学習者コーパス「なたね」の構築と応用の可能性

Construction of Learner corpus “NATANE” and possible application

曹紅荃 (Hongquan Cao), 西安交通大学
八木豊 (Yutaka Yagi), 株式会社ピコラボ
黒田史彦 (Fumihiko Kuroda), 早稲田大学
仁科喜久子 (Kikuko Nishina), 東京工業大学

要旨：

筆者らは、学習者作文支援システム構築の一環として、誤用種別のタグを付与した学習者コーパス「なたね」を作成している。そのデータは日本語学習者から収集した 310 作文（5, 462 文）からなり、誤用種別および訂正例がタグ付けされている。誤用種別は 4 階層に分けられ、最上位の階層は「誤用の内容、誤用の対象、誤用の要因・背景」に分類され、多方面から学習者言語を分析できるように誤用種別フレームの確立を目指している。また日本語教育および学習者言語研究に資するために、「なたね」のデータを利用した検索サイトを開発した。さらにデータを作文支援システムに取り入れることによって、展開できる作文支援および作文指導の新しい可能性を提示している。

キーワード：学習者コーパス、誤用種別、誤用分析、タグ付け、作文支援

1. はじめに

日本語学習者の作文や会話データなどの集積として、数多くの学習者コーパスが構築されている。誤用タグが付与されたコーパスを利用すると、誤用と正用の二つの視点から学習者言語を包括的に把握することが可能になる。学習者コーパスのうち、誤用タグが付与されているものとして、現在公開されている『外国人学習者の日本語誤用例集』（寺村、1990）と「オンライン日本語誤用辞典」（小柳他、2011）の二つが挙げられる。いずれも誤用種別が文法中心となっていると見られ、またその利用は学習者作文に含まれる誤用の指摘が中心となっている。筆者らは、日本語母語話者によるオーセンティックなコーパスと日本語学習者コーパスを合わせて利用し、異なる観点からの日本語作文支援を実現しようとしている。その一環として誤用種別のタグを付与した学習者コーパス「なたね」を構築し、また多方面から学習者言語を分析できるように、誤用種別フレームの確立を目指している。本論文では、「なたね」の概要を紹介し、誤用のタグ付け作業の詳細を説明した上で、「なたね」の応用によって展開できる学習者言語研究、作文支援および作文指導の可能性を提示する。

2. 「なたね」の概要

「なたね」は日本語学習者の課題作文を収集・電子化し、誤用だと思われる箇所に対して日本語教師がタグを付与した学習者コーパスである。次のプロセスを経て構築されてきている。

1）学習者作文の収集および電子化(310 作文、5, 462 文、185, 407 文字)[1]

2）Excel を用いた試験的誤用タグの付与

3）誤用種別の検討と確定

4）汎用アノテーションツールを用いた誤用タグの付与

上記１）にあたって、複数の日本語教師の協力のもとに、大学院、大学、または語学学校に在籍

[1] 数字は 2012 年 6 月時点のもの。

する日本語学習者が書いた作文を収集した。主な内容は、日本語の授業の中で設定したテーマについての作文である。また収集の際、プライバシー保護を遵守することで、作文データ利用許諾を学習者から得ている。作文データの他に、国籍、母語、学習歴などの学習者情報も合わせて電子化している。これまでに、210 人の学習者から 310 作文を収集したが、中国語を母語とする学習者が筆者の半分以上を占めており、学習者の母語分布には少し偏りがある状況である。

3. 誤用のタグ付け作業

学習者の大量の誤用を分析する際，誤用種別のフレームが必要不可欠である。既存の日本語学習者コーパスには、誤用種別、誤用の記述法や訂正法などを統合できるような基準がない。そのため、「なたね」のタグ付け作業を通して、具体的な誤用種別の内容と構造の在り方を考察し、日本語学習者コーパスに必要な誤用種別の根本的な基準を模索している。

0	1	2	3
誤用の対象			
	語	名詞	
		数詞	
		副詞	オノマトペ
			その他
		接続詞	
		助詞	格助詞
			並立助詞
			終助詞
			副助詞
			係助詞
			接続助詞
			助詞相当句
			その他
		動詞	
		形容詞	
		形容動詞	
		助動詞・助動詞相当句	
		接頭辞	
		接尾辞	
	句・節*		
	句読点		
	その他		

0	1	2	3
誤用の要因・背景	類似	意味	
		字形	
		音	
	母語干渉	中国語	
		韓国語	
		ベトナム語	
		その他	
	レジスター	話し言葉と書き言葉	
		その他	
	待遇表現		
	文体の不統一		
	論理性*		
	その他		

0	1	2	3
誤用の内容	脱落		
	付加		
	誤形成		
	混同		
	位置		
	接続	段落接続	
		文間接続	
		文内接続	
	統語的呼応		
	語の共起(コロケーション)		
	指示語		
	正書法からの逸脱		
	送り仮名		
	活用	未然形	
		連用形	
		終止形	
		連体形	
		已然形／仮定形	
		命令形	
	文法範疇	ヴォイス	受身
			可能
			使役
			授受(やりもらい)
			自他動詞
		テンス	
		アスペクト	
		モダリティ	
	文字種	漢字	
		ひらがな	
		カタカナ	
	音	清濁音*	
		長短音*	
		拗音	
		促音	
		撥音	
	その他		

図 1　誤用種別の項目表

曹他（2010）の作業を通して、誤用種別を定義していく上での問題点を解明した。その結果、曹他（2011）では、四階層からなる誤用種別フレームを確定し、上位の層ほど抽象度が高いものとし、最上位の層は「誤用の内容（どのような誤用か）」「誤用の対象（何に関する誤用か）」「誤用の要因・背景（何による誤用か）」の三種に大分類した。このような構成によって、異なる視点から誤用種別を判定し、系統立ててタグ付けができるようになる。誤用種別の各項目の詳細を図１に示した。その定義および具体例に関しては八木・鈴木（2012）を参照されたい。

タグ付作業を通して、誤用種別の更なる改善の必要性が明らかになったことから、まず、「誤用の対象」には「句・節」を追加した。これは、実際の誤用箇所は、「一人で」「暴力をふって」「なければならない」のような「句・節」となっているものがあるからである。次に、「誤用の内容」の下位の「音」に関する定義は、「発音に関する誤用を、誤用部分の音の種類にしたがって分類する」とあるが、現行の下位項目に不足と不整合があり、それを調整するために、「音」の下位分類を、「清濁音」、「長短音」、「拗音」、「促音」、「撥音」の項目に改善した。さらに、例えば「“中心となって、このうち3人は～” → “中心となって事務所を創立した。このうち3人は～”」のような、

内容を追加しないと成立しない文があり、これは書き手の論理性による誤用だと考えられるため、「誤用の要因・背景」に「論理性」という誤用種別を追加した。これらの追加修正項目も図１に示されている。

曹他（2010）の時点ではExcel を用いて試験的に誤用タグを付与していたが、その作業では、「書式の統一」、「文章全体の流れの把握」、「タグ付け状況の確認」などの面において困難であることが分かり、汎用アノテーションツールSlate（徳永他、2010）を取り入れた。Slate を利用することで、Excel 上での記述の不統一・不整合などの問題を解決し、一箇所に複数の誤用種別のタグを付けることが容易となった。また文章全体および誤用フレーム全体を俯瞰しながらタグ付けができ、効率的かつ精度の高いタグ付け作業を可能にしている。

4. 「なたね」の応用

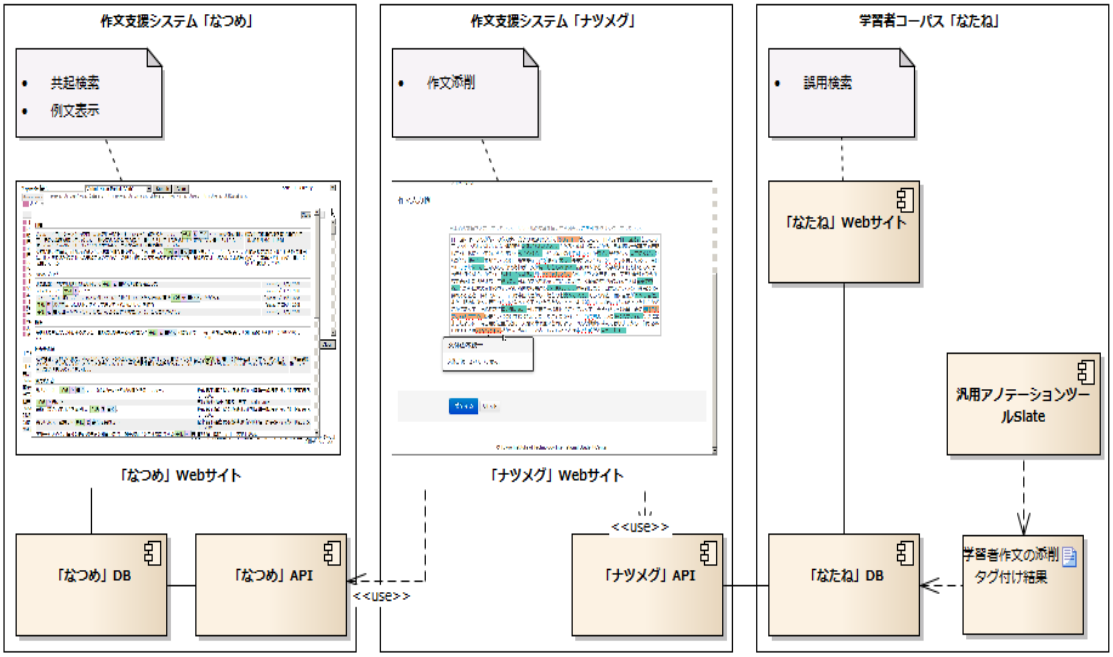

図 2 「なたね」の概況図および作文支援システムとの関係図

従来は日本語教師向けの誤用検索サイト試用版を利用者を限定して公開したが（曹他、2010）、誤用種別とタグ付け作業の改善に伴い、「なたね」データの公開および応用のために、筆者らは一般公開用の「なたね」検索サイト[2]を開発した。この検索サイトは誤用および正用から検索でき、誤用種別を指定することにより、使用者の目的に応じた検索ができる。例えば、練習問題や試験問題作成のために、典型的な誤用例と訂正例が提示される。また学習者言語研究のために、学習項目の誤用と正用を全体的に把握することができる。

さらに、「なたね」の誤用データを応用する可能性の一つとして、自動的に作文を添削する作文支援システム「ナツメグ」の開発を試みている（八木他 2012）。その関係図を図２に示す。仁科他（2011）では、日本語のコーパスから大量な共起情報を収集し、共起表現の検索や例文表示機能を作文支援システム「なつめ」として提供することで、日本語学習者への作文支援において一定の成果をあげている。この「なつめ」で収集した共起情報に、「なたね」の学習者誤用データや、概念体系辞書を

2 http://hinoki.ryu.titech.ac.jp/natane/

加えて統計的な処理を施すことで、日本語学習者が犯しがちな誤用パターンを自動獲得する。それに基づいて、学習者作文に含まれる誤用箇所を自動的に添削することを目指している。

5. 終わりに

本論文では学習者コーパス「なたね」の概要を紹介し、誤用種別およびタグ付け作業の説明を通して、「なたね」の構築プロセスを示した。また「なたね」の応用として、検索サイトの機能を紹介し、作文支援システムに取り入れることによる多方面からの作文支援の可能性について検討した。

現在「なたね」の作文は、中国語を母語とする学習者が大半を占めているため、今後は、さらに多言語母語学習者の作文データを追加収集する予定である。

一方で誤用分析の作業においては、まだ課題が多く残されている。例えば、「誤用箇所の選定」に関しては、語単位か文単位かによる誤用箇所の「単位」をある程度決めないと、タグ付けの結果が不統一になる恐れがある。また、誤用の判定基準が主観に影響されやすいという問題点があるため、その標準化も必要である。さらに「なたね」では、誤用種別のタグ以外に、訂正例も示しているが、これはアノテータ個人の見解であり、唯一の正解ではないという立場をとっている。しかし、学習者のレベルを考慮して、そのレベルに合った訂正法も検討しなければならない。今後も誤用種別の増減の可能性を含めた上で誤用分析を行い、誤用の判定および誤用種別のフレームの改善を進める必要がある。

謝辞：本研究は一部、日本文部科学省科学研究費補助金（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」、および中国の“SRF for ROCS, SEM.”の補助により行った。

参考文献

小柳昇、テレンス・シャア、望月圭子「オンライン日本語誤用辞典の作成と日本語教育への活用の可能性」『異文化コミュニケーションのための日本語教育②』高等教育出版社, 510-511, 2011

曹紅荃、黒田史彦、八木豊、鈴木泰山、仁科喜久子「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－」『世界日語教育大会論文集』1571-1-1571-9, 2010

曹紅荃、黒田史彦、八木豊、仁科喜久子「学習者作文コーパスのための誤用種別の整備と分析」『異文化コミュニケーションのための日本語教育②』高等教育出版社, 520-521, 2011

寺村秀夫,『外国人学習者の日本語誤用例集』, 1990（大阪大学：データベース版、国立国語研究所 2011）

徳永健伸、Dain Kaplan、飯田龍,「Slate-A multi-purpose annotation tool」,『情報処理学会自然言語処理研究会報告』NL-199, 19, 2010

仁科喜久子、村岡貴子、因京子、Joyce Terence、鎌田美千子、阿辺川武「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」『特定領域研究日本語コーパス平成 22 年公開ワークショップ予稿集』215-224, 2011

八木豊、鈴木泰山,「学習者作文コーパスの構築と誤用の分析」、『日本語学習支援の構築-言語教育・コーパス・システム開発-』凡人社, 249-273, 2012

八木豊、ホドシチェク・ボル、仁科喜久子「BCCWJ と学習者作文コーパスを利用した日本語作文支援 －表記と共起に関する誤用添削プロトタイプ構築－」『第１回 コーパス日本語学ワークショップ 予稿集』315-320, 2012