T2R2 東京工業大学リサーチリポジトリ
Tokyo Tech Research Repository

論文 / 著書情報
Article / Book Information

論題(和文)	日本語コーパスに基づいた日本語学習支援システムにおける語の提示
Title(English)	
著者(和文)	仁科喜久子
Authors(English)	KIKUKO NISHINA
出典(和文)	語彙・辞書研究会 第３８回研究発表会, , , pp. 9-16
Citation(English)	, , , pp. 9-16
発行日 / Pub. date	2010, 11

Powered by T2R2 (Tokyo Institute Research Repository)

第38回語彙・辞書研究会 2010年11月20日

仁科喜久子（東京工業大学）

日本語コーパスに基づいた日本語学習支援システムにおける語の提示

1．学習支援システム「なつめ」について

日本語学習者のための作文支援システム「なつめ」開発チーム[1]は2007年から開発を始め、Web上でシステムの公開している[2]。文章作成には全体の構成をどう組み立てるかという談話レベルから、語の選択、構文の整備、表記まで配慮すべき項目がある。現時点では談話レベルに関しては技術的にむずかしい点もあるが、談話も視野に入れつつ構文(文法)、語彙、表記を中心に開発を進めている。本稿はその中でも辞書編集に関連する事柄に焦点を当てて問題点とその解決策を検討することで、辞書学への貢献を願うものである。

本章では「なつめ」のシステムの概略を述べ、利用する正用データベースと誤用データベースの構築について概説する。2章以降では学習者が作文をする上での困難点として共起語の未習得、母語干渉などがあることを述べ、「なつめ」がその困難さを解消する可能性があることを示し、本システムのコンセプトを学習者辞書に反映させる可能性を述べる。

1.1 「なつめ」のデータベースと検索機能

作文支援システム「なつめ」は図1のような構造からなっている。

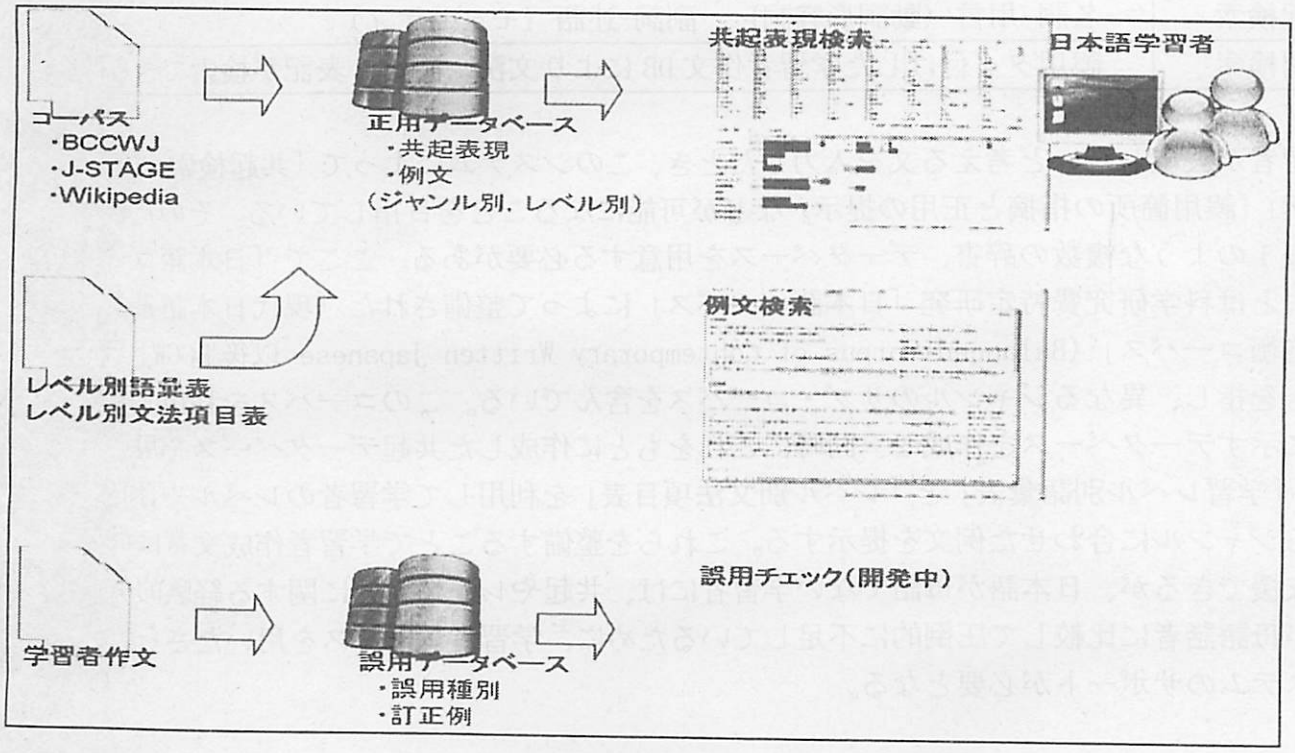

図1　「なつめ」の複数のデータベースと検索機能

開発チームは東京工業大学仁科研究室メンバー、阿辺川武（NII）、傅亮（フウズラボ）、鈴木泰山、八木豊（PICOラボ）からなる。
http://hinoki.ryu.titech.ac.jp

表1「なつめ」におけるコーパス

日本語コーパス	書籍(小説・雑誌記事・エッセイなど)	論文 科学技術	自然言語処理（許可有）
	Yahoo！知恵袋、Yahoo!ブログ		土木学会誌(許可有)
	国会議事録		日本医科大学論文誌(許可有)
	教科書(小中学校主要教科)		電気学会誌(交渉中)
	政府刊行白書	その他	Wikipedia

表2「なつめ」におけるデータベース（DB）の内容

正用DB	語彙辞書	
	レベル別語彙リスト	日本語能力試験レベル別に語彙を分類したリスト
	文法項目辞書	日本語能力試験レベル別に文法項目を分類したリスト
	例文DB	表1のコーパスをDB化
	共起DB	表1のコーパスから（名詞-格助詞-動詞、形容詞）（副詞-動詞-モダリティ）の共起をDB化
学習者作文DB		収集した作文コーパスに誤用タグをつけてDB化

表3　「なつめ」における検索API

共起検索	名詞-用言（動詞形容詞）　副詞-述語（モダリティ）
誤用検索	誤用タグ付けした学習者作文DBにより文法、語彙、表記が検索できる

学習者が表現したいと考える文を入力したとき、このシステムによって「共起検索」「例文検索」「誤用箇所の指摘と正用の提示」などが可能になることを目指している。そのためには表1のような複数の辞書、データベースを用意する必要がある。ここで「日本語コーパス」とは科学研究費特定研究「日本語コーパス」によって整備された「現代日本語書き言葉均衡コーパス」(Balanced Corpus of Contemporary Written Japanese 以後BCCWJと呼ぶ）を指し、異なるジャンルのサブ・コーパスを含んでいる。このコーパスを利用して表2に示すデータベースを作成し、同時にこれをもとに作成した共起データベースを用意する。「学習レベル別語彙表」と「レベル別文法項目表」を利用して学習者のレベルや目的とするジャンルに合わせた例文を提示する。これらを整備することで学習者作成文章はかなり支援できるが、日本語が母語でない学習者には、共起やレジスターに関する経験的な知識が母語話者に比較して圧倒的に不足しているために、学習者コーパスを用いたさらなるシステムのサポートが必要となる。

1.2　誤用分析

2000年以後、学習者作文を収集し、現在10ヶ国164名分の文章（5661文）のデータベースを作成している（曹ら2007)。他に大学院留学生の収集コーパスも利用可能である（村岡2009)。現在、前者のデータのうち約3500文に誤用タグを付けている。誤用の種類としては、文章の構造を談話レベルから単語の表記まで階層的に分類して記述している。その内訳は表4の通りである。本稿では辞書記述に関連する事柄として語彙の問題に焦点を絞

表4　誤用データベース内訳

第1層	件数	第2層	第3層
ディスコース	736	論理的整合性・段落接続・文接続　スタイル(文章としての統一性・場)　待遇表現※　統語的統一(文を超えたねじれ、)	序論、本論、考察、結論 接続語・指示語
構文	2、964	統語的統一(主述のねじれ、副詞呼応)動詞・助詞・助動詞・文末表現　形容詞・形容動詞助詞・助詞相当句副詞・疑問表現・体言※	用言活用・代名詞・数詞
語彙用法	1、412	不適切な語の共起、 語の余剰欠落語形 位相(register)	異義語、類義語、余剰、重複、冗長、欠落 敬語、専門語、
発音表記	279		仮名、漢字

り、学習者の語彙知識はどのようなものであるか、母語の知識がどのような影響を与えているかを検証しながら、語の共起に関連する辞書のあり方を考察する。

2．日本語学習者にとっての語の選択の困難点

学習者の誤用コーパスから語彙に関する例を示す。(LE は学習者誤用の略)

例 LE1　そのための基本的な問題が二つおこる。一つ目は、健康な問題である。(タイ)

例 LE2　男女の給料の差別はまだまだ大きいと思います。(ベトナム)

例 LE3　インタネットカフェは、24 時間に開いているので、ゲームをする機会が多くあり、長すぎる間にゲームをする。(タイ)

例 LE4　学習者の語の使用に結構個人差があることが観察された。(中国)

例 LE 1 から例 LE3 は日本語授業における課題作文であり、1 作文 20 文前後からなるそれぞれ異なる文章中の 1 文である。作文作成者は日本語能力試験 1 級から 2 級程度である。例 LE1「健康な＋（名詞)」は文法的には適合している。しかし「健康な人」は可能であるが、「問題」という名詞は意味的に合わない。例 LE2 も構文的には問題はないが、語の組み合わせが不適切である。「差別」が「格差」となれば、自然な文となる。例 LE3 は「ゲームをする時間が長すぎる」という内容を表現する語法が不自然な例である。「長い時間」「長時間」とはいえるが、「長すぎる間」という語の共起は不自然である。例 LE4 は修士論文の原稿の 1 文である。論文に使用できる語の領域から逸脱している、すなわちレジスターの問題を示す誤用といえる。学習者作文ではこのように様々な原因による誤用が見られ、「なつめ」はその防止を支援しようとするものである。

3．「なつめ」を用いた共起表現の学習

「なつめ」では日本語コーパス BCCWJ やその他のコーパスを参照することで様々なテキストにおける共起の用例に接することができる(阿辺川他 2010)。図 2 は「なつめ」における「名詞」と「動詞」の共起を示すインターフェースのスクリーン・ショットである。

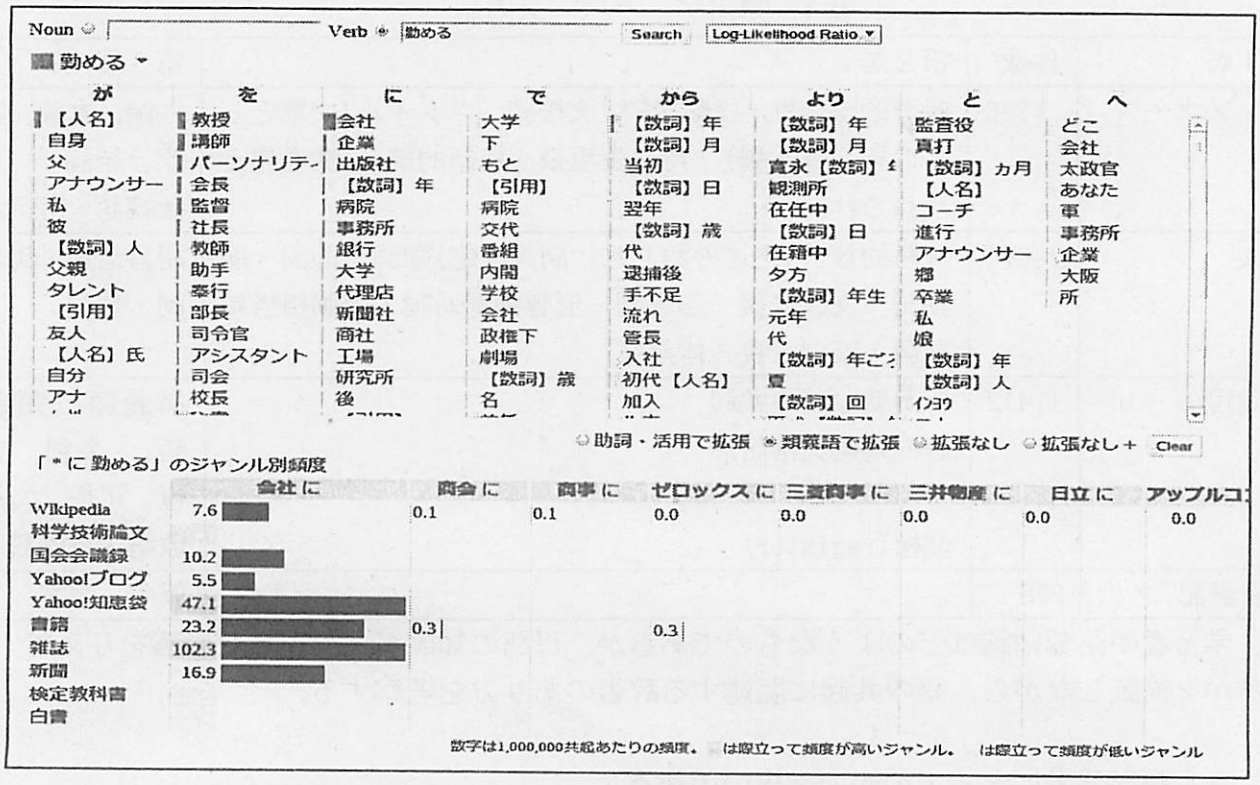

図２「なつめ」共起語検索インターフェース

それぞれの動詞の左に、頻度の高さを表す長短の四角形が表示されている。スクリーン上部には名詞と動詞のスロットがあり、その右側には「頻度、Dice 係数、Tscore、MI 値」など共起の指標が計算値の高い順に提示されるため、利用者が使用する意図によって、これらの指標を選択することができる。例えば学習者が、「コンピュータ関連の仕事に就く」あるいは「コンピュータ関連の会社に勤める」という意向を表現しようとして、「仕事」「勤める」という単語を想起し、「コンピュータ関連の仕事に勤める」という文を考えたとする。そこで検索したい名詞か動詞かいずれか１語をスロットに入力し、右側の指標を選択すると格助詞「が、を、に、で、から、より、と、へ」ごとに共起する名詞あるいは動詞の表示が見られる。ここで、この機能が学習者にはどのように利用されるかを一つの可能性として示す。学習者が LE5 のような文を入力したとする（LE5 は学習者コーパスからの実例である）。

例 LE5　今は大学で化学を専攻している。そのため 10 年後の将来は化学に関連がある{仕事}に勤めたい。(ベトナム)

まず名詞から「仕事」を入力すると、それぞれの格助詞の下には頻度順に共起する動詞が配置される。次にリストの右下にある「助詞・活用語で拡張、類義語で拡張、拡張なし」の中から、「類義語」にマークをつけた後、参照したい項目をクリックすると拡張情報を見ることができる（図２の下方）。「仕事」と共起する動詞として、どの格助詞の下の動詞にも「勤める」は見られず、高頻度で共起する語の組み合わせとしては「仕事ができる」「仕事をする」「仕事をこなす」「仕事に就く」「仕事に従事する」などであることがわかる。次に動詞「勤める」を見ると「私が勤める」「教授を務める」「会社に勤める」「企業に務める」が高頻度であることがわかり。学習者は、これらの情報から、自らが表現したい語の組み

合わせとして、「仕事に就く」あるいは「会社に勤める」が該当することを悟ることが想定される。

さらに、このインターフェースの下半分には類義語との共起が 10 項目のジャンル別に表示されている。これらのジャンルの最初の2項目は「なつめ」チームで準備したWikipedia、科学技術論文である。後の87項目はBCCWJのサブ・コーパスと呼ばれるものであり、その内訳は、書籍(小説、評論、ベストセラーなど)、検定教科書、白書、国会議事録、新聞、さらにYahoo！知恵袋、ブログなどのWeb文章からなっている。これらのジャンル別項目に頻度が棒グラフで示される。表中の数字は100万共起対あたりの頻度であり、際立って頻度が高い語はピンク色、際立って頻度が低い語は青色で表示される。それぞれの頻度を示す数字をクリックすると例文が現れる。「会社に勤める」は次のような正用例がある。

例 AU 1 「化粧品会社に勤めるなど、結婚/出産後も働きつづけた。」(加藤仁 Yomiuri Weekly2001) (AUは正用例の略)

一方、「仕事に」を検索すると、「就く、就ける、携わる、取り組む、専念する」が共起語として提示される。「仕事に就ける」の共起が最も高頻度であり、その一例を示す。

例 AU2 大学で地理学を専攻したいと思うのですが、卒業したらどのような仕事に就けるでしょうか？ (Yahoo！知恵袋)

「仕事に」の類義表現をみると「仕事に専念する/取りかかる/勤しむ」が出現する。「仕事に」の用例をさらにみると、「教育関係の、理数系の」など職種が前接していることがわかる。このように検索していくと、名詞「仕事」からも動詞「勤める」からも共起の例は見当たらない。これらの例を学習者が見て行った結果、「に勤める」は仕事の内容ではなく、組織を意味する語と共起することが理解できるはずである。このような学習方法は Data Driven Learning として知られており、英語教育などにおける効果は知られているが、日本語学習での研究も実践例もあまり見られない。

4. レジスターによる使い分け

レジスターという概念は、英国Firth派によって「社会的な拘束力をもつ言語学上の規範」として示されたものである。この流れを汲むHallidayらは機能文法を提唱し、次の3つの次元において言語の使用域による変異を提示している。

(1) コミュニケーションの目的と主題に関わる「フィールド」(Field of discourse)
(2) コミュニケーションを行うための手段に関わる「モード」(Mode of discourse)
(3) コミュニケーションパートナー同士の関係に関わる「テナー」(Tenor of discourse)

「フィールド」としては、学会での論文、官庁で記録される文書、情報を知らせるため新聞記事、日常生活での世間話、個人的な気持ちを述べる手紙などが考えられる。「モード」としては、書籍、新聞などの紙媒体、対面会話、電話、テレビラジオなどの画像および音声の有無によるヴァリエーションをもつ媒体が考えられる。「テナー」としては話者(書き手)と聞き手(読み手)の関係で、独話 (講義、講演を含む)、対話、会話など話者と聞き手の人数、両者の社会的、個人的関係、話者の特性 (男女、年齢、社会的地位) などがある。

さらにBiberらはこの考えを引き継ぎ、対象となるテキストに67の言語項目をアノテ

ーションとして付与し、各項目を変数として多変量解析をすることで、共起関係における主要因子を抽出し、レジスター間の比較分析を行っている(Biber 1988 , Biber&Conrad2009)。本稿では、これらのレジスター項目を念頭において学習者作文の分析を試みる。

例 LE9　21 a 001-007　子供のとき、中国で日本は男尊女卑の社会と聞いたが、具体的に何かよくわかりません。成長について(→するにつれて)だんだん深く了解(→理解)できた。今の社会では、男女平等と(→を)提唱しているけど、男尊女卑の現象は時々あります。今、女性は社会で重要な役割を演じている。いろいろな行業（→職業）と領域（→分野）で優れた成績が出ってきた（→をあげてきた)。以前の女性と比べて毎日家事したり、子供の世話したり、このような命（→人生）じゃなくて、自分が自分の夢を抱いて、それをかなう(→　叶える)ため女に(→は)社会で活躍しています。((　)内の→右は筆者の訂正案)(中国人)

例 LE9 は「男尊女卑」という課題作文の一段落である。教師はレポート作成を念頭に「である体」で、客観的な文章を書くように事前に指示している。Halliday の枠組みに当てはめると、フィールド：「レポート」、「モード」：文書テキスト、「テナー」：学生から教師、あるいは一般読者へというレジスターが考えられる。

実線下線部「聞いたが、」「よくわかりません」「提唱しているけど」「時々あります」「命じゃなくて」「活躍しています」は述語部であり、「文のスタイルの統一」をすべき箇所である。レポート文のレジスターとしては、「聞いたが」「よくわからない」「提唱しているが」「時々ある」「人生ではなく」「活躍している」という書き言葉が期待される。また。波線の語「だんだん」「ときどき」「女」は、レポート文のレジスターとして不適切な語選択と考えられる。すなわち、これらの語はアカデミックなレポートや論文のための書き言葉のレジスターとして、各々「次第に」「しばしば」「女性」などを選択するべきと考えられる。

例 LE10　39a01-15　日本語に{→<削除>}は、物のやりとりに関することばが実に詳しいと思われる。(中略)　物のやりとりを表わし{→表す}ため、授受動詞の正しく{→正しい}使い{→使い方/用法}に気をつけなければならないだろう。(中略)目上の人は尊敬すべきだ。だから、必ず尊敬語、または尊敬の意がある言葉を使う。(中国)

例 LE10 において最終行の「言葉を使う」は誤りではないが、レジスターの視点からみるとアカデミックな文章における表現としては、「語を用いる/使用する」などが適正な表現として考えられる。

例 AU3　重み付けされた語を用いてパラグラフごとに文書のクラスタリングを行ない、重要パラグラフを抽出する手法について述べる.（福本ら　自然言語処理　1996)

このように学会誌では「語を用いる」の用法多くみられる。一般に日本語学習では最初に会話表現を学び、その後書き言葉のジャンルを学ぶという流れの中で、それぞれの使い分けを習得するのは簡単ではないことがわかる。

5．母語干渉

外国人学習者にはレジスターの習得の他にも母語干渉に関連する様々な問題がある。例

LE15 は学習者が母語で習得している語の共起を反映したものと言える。

例 LE15　103a14 現在社会に、コンピューターは欠かせないツールになりました。103a15 毎日、世の中に大勢な人がインターネットを通じてそれぞれの目的が達成しています。103a16 ソフトを利用して、漢字が写すことが必要がありません。(中国)

例 LE 15 は中級レベルの中国人学部生である。「書く」は基本動詞であるにもかかわらず、「写字」という母語の表現があるために、その干渉により誤用が生じている。母語干渉としては他にも「経済面に先進する→経済面で発展する」、「親友を保護する→守る」、「将来の世界を連想する→想像する」などが観測されている（曹ら 2010)。現状では多言語における母語干渉に関する用例収集は不十分であるが、今後用例を大量に収集していけば、母語別の誤用防止策が可能となるであろう。学習者辞書には母語干渉に関する情報を記述する必要がある。

6．他のシステムおよび辞書との比較

文中の語と語との共起をセットで学習することにより学習効果があがるという考えに基づいて作成された辞書やテキストがこの数年間に出てきている。姫野(2004)は、動詞180、形容動詞類 364 語、計 1544 項目を収めた辞書である。共起語を意味分類ごとに例文提示し、関連語としての複合語や慣用語が項目の中に示されており、学習者にとっては有益な情報を提供している。また小野正樹ら（2009、2010）は、コロケーションに着目した語彙の学習書となっている。教師の視点で厳選されたコロケーションを意味分類して、その中で理解し、さらに発展的に学習できるようになっている。しかしながら、両者とも参照可能な語が少ないのが欠点である。

また、インターネットで利用可能なSketch Engine[3]の日本語版は「なつめ」と類似した機能をもつものであるが、Web上の 4 億語コーパスJpWaCを使用しシステムを構築している。このコーパスはジャンルを区別して表示できないため、学習者の使用目的に適切なものが必ずしも検索できないことが欠点である。

7．辞書記述への提言と今後の課題

本稿では日本語作文支援システムを構築する上で必要な語彙学習における困難点の解明を通して、外国人学習者にも利用可能な日本語辞書構築の在り方を考察した。そこでは、語と語の共起を意識した学習方法、レジスターを意識した記述、母語干渉を防ぐ何らかの対策の重要性が明らかになった。そこで本稿で議論したことから学習者辞書編纂への提言と今後の課題について述べる。

〈提言〉

)共起表現の記述

・名詞と助詞と用言（動詞、形容詞・形容動詞）の共起が理解できるようにする

・用言の連体修飾用法と述語用法の互換性についての情報を付与する。

「きれいな部屋」、「部屋がきれいだ」はともに言えるが、「頭がよい」に対して「よい

---

http://www.sketchengine.co.uk/

頭」は言えないなどを理解させる必要がある。

2) レジスターに関する記述

・学習者が作文をするに際して、目的に沿った表現が選択できることが重要である。それを可能にする手がかりとしての、フィールド、モード、テナーの情報を付与する。

3) 母語干渉に対する対策

・異なる母語の学習者作文の収集をし、タグ付けをする必要がある。

〈今後の課題〉

以上の提言をするとともに、さらに的確な例文が検索できるようにする。例文によって言語を学習する場合、authentic でありながら、出来るだけコンテクストなしで理解でき、用語は難解でなく、短く簡潔な例文提示を可能にすることを課題とする。

謝辞

本研究は科学研究費補助金特定研究「日本語コーパス」(研究代表者　前川喜久雄) 公募研究「バランス・コーパス利用による日本語作文支援システム「なつめ」の構築と評価」(研究代表者仁科喜久子;課題番号 21011002) 科学研究費補助金基盤研究 (B)「大規模コーパスを利用した日本語学習支援システム「ひのき」構築と評価」(研究代表者仁科喜久子;課題番号 21320092)、科学研究費補助金挑戦的萌芽研究「日本語学習者誤用コーパスを利用した作文システムの開発」(研究代表者仁科喜久子;課題番号 22652048) による支援を受けて行われた。

参考文献

阿辺川武・Hodoscek Bor・仁科喜久子(2010a)「日本語作文支援システム「なつめ」における共起語索方法の改訂」, 特定領域研究「日本語コーパス」平成 21 年度公開ワークショップ予稿集 pp243-244

阿辺川武・Hodoscek Bor・仁科喜久子(2010b)「日本語作文支援システム「なつめ」—利用者の視点—」特定領域研究「日本語コーパス」平成 22 年度全体会議予稿集 pp119-120

小野正樹・小林典子・長谷川守寿 (2009, 2010)『コロケーションで増やす表現 Vol. 1, 2』くろしお出版

曹紅荃・黒田史彦・八木豊・鈴木泰山・仁科喜久子(2010)「学習者作文支援システムのための誤用データベース作成-動詞の誤用分析を中心に-」pp. 1571-1～1571-9, 世界日語教育大会論文集　国立政治大学, 台湾

姫野昌子(2004)『日本語表現活用辞典』　研究社

Biber, Douglas (1988) Variation across Speech and Writing. Cambridge University Press

Biber, Douglas and Susan Conrad (2009). Register, Genre, and Style. Cambridge: Cambridge Textbooks

Halliday, M.A.K., and C. M.I.M. Matthiessen (2004). An Introduction to Functional Grammar. 3rd ed. London: Arnold

Hodoscek Bor (2010) Development of a Register-based Writing Assistance System for Academic Japanese (Master thesis at Tokyo Institute of Technology)

Irena Srdanovic Erjavec, Tomaz Erjavec, Kilgarriff (2008) A web corpus and word sketches for Japanese. Journal of NLP. pp.1-22