

## 科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 24 年 6 月 8 日現在

機関番号：27101

研究種目：若手研究（B）

研究期間：2009～2011

課題番号：21720190

研究課題名（和文）大規模コーパスを用いた日本語と英語イディオム・データベースの開発

研究課題名（英文）Development of a Japanese/English idiom database using large-scale corpus data

研究代表者

Danny MINN（ダニー・ミン）

公立大学法人北九州市立大学・基盤教育センター・准教授

研究者番号：60382412

研究成果の概要（和文）：

本研究の目的は大規模コーパスを用いた日本語と英語イディオム・データベースの開発することであった。慣用句リストの頻度調査のためのコーパスデータの収集を行った。そしてインターネット上の慣用句データベースを構築した（約1,500個の慣用句）。各慣用句について以下のような項目を整理した：読み方、ローマ字表記、意味、慣用句の種類、頻度情報、用例。コーパスから抽出した慣用句の用例の量が膨大のためまだ現在訂正中である。

研究成果の概要（英文）：

The aim of this research was to develop a Japanese/English idiom database using large-scale corpus data. In order to investigate the frequency of a list of idioms, corpus data was gathered. Then, an idiom database (with about 1,500 idioms) was constructed on the Internet. For each idiom, the following items were included: reading, Romanized spelling, meaning, idiom type, frequency information, and example sentences. Because the amount of example sentences extracted from the corpus data was so huge, the editing process still continues.

交付決定額

（金額単位：円）

	直接経費	間接経費	合計
2009年度	1,100,000	330,000	1,430,000
2010年度	800,000	240,000	1,040,000
2011年度	600,000	180,000	780,000
総計	2,500,000	750,000	3,250,000

研究分野：人文学

科研費の分科・細目：言語学・日本語教育

キーワード：コーパス・慣用句・イディオム・教育工学・教材

## 1. 研究開始当初の背景

本研究では、日本語の慣用句(idioms)について、日本語学習者に対する効率的な学習を促す教材データの構築と、ウェブを通じた教材提供サイトの開発を行う。なお、慣用句は慣用表現や熟語、広義には語法や慣用法を指

す。本研究では慣用句の定義を、それが複数の単語から構成され、個々の単語の意味の組み合わせから全体の語句の意味が直接に推察され得ない表現で、かつ文型特徴として述語を含む語連鎖であるとする。

日本語学習に限らず、語学学習の中でも、一般に慣用句の習得は、学習者にとって大き

な負担である。慣用句の表現を学習し、その意味を知り、正しい用法を習得するには、文型学習や語彙学習に比べて、多くの学習時間を必要とする。その理由は、(1) 慣用句を1つの意味を持った語連鎖として認識しなければならないこと、(2) 慣用句を構成する表現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、などである。

これらの指摘に対する学習方法は、語彙連鎖を句表現として認識するために記憶すること、そして、記憶した語彙連鎖を含む文表現あるいは段落表現に数多く接し、文脈を認識することで、正しい用法を習得することである。前者は、慣用句数が増えるほど学習負担が増すが、定量化された項目数によって、ある程度、学習時間を推し量ることができる。それに対して後者は、慣用句を含む文や段落のヒット率が不確定なことから、習得までの学習時間的な見積もりが不明確で、恐らく『語学学習が好きな』学習者を除き、学習行為を諦めるのではないかと推測される。

効率的で効果的な語学学習を実現するためには、定性分析され、表現範囲や使用領域が明瞭である学習目標と、学習目標ごとに定量化された学習項目と学習方法、そして見積もられた学習項目から積算される学習時間を明示することが必要である。

本研究は、学習項目の定量化と学習方法の改善に関する。その特徴は、(1) 分野ごとのコーパスを用いて慣用句の頻度情報を求め、(分野という表現範囲や使用領域を明示した上で)学習項目数を限定すること、(2) 慣用句学習において、単純記憶だけに頼るのではなく認知言語学的アプローチを用い、概念近接性に基づく類推方法を示し、慣用句が持つ意味を理解させることで学習効果を高めることである。

Danny MINN(研究代表者)は、大学院での研究(2000年~2002年)において、文学作品コーパス(『新潮文庫の100冊』、『青空文庫』のデータ)と、新聞コーパス(『毎日新聞』2000年度)を使った慣用句リストの作成を実施した。また、日本語学習者へのアンケートを通じて、慣用句習得状況を調査した。この調査によって、高頻度で現れる慣用句であっても学習者に習得されていない表現があることを確認している(例えば、「舌を巻く、板につく、裏目に出る」等)。

## 2. 研究の目的

日本語教育のための慣用句の学習教材を提

供する。電子化された教材は、ウェブアクセス可能なデータベースとして提供する。利用のためのインタフェースはウェブインタフェースを提供する。紙媒体やCD-ROM/DVD教材と異なり、こうした方法で成果を公開することで、インターネットに接続しているPC環境があれば、慣用句の学習教材にアクセスが可能である。インターネットを通じ、国内外の日本語教授者の教育、日本語学習者の自己学習に役立つ。

本研究では、分野と頻度情報が付いている日本語慣用句データベースの作成を目指している。この成果は、日本語学習者の学習効率の向上に寄与することも目指している。

## 3. 研究の方法

(1) 慣用句類型の拡張: 慣用句類型として、「名詞-助詞-動詞(NPV)」パターンと「名詞-助詞-形容詞述語(NPA)」パターンについて頻度調査を実施した。ただし、慣用句数は200弱で、未調査分がある。未調査分のNPV, NPAについて頻度調査を行った。入力する慣用句総数は約1,500個程度になった。

(2) コーパスデータの収集: 頻度調査は『新潮文庫の100冊』、『青空文庫』データと、毎日新聞データ(2000年度版)を対象に試験的に実施した。そして、調査対象のコーパスデータを増やした。『毎日新聞データ2008』『読売新聞データ2008』『朝日新聞データ2008』のコーパスデータを収集した。(8年の時間間隔を開けることで通時/共時に影響される慣用句項目の違いを調査できる)。

(3) 慣用句が現れる言語環境の調査: 慣用句抽出はコンピュータプログラムを使って行った。抽出時に、慣用句を含む文も抽出し、慣用句部分をタグ付けしておくことで、教材化したときに、(a)慣用句用例として利用できる、(b)KWIC(Key Word In Context)インタフェースを提供することで、学習者に慣用句が使用される言語環境を提示できる。利用頻度だけでなく、言語環境(文脈情報)を付加することで慣用句と共に頻出する単語や文末述語表現も収集する。

(4) 慣用句データベースの構築: 慣用句リスト及び頻度データと、コーパスから抽出し整理した慣用句用例をデータベース化した。従来型のCD-ROM媒体を利用するだけでなく、ウ

ウェブサイトを通じウェブブラウザを使った参照を可能にした。日本語教授者や日本語学習者に利用できるようウェブアプリケーションの開発に着手する。

(5) 慣用句データは、各慣用句インデックスについて以下のような項目を整理した：読み方、ローマ字表記、意味、慣用句の種類、頻度情報、用例。

#### 4. 研究成果

(1) 慣用句種類の拡張：入力する慣用句総数は約1,500個程度になった。今度の課題としてこの数を増やす。

(2) 慣用句が現れる言語環境の調査：慣用句抽出はコンピュータプログラムを使って行った。慣用句の用例をKWIC (Key Word In Context)形式で抽出した。

(3) 慣用句データベースの構築：ウェブサイトを通じウェブブラウザを使って参照と訂正できるようにデータベースを構築した。慣用句データは、各慣用句インデックスについて以下のような項目を整理した：読み方、ローマ字表記、意味、慣用句の種類、頻度情報、用例。日本語教授者や日本語学習者に利用できるようウェブアプリケーションの開発に着手した。

(4) コーパスから抽出した慣用句の用例の量が膨大のためまだ現在訂正中である。日本語の母国語話者に訂正してもらっている。

(5) 本来同様に英語のイディオム・データベースを構築する目的もあったが、これは今度の課題として残す。

#### 5. 主な発表論文等

本研究の成果は平成24年度に発表する予定である。

[その他]

ホームページ等

<http://ec2-46-51-232-188.ap-northeast-1.compute.amazonaws.com>

#### 6. 研究組織

#### (1) 研究代表者

Danny MINN (ダニー・ミン)

公立大学法人北九州市立大学・基盤教育センター・准教授

研究者番号：60382412