コンピュータと教育 62－8
（2001．12．14）

日本語学習者のための慣用句データベースの作成
–統計処理を用いた一手法の提案–

ダニー・ミン† 佐野洋‡
東京外国語大学 大学院地域文化研究科† 東京外国語大学 外国語学部‡

概要：語学学習の中でも，一般に慣用句の取得は学習者にとって難しい課題である．記憶の負担量に視点を置くと，日本語学習者が，イディオムを効率的且つ効果的に学習する方法は，頻繁に使われるイディオムを優先して習得することである．これを頻度順イディオムリストと呼ぼう．本稿は日本語学習者のための慣用句データベースの構築について述べる．日本語学習者の視点から従来の慣用句教本の問題点を指摘し，用例データを使う統計処理によるデータの抽出手法を提案し，頻度順イディオムリストの作成について述べる．

A Study of Japanese Idioms for Learners of Japanese
- A Statistical Approach to Making a Japanese Idiom Reference Database -

Danny MINN† SANO Hiroshi‡
Tokyo University of Foreign Studies Graduate School of Area and Cultural Studies†
Faculty of Foreign Studies‡

Abstract: For learners of second languages, the task of learning the proper usage of idioms is difficult. In order to learn idioms more efficiently and effectively, Japanese learners would benefit from starting with the most frequently used idioms. For this purpose, a ranking of idioms by their frequency would be beneficial. This paper will discuss how literary and newspaper corpora were used to develop an idiom database. Also, some problems with the currently available reference books about idioms, some proposed solutions, and the method of ranking idioms based on frequency are discussed.

1 はじめに

一般に，外国語の学習過程の中で，イディオムの習得は難しいとされる．イディオムは，慣用語や熟語，もしくは広義には語法や慣用法を指す．本稿では，イディオムを複数の単語から構成され，個々の単語の組み合わせの意味からは全体の語句の意味が推察され得ない単語連鎖(慣用語)と考える．

学習すべき外国語を日本語としよう．日本語学習者にとって，日本語イディオムを学習し，その意味と正しい用法を習得するには，多くの学習時間が必要である．(1) 単語連鎖を1つの意味の持つ語句として認識しなければならないこと，(2) 単語連鎖を構成する句が本来の構文機能を失っていて構成的に解釈が難しいこと，(3) 単語連鎖を構成する個々の単語の意味を組み合わせただけでは，語句全体の意味が分からないことなどが難しさの要因である．記憶の負担量に視点を置くと，日本語学習者が，イディオムを効率的且つ効果的に学習する方法は，頻繁に使われるイディオムを優先して習得することである．これを頻度順イディオムリストと呼ぼう．本稿は，コーパスを用いたイディオムリストの作成の試みについて述べ，我々が得た結果について報告する．

著者等は，イディオムリスト作成において2種類のコーパスを使った．文学作品コーパスと新聞データコーパスである．第2章では，研究目的について述べる．第3章は，3つのコーパスを使ったイディオムの頻度調査について説明する．種類の異なる2つのコーパスに共通に出現する語彙連鎖(イディオム)と，コーパス毎に違った分布をしている語彙連鎖(イディオム)があることが分かった．第4章では，日本語学習者に対するイディオムについての知識量調査を含めて，調査結果の考察を行う．5章で本稿をまとめる．今後，数多くの種類のコーパスを用いたイディオムリストの作成を実施し，日本語学習者のための学習素材としての基礎データとして整理する予定である．教材データバンクとして蓄積と流通を進める仕組みは将来の課題として残る．

2 本研究の動機と指針

2.1 背景となる動機

筆者の一人であるミンは，アメリカからの留学生である．2度の日本への留学を経験している．この経験を通じ，外国語として日本語を学ぶ過程の学習教材の評価とその検証を進めている．背景となる動機は，初・中級の日本語学習を終えた後の学習教材の貧困さに起因する．日本語の習得目標は個々に異なっている．初・中級の段階は，一般的に利用される基礎日本語の学習が有効である．しかしながら，次の段階では，個々の日本語の習得目標に適合する学習教材があることが望ましい．筆者自身，大学での学習と研究を遂行するために必要とする日本語学習教材を切望している．

日本語の学習教材の評価を行う基本の動機を次に示す．

1. 日本語を非母語 (Non-Native) とする日本語学習者[1]にとって，何故，イディオムは難しいのか．その要因を探る．
2. イディオムの学習に供される参考書には，どのような教材特徴があるか．
3. 必要とするイディオムを確認する方法があるのか．分野毎に利用頻度等の統計情報を利用して作成されたイディオムリストは存在するのか．
4. NNS の中・上級者はイディオムに関し，どの程度の知識を持っているのか．これは既存の学習教材の学習効果の測定を間接的に意図する．
5. イディオムの出現の生起数 (occurrence) は，文章の違い (ジャンルの違い) によって偏っているのか．もし偏っているとすれば，ジャンル毎にイディオムリストを作成するべきである．

[1]Non-Native Speaker. 以後，NNS とする．

2.2 関連調査

2.1節-1について考察した．学習の難しさを次に挙げる．

- (イディオムを構成する) 単語連鎖の個々の単語の意味が分かっても，イディオムの意味は不明である．従って，文全体の意味が不明瞭になる．
- イディオム構成の仕方やその表現する意味は，日本語の文法体系の仕組みに基づくのではなく，文化や社会習慣に基づく場合が多い．
- NNSは，どのようなイディオムが，(自分がいるコミュニティの中で) 頻繁に使われているか知らない．どのイディオムから学習を始めればよいか分からない．学習をガイドする教材がない．

2.1節-2について考察した．アメリカで流通するテキストブックをはじめ，日本に留学してから得た教材を調査した[宮地 82, Akiyama et.al 96, Garrison 90, Garrison et.al 94, Maynard et.al 96, Sasaki 93].

- イディオムの学習用参考書は，100個から2,000個ぐらいまでの慣用句が掲載されている．
- イディオムの意味が示され，英語の近似訳語や類語も説明されている．
- イディオムの意味の成り立ちが参考として掲載されている場合もある．
- イディオムを含む用例が掲載されている参考書もある．

テキストブックに共通することは，イディオム利用の特徴や傾向にジャンル毎の違いがある否か示されていないこと，及び，イディオムの利用頻度についてのデータが示されていないことである．学習者にとって，どのイディオムが最も学習する価値があるのか，学習順序を決める情報がないのである．

2.3 研究方針

上記の結果をもとに，2.1節-3の点を調査した．

- 日本人 (Native-Speaker) に尋ねる．
- 自ら必要となる分野の文章やテキストを対象に，目視確認によってイディオムを数える．

留学生として直感的には，日本人に聞けば，どのようなイディオムが頻繁に使用されているか知っていると思っていた．何人かのヒアリングを通じ，分かったことは，母語者であっても (1) 普段の読書量や認知レベルの違いによって頻繁に使うイディオムに偏りがあること，(2) 同じような意味を表すイディオムであっても，どちらのイディオムを普段用いているかは個人毎に違うこと，(3) 母語者は，必ずしも日本語文法を体系的に学んでいたり，外国語教育方法に精通している訳ではないので，留学生にイディオムについて適切な説明ができないことがある[2]．

次に，自ら必要となる分野の文章や論文，テキストを対象に目視確認でイディオムを探す試みを行った．個人目標に限定することで検索範囲は狭まるが，それでも，時間がかかり過ぎ，大変な作業となった．

上記の方法では，中・上級レベルからの学習効率が悪い．記憶の負担量に視点を置くと，日本語学習者が，イディオムを効率的且つ効果的に学習する方法は，頻繁に使われるイディオムを優先して習得することである．コンピュータを使ってコーパスデータを処理することで，イディオムの利用傾向と特徴が客観的に分析できる．処理速度も高速であるから，分野毎にイディオムリストを得ることができる．こうしてイディオムリストから，日本語学習者だけでなく教師も，学習者の言語能力や分野毎に，どのイディオムを習うべきか判断することができるだろう．効率的にイディオムを学習するためには，学習

[2]イディオムの利用について漠然とした認識はあるようだが，利用傾向や特徴を具体的に説明できる母語者は少ないようだ．

者は必要とする分野で頻繁に使用されているイディオムを優先して学習することがよい.

3 イディオムリスト

3.1 イディオムの類型と予備実験

イディオムテキスト [Sasaki 93] を参照して, 300 個のイディオムを, 頻度操作を加えていない (初期) イディオムデータとして用意した. 次に, この 300 個のイディオムをコーパスから抽出し利用傾向を調査する. コーパスは, 文学ジャンルとして新潮文庫 (CD-ROM 版) と青空文庫 (CD-ROM 版), 新聞ジャンルとして毎日新聞 (2000 年 CD-ROM 版) を用いた.

新潮文庫 (CD-ROM 版) は, 67 の作品が含まれている. データ量は, 約 1,050 万文字である. 青空文庫 (CD-ROM 版) は, 1,152 の作品を含み, データ量は, 約 2070 万文字である. 毎日新聞 (2000 年 CD-ROM 版) は, 2000 年 1 年分の記事が含まれている. データ量は, 約 6,514 万文字である (表 1 を参照). プログラム[3]を使い, (初期) イディオムデータ (300 個のイディオム) をコーパスから抽出した.

表 1: 使用したコーパスとその大きさ

略称	コーパス名	文字数
A	青空文庫 (CD-ROM 版)	20,699,461
S	新潮文庫 (CD-ROM 版)	10,528,384
M	毎日新聞 (2000 年 CD-ROM 版)	65,141,998
	合計	96,369,843

宮地 [宮地 82] は, イディオムの文法特徴を調べている. 概要を [宮地 82] から引用し, 表 2 に挙げる. 宮地が取りあげた 1,270 個のイディオム中 63 %は, 名詞 (Noun) +助詞 (Particle) +動詞 (Verb) パターン (NPV 型) である. その内訳は, 格助詞「を」によって結ばれているもの (N-を-V 型) は 57 %である.「に」と「が」によって結ばれているもの (N-に-V 型と N が V 型) は 20 %ずつである. その他のもの (N-で-V 型, N-から-V 型, N-と-V 型) は, いずれも少数である. 他の形は, 1,270 個のイディオム中 15 %は, 名詞+助詞+形容詞 (Adjective) パターン (NPA 型) である. その残りの 22 %は, 名詞+助詞+名詞パターン (NPN 型) とその他のパターンである.

表 2: イディオムの文法的な類型

イディオムの類型	1,270 個中	文法詳細	内訳
NPV 型	63 %	N を V	57 %
		N に V	20 %
		N が V	20 %
		N で/から/と V	3 %
NPA 型	15 %	N が A	
NPN 型とその他	22 %	N の N 等	

本稿では, 調査する文法構造として, 数が多いことが指摘されている名詞+助詞+動詞パターンを持つイディオムを選択した (NPV 型イディオム).

3.2 NPV 型イディオムとそのリスト

イディオムリストの有効性を検証する目的から, 日本語学習者の学習期間とイディオムに関する知識量を計測する (第 4 章を参照). アンケート調査を実施するためイディオム数を絞り込む. 300 個のイディオムデータ中, NPV 型イディオムは 127 個 (42.3 %) である.

次に, プログラムを用いて, 3 つのコーパスから NPV 型のイディオムを検索した. この調査からイディオムのコーパス内での分布 (利用頻度) を把握した. 紙面の都合上, コーパス別のトップ 20 リストを挙げる. 表 3 は, 毎日新聞 (2000 年 CD-ROM 版) から抽出した NPV 型イディオムのトップ 20 リストである. 尚, *記号は, 複数のトップ 20 リストに出現するイディオムである. 注には, コーパス名を略称で示している.

表 4 は, 新潮文庫 (CD-ROM 版) から抽出した NPV 型イディオムのトップ 20 リストである. 表 5 は, 青空文庫 (CD-ROM 版) から抽出した

[3]プログラムは, Perl 言語を用いて記述している.

表 3: 毎日新聞 (M) データトップ 20

順位	NPV イディオム	生起数	注
1	軌道に乗る*	395	M,S
2	手を出す*	194	M,S,A
3	手を打つ*	169	M,S,A
4	手を抜く	108	
5	頭角を現す	100	
6	赤字になる	90	
7	裏目に出る	89	
8	手が届く*	85	M,S,A
9	水を差す	75	
10	肝に銘じる	62	
11	口火を切る	58	
12	本腰を入れる	56	
13	舌を巻く*	51	M,S,A
14	首になる*	51	M,S,A
15	板につく	49	
16	舌鼓を打つ	47	
17	振り出しに戻る	41	
18	恥をかく*	39	M,S,A
19	峠を越す	38	
20	音頭を取る	35	

表 4: 新潮文庫 (S) データトップ 20

順位	NPV イディオム	生起数	注
1	手を出す*	74	M,S,A
2	恥をかく*	40	M,S,A
3	首になる*	35	M,S,A
4	相槌を打つ*	34	S,A
5	手を打つ*	30	M,S,A
6	有頂天になる*	30	S,A
7	首にする	23	
8	突拍子もない	22	
9	舌を巻く*	20	M,S,A
10	棚に上げる	17	
11	手が届く*	13	M,S,A
12	鼻にかける	12	S,A
13	軌道に乗る*	12	M,S
14	鼻につく	12	
15	愚痴をこぼす	12	
16	腑に落ちない*	11	S,A
17	けりをつける	11	
18	うやむやにする*	11	S,A
19	匙を投げる	11	
20	調子を合わせる*	10	S,A

NPV 型イディオムのトップ 20 リストである. 3 つの表から，文学ジャンルでは，共通するイディオムが多いことが分かる．それに対して，新聞ジャンルではイディオムの分布は違っている．ジャンルを問わず，共通して出現するイディオムがあることも見て取れる．これらのイディオムの特徴は身体表現が起源の意味を持つ語句である．内包する意味範囲が広い．そのため文章中での利用範囲も広く，コーパスのジャンルに依存せず出現頻度が高くなっているものを考えられる．

4 イディオムリストと学習知識

4.1 NNS 対象アンケート

2.1 節-4 は,「日本語学習者の中・上級者はイディオムに関し，どの程度の知識を持つのか」である．前章の結果，すなわち毎日新聞と新潮文庫のトップ 20 リスト[4]を使って，日本語学習者のイディオム習得についてのアンケート調査を実施した．

アンケートでは，留学生に対し，20 個のイディオムについて習得しているかどうかを次の形式で調査した．その形式は，調査用紙で提示する，それぞれのイディオムの横に 5 つの選択肢を書き，4 つの意味からイディオムの表現する意味として一番適当だと思われる番号に印をつけてもらうというものである．不明であるあるいは分からない場合には，残りの 1 つ (?記号の選択肢) を選ばせた．その他に，回答者の国籍，年齢，性別と日本語の学習期間も尋ねた．

NNS(国籍数 24) を対象に行ったこのアンケートの結果は次の通りである．

日本語学習期間 平均 5.45 年

学習期間のモード 4 年間

イディオム平均理解数 52 %

興味深いことに，調査で使った 2 つのコーパスのトップ 20 リストに高頻度で出現するイディオムでも理解されていない場合があった．例えば「手を出す」「手を打つ」「舌を巻く」は，平

[4]青空文庫のコーパスはアンケート調査途中に集計した．

表 5: **青空文庫 (A) データトップ 20**

順位	NPV イディオム	生起数	注
1	手を出す*	136	M,S,A
2	不意を打つ	58	
3	調子を合わせる*	55	S,A
4	恥をかく*	44	MSA
5	有頂天になる*	41	S,A
6	手を打つ*	35	MSA
7	手が届く*	34	MSA
8	腑に落ちない*	32	S,A
9	合点が行く	30	
10	相槌を打つ*	30	S,A
11	暇を潰す	29	
12	物にする	27	
13	焼餅を焼く	27	
14	首になる*	27	M,S,A
15	図に乗る	26	
16	うやむやにする*	25	S,A
17	愛想が尽きる	25	
18	舌を巻く*	25	M,S,A
19	鼻にかける*	21	S,A
20	涙を呑む	21	

均 49 %の理解数であった．この 3 つのイディオムは新聞と文学両方のジャンルに頻繁に出現する．

4.2 教材内容と習得仮説

ある教材を選択した学習者は，その教材内容に従って，長く勉強したり学習したりすればするほど，習得知識は向上すると考える．外国語学習であれば，学習時間に比例して理解できる事柄が増えることを期待する．これは一般的で且つ原則的な習得仮説である．

前節で示したアンケート結果は，日本語学習者のイディオム学習に関して，この習得仮説が成り立っていないことを示している．回答者の中には，比較的短い勉強期間 (4 年間以下) にも関わらず，20 個のイディオムをほぼ全て知っている者がいた．一方，比較的長く (6 年間以上) 勉強しているにも関わらず，知っているイディオム数が少ない者もいた[5]．

図 1 は，一般的で且つ原則的な習得仮説に基づき，学習期間を横軸にとり，イディオム知識量 (スコア) を縦軸にとった学習直線である．

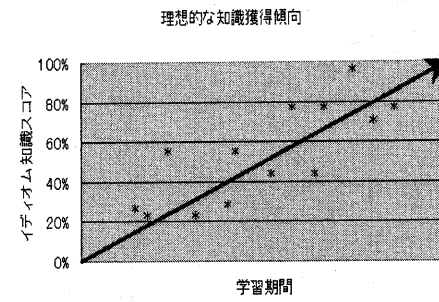

図 1: **理想的な知識獲得傾向**

アンケート結果から得られたイディオム知識量 (スコア) を図 2 に示す．図で示すように図 1 に示すような分布を示さない．回答者のイディオム知識量 (スコア) は，幅広く分散している．漢字圏からの留学生を限定して調べたところ，やはりイディオム知識量 (スコア) は幅広く分散している．さらに，アンケートで明らかになったことは，最も頻繁に使われているイディオムを知っている中・上級者は少ないということである．

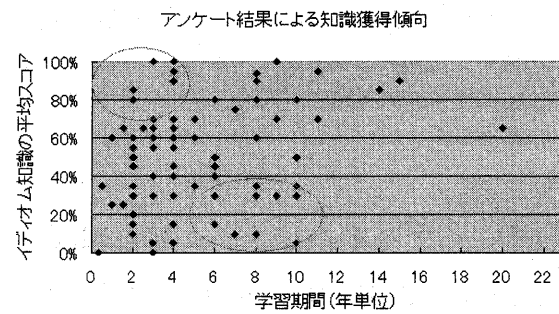

図 2: **アンケート結果による知識獲得傾向**

日本語を学習する中・上級者にとって，イディオムの習得は重要である．2.3 節では，日本語母語話者に尋ねてもあまり学習効果がないこと，自らがイディオム学習帳を作成することは非効

[5]勿論，学習ペースや時間あたりの学習量には，個人差があるし記憶力も個々で違う．

率であることを示した．更に，前節で示したアンケート調査の結果は，イディオム学習教材の内容が，日本語におけるイディオムの実際の運用の様態を明確に反映していないことを示していると考えられる．

我々の主張は，イディオムの実際の運用の有様を的確に反映した教材を使えば，学習時間に比例してイディオム知識量を (理想的には) 直線的に伸ばすことができるということである．

少なくともトップ20リストに挙がるイディオムは，新聞または文学的作品によく出現するものである．但し，日常会話に，こうしたイディオムが使われるか否かは，この結果からは分からない．しかし，成人の日本語母語話者であれば，例外なくトップ20リストに挙がるイディオムを全て知っている (だろう)[6]．

4.3 ジャンルの細分化

イディオムの出現の有様が細かなジャンル毎に偏っているならば，イディオムリストの作成は，ジャンルに分けて実施する必要がある．これは前節で主張したように，イディオムの実際の運用の有様を的確に反映するためで，その結果，学習効率が向上することが期待される．

表6と表7は，新聞ジャンルと文学ジャンルに頻出するイディオムである．

表 6: **新聞ジャンルの頻出イディオム**

手を抜く，頭角を現す，赤字になる，裏目に出る，水を差す，肝に銘じる，口火を切る，本腰を入れる，板につく，舌鼓を打つ，振り出しに戻る，峠を越す，音頭を取る

新聞のデータをさらに詳しく分類した．例えば，イディオムリストの内，最も頻度の高い「軌道に乗る」の場合，毎日新聞コーパスでは395回生起している．この出現回数の内，34％は経済ニュースに出現している．14％は国際ニュース，9％は一面記事ページに見られた (表8を参照)．日本語の新聞を読むことの多い日本語学習者，もしくは経済ニュースを好む日本語学習者は「軌道に乗る」というイディオムを学習の早い段階で習得した方が良いと推論できる．

表 7: **文学ジャンルの頻出イディオム**

有頂天になる，調子を合わせる，相槌を打つ，不意を打つ，腑に落ちない，うやむやにする，鼻にかける，合点が行く，暇を潰す，物にする，焼餅を焼く，図に乗る，愛想が尽きる，首にする

イディオムリストに挙がる「頭角を現す」の分布を見ると，毎日新聞コーパスに100回の生起があった．この出現回数の内，32％はスポーツニュースに出現し，14％は総合ニュース，10％は二面記事の中にあった (表9を参照)．このイディオムは，スポーツニュースや総合ニュースに出るので，日本語の新聞でも一般記事を読む際に有用である．こうしたジャンル毎の分析は，コンピュータとコーパスを利用すれば，イディオム1つずつに対して，簡単に，しかも短時間で適用できるものである．

表 8: **「軌道に乗る」のジャンル分布**

	経済	国際	一面
軌道に乗る	34％	14％	9％

表 9: **「頭角を現す」のジャンル分布**

	スポーツ	総合	二面
頭角を現す	32％	14％	10％

5 おわりに

本稿では，3つのコーパスとテキスト処理プログラムを使用し，127個のNPV型イディオムについて，頻度をもとにしたイディオムリストを作成した．従来型のイディオム参考書とは違う資料を作成した．また，アンケートの結果に拠ると，長期間，日本語の勉強をしているNNSでさえ，高頻度で使用されているイディオムを

[6]研究室内の日本人学生はトップ20リストの意味を全て知っていたし，一般の成人日本人にもその意味が分かることは十分推察できる．

知らないことが分かった．中・上級の日本語学習は効率的な学習手段と，自らの目標となる分野の学習教材を求めている．本稿で示した結果は，こうした要求に応えようとするものである．さらにジャンルを細分化すると，イディオム分布が変化することを確認した．

今後は，イディオムの種類を増やし，ジャンル別に分類された本格的なイディオムリストを作成する予定である．2000 個を越えるイディオムの量は，その入力作業にも時間を要する．コンピュータによる検索は高速だが，検索後のデータをチェックするのに時間を要するだろう．イディオムの電子化リストの入手の努力や，コンピュータプログラムの精緻化による検索制度向上などの手段によって，イディオムリスト作成作業を効率化したい．

次の課題は，検索対象のイディオムの数と種類，そしてコーパスの大きさと種類を増やすことである．こうした取り組みによって，総合的なイディオムリストデータが作成できるだろう．本研究は，日本語学習者と教師のための優れた日本語のイディオムデータベース，参考書を提供する試みである．

参考文献

[宮地 82] 宮地裕，『慣用句の意味と用法』，明治書院，1982 年.

[Akiyama et.al 96] Akiyama, Nobuo and Carol Akiyama,"2001 Japanese and English Idioms", Barron's Educational Series, Inc. (New York), 1996.

[Garrison 90] Garrison, Jeffrey G,""Body" Language", Kodansha International (Tokyo), 1990.

[Garrison et.al 94] Garrison, Jeff and Kayoko Kimiya, "Communicating with Ki, The "Spirit" in Japanese Idioms", Kodansha International (Tokyo), 1994.

[Maynard et.al 96] Maynard, Michael L. and Senko K. Maynard, "101 Japanese Idioms", Passport Books, NTC Publishing (Chicago), 1996.

[Sasaki 93] Sasaki, Mizue, "The Complete Japanese Expression Guide", Charles E. Tuttle Co. (Tokyo), 1993.