Lorge & Thorndike 1938; West 1953) や複数の語彙表を合成して1つの大型の語彙表を作成する (Horn 1926) などの試みがなされていることも付記しておく。

2.1.2. 電子化以後

英語の語彙表に革命をもたらしたのは、ブラウン大学で開発された 100 万語のアメリカ英語のコーパス、Brown Corpus であった。これをもとにした語彙表が Kucera & Francis (1967)である。この時点からコンピュータ分析が本格的に導入され、コーパスに出現した全単語について 15 分野 500 のテキストに現れた出現分布と絶対頻度が記録された。この後、米国は Chomsky による行動主義心理学の否定、認知心理学・生成文法の隆盛により、言語統計に関する研究的興味が失われる「冬の時代」を迎えることになる。これによりコーパス言語学の中心はヨーロッパに移っていくことになる。唯一、米国で Thorndike ら

[2] Bontrager (1991) に基づく

の言語統計を受け継いで認知心理学までその仕事を継承したのは、John B. Carroll であった。彼は60年～70年代に旺盛に精神測定の分野で活躍し、言語能力の認知モデルや言語統計などでも多くの貢献をした。特に American Heritage の一連の辞典のための語彙表作成は有名である（Carroll et al. 1977)。これは主として本国人の国語教育のためであったが、分野別の頻度を分布やコーパス・サイズなどを勘案して対数表示をすることにより、異種サイズのコーパス間の頻度を相対的に示す方法を考案するなど、現在でもその手法は広く応用されている[3]。

ヨーロッパでは Brown Corpus の成果を受けて、イギリス英語の100万語のコーパス LOB Corpus が1970-78年の間に作成され、Hofland & Johansson (1982) が LOB と Brown という2つのコーパスを比較してイギリス英語とアメリカ英語の頻度比較を行った。このように大規模なコーパスでの変種英語の頻度情報がこの後さまざまなコーパス構築の成果を受けて利用できるようになる。また Brown Corpus は単語に品詞を判定して自動的に付与するという品詞タグ付与プログラムを最初に開発したことでも有名で、その後、LOB はランカスター大学のタグ付与プログラム CLAWS の開発により、品詞タグ付の頻度リストが上掲されている（Johansson & Hofland 1989)。

リスト名	語数	人手(M)/コンピュータ(C)	利用した資料の規模
Kucera & Francis (1967)	全体	C	100万語; Brown
Carroll, Davies & Richman (1971)	86,741	C	500 万語; SFI, U 統計値
Hofland & Johansson (1982)	全体	C	LOB & Brown; カイ2乗値
Johansson & Hofland (1989)	全体	C	LOB; h collocation freq
Zeno, et al. (1995)	154,000	C	1700万語; SFI, U, 教育用
Leech, Rayson & Wilson(2001)	全体	C	1億語, BNC; Log-likelihood

表 2： 電子化以後の英語の主要語彙表

その後、これらのコーパス・データからの情報をもとに最初の英語辞典を開発したのが、バーミンガム大学の COBUILD Project であった。彼らは外国の英英辞典では初めてコーパスの頻度をもとに5段階の重要度表示を見出し語に対して行った。続くロングマンの英英辞典（LDOCE）では、1億語の British National Corpus（BNC）などからの頻度情報をもとに、重要語3000語を選定し、話し言葉・書き言葉別にランク表示を行った。

これらの例に見られるように、現在はコーパスからの頻度データを学習語彙表に利用する試みが急速に浸透している。日本でも、アルクが独自の基準で BNC の頻度リストを既存学習語彙表、ネイティブの直観、日本語コーパスのデータなどで重み付けし、Standard Vocabulary List 12000 を提唱している。また同様の試みで大学生用語彙を北海道大学でも提案している（北大語彙表)。また大学英語教育学会でもコーパスを用いた新しい手法の語彙リスト作成を行い、JACET8000 として提案する予定である。

[3] 例えば最新の米国の教育語彙表としては Zeno et al. (1995) を参照

以上のように、英語の学習語彙表はその種類においても作成方法においても、他の言語に先んじて多くの工夫や方法論上の改革があったと言えよう。言語統計の取り方に関する厳密さ、分布、相対頻度などの客観的指標の提案などは、他の言語の語彙表作成にも共通する重要な概念である。それらが、20 世紀前半の教育測定運動などの流れと一致して影響を受けていることも興味深い。しかし、日本語・中国語の例を比較してみるとわかるように、英語はスペースで単語と単語が初めから区切られているゆえに、語の単位認定に関する概念が非常に曖昧であったことは否めない。逆に後発の日本語・中国語の方が、単語認定の基準と真正面から取り組まねばならなかった、という点では、語彙表作成の重要課題をより深く取り扱ってきたと言えるかもしれない。

2.2. 日本語の場合

日本語の学習語彙に関する調査・研究は、大正頃より、欧米での研究の影響のもと、国語教育の流れの中で行われ始めたようである。初期のものに、澤柳政太郎他『児童語彙の研究』同文館（1919 年）などがある。昭和 10 年代になると占領下にあった地域での「日本語教育」を目的としたものが現れるが、これが日本語を母語としない学習者のための語彙調査のはじめだろうと思われる。これら戦前の研究・調査の中には語彙表を持つものもあるが、客観的な大規模調査に基づくものではなかった。

戦後になると、小学校をはじめとする国語教育のための語彙調査表がさかんに出版される。これは国定教科書の時代が終わり、新しい教科書が必要とされたためだと思われる。この後も引き続き、教科書作成のための語彙調査が行われているが、教科書を調査対象としたもので、比較的小規模なものが多い。

一方、戦後間もない 1948 年に設立された国立国語研究所によって、これまでにない大規模な語彙調査が行われはじめた。『語彙調査－現代新聞用語の一例－』（1952 年）、『総合雑誌の用語－現代語の語彙調査－』（1957, 1958 年）、『現代雑誌九十種の用語用字』（1962, 1963 年）などである。日本語を対象とした初の本格的な語彙調査であり、多くの新しい知見をもたらしたが、ほとんどがサンプリングによる調査であった。

1970 年代になると、国立国語研究所ではコンピュータを導入して語彙調査に利用している。『電子計算機による新聞の語彙調査』（1970 年）が最初であるが、これはきわめて早い時期のコンピュータ導入であり、コンピュータで日本語を扱えるようにする試みそのものが、この研究の中で進められていた。この後も『高校教科書の語彙調査』（1983 年）、『中学校教科書の語彙調査』（1995 年）など、継続的にコンピュータを利用した語彙調査が行われている。ただし、これらの調査におけるコンピュータ利用は、“調査対象テキスト全文を電子化して集積し、コンピュータで処理する”といった形態ではなく、むしろ“単語カードを電子化してコンピュータに集積、処理する”といったものであった。当然、単位切りも人手によるものである。

ちょうどこのころから日本語教育が盛んになるのを反映して、新たに日本語教育を目的とした語彙表が作られるようになった。早いものでは、樺島忠夫・吉田弥寿夫「留学生教育のための基本語彙表」『日本語・日本文化』大阪外大（1971 年）などがある。規模の大きなものとしては、国立国語研究所『日本語教育のための基本語彙調査』秀英出版（1984 年）、『品詞別・レベル別 1 万語彙分類集』専門教育出版編集部テスト課 専門教育出版（1991

年）などが挙げられよう。内容的には、日本語教育に使用されている教科書の語彙調査のほか、専門家による既存の語彙表からの選定が行われている。

これまでに挙げた語彙表は、基本的に人手によるものであった。コンピュータを利用していても、電子化テキストを処理して作成された語彙表ではなく、単位切り作業などは人手に頼っていた。しかし近年、自然言語処理の発達により、自動形態素解析による語彙表作成が可能になってきた。NTT コミュニケーション科学基礎研究所による『NTT データベースシリーズ　日本語の語彙特性—朝日新聞の語彙・文字頻度調査』などがその例である。人手による解析結果の見直し作業が必要であるとはいえ、形態素解析により、今まで不可能だった大規模な語彙表が作成できるようになってきた。今後、こうした新しい成果が学習語彙表づくりに反映されるようになるのが期待される。

2.3.　中国語の場合

2.3.1.　文字表の作成：20～40 年代

中国語では語の単位より文字の単位のほうがはるかに認識しやすいため、基本文字の研究やその文字表の作成は比較的早くからおこなわれていた。基本語彙のほうはいくつか表が作成されたようである。

2.3.2.　標準語の普及と語彙リスト：50～60 年代前半

新体制の中国として国造りが始まった時期であり、標準語の規範化とその普及、識字率の向上が緊急課題のひとつであった。それを背景にソ連からの影響を受けつつ、基本語彙についての研究が本格化した。識字率の向上のためには、語と結びついた文字学習が必要となる。1955 年 10 月に北京にて≪现代汉语规范问题学术会议≫が開催されたのもこの流れを受けたものである。標準語の中の外来語・隠語・方言的要素・文言的要素・新語の取り扱いが議論され、また、方言話者・少数民族への標準語の普及、外国人への中国語教育などが取り上げられ、それらを目的とした語彙表が、既存の辞書から選択するなどの経験的手法により作成された。

その後 80 年代までは、反右派闘争から文化大革命にいたる政治の混乱の中で、中国国内においては目立った成果もないままとなる。

2.3.3.　コンピュータの利用：80 年代～

中国が再び国際社会に眼を向けた時、そこはコンピュータ利用の時代に入っており、中国語をコンピュータに載せるという課題が突きつけられた。より効率的な入力システムを開発するために必要な語彙情報の研究が進んだ。また、著名な文学作品などをデジタル化して CD-ROM で販売、またはネット上に公開するなどの形でコーパスの蓄積が始まったが、それらは著作権無視の、信頼性にも欠けるものであった。その後、主にアメリカでの研究成果を取り入れて、言語資料の自動解析による統計的処理が行われるようなり、その成果が一部公開され、機械翻訳の精度向上に利用されるなど、コンピュータ処理の研究が本格化している。表 3 に主な語彙表を目的別・作成法別に整理して挙げる。

		主要目的			
		標準語の普及	初等教育	対外教育・検定試験	コンピュータ処理
作成法	語の区切り・選定が経験的手法によっているもの	普通话三千常用词表[1959，文字改革出版社]		高校中国語教育のめやす[1999，全国高等学校中国語教育研究会] コミュニケーションのための中国語基本単語 1400[2000，相原茂，東方書店]	
	語の区切り・選定がコンピュータ処理によっているもの		常用字和常用词[1985，北京语言学院]		现代汉语常用词词频词典[1990，北京航空航天大学等，宇航出版社]
			现代汉语频率词典[1985，北京语言学院]		
	既存の辞書・語彙表をもとにしたもの			中国語検定試験使用基準単語表[1989，日本中国語検定協会] 汉语水平词汇与汉字等级大纲[1992，北京语言文化大学] 基礎段階における語彙ガイドライン策定の試み[1999，山田眞一]	现代汉语语法信息词典详解[1998，清华大学出版社]

表 3： 中国語の主要語彙表

3. 語彙表作成の実際

3.1. 全体的なプロセス

語彙表作成には各国語で共通するプロセスがあるので、まずそれを概観しておく。現在では、語彙表は概略以下のような工程で作成される：

(a) コーパス・データの選定・電子化
(b) 単語の単位認定（固有名詞・複合語の処理）
(c) 形態素解析 （日本語・中国語は分かち書き；英語の場合は品詞タグ付与）
(d) 見出し語化
(e) 見出し語ごとの頻度集計
(f) 関連統計の算出（順位・使用率・累積頻度・累積使用率・総体頻度・分布）

ここではまず日本語の例をとりながら、形態素解析の概要を説明し、それらを中国語でどのように実現するかを紹介してもらい、最後に英語の処理に関して述べたい。

3.2. 日本語の処理

3.2.1. 「単位切り」の問題

単語ごとに分かち書きされる英語などとは異なり、一般に、書かれた日本語には切れ目がない。そのため、日本語の語彙表を作成するためには、まず単位ごとに切れ目を入れる

という作業が必要になる。電子的なテキストデータが用意されている場合であっても、語彙表を作るためには「単位切り」という人手による作業が必須だったのである。この単位切り作業には膨大な人手と時間を要する。とりわけ、漢字かな交じり文という複雑な表記体系を持つ日本語ではテキストの入力の段階から大きな困難を伴う。そのため語彙の使用頻度表作成は研究者個人でおいそれと始められるものではなかった。用例調査に基づく語彙表のほとんどが国立国語研究所のような機関において作られた背景には、このような作業コストの問題が大きかったと思われる。

切れ目がない日本語のテキストを切っていくという単位切り作業は、単にコストがかかるというだけではなく、いかに切るかという理論的な問題も抱えている。英語などの分かち書きされる言語においては、ごく一部の複合語など例外的なものを除けば、どのように切るべきか悩むことはほとんどない。ひとかたまりで書き表すべき語が社会的に定まっており、その単位が持つ言語学的な意味は別として、既定の単位が一般に共有されているからである。ところが、分かち書きされる習慣のない言語においては、いかに切るかが大きな問題となる。分かち書きのない言語で単位切りを行うということは、新たにひとまとまりと認めるべき言語学的、理論的背景をもつ単位を定めるところから開始しなければならない。また、理論的な問題とは別に、作業上の便宜という実際上の問題も考慮しなければならない。最適な単位は研究目的によっても変わってくる。いずれにしても、単位の大きさとして様々なものがあり得ることになる。実際、国語研究所の調査報告書だけでもM単位・W単位、長単位・短単位、β単位・$\alpha 0$ 単位・α単位といった各種の単位が使われており、それぞれ数ページに及ぶ切り方の説明があるほどである。

3.2.2. 形態素解析

コンピュータの利用が広がるとともに、このようにコストがかかり問題の多い日本語テキストの単位切りを、コンピュータを利用して自動で行おうという動きが広がった。分かち書きがなされない日本語では、かな・漢字といった字種の別が、読み取りにおいて区切りの識別に役立っているが、一時期はこれを単位切りに応用した試みなどもあった。

1990年代になると自然言語処理の分野で「形態素解析」と呼ばれる技術が発達し、本格的に利用することができるようになってきた。形態素解析とは、あらかじめ用意した辞書と文法的な規則をもとにして、文章データを形態素単位に分解し、同時に品詞などの情報を付加するソフトウェア技術である。テキストデータを与えれば「単位切り」を行うと同時に個々の単位に品詞情報を付与することができる。これは語彙表作成にきわめて大きな力となるものである。分かち書きされる英語の場合には、「単位切り」については大きな問題はなかったが、品

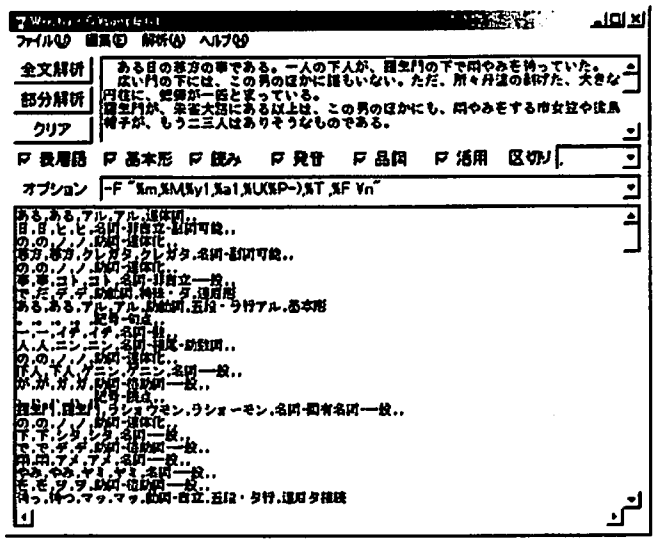

図 1

詞情報の付与はやはり手間のかかる作業である。これが機械によって処理できることになる。

今日では各種の形態素解析のためのソフトウェアが利用可能になっている。日本語の形態素解析ソフトウェアにもいくつかあるが、ここでは最も広く利用されていると思われる「茶筌」を紹介したい。「茶筌」(Chasen) は奈良先端科学技術大学院大学自然言語処理学講座で開発され、フリーウェアとして公開されているものである。今日では Windows をはじめとする多くの OS 上で、比較的簡単に利用することができるようになっている。現在、http://chasen.aist-nara.ac.jp/chasen/ から入手可能である (図 1)。

形態素解析はきわめて有用な技術であるが、いくつかの問題もある。まず、形態素解析の解析結果は完全ではないこと。辞書に載っていない言葉は処理できないうえ、誤った解析結果を返すこともある。新聞記事などの一般的な文章ではかなり高い精度で解析するが、文学作品などになると精度は下がってくる。したがって、研究の目的にもよるが、解析結果を人手で見直すことが必要な場合もある。もう一つは、形態素といってもその単位が必ずしも整ったものではないこと。日本語の「単位切り」の問題として、単位の認定基準の問題を挙げたが、形態素解析を利用したからといってこの問題が解決するわけではない。結果の利用目的に応じて、解析結果を分析し直すことが必要である。

3.2.3. 解析結果の利用

形態素解析の結果は、表のような形式のテキストデータとして与えられることが多いが、これだけでは語彙表として有意義に利用することができない。このデータを集計し、分析するためのツールが必要となる。英語を対象としたものであれば語彙表作成に便利な専用ツールが出ているようであるが、現在のところ日本語の語彙表作成に使えるものはないようである。こうした場合には Perl などのテキスト処理言語を用いて専用のプログラムを自ら作成する場合が多いが、ここでは汎用のリレーショナルデータベースソフト・MS Access を利用して簡単に集計する方法を紹介しよう。「茶筌」が出力したデータを表として Access のデータベースファイルに読み込むと、そのデータベースは一つのコーパスとして機能させることができる。このデータベースの「選択クエリ」で見出し語を基準として「グループ化」を行い、見出し語ごとのカウントを表に加えれば、見出し語ごとの頻度表が作成できる。あるいは、クエリで ID を 1 つずつずらして出現形を結合させてゆくと、ある見出し語の前文脈、後文脈を出力させることができる。この表示形式を工夫すれば KWIC として利用できる (図 2)。

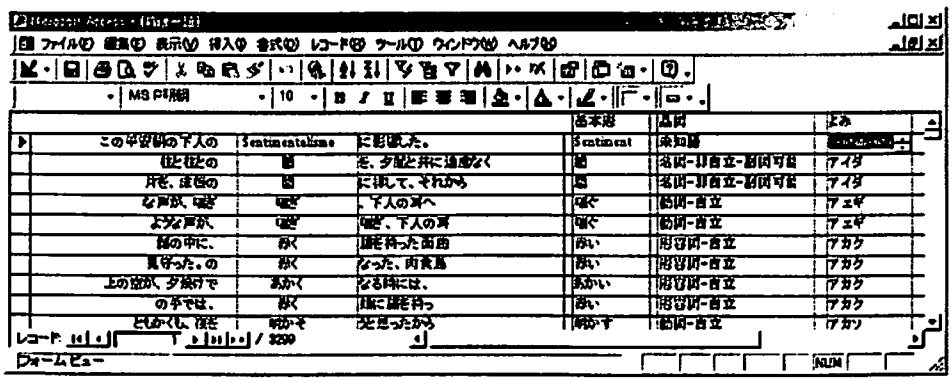

図 2

具体的な内容は Web 上で公開中のサンプル[4]をご覧いただきたいが、プログラミングを行わなくても、工夫をすればこのような簡単な操作によって望む結果を得ることができる。

3.3. 中国語の処理

中国語において同様の過程がどのように実現されるかを考える。

3.3.1. 中国語の特徴と形態素解析

中国語のテキストの大きな特徴は、語と語の区切りがなく、すべて漢字で表記する点である。形が品詞を表さず、形態変化もないために、それらを手懸りに語を取りだすことができない。また、固有名詞と一般の語との区別がなく、外来語、海外の地名などもすべて漢字を充てるため、表面上は区別がつかない。従って中国語を自動で形態素解析しようとする場合には、事前に、品詞、出現頻度、前後の組み合わせの可能性など、多くの情報を備えた辞書を作成しておく必要がある。何を以って語とするかという問題も解決されなければならない。その上でさらに、例えば“发展中/国家(発展途上国)”の中の“中国”を取りださないようにする、“高兴(うれしい)”→“高高兴兴(うれしい)・高兴高兴(うれしくさせる)”のように音節を重ねて作られる「重ね形」と呼ばれる語形と“汉语/语法(中国語文法)”の中の“语语”とを区別する、“跳舞(踊る)”→“跳起来舞(踊りだす)”“游泳(泳ぐ)”→“游一会儿泳(しばらく泳ぐ)”などのように1語でありながら他の語句の割り込みを許す「離合詞」と呼ばれる語の処理を的確に行なう、などの問題の解決が求められる。

このような辞書を備えた中国語解析ツールとして公開されているものに、北京大学计算语言学研究所汉语切分与标注软件[5]がある。このサイトで利用できるのはサンプル版で、一度に100文字までの解析しかできないが、早稲田大学などではサイト契約を結び本格的に利用しているようである。ただし、このサンプル版で、重ね形、離合詞を含む文を解析した結果は以下のとおりで、学習語彙リスト作成のために理想的に解析されているとはいえない。なお品詞記号は表4に示す。

Ag	形语素	g	语素	ns	地名	u	助词
a	形容词	h	前接成分	nt	机构团体	Vg	动语素
ad	副形词	i	成语	nz	其他专名	v	动词
an	名形词	j	简称略语	o	拟声词	Vd	副动词
b	区别词	k	后接成分	p	介词	vn	名动词
c	连词	l	习用语	q	量词	w	标点符号
Dg	副语素	m	数词	r	代词	x	非语素字
d	副词	Ng	名语素	s	处所词	y	语气词
e	叹词	n	名词	Tg	时语素	z	状态词
f	方位词	nr	人名	t	时间词		

表 4：北京大学计算语言学研究所汉语文本词性标注标记集

[4] http://tonolab.meikai.ac.jp/~goilist/ogiso/goi.mdb
[5] http://icl.pku.edu.cn/nlp-tools/segtagtest.htm

输入的句子为：

他高高兴兴地跳起舞来了。

切分与标注的结果为：

他/r{ta1} 高/v{gao1} 高兴/a{gao1xing4} 兴地/ns{xing1de5} 跳/v{tiao4} 起/v{qi3} 舞/v{wu3} 来/v{lai2} 了/y{le5} 。/w{。}

サンプルとして、≪狂人日记≫(鲁迅)全文を解析したのが以下の結果である。≪狂人日记≫の全文は≪新语丝 鲁迅专辑≫[6]より入手した。

输入的句子为：

狂人日记某君昆仲，今隐其名，皆余昔日在中学时良友；分隔多年，消息渐阙。日前偶闻其一大病；适归故乡，迂道往访，则仅晤一人，言病者其弟也。劳君远道来视，然已早愈，赴某地候补矣。

切分与标注的结果为：

狂 人 /n{kuang2ren2} 日 记 /n{ri4ji4} 某 /r{mou3} 君 /Ng{jun1} 昆 /Ng{kun1} 仲 /Ng{zhong4} ，/w{，} 今/Rg{jin1} 隐/Ng{yin3} 其/Ng{ji1} 名/Vg{ming2} ，/w{，} 皆/d{jie1} 余 /m{yu2} 昔 日 /t{xi1ri4} 在 /p{zai4} 中 学 /n{zhong1xue2} 时 /Dg{shi2} 良 友 /n{liang2you3} ；/w{；} 分隔/v{fen1ge2} 多年/m{duo1nian2} ，/w{，} 消息/n{xiao1xi5} 渐 /d{jian4} 阙/Ng{que1} 。/w{。} 日前/t{ri4qian2} 偶/d{ou3} 闻/v{wen2} 其一/r{qi2yi1} 大 /a{da4} 病/n{bing4} ；/w{；} 适/Ag{shi4} 归/p{gui1} 故乡/n{gu4xiang1} ，/w{，} 迂/a{yu1} 道/n{dao4} 往/p{wang3} 访/v{fang3} ，/w{，} 则/d{ze2} 仅/d{jin3} 晤/Vg{wu4} 一/m{yi1} 人/n{ren2} ，/w{，} 言/Vg{yan2} 病者/n{bing4zhe3} 其/u{qi2} 弟/n{di4} 也/y{ye3} 。/w{。} 劳/nr{lao2} 君远/nr{jun1yuan3} 道来/nr{dao4lai2} 视/Vg{shi4} ，/w{，} 然/r{ran2} 已 /d{yi3} 早/a{zao3} 愈/d{yu4} ，/w{，} 赴/v{fu4} 某/r{mou3} 地/u{de5} 候补/b{hou4bu3} 矣/U{yi3} 。/w{。}

3.3.2. 解析結果の利用

日本語にならい上の結果を、Access に入れてみる。Access は 2000 バージョンから多言語対応を強化しており、テーブル全体のフォントを指定することで文字化けの問題なく処理を行なうことができる。

図 3 は単純な出現頻度を示したもの、図 4 は後の語の組み合わせ別に頻度を示したもの、図 5 は前の品詞の組み合わせ別に頻度を示したものである。

	PINYIN	WORD	POS	FRQ
▶	，	，	标点符号	367
	。	。	标点符号	156
	de5	的	助词	131
	wo3	我	代词	117
	ren2	人	名词	81
	chi1	吃	动词	72
	ye3	也	副词	58
	bu4	不	副词	57
	shi4	是	动词	55
	ta1	他	代词	51
	；	；	标点符号	51
	le5	了	助词	44

図 3

[6] http://www.xys.org/pages/luxun.html

	PINYIN	Key	->	POS	->POS	FRQ
▶	ren2	人	，	名词	标点符号	29
	chi1	吃	人	动词	名词	28
	de5	的	人	助词	名词	24
	ren2	人	的	名词	助词	22
	shuo1	说	，	动词	标点符号	14
	yi1	一	伙	数词	量词	11
	le5	了	。	语气词	标点符号	9
	talmen5	他们	的	代词	助词	9
	liao3	了	，	动词	标点符号	9
	wo3	我	。	代词	标点符号	8
	de5	的	，	助词	标点符号	8
	liao3	了	。	动词	标点符号	8

図 4

	POS<-	POS	Key	PINYIN	FRQ
▶	动词	助词	了	le5	39
	动词	动词	了	liao3	16
	动词	语气词	了	le5	8
	形容词	语气词	了	le5	7
	名词	动词	了	liao3	3
	动语素	助词	了	le5	1
	量词	语气词	了	le5	1
	量词	助词	了	le5	1
	名语素	助词	了	le5	1
	方位词	语气词	了	le5	1
	代词	动词	了	liao3	1
	成语	助词	了	le5	1

図 5

この他に Visual Basic .Net を利用したプログラムも試作したが[7]、文字コードを指定した文字処理が行なえるので、中国語も問題なく扱うことができる。

3.4. 英語の処理

3.4.1. 英語語彙表作成上の技術的な観点

コンピュータが導入される以前は、すべてテキストを人手で検査し、個々のアルファベット毎に担当員を決めて、雑誌や新聞などの記事をにらみながら、該当する単語のカウントを行うという気の遠くなるような作業をしていた。コンピュータが利用されるようになってからは、主として電子テキストを加工して出来るだけ大量に自動処理する方向が主流になってきている。大きな流れとしては、3.1 の (a)~(f) と同じであるが、以下に若干の補足を加えてまとめておく：

(a) 入力 (input)
コーパスの中身によってかなり出てくる語彙は異なってくる。その中身が適切に取捨選択されたかを知る必要がある。特に話し言葉のように書き起こしのガイドラインがきちんと選定されていない資料の場合には、データ間の不統一が問題にならないようにする。

(b) 単語認定 (tokenization)
英語の場合、単位切りは空白および句読点で行えるといっても実際は種々の問題点が生じてくる。例えば、it's などは it と is の短縮形('s) とに区切りたいが、同じアポストロフィの入った o'clock は 1 単語として認識させたい。ピリオドを文の区切り目と定義すると、U.S. のような例はどうするのか、といった問題がある。また New York は 1 かたまりの地名として定義したいが、通常の分割処理では new と York に別れてしまい、イギリスの York と区別がつかなくなってしまう。こういった諸処の問題点を判別して、単語認定を行う一連の作業が必要になる。

(c) 品詞タグ付与 (POS tagging)
単語認定が終わったら、次に個々の単語に品詞情報をタグ付与する。これは spring という単語があったら、spring_NN1(単数名詞); spring_VV1(自動詞); spring_VV2

[7] http://tonolab.meikai.ac.jp/~goilist/kato/sem02_kato03.htm

(他動詞) のように文脈によって異なる品詞になる場合を判定し、それに必要なタグを付与する。これは現在では日本語の形態素解析とほぼ同じように自動で行われるプログラムが存在し、精度は95%程度にまで及んでいる。

(d) 見出し語化 (lemmatization)

語彙表を作成する際には、spring の動詞の活用が spring, sprang, sprung, springing とあれば、それをみな spring の活用形である、と認定して、spring の頻度に合算させて頻度集計することができる。これを見出し語化という。この作業をした語彙表としていない語彙表が英語には存在する。後者は変化形の頻度も表示しておいた方が教育的である、とするような判断がある場合が多い。

(e) 頻度集計 (frequency count)

頻度集計にはいろいろな方法がある。ここでは専門的になるので触れないが、単純頻度、コーパス・データのサイズを100万語などに標準化した相対頻度、コーパスの語彙の何割に当たるかなどを示す累積頻度をはじめ、John Carroll などが開発した単語間の頻度の極端な差異を考慮に入れて対数などを利用した頻度指標 (U-score) のように複雑な頻度情報の表し方まで多様な方法が提案されている。

(f) 分散 (dispersion)、レンジ (range)

頻度と同時にテキスト間の頻度の分布を考慮することも大事である。例えば、頻度が100あっても500あるテキストのうち2つだけに50回ずつ出てきていて、他の498 のテキストには出てこないような単語では頻度の高さは違った意味を持ってくる。50～60年代、Julliand らによって行われた一連のロマンス語族の頻度辞典の刊行はこの分散の概念を非常に精密に定義した研究として有名である。

3.4.2. 英語の代表的な語彙表作成ツール

3.4.2.1. 汎用コンコーダンサ

英語は日本語、中国語に比べると、外国語としての学習者人口もはるかに多いし、また教育産業も盛んな分野なので、語彙表を作成するツール類に関しても進んでいる。主として一般の英語教師のレベルが扱えるような汎用プログラム類が充実しているといえる。代表的なものとしては、WordSmith (Mike Scott 作成)、MonoConc (Michael Barlow 作成)、WordPilot (John Milton 作成)といった Windows ベースの商用コンコーダンサが1万円程度の費用で購入できる。これらを用いれば、自分の持っているコーパス・データをもとに瞬時にして語彙リストを作成することが可能になる。また品詞タグなどが付与されたデータを用いれば、品詞ごとのリストなども簡単に作ることが出来る。

3.4.2.2. より大規模な語彙表作成システム

英語では一般利用者への垣根を低くしつつ、研究者向けには大掛かりな語彙表作成の統合システムが開発されている。例えば、英国ランカスター大学では Wmatrix という自作のコーパスを読み込ませると自動処理で入力 → 品詞タグ付与 → 見出し語化 → 意味タグ付与 → 語彙リスト作成 → 語彙リスト比較、という一連の作業を行ってくれるシステ

ムがある[8]。これらは、分析する者のテキスト処理の負担を大幅に軽減してくれ、リストの内容に集中することが出来る。

4. 学習語彙表のあり方

最後に日中英のそれぞれから見た学習語彙表のあり方についてまとめて本稿を閉じたい。

4.1. 語彙表の元となるテキストの中身

学習語彙リストとしては、まず、語彙に偏りがないことが条件となるであろう。英語の場合は、古くは Thorndike, また Brown Corpus 以降のテキスト標本を一定の科学的な方法で客観的に選択し、分野ごとにバランスを保つように工夫をする、という伝統が強くある。中国にも原資料を分野別に分け、一定の割合にしたがって頻度統計を取り、使用度とともに分布を示した資料が存在する。≪现代汉语常用词词频词典≫(1990, 北京航空航天大学等, 宇航出版社)などがそうである。しかし、中国語の場合、正式な印刷物とする場合の言葉と、外国人学習者が初・中級の段階で学ぶ言葉とには、もともとかなりの開きがあり、たとえ原資料に戯曲・漫才などが含まれていても、学習者向けの資料とするにはやや無理がある。また、例えば市場での会話を録音し、文字に起こした『コンピュータによる北京口語語彙の研究(第一冊資料編)』(1995, 中嶋幹起, 内山書店)のような資料もあるが、これもまた最初に学ぶ言葉として適当かどうか疑問である。

こういった問題に対処するためか、教材をテキストとして語彙表を作成することも広く行われている。その場合でも、中国を舞台とする場合と日本を中国語で紹介する場合とでは出現語彙が当然異なってくる。また、その結果から再び教材を作成するのでは「循環論」に陥ってしまうであろう。この点は日本語、英語でも同様である。英語では特にこの教科書で用いられる表現と実際の言語使用のデータを比較する研究が最近増加傾向にある。

学習語彙表は、学習によって達成が期待される、一定レベルの文章の語彙を目標として、それを徐々に獲得してゆくことができるように配慮されるべきであろう。たとえば中学・高校の国語教育であれば、新聞や新書などの一般向けのテキストが読解できる能力を身に付けることがひとまずの目的となるだろう。そのためには新聞などの一般向けのテキストを学習語彙表づくりに活かしていく必要がある。これには新聞社や放送局で用意されている用字・用語集の類も有用であろう。

また、言葉の変化にあわせて、俗語的にならない程度に、新しい語彙が導入されるように配慮する必要がある。そのためには広く読まれている雑誌など幅広いジャンルのテキストを元としなければならない。

4.2. コンピュータ処理の妥当性：頻度だけでよいのか？

頻度表をコンピュータで作成する技法は3節で紹介したように急速に進歩している。しかし、学習語彙表という性格を考えた際に、頻度だけでよいのかという素朴な疑問がある。英語の場合には、第二次大戦前の Thorndike らの業績はもっぱら使用頻度が最重要という観点から語彙表作成が行われていた。しかし、1934 年に行われた最初の語彙制限に関する国際会議である Carnegie Cooperation Conference において、Thorndike は Harold E.

---

[8] http://www.comp.lancs.ac.uk/computing/users/paul/publications/icame01.pdf

Palmer, Michael West らの作成する語彙表作成の基準を監修するような立場になったのだが、Palmer らによってその頻度絶対主義を厳しく批判され、経験による語彙表補正の観点を受け入れるという一幕があった（Palmer 1936)。Thorndike は心理学者であり、Palmer は外国語教育の専門家だったことは示唆的である。

コンピュータ処理を利用すれば、手作業ではとても不可能な大量のデータを扱うことができることはいうまでもない。ただしその場合でも、処理の前提となる品詞情報などを含む辞書は、主に手作業で作成しておく必要がある。その辞書も学習語彙リストの作成のためには、外国人学習者に必要な情報を持つものでなければならい。例えば、中国語についていえば、音節を重ねて作られる特殊な語形についての情報、1 語でありながら他の語句の割り込みを許す語についての情報など、学習者に不可欠な情報が現在必ずしも充分ではない。

一方、理想的な辞書、理想的なサンプリングができたとして、それだけで十全な語彙リストが完成するであろうか。中国で学習語彙リストを作成する際は、まず頻度リストを作成した後、必ず意味連想法による増減を加える。「天」に対して「地」、「天」から「太陽」「月」「星」などのように挙げていくものであるが、複数の人が同時に思いついたものを採用するという方法を取る。語彙は体系をなすものであり、この手法によりリストの完結性を高めるのは、語彙のこの側面にも配慮したものとして妥当なものであろう。

使用頻度による客観的方法と、専門家による選定という主観的方法のそれぞれの長所・短所は日本語の語彙表作成の歴史の中でも古くから指摘されていた。しかし作業量の問題から、客観的方法によって語彙表が作成されることはまれであり、実際にはほとんどは主観的方法によって作成されたものであった。たとえば国立国語研究所『日本語教育のための基本語彙調査』秀英出版（1984 年）などは、専門家によって選定されたものである。

たしかに頻度だけでははかれない問題も多い。しかし、十分な量の新しいデータなしに、主観的な選択のみで作られると、その語彙表は頻度のみによって作られたもの以上に問題の多いものになる。これまでの日本語の学習語彙表の多くは主観的な方法に頼りがちであった。その弊害は、たとえば現在ではほとんど使用されなくなった語が相変わらず基本語彙として選定されるといった形で現れてきているようである。

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思われる。今後、コンピュータの利用により頻度データが語彙表作成に活かされることで、両方のバランスがとれたものになることが期待される。

4.3. 作成した語彙表の具体的利用：どのように利用するのか？

この分野ではやはり英語が最も具体的な活用が進んでいるといえよう。最も利用が盛んなのは学習辞典の分野である。日本の英和辞典は早くから学習語彙の重要度表示に関心を持っており、Thorndike の語彙表をもとにした見出し語ランク表示を導入している。欧米で本格的に学習語彙表の利用がなされたの辞典は 70 年代の American Heritage と 80 年代後半の COBUILD であろう。American Heritage は J.B.Carroll らの教育的語彙頻度統計をもとにカレッジ版の辞書の語彙を選定しているし、COBUILD は独自に構築した Bank of English のデータをもとに 5 段階の頻度表示を開発した。その後、COBUILD プ

ロジェクトでは大規模コーパスからの頻度データや言語統計をもとに、辞典、文法書、類語辞典、会話辞典、単語・熟語集、英会話教材、などを包括的に作成している。COBUILDの中心地であったバーミンガム大学では、David Willis が語彙を中心としたシラバス（Lexical Syllabus）を提案し、コーパスからの単語やコロケーションの頻度情報を活用した指導法や教材開発が盛んになってきている（Willis 1990）。

コーパスからの頻度データは海外の英英辞典の定義文に使用される定義語彙（defining vocabulary）にも影響を与えた。以前は定義語彙に語彙制限がなかったが、1978 年に出版された Longman Dictionary of Contemporary English (LDOCE)で初めて定義語彙が2000 語と限定され、その後、各出版社（CUP, OUP, Macmillan など）もそれに追随した。これらの定義語彙は現在はコーパスからの頻度情報を重要な柱にしている。

またインターネット普及の波に乗って、語彙表を活用した語彙・リーディング学習支援プログラムがマルチメディアで展開してきている。例えば、アルクでは自社開発のSVL12000 という語彙表をもとに、web 上に単語辞書エンジンを設置し、それによる自動単語検索システムを組んでいる[9]。またこれらの単語リストをもとにした i-mode の単語学習プログラムを配信するなど、積極的に語彙表を学習教材化している。本学でも、英米語学科が中心で 1 万語プロジェクトと題して学習語彙表を選定し、i-mode による自学自習プログラムを開発中である。

日本における中国語教育には現在、学習指導要領的なものがない。中国語の学習語彙表はその作成の基礎資料、教材作成や習熟度の測定のよりどころとしての利用がまず考えられる。また、語と語の組み合わせ情報などを含む学習辞典の作成にも生かすことができる。中国語の場合「～を食べる」も「～で食べる」も同じ動詞＋名詞の形で表されるため、このような情報の提供は不可欠である。また、一定の方法論に従った語彙表ではないものの、「成蹊大学中国語音声教育データベース」のように、インターネット上に単語学習サイトを公開している例もある。高電社による自動翻訳サイト「J-Server」にも中国語が加わった。

日本語には、いまだ語の使用頻度や重要度を明示した辞書がほとんどない。頻度データを元にした語彙表は、学習者にとって使いやすい辞書づくりに活かすことができそうである。

5. まとめ

本稿では、学習語彙表の作成に関して、日・中・英それぞれの視点から歴史と現状を概観し、考察を加えてみた。多言語で比較してみるとわかる言語独自の特徴や分析の難しさ、また共通して言える特徴の面白さなどを体験した。何よりも、三言語それぞれである程度のコンピュータ処理が可能であり、その作成工程はノウハウを共有できる部分が多くあったので、本学でも学習語彙表を設計・作成し、授業に活用していこうという機運のようなものが出てきてもよいと痛感した。

---

[9] 「どこでも辞書 SVL12000」（http://210.239.181.110/alc/）参照

参考文献

〔英語〕

Anderson, W.N. (1921). Determination of a spelling vocabulary based upon written correspondence. *University of Iowa Studies in Education*, 2(1), 1-66.

Bontrager, T. (1991). The development of word frequency lists prior to the 1944 Thorndike-Lorge list. *Reading Psychology: An International Quarterly* 12: 91-116.

Carroll, J.B., Davies, P., and Richman, B. (1971). *The American Heritage Word Frequency Book.* Boston: Houghton Mifflin.

Elbridge, R.C. (1911). *Six thousand common English words.* Buffalo, NY: The Clement Press (Ann Arbor, MI: University Microfilms International).

Fries, C.C. (1950). *English word lists: A study of their adaptability for instruction.* Ann Arbor, MI: George Wahr Publishing.

Hofland, K. and Johansson, S. (1982). *Word Frequencies in British and American English.* Longman.

Horn, E. (1926). *A basic writing vocabulary: 10,000 words most commonly used in writing* (University of Iowa Monographs in Education, No.4). Iowa City: University of Iowa.

Johansson, S. and Hofland, K. (1989). *Frequency Analysis of English Vocabulary and Grammar.* Oxford: Clarendon Press.

Kucera, H. and Francis, W.N. (1967). *Computational analysis of present-day American English.* Providence: Brown University Press.

Lorge, I. and Thorndike, E.L. (1938). *A semantic count of English words.* New York: Columbia University, Teachers Collge.

Palmer, H.E. (1936). "The history and present state of the movement towards vocabulary control." *IRET Bulletin* 120, 14-17; 121:19-23.

Thorndike, E.L. (1921). *The teacher's word book.* New York: Teachers College Press.

Thorndike, E.L. and Lorge, I. (1944). *The teacher's word book of 30,000 words.* New York: Teachers College Press.

West, M. (1953). *A General Service List of English Words.* Longman.

Willis, D. (1990). *Lexical Syllabus.* Collins ELT.

Zeno, SM, Ivens, SH, Millard, RT, & Duvvuri, R. (1995). *The educator's word frequency guide.* Brewster, NY: TASA

〔日本語〕

岡本禹一 1944『日本語基本語彙』国際文化新興会

国立国語研究所 1962『現代雑誌九十種の用語用字』秀英出版

国立国語研究所 1970『電子計算機による新聞の語彙調査 I』秀英出版

国立国語研究所 1984『日本語教育のための基本語彙調査』秀英出版

阪本一郎 1984『新教育基本語彙』学芸図書

国立国語研究所 1986『中学校教科書の語彙調査』秀英出版
国立国語研究所 日本語教育研修室・日本語教育研究会 1997「学習基本語彙研究文献」(http://202.245.103.49/resources/goi.htm)
工藤真由美 1999『外国人生徒に対する日本語教育のための基本語彙調査』ひつじ書房

〔中国語〕
陈小荷 2000《现代汉语自动分析--Visual C++实现》北京语言文化大学出版社
国家对外汉语教学领导小组办公室汉语水平考试部 1992《汉语水平词汇与汉字等级大纲》北京语言文化大学出版社
輿水優 2001「語彙の選定，語彙表，語彙教育」中国語教育研究会　配布資料
山田眞一 1998「基礎段階のガイドライン（語彙）」全国中国語教育協議会 98 年度第Ⅱ期セミナー　配布資料
山田眞一 1999「基礎段階における語彙ガイドライン策定の試み」全国中国語教育協議会セミナー報告 No.1
周荐 1995《汉语词汇研究史纲》语文出版社