

平成 21 年度研究進捗状況中間報告：作文支援システム班 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

仁科喜久子（班長：東京工業大学留学生センター）[†]

Progress Report of the Year 2009: 'Writing Support System' Group

Kikuko Nishina (International Student Center, Tokyo Institute of Technology)

1. 本年度の目標

前回の公募班に引き続き、今年度、再度採択されたことから、課題を継続発展することとする。最終年度(22年度)までには、BCCWJを利用した日本語作文支援システムを構築し、それとともにBCCWJの評価をすることを最終目標とする。

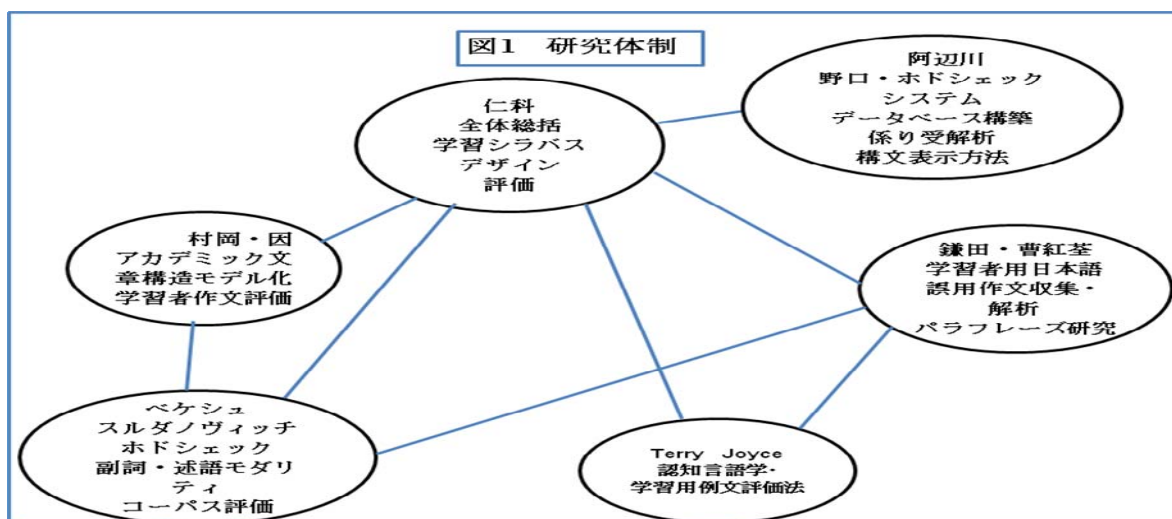
その中で、本年度は、特に次の1)から7)の項目を重点的に進める予定である。

1) 科学技術作文に関するフレームワークの分析、2) 共起関係特にモダリティを含む述部処理、3) 例文表示の拡張機能、4) ジャンル別表示、5) レベル別表示の改善、6) 学習者コーパス構築と「なつめ」での利用、7) BCCWJの評価-科学技術論文との対照

本稿では、この中から1)、2)、7)に関連するものに焦点を当てて、進捗状況を報告する。

2. 研究体制と分担

上記の研究を遂行するための研究体制を下記に示す。



本班は国際的、学際的なメンバーで構成され、次の5つのサブグループから成っている。

以下に各メンバーの本年度の活動と計画について述べる。

1) 村岡・因：理工系を中心とするアカデミック・ライティング学習支援方策の検討

学習者の作文データを分析し、作文支援をするためのデータ収集を行い、モデル化を検討している。「なつめ」システムにおいて学習者が作文する上でのヒントを与える機能、

[†] knishina@ryu.titech.ac.jp

また言語研究者・教師支援ともなることを目指している。

2) 曹・鎌田：学習者コーパスの構築と分析

日本語学習者が誤りやすい語彙・文法・スタイル・文書形成の問題を学習者コーパスから分析を行う。母語転移、第二言語習得過程の問題を検討するとともに、誤用コーパスを構築することで、1)と同様に効果を目指す。

3) テリー・ジョイス：認知言語学からのコーパス分析

1) 2)の研究に関連する学習者の認知的な学習過程を把握して、利用しやすい機能向上に資することをはかる。

4) アンドレ・ベケシュら：BCCWJ およびその他のコーパス分析

BCCWJを中心に大規模なコーパスを利用して、日本語の構文、語彙、語用論的な考察も視野に入れ、言語現象の解明を行う。その結果を「なつめ」に反映させ、学習に利用できるように目指す。

5) 阿辺川ら：「なつめ」システムの構築

「なつめ」のデータベース整備、共起および例文表示検索プログラム、インターフェースの構成設計、デザイン、学習履歴の収集などを担当している。これらの機能の新規追加とともに昨年度からの改良を行う。

3. BCCWJ と他の比較コーパス収集

表1は本年度までに利用可能になったシラバスを示している。BCCWJ以外のコーパスを別途準備した。これらは共起関係の頻度計算に用いられるデータとなる。例文表示はこの中から、Web上に公開可能なもののみが用いられる。「科学技術文」はJ-StageのWeb上にある日本語論文PDF版からテキストファイルに変換し、現在までに約400論文を取り込んだ。また「自然言語処理」のテキストファイル化をし、理系論文としてまとめると1300万字近いデータとなる。このうち「自然言語処理」は例文提示の許可を得ている。J-Stageのデータは、今後さらに増やす予定である。

表1 分析対象コーパス一覧 (野口慎一郎作成)

コーパス	分量 (文字数)	コーパス	分量 (文字数)
BCCWJ-白書	22,987,412	毎日新聞(10年分)	616,235,775
BCCWJ-書籍	210,675,620	日本語版 Wikipedia	1,014,070,429
BCCWJ-Yahoo!知恵袋	27,847,731	科学技術論文 (J-Stage)	6,175,566
BCCWJ-国会会議録	26,201,152	自然言語処理 (ジャーナル)	6,690,320

4. 論文作成スキーマ形成と学習者コーパスに関する研究

本年度前半までに村岡貴子、因京子、仁科喜久子らは留学生の論文作成スキーマの知識と日本語能力との関係を分析し、専門文書作成スキーマ獲得のために必要な作文能力についてアンケートおよびインタビュー調査を行った結果、日本語能力の高低に関わりなく、スキーマの有のある学習者は、論文として流れの妥当性や論理性が指摘できることがあき

らかになった。例えば、「話題別・段落別に中心文を書く」「パラグラフ内のトピックは複数混在させない」などとインタビュー調査で言及している。この結果に基づいてスキルに関するメタ知識のリソース化を検討している(村岡ら 2009)。また、曹らは、学習者作文コーパス中から漢字を含む動詞を分析することで、母語転移による誤用の問題を考察した(曹ら 2009)。鎌田は、日本語学習者に話し言葉を要約する課題を課し、パラフレーズによって起こる誤りのタイプを観察し、認知言語学的な分析を行った(Kamada, Nishina. 2009)。これらの結果を誤用データベースとしてまとめ、誤用検索インターフェースの試行版を作成し、日本語教師に試用を依頼している。

5. BCCWJ コーパスの分析

BCCWJ を「なつめ」の共起表現や例文表示システムに利用するために、BCCWJ 以外のコーパスデータとも比較しながら分析を行ってきた。本年度前半までには、他のコーパスとしては Web コーパス、日本語教科書、科学技術文書を比較した。この中で BCCWJ の書籍コーパスと Web コーパスが均衡のとれたコーパスであることを明らかにした。これによりジャンル別のコーパスの特徴も示すことが可能になり、ジャンル別の作文支援の基礎ができた。

5.1 BCCWJ と日本語教科書シラバスとの比較

スルダノヴィッチ・ベケシュ・仁科(2009)では、13種のコーパスを分析し、BCCWJ の中の「書籍」と Web コーパスが現代日本語の様相を反映した、最も均衡が取れたコーパスであることを明らかにし、日本語教科書に現れる副詞とモダリティの共起とそれぞれのコーパスをクラスター分析などの統計的方法で比較した。その結果、現実の用法を反映している均衡コーパスとのずれを指摘し、学習者辞書や教科書などの教材作成の改善への示唆をした。

5.2 文末モダリティ検索の自動化のための基礎研究

日本語におけるモダリティとは「かも/しれ/ない」「と/思わ/れる」など複数の形態素の連なりからなる文末複合辞およびそれと共起する副詞などの語彙を含むものとする。

日本語におけるモダリティの例：かしら、かな、かもしれない、ざるを得ない、そうだ、だろう、てはならない、といえない/といえる、とは限らない、と思う、と考える、なくてはいけない、なくてはならない、なければいけない、なければならぬ、に違いない、のか、のだ/のではない、はずだ、べきだ、みたいだ、ようだ、ように思う、らしい、わけだ/わけではない、わけにはいかない、気がする/気がしない

スルダノヴィッチら「ウェブコーパスと検索システムを利用した推量副詞とモダリティ形式の遠隔共起抽出と日本語教育への応用」においては、モダリティの切れ続きを手で見分けていたが、現在それらを自動化することを試みている(スルダノヴィッチら 2009b)。機能表現辞書「つつじ」は特に複数の形態素からなる文法化された(されつつある)モダリティ表現を多く含んでおり、構造が XML で記述されているため、機能表現の階層的な記述が可能である。この辞書中には言い換えのための意味的な関連の情報を含んでいるが、それが

「推量」などの意味情報のみで、モダリティのカテゴリで体系化されていない。本研究ではこれらの項目の中から新しい attribute としてモダリティに関する情報を付与することを目指している。

さらに今後は南(1974)による階層構造における提題、時間、場所を表す修飾句との共起、程度副詞と述語の遠隔共起を調査する予定である。

6. 共起表現に関する研究

システムに関する詳細は、阿辺川・Hodoscek によるポスター発表原稿に譲る。ここでは、テキストデータの分析に関する内容について述べることにする。現時点で表示される名詞+格助詞+述語の組み合わせでは、格助詞に「ガ、ヲ、ニ、デ、カラ、ヨリ、へ、ト」を取り上げている。本システムでは能動文と受動文の共起を区別するために述語は受身形、使役形のまま表示するようにしている。同一深層格の文において受動文と能動文ではガ格とヲ格の交替がみられる。

- ・経済の発展を保証する（能動態）
- ・経済の発展が保証される（受動態）

BCCWJ 中の書籍、白書と論文における受動文と能動文が、それぞれのコーパスでどのような比率で用いられているかを調べた結果を表 2 に示す（この調査では白書と書籍はモニター版のデータを使用したため表 1 の数字と異なっている）。白書は論文の約 3 倍、書籍は論文の約 10 倍ではある。また各コーパス内での受動態と能動態の比率（述べ語数）をみると、論文では受動態は能動態の 7.1%、白書では 5.8%、書籍は 4.7% となり、文章中の受動文は論文、白書、書籍の順で多いことがわかる。

さらに文中のひとつの述語が、受動文と能動文で何項の格をもつか調べたところ、表 3 のような結果となった。3 種のコーパスとも、能動文ではヲ格を最も多く取り、次に二格を取るということでは共通しているが、能動文でヲ格を取る比率は、論文、白書、書籍の順で少

表 2 論文・白書・書籍中の能動態と受身態の分布(野口慎一郎作成)

		論文	白書	書籍
データ量 (文字数)		2,052,708	5,681,456	22,063,844
文数		37,027	73,866	628,489
共起数 (のべ)		90,851	215,219	899,565
動詞 延べ(異なり)	能動態	84923 (3,270)	203,289 (4,665)	858,580 (23,177)
	受動態	5,928 (639)	11,930 (889)	40,985 (4,112)

表3 論文・白書・書籍コーパス中の格の分布(野口慎一郎作成)

		ガ	ヲ	ニ	デ	カラ	ヨリ	ヘ	ト	計
論文	能	16.01%	46.28%	19.00%	6.57%	3.36%	0.27%	0.29%	8.22%	100%
	受	40.71	2.42	25.76	16.53	7.25	0.62	0.19	6.50	100
白書	能	16.46	43.28	24.47	4.09	2.40	0.37	0.23	8.70	100
	受	51.56	2.88	27.24	8.89	5.42	0.74	0.17	3.10	100
書籍	能	18.75	39.29	25.23	6.35	3.13	0.15	1.25	5.85	100
	受	31.06	9.53	34.25	10.38	5.73	0.24	0.69	8.12	100

なくなっている。一方、受動文では、論文と白書では、ガ格、二格、デ格という順であり、書籍は二格、ガ格、デ格の順となっており、他の2種のコーパスとは異なっている。書籍では、深層構造ではガ格を取るものが、小説などにおいては主語省略、あるいは堤題を示す係助詞「ハ」との交替があるためにガ格の割合が少なくなっていることが推測される。構文中の格構造において論文は白書に似た傾向をもち、均衡コーパスに近い書籍とは異なる傾向を示していることが分かる。さらに論文と白書を比較すると、論文におけるガ格が10%程度少なく、書籍と白書の間位置することがわかる。論文におけるデ格は白書および書籍の2倍近くとなっている。理由としては、白書ではデ格は場所を表すことが多く、一方、論文では場所のほかに手段などを表す場合にデ格を用いており、「式で表わされる」など定型的表现が多い。このため、論文ではデ格の頻度が多くなっていると推測できる。

また、受け身の助動詞「れる」と同様に態の変化とともに変動するアスペクト（<目的格名詞+ガ+他動詞+てある>）、複合動詞（突き+出す、突き+出る、押し+出す）において自動詞と他動詞の組み合わせをどう扱うか検討する必要がある。これらは合わせて今年度の今後の課題とする。

7. まとめ

平成21年度の新規採択公募班として、研究体制の概要を述べ、各サブグループの進捗状況を報告した。

- 1) 日本語学習者の作文支援をするための基礎的な研究として、学習者のスキーマと日本語能力の関連性を調査した。今後は論文作成スキーマのリソース化の実現を検討する。
- 2) 同じくシステムに蓄える知識として、学習者の漢字知識と運用に関する調査から母語移転の問題を検討した。また、話し言葉から書き言葉への要約時に見られる誤用を含む注目すべき問題点を明らかにした。
- 3) BCCWJ 評価分析のために、比較対照用に科学技術論文のコーパスの収集をした。
- 4) ジャンル別コーパスの比較として、構文中の格構造と受動文、能動文を観察した結果、科学論文コーパスは白書に類似する傾向がみられ、最も均衡の取れたコーパスである書籍とは異なる傾向があることが分かった。今後は、この結果をふまえて、「なつめ」におけるジャンル別の特徴を活かした表示方法を検討する予定である。
- 5) 「モダリティ」の定義を考察し、モダリティが検索可能なシステムを完成するために機

能語辞典「つつじ」を参照して、モダリティ項目を含む辞書を検討している。

6) 平成21年4月から現在までの研究成果は、下記の通りであり、国際会議などでの発表予定を含んでいる。

参考文献

佐尾ちとせ、江口萌、松吉俊、乾健太郎 (2009) 日本語文のモダリティ・極性情報を捉えるために 言語処理学会第15 回年次会発表論文集pp. 793-796

南不二男 (1974) 『現代日本語の構造』大修館

Bekeš, A. (2006) Japanese suppositional adverbs in speaker-hearer interaction Proc.of the 3rd conference on Japanese language and Japanese language teaching. Venezia: Cafoscarina, 34-48

関連 URL

「あすなる」「なつめ」ホームページ: <http://hinoki.ryu.titech.ac.jp>

「つつじ」機能表現辞典: <http://kotoba.nuee.nagoya-u.ac.jp/tsutsuji/>

研究成果

・ Bor Hodoscek, Nishina Kikuko(2009)日本語コーパスを活用したオノマトペの計量言語学的アプローチジャンルと型の相関 Asian and African Studies Volume 13,pp.169-178

・ スルダノヴィッチ・イレーナ、ベケシュ・アンドレ、仁科喜久子(2009a)コーパスに基づいた語彙シラバス作成に向けてー推量的副詞と文末モダリティの共起を中心としてー

日本語教育 142 号, pp. 69-80

・ スルダノヴィッチ・イレーナ、ホドシチェク・ボル、ベケシュアンドレイ、仁科喜久子(2009b) ウェブコーパスと検索システムを利用した推量副詞とモダリティ形式の遠隔共起抽出と日本語教育への応用 自然言語処理(印刷中)10 月刊行予定

・ 村岡貴子、因京子、仁科喜久子(2009)多様な学習者に対する専門文書作成スキーマ形成のためのリソース集構築に向けて 「専門日本語教育研究」(投稿中)

・ 曹紅セン、仁科喜久子 漢字表記のある基本動詞の意味上の母語転移ー中国人学習者の場合ー「日本語教育法研究会」論文誌(印刷中)

国際会議

・ Michiko Kamada, Kikuko Nishina (2009) Japanese Language Learners' Use of Paraphrasing in Summarizing from Spoken Discourse to Written Discourse JSAA-ICJLE 国際研究大会 July 13-16, 2009, Australia

・ Irena Srdanović, Andrej Bekeš, Kikuko Nishina (2009) Classifying corpora based on adverbs distribution 《Text and Language: Structures, Functions, Interrelations Qualico 2009 September. 2009, Graz (発表予定)

学会発表

・ ボル・ホドシチェク、アンドレ・ベケシュ、仁科喜久子 モダリティ表現辞書の構築とその日本語作文支援への応用『NLP 若手の会 第4回シンポジウム』2009年9月発表予定