〔2017 年度日本語教育学会秋季大会（朱鷺メッセ，2017.11.26）発表・ポスター発表⑮〕【予稿集】

日中対照漢字語データベースの開発と応用

松下達彦（東京大学）・陳夢夏・王雪竹・陳林柯（一橋大学大学院生）

1. はじめに

1970 年代以降，文化庁(1978)をはじめ，漢字語の対照研究が数多く行われてきた。1990 年代以降, 中国語母語学習者の日本語漢字語の認知や習得に関する研究(茅本 1996, 邱 2002, 加藤 2005）も増えてきたが，成果が十分に教育現場で応用されているとはいえない。その理由として，教育現場で教材中の漢字語の中国語との違いなどを逐一検索することが困難だということが考えられるだろう。

日本語と中国語の同一表記による漢語の意味を対比する日中両国初めての辞典である張（1987）をはじめ，意味・用法の異同の記述的研究は数多くある（唐 1993，王・許・小玉 2007 など)。ただし，これらの研究はすべて記述のみであるため，辞書としての価値がある一方，意味対応のパターンを量的に詳しくとらえたものではない。また，文化庁（1978）は意味の対応関係に基づいて分類したものであるが，その分類は現在の使用事実と対応しないところが見受けられる（荒川 1979，飛田・呂 1986，周 1986，岡 2002)。加えて，これらの研究は電子的な検索が可能なデータベースになっていない。

朴・熊・玉岡（2014）は旧 JLPT の２字漢語に限定し，約 2000 語を取り上げ，すでにある辞典にしたがってその意味，品詞性などをデータベースにしたが，主に品詞情報を掲載し，エクセルデータも 2017 年 9 月時点で公開されていない。

同形語の数や同形語の意味対応を数量的に扱っている研究もある。例えば，中国語教育の立場から，曽根（1988）は中国語の語彙頻度表と日本語辞書を参照し，上位 8441 語から単音節語を除いた 6112 語中，56％を同形語と認定している。同様に高野・王（2002）は中国語頻度表と日本の高校教科書の語彙を比較し，日本語上位 3000 語中，41％を同形語と認定している。日本語教育のほうでは，松下（2009; 2011），Matsushita (2012) が日本語の雑誌と書籍のコーパスの上位 5000 語中, 雑誌で 38%, 書籍で 43％を同形語と認定している。しかし，これらはコーパスの規模も現在の水準に比べると小さく，対象の異なり語数も 1 万語に満たない。

本研究では，日本語の学習・教育・研究により直接的に役立てることを目指して日中対照漢字語データベースを開発した。本データベースでは意味対応の判定結果のほか，字体の違いや意味のズレなども検索でき, 授業の予復習, テスト作成などに活用できる。また，本研究では，日本語の語彙における，漢語（字音語）の日中両語の意味対応パターンの量的分布を調査した。文化庁（1978），三浦（1984）に基づいて同形語の意味対応のパターンを同形同義，同形異義，同形類義 3 種類（日本語の意味範囲が広い語，中国語の意味範囲が広い語，日中両語に独自義のある語）の 5 種類に分け，非同形を加えた計 6 種類に分類し，その対応パターンの語彙頻度レベル別，品詞別分布などを報告する。

2. データベースの内容および作成方法

本研究で開発したデータベースは，2017 年 9 月現在，『日本語を読むための語彙データベース』（松下 2011）の留学生用語彙ランクの上位 2 万語に含まれる漢語 10054 語を含めており，以下の項目からなる。

ID，留学生用語彙ランク（松下 2011），見出し語，標準的（新聞）表記，文字数，標準的読み方（カタカナ），品詞，語種，中国語表記，意味対応，日本語と中国語に共通の意味／用例，中国語のみに存在する意味・用法，日本語のみに存在する意味・用法，日本語独自の意味の類推可能性。

データベースの開発に当たり，字体の対応は天沼（1981）を参照し，共通の康熙字典体（中国語の繁体字とも共通の，いわゆる旧字体を含む字体）から派生した字体（例えば康熙字典体「澤」から派生した日本語漢字「沢」と中国語漢字〈泽〉）は同形として扱った。（中上級学習者の脳内で対応字体が概ねリンクされていることは茅本（1996）等で明らかになっている。）同形語の意味対応は同形同義，同形異義，同形類義 3 種類（日本語の意味範囲が広い，中国語の意味範囲が広い，日中両語ともに独自義あり）とし，非同形（中国語に存在しない）を加えた 6 種類に分類した。

判定作業は以下のように行った。1) 日本で日本語学や日本語教育を専攻する中国語母語の博士課程大学院生 3 名 A, B, C のうち，2 名が独立して第 1 次判定を行った。2) 判定が一致したものは確定し，判定が一致しないものは独立した第 3 判定者が第 2 次判定を実施し，第 1 次判定の一方と一致した場合はそれで判定を確定した。3) 第 2 次判定で 3 名の判定がすべて異なった場合は, ABC の 3 名のうちの 2 名以上が合議の上, 第 3 次判定を行い，判定を確定した。

なお，非同形語および日本語に独自義のある同形類義語の計 3 パターンについては，中国語母語話者による日本語の意味推測の可能性について，高（文脈なしで正しく意味推測できる可能性が高い），中（意味がわかれば覚えやすいが，文脈がない場合，正しい意味推測が難しい），低（意味推測は難しく，中国語の意味は日本語の意味を覚えることに貢献しにくい）の 3 段階で判定を実施した。

現代中国語において, 漢字語の存在や意味が現代日本語や台湾･香港などの一部の中国語地域の影響を受けることもあり, 揺れが存在するため, 中国語の意味の判定に当たっては，現代の平均的大学生の語彙知識を基準とするように意識することにした。また，意味･用法の確認の必要がある場合には検索サイト Goo の辞典を主に参照したが，これらのもとになっているデータは，『デジタル大辞泉』（松村明監修，小学館），『デイリーコンサイス中日辞典』（第 2 版）（杉本達夫・牧田英二・古屋昭弘編，三省堂）である。

3. データベースの分析結果と考察

3-1. 作業過程での一致率

表 1 は第一判定の 3 組の判定者ペアの一致語数，一致率, Kappa 係数である。一致率は 76.0％で, Kappa 係数（-1~1 に分布し，1 に近いほど一致度が高いと

表 1　第1判定者ペア別一致語数･一致度

	A-B	A-C	B-C	合計
判定語数	3352	3351	3351	10054
一致語数	2567	2580	2490	7637
一致率 (%)	76.6	77.0	74.3	76.0
Kappa 係数	0.55	0.58	0.52	

表２　第１判定で不一致だった語の不一致パターン別／判定者ペア別の語数・割合　＊A～C は判定者を表す

	</=	</≠	</>	</><	</φ	=/≠	=/>	=/><	=/φ	≠/>	≠/><	≠/φ	>/><	>/φ	></φ	計
A･B不一致	86	10	8	5	17	61	103	23	365	7	7	44	6	34	9	785
A･C不一致	95	12	10	5	12	65	108	12	341	19	8	50	6	27	1	771
B･C不一致	70	9	9	4	15	82	157	22	357	23	9	52	12	32	8	861
不一致合計	165	21	19	9	27	147	265	34	698	42	17	102	18	59	9	1632
不一致語数に占める割合 (%)	10.1	1.3	1.2	0.6	1.7	9.0	16.2	2.1	42.8	2.6	1.0	6.3	1.1	3.6	0.6	100.0

される）は，0.5~0.6 の間のレベルにあり，ある程度一致している水準であるが，判定が易しくないことを示している。三つの Kappa 係数に差はなく（Z=-1.67, Z=-1.87, Z=-3.58，いずれも有意差なし）(1)，特定の判定者の影響はないと考えられる。

第１判定で一致した語の 68.1%はいわゆる同形同義語で，同形同義語が漢語に占める割合 60.1%（表 3）よりも多い（χ=26.830, *df*=1, *p*<.001）ので，判定が一致しやすいと言えるであろう。第１判定で一致した語の 29.5%は非同形語であり，これら２種類で 97.6%を占める。同形語は 86%が，非同形語は 75.0%が第１判定で判定が確定している。表 2 は第 1 判定で不一致だった語の不一致のパターンを示したものである。不一致のうち，56.4%（表中のφの割合の合計）が，判定者のどちらかが非同形語と判断した語である。そもそも中国語に存在するかどうかでかなり判断が分かれることがわかる。日中両語のどちらかの意味が広い語は，第２判定で確定するものが多いが，それぞれに共有義と独自義を持つパターン（＞＜）は最終判定までもつれ込んだものが多い。これはかなり上級レベルの判定者であっても両語に存在する意味を認定することが容易でないことを示す。

3-2. 日本語語彙に占める漢語，同形語の割合，および同形語に占める意味対応の分布

判定の結果，頻度上位２万語のうち，50%が漢語で，漢語の 70%（全体の 35%）が同形語で，30%（全体の 15%）が非同形語であることがわかった。中国語起源でありながら現代中国語に存在しない非同形語が意外に多い。同形語 7047 語のうち，86%（全体の 30%，

表３　上位２万語以内の漢語の中国語との意味対応パターン別／確定段階別の語数･割合（対象語数: 10054）

＊「日＞＜中」は共通義のほかに日中両語に独自義があることを示し，「φ」は非同形語であることを示す。

	同形語					非同形語			
[日本語__中国語]の意味対応 (*)	日<中	日＝中	日≠中	日>中	日><中	φ			
語例	出身	不足	裁判	世間	単位	苦労	計	確定率	累積計 (%)
第1判定確定	45	5195	98	41	4	2254	7637	76.0	76.0
第2判定確定	105	769	144	133	23	662	1836	18.3	94.2
第3判定確定	72	75	133	122	88	91	581	5.8	100.0
計	222	6039	375	296	115	3007	10054	100.0	100.0
上位2万語に占める割合 (%)	1.1	30.2	1.9	1.5	0.6	15.0	50.3	--	--
漢語に占める割合 (%)	2.2	60.1	3.7	2.9	1.1	29.9	100.0	--	--
同形語に占める割合 (%)	3.2	85.7	5.3	4.2	1.6	--	100.0	--	--

表 4　上位２万語以内の漢語の中国語との意味対応パターン別／頻度レベル別の語数・割合（対象語数: 10054）

	同形語					非同形語			
[日本語＿中国語]の意味対応 (*)	日<中	日＝中	日≠中	日>中	日><中	φ	全体		
語例	出身	不足	裁判	世間	単位	苦労	計		累積計 (%)
初級(0001-1315位)	19	215	43	19	21	94	411	4.1	4.1
中級前半(1316-4000位)	71	1082	60	70	35	178	1496	14.9	19.0
中級後半(4001-7000位)	21	1018	60	47	22	418	1586	15.8	34.7
上級(7001-20000位)	111	3726	212	160	37	2317	6563	65.3	100.0
計	222	6041	375	296	115	3007	10056	100.0	100.0
初級(0001-1315位) (%)	4.6	52.3	10.5	4.6	5.1	22.9	100.0		
中級前半(1316-4000位) (%)	4.7	72.3	4.0	4.7	2.3	11.9	100.0		
中級後半(4001-7000位) (%)	1.3	64.2	3.8	3.0	1.4	26.4	100.0		
上級(7001-20000位) (%)	1.7	56.8	3.2	2.4	0.6	35.3	100.0		
計	2.2	60.1	3.7	2.9	1.1	29.9	100.0		
初級(0001-1315位) (%)	6.0	67.8	13.6	6.0	6.6	--	100.0		
中級前半(1316-4000位) (%)	5.4	82.1	4.6	5.3	2.7	--	100.0		
中級後半(4001-7000位) (%)	1.8	87.2	5.1	4.0	1.9	--	100.0		
上級(7001-20000位) (%)	2.6	87.8	5.0	3.8	0.9	--	100.0		
計	3.1	85.7	5.3	4.2	1.6	--	100.0		

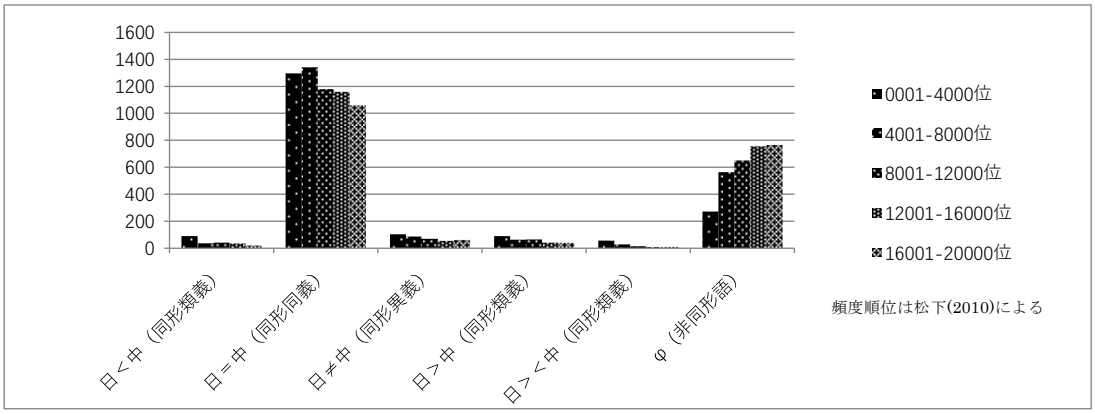

図１　上位２万語以内の漢語の中国語との意味対応パターン別／頻度レベル別の語数（対象語数: 10054）

漢語の 60%）が同形同義で，同形類義や同形異義といった要注意の語が約 14%（全体の 5%＝20 語に 1 語，同形語の 7 語に 1 語）であることなどが明らかになった（表 3）。

表 4 と図 1 は意味対応パターンを頻度レベル別に見た語数や割合である。同形同義は初級には相対的に少なく（52.3%），中級前半で一番多い（72.3%）。逆に，非同形語は初級で 20%を超えるが，中級前半で 12%程度となり，中級後半以降，再び増加し，上級では 3 分の 1 を超える。同形語だけを見ると，同形同義が中級以上では 82～88%程度で安定しているが，同形類義は減少する傾向がある。低頻度語ほど単義的になるためであろう。

表 5　上位２万語以内の漢語の中国語との意味対応パターン別／品詞別の割合（%）・語数（対象語数: 10054）

品詞／[日本語＿中国語]の意味対応 (*)	語例	日<中	日＝中	日≠中	日>中	日><中	φ	総計(%)	総語数
副詞	一旦	1.4	31.5	12.3	6.8	2.7	45.2	100.0	73
名詞-普通名詞-副詞可能	多数	4.1	49.5	6.1	8.2	1.5	30.6	100.0	196
接頭辞	超	14.9	73.6	4.6	1.1	3.4	2.3	100.0	87
接尾辞	化 団 歩	11.5	64.4	0.8	5.5	8.3	9.5	100.0	253
形状詞-一般	単純 漠然	2.8	43.8	8.1	4.0	1.2	40.1	100.0	322
名詞-普通名詞-形状詞可能	面倒	1.6	62.3	5.7	3.1	2.3	24.9	100.0	385
名詞-普通名詞-サ変可能	集中	1.4	58.3	4.3	3.0	0.7	32.2	100.0	3203
名詞(普通名詞、固有名詞、数詞)	医療	2.0	62.3	2.9	2.5	0.9	29.4	100.0	5528
その他(接続詞、代名詞など)	乃至 僕	0.0	14.3	28.6	0.0	0.0	57.1	100.0	7
総計		2.2	60.1	3.7	2.9	1.1	29.9	100.0	10054

表 5 は意味対応パターン別／品詞別の割合を示している。ここで目立つのは，「一旦」のような副詞や「漠然」「単純」のような形状詞（いわゆるナ形容詞や「～とした」の形で形容語的に使う語）に非同形語が多いということである。接頭辞や接尾辞に中国語のほうが意味範囲の広いものが多く，非同形語が少ないということも言えそうである。

4. まとめと今後の課題

漢字語の中国語との対応を定める作業は，一部の語彙については容易でないが，所定の手続きに従って判定した結果，学習・教育上，要注意の同形類義や同形異義が全体の約 5%であることなど，意味対応パターンの割合や頻度レベル別，品詞別傾向が明らかになった。

本研究で開発した日中対照漢字語データベースの意義は，以下の 2 点にまとめられる。第一に，これまで同形語の意味記述に関する書籍等はあったが，データベースは少なかった。本データベースは教師，学習者，研究者が語の検索など直接利用できるほか，J-LEX（菅長・松下 2014）のような語彙頻度プロファイラーに搭載し，文章の語彙的負荷を中国語母語と非母語に分けて表示する機能や，中国語母語学習者にとっての要注意点を画面表示する機能に応用できる。第二に，漢語における意味対応パターンの量的側面の先行研究はいずれも数千語レベルにとどまっており，2 万語レベルの規模で明らかにした資料はこれが初めてだと思われる。

データベースの判定や記述には，まだ改良の余地がある。随時，アップデートする予定であるが，いつでも，以下の URL から利用できるようにしておく予定である。

URL: http://www17408ui.sakura.ne.jp/tatsum/database.html#cvd

謝辞 : 本研究は日本学術振興会科学研究費，基盤研究Ａ（課題番号 25244022，研究代表者：庵功雄），および基盤研究Ｂ（課題番号 17H02350，研究代表者：庵功雄）の助成を受けたものです。

注

（1） 村上拓彦氏作成の Kappa 係数マクロ（http://www.agr.niigata-u.ac.jp/~muratac/）による

参考文献

(1) 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, p.1-28

(2) 岡益己（2002）「日本経済語彙における日中両国語間でのずれについて」『日本語教育』113, p.63-79

(3) 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, p.96-105

(4) 茅本百合子（1996）「日本語漢字と中国語漢字の形態的･音韻的差異が中国語母語話者による日本語漢字の読みに及ぼす影響」『広島大学教育学部紀要』45, p.345-352

(5) 邱学瑾（2002）「漢字圏・非漢字圏日本語学習者における漢字熟語の処理過程―意味判断課題を用いた形態・音韻処理の検討―」『教育心理学研究』50(4), p.412-420

(6) 周錦樟（1986）「日中漢語対応の問題―文化庁『中国語と対応する漢語について』―」『日本語日本文学』12, p.69-89

(7) 菅長陽一・松下達彦（2014）「オンライン日本語テキスト語彙分析器 J-LEX」URL：http://www17408ui.sakura.ne.jp/tatsum/webtools.html#jlex（2017 年 9 月 30 日参照）

(8) 杉本達夫・牧田英二・古屋昭弘編（2005）『デイリーコンサイス中日辞典』第 2 版，三省堂，URL: https://dictionary.goo.ne.jp/cj/（2017 年 9 月 30 日参照）

(9) 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, p.61-96

(10) 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』p.118-139　勁草書房

(11) 張淑栄（1987）『中日漢語対比辞典』ゆまに書房

(12) 飛田良文･呂玉新（1986）「『中国語と対応する漢語』を診断する」『日本語学』5(6)(44) , p.72-84

(13) 唐磊（1993）『現代日中常用漢字対比詞典』北京出版社

(14) 文化庁（1978）『中国語と対応する漢語』大蔵省印刷局

(15) 松下達彦(2009)「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―」『桜美林言語教育論叢』5, p.117-131

(16) 松下達彦（2010）「日本語を読むための語彙データベース」URL：http://www17408ui.sakura.ne.jp/tatsum/database.html#vdrj（2017 年 9 月 30 日参照）

(17) 松下達彦（2011）「複数の語彙リストの比較による日本語の常用語に含まれる日中同形漢語の量的検証」第 3 回北東アジア言語教育学会発表資料

(18) 松村明監修『デジタル大辞泉』小学館，URL: https://dictionary.goo.ne.jp/jn/（2017 年 9 月 30 日参照）

(19) 三浦昭（1984）「日本語から中国に入った漢語の意味と用法」『日本語教育』53, p.102-112

(20) 朴善媚・熊可欣・玉岡賀津雄（2014）「同形二字漢字語の品詞性に関する日韓中データベース」URL：http://kanjigodb.herokuapp.com（2017 年 9 月 30 日参照）

(21) 王永全・許昌福・小玉新次郎（2007）『日中同形異義語辞典』東方書店

(22) Matsushita, T. (2012) In What Order Should Learners Learn Japanese Vocabulary? A Corpus-based Approach. PhD Thesis, Victoria University of Wellington. URL: http://hdl.handle.net/10063/4476（2017 年 9 月 30 日参照）