―研究論文―

マクロに見た常用漢字語の日中対照研究
―データベース開発の過程から―

松下 達彦

要　旨

稿者は「日中対照常用漢字語データベース」の開発を進めており、その目的、作成手順と概要を述べた上で、以下の諸点を明らかにした。1）現在有力な二つの語彙頻度表、国立国語研究所（2006）と天野・近藤（2000）を比較すると、上位約1000語での一致語数は41.6%しかなく、相関も高くない。2）漢語は5000語までの頻度レベルにおいても40%以上を占め、安定して用いられている。3）5000語までの頻度レベルにおいて、表記上の日中同形語は全体の3分の1以上を占め、上位2000語では40%を超える。漢語においては85%以上を占め、上位2000語では90%を超える。4）5000語までの頻度レベルにおいて、日中同形語の約4分の3が二字漢語であり、約4分の1が一字漢語である。5）日中同形語の日中両語における頻度の相関は全体として.334であり、ある程度の相関がある。

【キーワード】漢語、日中対照、同形語、データベース、頻度

1．はじめに

これまでいわゆる日中同形語の意味・用法の異同や習得について多くの論考があるが、その量的な分布をマクロに捉えた研究は非常に少ない。本稿では、「日中対照常用漢字語データベース」の開発目的および概要について述べた上で、頻度を参照する日本語の語彙頻度表につき、比較に基づく選定理由を説明する。その後、現在の日本語の書き言葉に於ける漢語や日中同形語の分布、日中同形語のそれぞれの語彙表に於ける頻度の相関などについて明らかにする。

2．「日中対照常用漢字語データベース」開発の目的

本研究で開発を目指しているデータベースは、一義的には、妥当化（validate）された日本語語彙テストを開発するための基礎的データの整備を目指すものである。開発を予定しているテストは、書きことばの日本語語彙の受容的知識（読み）について1万語以上の頻度レベルまで測定でき、背景の異なる学習者によって発達がどのように異なるか、特に漢字知識の影響によって中国語系学習者と非中国語系学習者でどのように異なっているかを、質と量の両面から検証するためのテストである[1]。それによって、語彙力が他の諸側面とどのように関わるかを検証し、第二言語としての日本語習得の研究、ならびに日本語教育に貢献しようとするものである。その基礎的なデータの整備として、各種コーパスを利用して日中対照常用漢字語のデータベースを構築し、テスト項目の選定のための基礎的な資

料を提供することを目的とする。

広く知られているように、漢字圏の学習者と非漢字圏の学習者では日本語学習の条件が大きく異なっているが、具体的にどの程度、どのように異なっているか、把握することが難しい。言語教育プログラム運営においても、漢字圏学習者にとって易しい項目と難しい項目が理解されていないために無駄な学習をさせていたり、非漢字圏学習者が漢字圏学習者に混ざって無理な学習に挑戦したりするなどの状況がある。このような問題を解決し、よりよいカリキュラムやアプローチを開発するためには、各学習者の語彙力の諸側面を、簡便かつある程度の精度で測定する必要がある。

日本語の語彙テストはすでにいくつか存在するが、十分な検証を経て妥当化され、かつ公開されているテストは管見の限り存在しない。例えば、日本語能力試験には文字・語彙の知識を測定する部門があり、テスト結果の分析が毎年公刊されているが、漢字を多く含むテストであるにも関わらず、中国語と日本語の相違のタイプはまったく考慮されておらず、内容的妥当性を欠く可能性を否定できない。各種の研究目的で採用された語彙テストもあるが（山本1994、松本・堀場2007など）、これらは公開されておらず、テストの妥当化の過程も明らかではない。

日本語語彙テストが漢字を多く含むテストである以上、中国語と日本語の相違のタイプを適切に反映することによって妥当化されるべきである。第二言語の語彙習得への第一言語影響についてはこれまで数多く論じられ、（総括的なものとしてSwan 1997, Ringbom 2006 など）、語彙テストにおける第一言語の影響への考慮についても英語テストの語彙項目において以前から指摘されている（Elder 1997にまとまった考察がある）。例えばChen and Henning（1985）は、中国語母語話者とスペイン語母語話者が英語のテストを受けた際に、スペイン語に同根語（cognate）のある英語語彙のテスト項目はスペイン語話者の正答率が有意に高かったことを報告している。また、Cobb（2000）はカナダのケベック州のフランス語使用者が英語のVocabulary Levels Test（Nation, 1990: 261-272）を受ける場合にギリシア・ラテン系の語彙とアングロサクソン系の語彙では正答率が異なることを実証している。

中国語からの借用語である漢語は、日本語語彙の総体においてどの調査においても異なり語彙で4割以上を占めており[2]、中国語系学習者には有利な点と不利な点の両方がある（陳2003、加藤2005）ため[3]、テスト項目が適切な割合で有利な点と不利な点を反映しなければ、第二言語としての日本語のテストとして妥当性を欠く可能性がある[4]。また、漢字は中国語と日本語で広く共有されている表意文字であり、中国語系学習者は漢字語を発音できない場合でさえその意味を理解できる可能性があるため、中国語系学習者の日本語語彙の発達は書きことばと話し言葉で大きく異なっていると予測される。

このような相違は、これまでの日本語語彙テストにおいてはほとんど無視されてきたが、中国人学習者が、同じく中国語からの借用語を大量に有する言語を第一言語とする韓国人、ベトナム人などと並んで日本語学習者の多くを占める事実[5]を踏まえて考えれば、日本語の語彙テストの項目は、特に中国語系語彙の視点で分類・層化（stratify）されるべき

である。

本研究では、そのような分類・層化のために、日本語の基礎的な漢字語について、音声形態、書字（表記）形態、意味、社会言語学的特性などの諸側面について、中国語との相違のパタンおよび例文を記述したデータベースの構築を始めた。まだ作成の途上であり、特に意味や例文に関わるフィールドの内容の検討とデータ入力はまだ残されているが、今後は大規模コーパスを使用することによって、各語の最も典型的な用法を検索し、連語の使用頻度まで考慮に入れた上で適切な例文を入力する予定である。このようなデータベースは通常のコンピュータで使用可能なフォーマットで簡便に提供することができ、試験の開発のみならず、日常の授業運営や教材開発、各種の習得研究などについても、基礎的なデータを提供するものであり、実用的な価値を持つであろう。

3.「日中対照常用漢字語データベース」の作成手順および概要

以下、開発中のデータベースの作成手順および概要を合わせて述べる。

1）国立国語研究所（2006）の「度数順語彙表（自立語）」をダウンロードする[6]
2）「シートの保護」を解除する
3）語種・順位（昇順）・見出し（昇順）でソートして、漢語以外の語種（和語、外来語、混種語、人名・地名などの固有名詞）を削除する[7]
4）以下の六つのフィールドを残し、他のフィールドをすべて削除する
5）順位、見出し、品詞、表記〔注記〕、使用率（‰）、出現形の内訳
6）以下の順にフィールドを作成し、フィールドのサイズを整える（下線のフィールドは国立国語研究所（2006）のデータを残す）
7）ID、国研順位[8]、日本語能力試験級別、見出し、品詞、表記〔注記〕、日本語使用率（‰）、中国語表記、表記の類似度、中国語頻度、意味対応、使用域対応、日本語と中国語に共通の意味、中国語のみに存在する意味・用法、使用域対応注記、中国語の常用対応語、中国語ピンイン表記、音声的類似度、日本語の常用共起成分、類義語[9]、中国語の常用共起成分、日本語の常用例文、出現形の内訳
8）フォントを日本語は「MSP ゴシック」に、中国語は「SIMHEI」に統一する。
9）IDを漢語のみの頻度順でつける。各データベースに共通の通し番号にする。（例：2000番台の最初の番号は1000番台の最後の番号の続きとする）
10）「中国語表記」は中華人民共和国の標準字体を入力する。
11）「中国語頻度」については北京语言学院语言教学研究所（1986）を参照して入力する。
12）「中国語ピンイン表記」は中華人民共和国の定めるものを入力する。
13）「音声的類似度」については茅本（1995）を参照して入力する。（ただし教育漢字のみのデータである。）
14）「類義語」については後述するSketch EngineのThesaurus情報を利用し、共起成分の類似する語のうち、学習に役立ちそうなものを判断して選んで入力する。
15）その他のフィールドについても可能な範囲で客観性の高い基準を立てて入力する予定

であるが、現在、検討中である。意味や例文など、ある程度、主観的判断に頼らざるを得ない項目もある。可能な限り、表記、音声、意味、使用域など、類似度を評価・分類し、その際、基本義と使用域が重なる場合は、重なりのタイプを分類する。

16）以上のほか、ノイズと思われる語は基準を立てて削除することを検討しているが、基準はまだ検討しているところである。例えば、使用率の順位3029番目の「三和」のように固有名詞などは一律に除去したほうがよいように思われる[10]。

共起成分の頻度を数えるコーパスについては、現時点ではイギリスで開発されたコーパス‘Sketch Engine’（延べ語数4億語以上[11]）が最も簡便で本研究の目的に合ったコーパスのようである[12]。他のコーパスと異なり、すでに連語頻度を数えるためのタグがついており、簡単な入力で頻度を数えられる。また、中国語のコーパスもあるため、連語頻度の両言語の対照も可能であるという利点がある。Web上の言語を資料としたコーパスのため、多少の偏りやノイズはやむをえないが、本研究には向いていると現時点では判断している。

以下、データベース開発の過程で現時点までに得られたいくつかの知見について、主に日本語の語彙頻度表の問題、漢語や日中同形漢語の位置づけについて記述する。

4．日本語の語彙頻度表の比較と選定 ―採録語彙の比較および使用頻度の相関―

データベースを作成するに当たっては、まず頻度を参照する日本語の語彙表を定めなければならない。現在、学術研究において使用される代表的な語彙頻度資料として国立国語研究所（2006）と天野・近藤（2000）がある[13]。前者の利点としては、雑誌コーパスに基づくため分野の偏りが少ないこと、無料でweb上に公開されており誰にでも入手できること、加工して公開する場合に著作権上の制限が少ないことなどが挙げられる。

一方、天野・近藤（2000）は新聞のみをコーパスとしていること、価格が2万円以上のため誰にでも使えるものではないことなどの欠点があるが、もとになるコーパスが国立国語研究所（2006）とは桁違いの巨大な規模であり[14] 低頻度語彙の使用率に比較的信頼度が高いこと、漢字語彙に限れば徳弘（2008）ですでに加工されたものがあり、天野・近藤（1999）の単語親密度や日本語能力試験の級別などとの対照が容易であることといった利点がある。データベース作成に当たってはどちらかの語彙表を基準として採用しなければならないため、まず、この両語彙表の特性を比較検証することにした。具体的には、それぞれの語彙表の各上位1000語のうち、どのぐらい一致しているか、また、一致している語の頻度はどの程度相関しているかを調べた。

4.1　方法

国立国語研究所（2006）の上位1002語（1000位に相当する同順位の語が複数あるため）の各語が天野・近藤（2000）の上位1000語に含まれるかどうかを調べた。両語彙表の上位1000語を並列し、仮名表記でソートして一致語数を数えた。天野・近藤（2000）のデータは、徳弘（2008）のCD-ROMのデータを用い、頻度順の上位1000語を選んだ。徳弘（2008）は総

採録語彙のうち和語や混種語も含めて漢字表記されうる語彙のみを選んで採録している。通常は仮名表記されることの多い語も、漢字かな混じり表記できる場合は採録しているようである[15]。仮名表記しかできない語を除いた上位1000語であるため、実際には上位1000語より多い語彙から漢字表記できる語だけを選んでいることになるが、頻度の相関を調べるには問題ないと考えた。同訓異字（例：「物」「者」は国立国語研究所（2006）では同一語だが、天野・近藤（2000）を採録した徳弘（2008）では別語）と一部の数詞（例：「二」「二十」は国立国語研究所（2006）では別語だが、徳弘（2008）は「二十」は「二」と「十」に分けて集計）など、それぞれの語彙表における採録方法が異なる114語は分析から除外した。除外した結果、比較の対象となった語数は国立国語研究所（2006）の上位1002語のうち888語である。リストアップした一致語につき、両語彙表における頻度の相関を算出した。

4.2　結果

このようにして上位888語における一致語数を調べた結果、完全に一致している語は369語、41.6%しかなかった。天野・近藤（2000）のほうは漢字表記できる語のみを対象としており、仮名表記しかできない語も含めれば上位1000語以上まで含めていることになると考えられるため、純粋な上位1000語を比較すればさらに一致語数はさらに少なくなると考えられる。仮名表記しかできない外来語や一部の和語が数多く一致する可能性もないとはいえない。しかし、一致している369語について、両語彙表における頻度の相関（スピアマンの順位相関[16]）を調べたところ.259（$p < .01$）しかない。なお、両語彙表の共通上位369語の記述統計量は表1の通りである。

表1　国立国語研究所(2006)と天野・近藤(2000)の共通上位369語の度数

	N	最小値	最大値	平均	標準偏差
国立国語研究所(2006) 度数*	369	93	5854	272.35	399.08
天野・近藤(2000) 度数*	369	17850	636980	56346.90	59136.15

*国立国語研究所(2006)の延べ語数738377語における度数
**延べ語数は不明だが、1億語以上であると推計される

4.3　考察、および語彙表の選定

高頻度語彙は低頻度語彙に比べてコーパス間の相違は小さいはずだが、それでも上位1000語の一致語数は半分にも満たない。両語彙表の頻度間には弱い相関があるといえるが、ともに書き言葉の大きなコーパスであり、しかもそれぞれの語彙表で上位1000語に含まれる語彙であることを考えると、低い相関であるといわざるを得ない。相関が低い理由は、大きく二つ考えられる。

一つは、国立国語研究所（2006）の対象としたコーパスの延べ語彙数が十分に大きくなく、安定したコーパスとはいえないことである。Chung（2003）は、低頻度の専門語彙を特定するために専門テキストのコーパスと一般テキストのコーパスの語彙を比較するアプロー

チを検討しているが、コーパスに相当の語数がないと専門語彙を特定するのは難しいようである。専門テキストのほうは、分野にもよるが、一般的には数万から数十万語程度のコーパスが使用される（Chung 2003: 225, 233）のに対し、一般テキストのほうは低頻度の専門語彙について安定した数値を得るため、約189万2千語からなるコーパスを使用している（Chung 2003: 233）。ところが、国立国語研究所のデータは延べ語数で1,065,617語しかなく、付属語327,240語を除いた自立語のみだと738,377語しかない。これは付属語を含めてもChung（2003）が使用したコーパスの延べ語数の半分程度しかないことになる。

より大きなもう一つの理由はコーパスの採録分野による語彙の偏りである。天野・近藤（2000）は新聞のみをコーパスとしているため、語彙が新聞語彙に偏っている。二つの語彙表で一致しない高頻度語のうち、天野・近藤（2000）の上位10語は「米国」「政府」「首相」「側・傍」「出る」「後ろ」「大統領」「開く」「億」「午前」である。これらは新聞的な語彙であるということが言えるであろう。ちなみに二つの語彙表で一致しない高頻度語のうち、国立国語研究所（2006）の上位10語は数詞を除くと「様（ヨウ）」「的」「御」「者（シャ）」「店（テン）」「区」「特価」「号」「都」「車（シャ）」である。

さらに共通上位369語の両語彙表のそれぞれの頻度のz得点（各度数から平均を引いて標準偏差で割った値）を求め、国立国語研究所（2006）の頻度のz得点から天野・近藤（2000）の頻度のz得点の差を求めてみた。この差は、数字が大きいほど共通369語の中では国立国語研究所（2006）に特徴的であることになるが、その差の数字の大きかった上位10語は、「言う」「月」「来る」「何」「時」「方」「市」「県」「彼」「株式」であった。逆に、差の負の値が大きいほど、共通369語の中では天野・近藤（2000）の特徴が強いことになるが、その上位10語は「人」「難しい」「氏」「問題」「約」「年」「同」「側」「昨年」「企業」であった。つまりこれらの語は両語彙表で共通に高頻度であるが、これ以上に分野が偏ると共通上位1000語に入らないという限界ギリギリの語彙ということになる。

茂木ほか（2005）によれば、新聞（毎日新聞9年分）の語種構成は雑誌のジャンルの中では専門雑誌に近く、延べ語数において「家庭」「大衆」「趣味」「文芸」などを含めたすべてのジャンルの和語の平均が51.8%、漢語の平均が37.5%であるのに対し、専門雑誌は和語39.1%、漢語53.1%、新聞はそれぞれ39.4%、54.1%で、いずれも平均とはかけ離れている（p.344）。語種構成の点からも新聞は均衡の取れた書き言葉コーパスとは言えない。

「日中対照常用漢字語データベース」が、一般的な語彙テスト開発という目的をもっていることを考慮すると、各分野に均衡していることが必要な条件であると考え、信頼性に欠ける面があることは致し方がないものと考え、国立国語研究所（2006）を語彙頻度に関する当面の基礎データとすることにした。将来、より妥当かつ信頼度の高い語彙頻度表が出てきた場合には、その頻度によってデータベースを更新することが可能である[17]。

5．高頻度自立語5000語において漢語はどのぐらい存在するか

まずは国立国語研究所（2006）において漢語（古代中国語からの借用語、字音語）がどの程度の数・割合を占めるかを1000語ずつに区切った範囲[18]で調べてみた（表2、表3）[19]。

表2 高頻度自立語5000語に占める語種別の語数・割合（固有名詞を含む）

レベル		語数						割合（%）					
レベル	頻度順位*	全体	漢語	外来語	和語	混種語	その他**	全体	漢語	外来語	和語	混種語	その他**
-1000	0001-992	1002	461	110	389	16	26	100.0%	46.0%	11.0%	38.8%	1.6%	2.6%
-2000	1003-1964	999	454	150	338	13	44	100.0%	45.4%	15.0%	33.8%	1.3%	4.4%
-3000	2002-2955	1027	451	204	280	26	66	100.0%	43.9%	19.9%	27.3%	2.5%	6.4%
-4000	3029-3903	1034	417	245	270	24	78	100.0%	40.3%	23.7%	26.1%	2.3%	7.5%
-5000	4063-4794	960	397	216	235	20	92	100.0%	41.4%	22.5%	24.5%	2.1%	9.6%
全体	0001-4794	5022	2180	925	1512	99	306	100.0%	43.4%	18.4%	30.1%	2.0%	6.1%

*同一順位に同頻度の複数の語があるため順位と累計語数は必ずしも一致しない。
**高頻度自立語5000語に含まれる「その他」はすべて固有名詞（地名・人名）である。

表3 高頻度自立語5000語に占める語種別の語数・割合（固有名詞を除く）

レベル		語数					割合（%）				
レベル	頻度順位*	全体	漢語	外来語	和語	混種語	全体	漢語	外来語	和語	混種語
-1000	0001-992	976	461	110	389	16	100.0%	47.2%	11.3%	39.9%	1.6%
-2000	1003-1964	955	454	150	338	13	100.0%	47.5%	15.7%	35.4%	1.4%
-3000	2002-2955	961	451	204	280	26	100.0%	46.9%	21.2%	29.1%	2.7%
-4000	3029-3903	956	417	245	270	24	100.0%	43.6%	25.6%	28.2%	2.5%
-5000	4063-4794	868	397	216	235	20	100.0%	45.7%	24.9%	27.1%	2.3%
全体	0001-4794	4716	2180	925	1512	99	100.0%	46.2%	19.6%	32.1%	2.1%

*同一順位に同頻度の複数の語があるため順位と累計語数は必ずしも一致しない。
**高頻度自立語5000語に含まれる「その他」はすべて固有名詞（地名・人名）である。

表2は固有名詞を含んだ集計であり、表3は固有名詞を除外した集計である。

漢語はいずれの集計においてもすべて40%台であり、上位3000語レベルまでのほうが4000語レベル、5000語レベルよりも若干割合が高い。

その代わりに頻度レベルが下がるについて外来語が増えている[20]。漢語は新聞や専門雑誌などのフォーマルな文脈で高い割合を占めるため、低頻度語彙に多いように思われやすいが、現代雑誌の書き言葉においてはむしろ頻度が下がると漢語の割合は減る傾向にあるといえよう。それでも全体的にはすべての頻度レベルで40%台となっており、安定した割合で用いられている。頻度が下がるにつれて外来語の割合が増えて和語の割合が著しく下がるのとは対照的である。このことは漢語が幅広い使用域で用いられていることを示しているように思われる。

6．高頻度自立語 5000 語において表記上の同形語はどのぐらい存在するか

同形語はこれまで意味や用法の異同について数多く論じられてきており（文化庁1978、荒川1979など多数）、中国語語彙全体における位置づけを量的に論じたものはいくつかあるものの（荒屋1983、曾根1988、高野・王2002）、日本語語彙全体における位置づけに関しては論じられてこなかった。そこでまず同形語が日本語の高頻度語彙においてどれぐらいあるかを調べてみた。付属語には漢語が存在しないため、自立語だけを調べればよい。

これを数えるにはいくつかの前提が必要である。まず、日中両語の字体の異なりはここでは考慮しない。対応字体で書いて同じになるもの（旧字体＝康熙字典体が共通のもの）は

同形語と判断する[21]。例えば日本語の「東」と中国語の〈东〉はここでは同形であるとみなす。中国語に存在する表現かどうかは、以下の二つの方法で集計してみた。

1）《现代汉语频率词典》(北京语言学院语言教学研究所1986) に頻度が掲載されている語のみを数える。

2）《现代汉语频率词典》で認定される語に加えて、《现代汉语频率词典》にない語については日本語教育を専攻する中国語母語の大学院生2名による主観的判定を加えた[22]。

表4 高頻度自立語5000語に占める日中同形語の語数・割合

レベル		語数		日中同形語の語数／割合					
				《现代汉语频率词典》のみで認定			《现代汉语频率词典》に主観的判断を加えて認定		
レベル	頻度順位*	全体	漢語	同形語(n)	n／漢語	n／全体	同形語	n／漢語	n／全体
-1000	0001-992	1002	461	372	80.7%	37.1%	423	91.8%	42.2%
-2000	1003-1964	999	454	344	75.8%	34.4%	413	91.0%	41.3%
-3000	2002-2955	1027	451	292	64.7%	28.4%	386	85.6%	37.6%
-4000	3029-3903	1034	417	234	56.1%	22.6%	342	82.0%	33.1%
-5000	4063-4794	960	397	224	56.4%	23.3%	323	81.4%	33.6%
全体	0001-4794	5022	2180	1466	67.2%	29.2%	1887	86.6%	37.6%

*同一順位に同頻度の複数の語があるため順位と累計語数は必ずしも一致しない。

結果は表4の通りである。上位1002語のうち漢語（音読漢字語）は461語で、そのうち主観的判定を加えた場合、漢語の9割以上、語全体の4割以上に相当する423語が日中同形語であり、中国語にない漢語は38語しかない[23]。同形語の割合は頻度レベルが下がるにつれて若干減少するが、4000-5000語のレベルでも漢語の8割以上、語全体の約3分の1が同形語で、5000語までの全体としてみると漢語の86.6%、語全体の37.6%が同形語である。これは中国語母語の学習者対象の日本語教育を考える上で、無視できない数字であるといえよう。

稿者はこの調査をするまでは同形語は頻度が下がるほど多くなると考えていた。5000語を超えるレベルまで調べればそうなる可能性もあるが、少なくとも5000語までは減少している。これは意外な結果である。その原因については今後の課題としたい。

7．同形語の日中両語での使用頻度の相関と比較

日中同形語には意味や用法の異同のほか、文体差もある（宮島1994）。例えば、中国語では日常語である語が、日本語ではフォーマルな語である、あるいはその逆という場合である。文体差を考える上では、頻度が一つの参考になる。高頻度語彙は一般に日常語彙であると考えられるからである。そこで日中同形語の頻度の相関や各語の頻度のz得点、頻度レベルごとの平均等を算出し比較してみた。また、一字漢語の同形語と二字漢語の同形語についても同様の比較を試みた。

7.1 方法

中国語の頻度は《现代汉语频率词典》（北京语言学院语言教学研究所1986）を参照した。この資料は1979年から1985年にかけて行なわれた調査の結果をまとめたもので、政治〈报

刊政论〉(24.39%)、科学〈科普〉(15.73%)、生活口語〈生活口语〉(11.17%)、文学〈文学作品〉(48.71%) の四つの分野の179種の文章に小中学校の教科書を加え、延べ約181万字、131万語を対象として行なわれた調査である。やや古い資料なので例えば経済、現代大衆文化、国際関係などの用語は変化している可能性が高いが、これ以外に現時点では適当な資料が見当たらないため、これを使用する。

一部に日本語では複数の語として扱われているもの（例：「分」（フン・ブン））が中国語では一語であったり、その逆であったりする場合があるが、文字表記が日中で対応していて合算できる場合は合算してから頻度を算定した。多様な組み合わせがある数詞46語だけは頻度の計算から除外した。

頻度はノンパラメトリックであると考えられるので、相関の計算にはスピアマンの順位相関を用いた。z得点は日本語と中国語のそれぞれの頻度について算出し、前者と後者の差を算出して、それぞれの言語につき頻度が相対的に高い語を調べてみた。

7.2 結果

以下の表5、表6の通りである。

表5 日中同形語の頻度の平均・標準偏差と日中両語における頻度の相関（レベル別）

レベル		語数				頻度（日本語）(‰)		頻度（中国語）(‰)		日中両語の使用頻度の相関	
レベル	頻度順位*	全体	漢語	同形語**	頻度計算の対象(N)***	平均	標準偏差	平均	標準偏差	スピアマン順位相関	有意確率
-1000	0001-992	1002	461	423	327	0.3297	0.4623	0.0054	0.0330	0.297	*p<.01*
-2000	1003-1964	999	454	413	344	0.0909	0.0172	0.0015	0.0031	-0.069	*n.s.*
-3000	2002-2955	1027	451	386	291	0.0515	0.0069	0.0014	0.0047	0.081	*n.s.*
-4000	3029-3903	1034	417	342	234	0.0344	0.0034	0.0010	0.0026	0.081	*n.s.*
-5000	4063-4794	960	397	323	224	0.0254	0.0018	0.0006	0.0012	-0.014	*n.s.*
全体	0001-4794	5022	2180	1887	1420	0.1182	0.2512	0.0021	0.0162	0.337	*p<.01*

*同一順位に同頻度の複数の語があるため順位と累計語数は必ずしも一致しない。
**主観的判断を加えて認定したもの
***《现代汉语频率词典》にデータのある同形語1466語から数詞46語を除いたもの

表6 日中同形語の使用頻度の平均・標準偏差と日中両語における頻度の相関（字数別）

漢語のタイプ	頻度計算の対象(N)*	頻度（日本語）(‰)		頻度（中国語）(‰)		日中両語の使用頻度の相関	
		平均	標準偏差	平均	標準偏差	スピアマン順位相関	有意確率
一字漢語	366	0.2095	0.4579	0.0057	0.0315	0.248	*p<.01*
二字漢語	1054	0.0865	0.0919	0.0009	0.0018	0.327	*p<.01*
全体	1420	0.1182	0.2512	0.0021	0.0162	0.337	*p<.01*

*《现代汉语频率词典》にデータのある同形語1466語から数詞46語を除いたもの

また、各語の日本語および中国語における使用頻度につき、それぞれ対象とした同形語1420語におけるz得点を算出し、日中両語間の差を求めた結果、最も正の値の大きかった10語（相対的に日本語において使用頻度の高い同形語）は順に「円〔通貨単位〕」「年」「様」「月」「市」「第」「県」「店」「者」「時間」で、最も負の値の大きかった10語（相対的に中国語において使用頻度の高い同形語）は「的」「不」「他」「説」「個」「上」「好」「主義」「国」「両」であった。

7.3 考察

全体の相関（.337）は高いとは言えないが、著しく低いわけでもない。ある程度の相関があると言える。レベル別に見ると最も頻度の高い1000語のレベルである程度の相関が認められるほかには、レベル別には相関が見られない。日中間の同形語の使用順位のズレは、概して1000語のレベルよりも大きいが5000語のレベルを超えるほどには大きくないということであろう。また、同形語における一字漢語と二字漢語を比べると、まず頻度順5000語までにおいては、同形語のうち二字漢語がほぼ4分の3を占めていることを指摘できる。相関はいずれもそれほど高くはないが、一字漢語における相関は.248と全体の相関よりもやや低い。これは一字漢語には極端に頻度の大きい語が日中の双方にいくつかあるためであろうと思われる。

宮島（1994）は「日中間で文体的ランクのちがう同形語がある」としながらも「大部分の同形語では、意味や文法的性質とともに、文体的性質も、むしろ似ている」（p.291）と述べ、その実例を挙げている。文体差に関しては、宮島の指摘は概ね正しいと思われるが、頻度の相違には文体差以外の原因も考えられる。日本語の相対使用頻度の高い語には際立った特徴はなく、「様（よう）」のように日本語において中国語にない用法を持つ語があることや、「円」「県」のように日本社会独自の面を反映する語を指摘できる程度である。しかしながら、中国語の相対使用頻度の高い語（z得点の負の値が大きい語）には著しい特徴が認められる。おおよそ以下の4種類に分けることができよう。

a．政治性の強い語･･･その言語社会の特徴を表す語
　例）「主義」「革命」「発展」「国家」
b．中国語においては機能語や代名詞の用法を持つもの
　例）「的」「不」「他」「個」「上」
c．中国語において日本語よりも高頻度の意味・用法をもつもの
　例）「好」（「よい」の意あり）、「説」（「言う」などの意あり）
d．日本語においてはややフォーマルだが、中国語においては日常的である語
　例）「継続」「出現」「増加」

上記のbはcの一種であるとも考えられるが、日本語漢字にはない特徴であるため別項目として考えたほうがよいと思われる。また、これらの複数の要素をかねている場合もある。「自己」は中国語では「自分（が／を／で･･･）」という意味･用法もあり、日本語の「自己」よりも日常語的であり、意味も広く、かつ機能語的な用法もあると考えられる。dに関しては、「継続」に対する「続ける」、「出現」に対する「現れる」、「増加」に対する「増える」「増やす」のように、日本語においては和漢対応がある語の対が中国語では一語であるケースが多いことを付け加えておきたい。日本語教育の現場においては、これらの要素に注意を払うことが必要であろう。

8．結論と今後の課題

本稿では以下の諸点について述べた。

・「日中対照常用漢字語データベース」の開発目的、作成手順と概要
・現在有力な二つの語彙頻度表、国立国語研究所(2006)と天野・近藤(2000)を比較すると、上位約1000語での一致語数は41.6%しかなく、相関(スピアマン)も.256しかない。その原因は前者の調査語数が十分に多くないことと、後者が新聞に偏ったコーパスを元にしていることに由来すると考えられる。
・漢語は5000語までの頻度レベルにおいても40%以上を占め、安定して用いられている。和語や外来語の割合が頻度レベルによって大きく異なるのと対照的である。
・5000語までの頻度レベルにおいて、表記上の日中同形語は全体の3分の1以上を占め、上位2000語では40%を超える。漢語においては85%以上を占め、上位2000語では90%を超える。
・5000語までの頻度レベルにおいて、日中同形語の約4分の3が二字漢語であり、約4分の1が一字漢語である。
・日中同形語の日中両語における頻度の相関は全体として.334であり、ある程度の相関がある。相関を下げる要因としては、中国語のほうに特徴的な意味・用法・使用域をもつ語があることが挙げられる。

今後はさらにデータベースの構築を進め、使用頻度と使用範囲の関係を見ながら専門性(特定分野への偏り)の高い語彙を特定すること、それらが同形語においてどのような分布を成しているか調べたい。その上で音韻・文字表記、意味・用法、使用域などについて類似度の分類を進め、中国語系と非中国語系の学習者の語彙発達の相違を明らかにするための語彙テストの開発の準備を進めたい。

謝辞

本稿は2008年度桜美林大学言語教育研究所より研究運営助成を受けて行なった研究によるものです。同形語の判断、データ入力に当たっては名古屋大学大学院生の謝文儀さんと劉佳珺さんにお世話になりました。記して感謝の意を表します。当然のことながら、本稿の内容については稿者が責任を負うものです。

注

[1] 英語教育においては、Nation(1990)のVocabulary Levels Test(p.261-272)を初めとする数種類の語彙テストが広く共有され、教育と研究に活用されている。
[2] 国立国語研究所(1962)によると1956年発行の雑誌90種で漢語の割合は異なり語彙で47.5%、延べ語彙でも41.3%を占め、茂木ほか(2005)は2003年発行の月刊雑誌50種につき漢語の割合は異なり語彙で44.9%、延べ語彙で37.5%と報告し、本稿で調査した国立国語研究所(2006)の頻度上位5000語における漢語の割合は異なり語彙で43.4%である。
[3] 漢字語の日中対照研究の成果として文化庁(1978)が先駆的な成果として知られているが、意味に関する大分類のみであり、その分類にも多くの問題が指摘されている(荒川

1979など)。

4 同様の指摘がChen and Henning (1985: 162) にも見られる。

5 日本学生支援機構 (JASSO) (2008) の統計によると、2008年5月1日現在、日本国内の外国人留学生のうち、中国と台湾の出身の学習者が合わせて62.9%を占めている。

6 基礎となる語彙表を国立国語研究所 (2006) とした理由は4.で後述する。なお、付属語には漢字語はないので付属語の語彙表は使用しない。

7 今回は漢語のみを採録しているが、一定の頻度レベルまで到達したら、漢字表記の和語や混種語についても採録していきたい。その際は語種のフィールドを追加する。

8 国立国語研究所 (2006) に基づくすべての語の中の頻度順位

9 共起成分の類似する語。後述するSketch Engine (Kilgarriff ほか) のThesaurus情報を利用する。

10 国立国語研究所 (2006) では、語種の分類として漢語·和語・外来語・混種語のほかに、人名・地名・その他が立てられているが、人名・地名以外の固有名詞は漢語起源であれば漢語に分類されるため、「三和」も漢語となっているものと思われる。

11 正確には409,384,405語

12 情報通信研究機構の李在鎬 (イ・ジェホ) 氏にご教示いただいた。

13 天野・近藤 (1999) の単語親密度は単語の基本度を測る上で重要な指標であり (川村2006: 76)、現実の頻度を反映していると考えられるが、使用実態の調査に基づく語彙表ではなく、書き言葉と話し言葉の頻度の両方を反映していると考えられる上、辞書の語彙のみを対象としており、高頻度の新語、俗語等を採録していないと考えたため検討しなかった。

14 朝日新聞14年分を収録したコーパスに基づいており、延べ語数は明らかにされていないが、異なり語数だけで約34万語あり、徳弘 (2008) をもとにして漢字表記できる語のみ頻度を合計しても8000万語を超えるので、助詞などの機能語を含めれば延べで1億語は軽く超えるものと考えられる。これは国立国語研究所 (2006) の1,065,617語の100倍程度に相当する。

15 徳弘 (2008) には、CD-ROMデータの説明として、天野・近藤 (2000) と他の参照資料の共通しない語が抜けていることがある (p.442) と説明されている。

16 頻度はノンパラメトリックであると考えられるので、相関の計算にはスピアマンの順位相関を用いた。

17 国立国語研究所 (2008) (通称BCCWJ2008モニター版) は約2800万語という大規模の均衡コーパスであるが、まだ語彙頻度表は公開されてない。語彙頻度表が公開されれば有力な資料になるものと思われる。

18 実際には同一順位に同頻度の複数の語があるため順位と累計語数は必ずしも一致しない。

19 山崎・小沼 (2004) にも類似のデータがあるが、中間発表の集計であり、本データとは若干異なるうえ、グラフのみで数値がないので、改めて本稿で算出した。なお、同調査の

語種の判断は「かたりぐさ」というソフトウェアによっている（茂木ほか2005）。

[20] 山崎・小沼（2004）にも同様の指摘がある。また、同論文は1956年発行の雑誌90種を調査対象とした国立国語研究所（1959）と比較して外来語の割合が著しく増えていることを指摘している。一番外来語の割合が低い上位1000語に限定しても、1.2%しかなかった外来語が1割を超えるまでになっている。国立国語研究所（1962）では含まれていなかった広告が調査対象に含まれていることを差し引いても外来語が、少なくとも雑誌において著しく増えているということがいえるであろう。異なり語数における漢語と和語の比率は国立国語研究所（1962）から大きく変わっていない。延べ語数では漢語の使用割合が和語に比べて若干増えている（山崎・小沼2004）。

[21] 天沼（1981）の作成した日中同形の漢字の表によれば常用漢字1945字のうち、1090字（56.0%）が日中同形である。兒島（2003）は1165字（59.9%）を同形としている（p.72）。字体の異なるものも糸偏のように形の似ているものも多く、また同じパタンが多くの字に現れる場合も多いため、一部の著しく字体の異なるものを除けば字体の相違は少なくとも上級学習者には認知処理に影響を及ぼさない（玉岡・松下1999）。

[22] 認定に迷うものもかなりあり、最終判断が稿者と異なるものもあり、正確に認定するためにはより広く母語話者の判断の調査を行なう必要があるが、今後の課題としたい。

[23] 上位1000語に同形語が多いことには、語単位の区切り方の影響もある。同語彙表では数詞は例えば「二」と「二十」は別語となっている。これらの数詞46語を構成要素に分解して中国語の単位とそろえた場合：一〜十までの10字と零、百、千、万をあわせた14語となる。そのほかにも「分」（フン・ブン）「中」（チュウ・ジュウ）など五つの接辞・語は複数に数えられており、これらの同一表記の別語11語は中国語の単位にそろえると5語となる。したがって、中国語の単位にそろえて数えた場合、総語数は964語となる。その場合、上位964語（100%）において、主観的判断による同形語は、漢語423語（43.9%）のうち385語（39.9%、漢語423語の91.0%）ということになる。同語彙表の数え方に比べると若干少なくなるが、それでもかなり多いことに変わりはない。

参考文献

Chung T. M. (2003). A corpus comparison approach for terminology extraction. *Terminology,* 9 (2), 221-246.

Chen and Henning (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing,* 2 (2), 155-163.

Cobb, T. (2000). One Size Fits All? Francophone Learners and English Vocabulary Tests. *Canadian modern language review,* 57 (2), 295-324.

Elder C. (1997). What does test bias have to do with fairness? *Language Testing,* 14 (3), 261-277.

Kilgarriff A., Rychlý P. and Pomikálek J. *'Sketch Engine.'*
http://www.sketchengine.co.uk/ より利用可（2009年1月29日確認）

Nation, I.S.P. (1990). Teaching and Learning Vocabulary. New York: Newbury House.

Ringbom, H. (2006) *Cross-linguistic Similarity in Foreign Language Learning.* Ebook Library: Multilingual Matters.

Swan, M. (1997). The influence of the mother tongue on second language vocabulary acquisition and use. In N. Schmitt and M. McCarthy (Eds.) *Vocabulary* (pp.156-180). Cambridge: Cambridge University Press.

天沼　寧（1981）「中日漢字字体対照表」『大妻女子大学文学部紀要』13, 59-82

天野成昭、近藤公久編著（1999）NTTデータベースシリーズ『日本語の語彙特性』第1期（第1巻～第6巻）、書籍＋CD-ROM版、三省堂（CD-ROM版2003）

天野成昭、近藤公久編著（2000）NTTデータベースシリーズ『日本語の語彙特性』第2期（第7巻）、書籍＋CD-ROM版、三省堂（CD-ROM版2003）

荒川清秀（1979）「中国語と漢語　―文化庁『中国語と対応する漢語』の評を兼ねて」『愛知大学文学会文学論叢』62、1-28

荒屋　勤（1983）「日中同形語」『大東文化大学紀要　人文科学』21, 17-29

加藤稔人（2005）「中国語母語話者による日本語の漢語習得 ―中国語との対応のしかたによる漢語習得過程の違い―」『日本語教育』125, 96-105

茅本百合子（1995）「同一漢字における中国語音と日本語の音読みの類似度に関する調査」『広島大学日本語教育学科紀要』5, 67-75

川村よし子（2006）「日本語学習者のための基本語選定の一試案」『ヨーロッパ日本語教育』11, 72-78

国立国語研究所（1962）国立国語研究所報告21『現代雑誌九十種の用語用字 第一分冊 総記および語彙表』秀英出版

国立国語研究所（2006）『現代雑誌200万字言語調査語彙表』公開版（ver.1.0）
http://www2.kokken.go.jp/goityosa/index.html よりダウンロード可（2009年1月26日確認）

国立国語研究所（2008）『現代日本語書き言葉均衡コーパス』モニター公開データ

兒島慶治（2003）「日本・中国・台湾・香港における漢字字体の共通性と相違性」『比較文化研究』62, 63-74

曾根博隆（1988）「日中同形語に関する基礎的考察」『明治学院論叢』424, 61-96

高野繁男・王　宝平（2002）「日中現代漢語の層別　―日中同形語に見る―」神奈川大学人文学研究所編『日中文化論集』118-139、勁草書房

玉岡賀津雄・松下達彦（1999）「中国語系日本語学習者による日本語漢字二字熟語の認知処理における母語の影響」第4回国際日本語教育・日本研究シンポジウム「アジア太平洋地域における日本語教育と日本研究：現状と展望」（香港理工大学）、配布資料

陳　毓敏（2003）「中国語を母語とする日本語学習者の漢語習得について　―同義語・類義語・異義語・脱落語の4タイプからの検討―」『2003年度 日本語教育学会秋季大会 予稿集』174-179、日本語教育学会

徳弘康代編著（2008）『日本語学習のためのよく使う順漢字2100』三省堂

日本学生支援機構（JASSO）（2008）「平成20 年度　外国人留学生在籍状況調査結果」
http://www.jasso.go.jp/statistics/intl_student/documents/data08.pdfよりダウンロード可（2009年1月31日確認）
文化庁（早稲田大学語学教育研究所日本語科）（1978）『中国語と対応する漢語』大蔵省印刷局
北京语言学院语言教学研究所（1986）《现代汉语频率词典》北京语言学院出版社
松本順子・堀場裕紀江（2007）「日本語学習者の語彙知識の広さと深さ」『第二言語としての日本語の習得研究』10,10-27
宮島達夫（1994）「日中同形語の文体差」『語彙論研究』283-299、むぎ書房（初出は『阪大日本語研究』4）
茂木俊伸・山口昌也・丸山岳彦・田中牧郎（2005）「語種辞書『かたりぐさ』の開発と月刊雑誌の語種構成分析」『言語処理学会第11回年次大会発表論文集』341-344
山崎　誠・小沼　悦（2004）「現代雑誌における語種構成」『言語処理学会第10回年次大会発表論文集』670-673
山本富美子（1994）「上級聴解力を支える下位知識の分析―その階層化構造について―」『日本語教育』82, 24-46