言語処理技術を利用した
日本語学習者作文コーパスの開発

林 炫 情*
李 在 鎬** ・宮岡 弥生*** ・柴崎 秀子**** ・趙 堈 熙*****
(e一mail: hjlim@yamaguchi-pu.ac.jp)

目 次

1. はじめに

コーパスとは、言語処理及び言語学の研究のために作られた言語資料の集合体のことで、一般的にはコンピューター上で閲覧及び編集が可能な大規模データベースを指す。そのなかでも、学習者の産出データを収集したものを学習者コーパスと呼ぶ。学習者コーパスは、中間言語の研究、そして第2言語の学習メカニズム研究のための資源としての有用性は認められているが、開発に多くのコストがかかることから、研究ニーズに答えるだけの十分な量・質のデータが存在するわけではない (cf. 石川, 2008) 。こうした状況を踏まえ、本研究は、時代のニーズにあった学習者コーパスを設計し、それを共有することで

* 山口県立大学 准教授 第二言語習得
** 筑波大学 准教授 日本語教育
*** 広島経済大学 教授 日本語教育
**** 長岡技術科学大学 教授 日本語教育
***** 釜山大学 教授 日本語教育

日本語教育全体に資することを目的とする。本稿では、李・林・宮岡・柴崎(2012)で示した自然言語処理の技術を利用した作文コーパスの開発プロセスおよび現状について紹介する。

2. 日本語学習者コーパスをめぐる現状

李・石川・砂川(2012)では、学習者コーパスが持つ研究資源としての意義について、次の2点を指摘している。一つ目は言語データベースとしての意義、二つ目は習得研究上の意義である。一つ目の言語データベースの意義としては、次のことが考えられる。一般的にコーパスと言えば、母語話者の産出データを集めたものを指すことが多く、日本語学分野では国立国語研究所が中心になり、様々なデータの整備が行われている1)。一方、言語使用を広義に捉えた場合、日本語を産出するのは母語話者だけではない。日本語学習者も日本語の産出主体の一人であると考えられる。このことを前提にするなら、日本語という言語を総体として捉えた場合、やはり学習者の言語使用についてもデータベース化が必要であると言える。

二つ目の習得研究上の意義として、次のことが考えられる。学習者コーパスは、日本語学習者の生きた言語使用をダイレクトに反映したデータベースである。そのため、どのような学習者が、どのような言語使用を行っているかを体系的に捉えることができる。石川（2008）では、学習者コーパスを使った研究の視点として、次の 5 つをあげている。①過剰使用語や過小使用語など、学習者が使用する目標言語の言語的特徴を探る。②目標言語の使用における母語転移の程度を調べる。③目標言語において言いたいことが言えない場合に、使用される回避方略を調べる。④母語話者的な言語運用が行われる言語領域と非母語話者的な言語運用が行われる言語領域を特定する。⑤学習者が苦手とし、援助を必要とする言語領域を特定する。つまり、学習者コーパスを用いて、学習者ならではの不自然な過剰・過少使用のパターンを特定し、初級学習者にとって習得が難しい項目は何か、韓国語母語話者にとって難しい項目は何かといった問題を具体的に確認し、検討することができるのである。そして、その結果をシラバスに反映させたり、テスト問題の作成や教材の開発に応用することによって、教育現場へのフィードバックにつながるのである。

次がその一例であり、これは、日本語学習者の話し言葉コーパスであるKYコーパスを用いて助詞の誤用と動詞の誤用を調べたものである。

1)書き言葉で言えば、「現代日本語書き言葉均衡コーパス」(http://www.ninjal.ac.jp/corpus_center/bccwj)
話し言葉で言えば、「日本語話し言葉コーパス」 (http://www.ninjal.ac.jp/csj)が広く知られている。

表1．KYコーパスにおける学習者誤用例

誤用パターン	学習者	初級	中級	上級	合　計
助詞の誤用	韓国語話者	11	77	60	148
	中国語話者	19	204	168	391
動詞の誤用	韓国語話者	24	111	93	228
	中国語話者	16	129	86	231
合　計		70	521	407	998

表1の助詞と動詞の誤用例の分布に共通する傾向として、初級学習者は母語に関係なく誤用が少ないという事実が観察される。そして、中級において誤用がもっとも多く産出され、分布上の山を形成している。次に母語別に見ると、韓国語母語話者は、助詞の間違いが全体的に少ないのに対して、動詞の間違いが多いことが分かる。また、中国語母語話者に関してはどのレベルにおいても韓国語母語話者に比べ、助詞の誤用が頻出しているという事実が確認される。これは、一般に言われている母語の文法体系が類似しているかどうかが関係している問題である。このように、習得研究者が経験則から得た予測をデータでもって検証することができる点で、学習者コーパスが持つ意義は大きい。なお、コーパスを利用した日本語教育シラバスについては寺嶋(2011)を、テスト問題作成に関する応用については李(2011)を参照してほしい。

日本語学習者コーパスは、データの種類や構築の方法の違いによって、話し言葉コーパスと書き言葉コーパスの2種類に分類される。まず、話し言葉コーパスは、学習者の話し言葉の特徴を反映したもので、会話などの音声言語を録音し、それを文字起こしして構築される。とりわけ、OPI（Oral Proficiency Interview）に特化して構築された「KYコーパス」が広く使われている。一方、書き言葉コーパスは、学習者の書き言葉の特徴を反映したものであり、国立国語研究所が開発した「日本語学習者による日本語作文と、その母語訳との対訳データベース（作文対訳DB）」が有名である。その他に、海外の教育機関で収集した日本語学習者作文をデータベース化した「日本語学習者言語コーパス」、寺村(1990)の『外国人学習者の日本語誤用例集』をデータベース化した「寺村誤用例集」、日本語を母語とする大学生（134名）と日本語を学ぶ大学生（台湾57名、韓国55名）が日本語で執筆した意見文を収録した「日本・韓国・台湾の大学生による日本語意見文データベース」などがある。

表2は、日本国内で利用されている主要な書き言葉学習者コーパスを比較したものである。これらはそれぞれ収録データ数や検索仕様などの表面的な違いのほか、テーマの設定やタグセットなども異なる。そのため、一概には言えないが、これらの先行コーパスは、

1) 言語情報のアノテーションがされていないため、文字列検索以上のことはできず、多様なニーズに答えるには不十分である。2) 誤用例に対する添削やタグ情報が標準化されていない。3) 学習者の日本語レベルに関する情報が入っていない。しかしながら、学習者のデータをコーパス化するためには、データの書式などに関する標準化がなされていること、データ形式が汎用的であることなどが求められる。また、公開を前提とする場合、学習者個人の特定につながる情報やプライバシーに関する情報は、削除しておく必要がある。

そこで、本研究では、これらの問題を解決すべく、誤用例も含めたすべてのデータにアノテーションを施し、ウェブベースの検索システムを構築し、一般公開する予定である。また、作文のほかに、語彙テスト、文法テストを実施し、学習者の日本語能力を体系的に測定することを目指した。

表2. 日本国内における学習者作文コーパスの現状[2)]

コーパス名	収録データ数	全文検索	誤用タグ	検索仕様
作文対訳DB	1575編	×	△	
日本語学習者言語コーパス	1756編	○	×	文字例から検索
寺村誤用集	4601文	×	○	誤用の種類から検索
日本・韓国・台湾の大学生による日本語意見文データベース	246編	×	×	

3. 日本語学習者作文コーパス（JCコーパス）の概要

日本語学習者作文コーパス（Japanese Learner's Written Composition Corpus、 以下JCコーパス）とは、日本語学習者の作文をデータベース化したものである。JCコーパス開発にあたっては、1) 「誤用分析に影響を及ぼしうる外的要因をできる限り排除すること」、2) 「使いやすい検索ツールの開発」、3) 「コーパス構築＝データ公開」の三つの点を基本方針とした。まず、1) については、学習者の誤用分析に

2)迫田・木下・小西・李 (2012) 調べ。

おいて学習者の母語、日本語レベル、学習環境、作文テーマは重要な変数要因になりうる。そのため、JCコーパスでは、学習者の母語を「韓国語、中国語、英語」、日本語レベルを「初級、中級、上級」、学習環境を「国内、国外」、トピックを「外国語が上手になる方法」として統制を行った。これによって、多少偏りはあるが比較の目的に最適かつ分析結果の解釈が容易なデータを集めることができたと考えている。2）については、日本語教育の共有資源化を開発の目的としているだけに、利用上の制限は設けておらず、パソコンスキルがなくても簡単に用いることができる検索ツールを開発した。3）については、作文収集時においてデータ提供者(日本語学習者)に使用許可をとっており、データ公開における著作権問題をクリアしている。とくに、タグ付きコーパスであること、ウェブブラウザ上で様々なオプション付きでデータ検索ができるUIを提供していること、語彙・文法テストを実施し、学習者の日本語能力を測定していることは、JCコーパスのもつ大きな特徴である。

図1は、JCコーパスの構築手順を示したものである。JCコーパスの構築手順は、大きく「データ収集」、「電子化」、「公開」の3つの段階に分けることができる。以下では、データ収集の方法、タグセット設定、レベル判定に絞って、報告する。

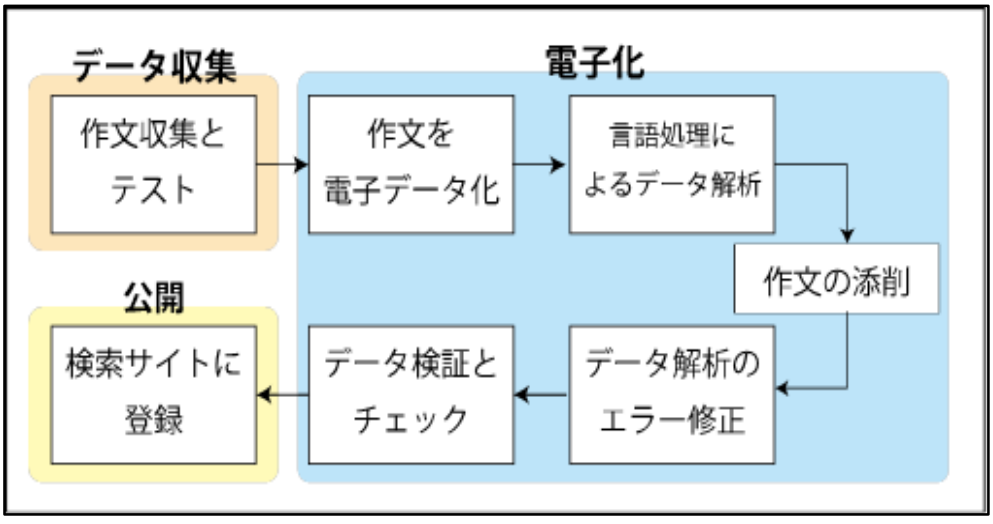

図1．JCコーパス構築手順

3.1　データ収集について

作文執筆にあたり、テーマは「外国語が上手になる方法」とし、辞書を使用せず書くよう指示したうえで、テスト形式で実施した。また、作文執筆後に学習者レベル判定のための語彙テスト（36問）と文法テスト（36問）3)を実施した。データ収集の手順と作文執筆

3)宮岡・玉岡・酒井（2011）が開発した語彙・文法テストを修正して、今回新たに作成した語彙・文法テスト。

時の条件は表 3 のとおりである。

表3. データ収集の手順と作文執筆における条件

データ収集手順	①内容説明・誓約書作成：調査の趣旨の説明。 誓約書を読み、確認（10分） ②作文収集（日本語）：日本語で作文を執筆 （50分：上限設定） ③作文収集（母語）：母国語で作文を執筆（30分） ④語彙テスト・文法テスト：テストを受ける （50分：上限設定）
作文執筆における条件	①辞書なし ②時間制限：50分 ③語数制限：300字－800字（上限設定） ④テーマ：「外国語が上手になる方法」

現在、JCコーパスには、韓国語母語話者の作文が92名文、中国語母語話者の作文が86名分、合計178名分のデータが収録されており、2012年11月末時点での作文執筆者の学習歴、性別、母語別の内訳は、表4のとおりである。

表4. JCコーパスの収録データ

母　語	学　習　歴	女　性	男　性	合　計
韓国語母語話者	2年未満	10	10	20
	2年以上5年未満	34	26	60
	5年以上	2	10	12
中国語母語話者	2年未満	2	2	4
	2年以上5年未満	56	24	80
	5年以上	2	0	2
合　計		106	72	178

コーパスのサイズは、全体の語数の述べ頻度が58,447、文字数の述べ頻度が96,510である。また、一作文あたりの平均語数は328.4、文字数は542.2語で構成されている。

3.2 タグセットについて

いわゆるテキストデータに対する文字列検索では、限られた範囲の抽出しかできない。例えば、基本形によるキーワード検索や品詞情報によるキーワード検索をしようとした場合、何らかの言語的情報を付与する必要があり、一般にはタグというものをつけ、語に対する言語情報を与える作業を行う（近藤・小森（編），2012）。こうしたタグがついているコーパスをタグ付きコーパスと言い、日本語学習者コーパスにおけるタグ付きコーパスは、李（2009）による「タグ付きKYコーパス」がある。本研究では、李（2009）の方法にならって、タグ付きコーパスとしての学習者コーパスの開発を試みた。

本コーパスの構築においては、二つのタグセットを使用している。一つ目は、形態素解析[4]に基づく品詞情報や基本形に関するタグセットである。二つ目は、人手による誤用例に関するタグセットである。形態素解析においては、解析エンジンとしては、MeCabを使用し、解析辞書としてはUniDicを使用した（MeCabおよびUniDicの詳細は、李・石川・砂川（2012）参照）。形態素解析エンジンで処理したデータを目視で確認し、解析誤りなどは人手で修正した。二つ目の誤用タグとしては、三種類のものを使用した。文字や語彙の誤用に関するタグ、文法的な誤用に関するタグ、スタイルの誤用に関わるタグである。それぞれの誤用タグを形態素単位で付与した。そして誤用タグとともに、作文意図を踏まえた添削情報も追加した。誤用タグの度数は、表５のとおりである。

表５．誤用の集計

	語彙の誤用	文法の誤用	スタイルの誤用	合計
韓国語	970	3,683	645	5,298
中国語	521	4,450	248	5,219
合計	1,491	8,133	893	10,517

それぞれの誤用例を挙げてみると、まず語彙の誤用とは、(1)のような文字レベルの誤用、(2)のような語彙選択に関する誤用である。

(1) 例え、外国語の**トラマ**とかアニメとか、小説とか見る。（中国語，2年3ヵ月）
(2) そして**分別書き**とか綴学法のような細かいところまで気にすることは色々あると思いますがよく見てください。（韓国語，7年）

4)形態素分析（Morphological Analysis）はコンピューターを利用し、機械的な方法で文を形態素に切る自然言語処理の技術の一つである。形態素分析を行うことによって、コーパスのサイズを明らにすることができる。また、頻度に対する定量的分析やデータ間の比較が可能になる。

次に、文法の誤用は、(3)のような助詞の誤用や(4)のような活用の誤用である。

(3) 日本人**を**接して話をして、ドラマを見る。（中国語，2年5ヵ月）
(4) 友だちとコミューミケーション**すれば**外国語で話たこといがいはないし、
（韓国語，3年9ヵ月）

また、スタイルの誤用は、文体に関するもので、丁寧体と普通体の混在による誤用が主である。こうした誤用情報を活用することで、学習者における習得の困難さを直接的に確認することができる。

図２は、一つの作文における誤用の平均値を示す。図２を見ると、母語の相違に関係なく、共通する傾向として、文法の誤用がもっとも多いということである。一方、母語による差として、韓国語母語話者の場合、語彙に関しては一つの作文に平均10.78回、スタイルに関しては一つの作文に平均7.17回、誤用を産出しているのに対して、中国語母語話者の場合、語彙に関しては5.93回、スタイルに関しては2.76回を産出しており、相対的に少ない。しかし、文法の誤用に注目した場合、韓国語母語話者は平均40.92回、中国語母語話者は51.36回と産出しており、中国語母語話者の誤用率が相対的に高いことがみてわかる。

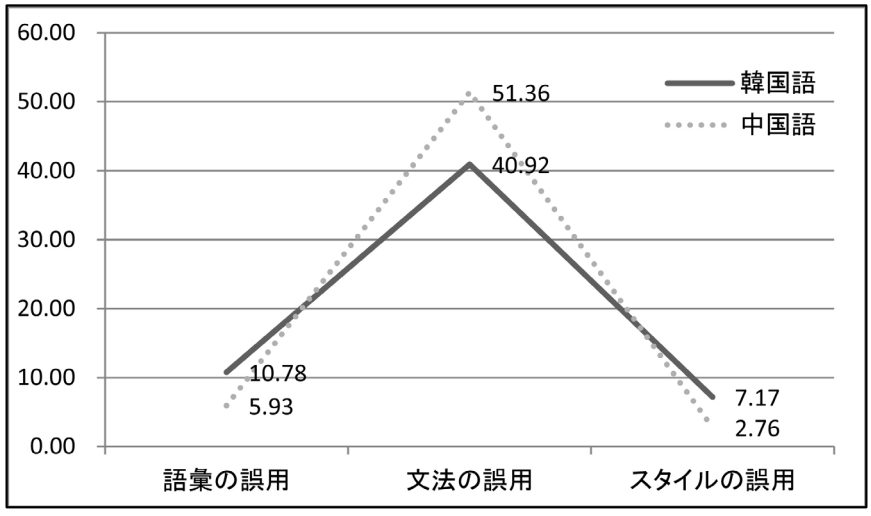

図２．誤用の平均値

次に、文法的誤用の主要項目として、動詞、助詞、形容詞および名詞の誤用例を産出した結果、図３の分布が確認された。具体的にみると、文法的誤用のもっとも主要なものは助詞であること、それに次いで、形容詞、名詞の誤用、動詞の誤用の順になっていることが明らかになった。

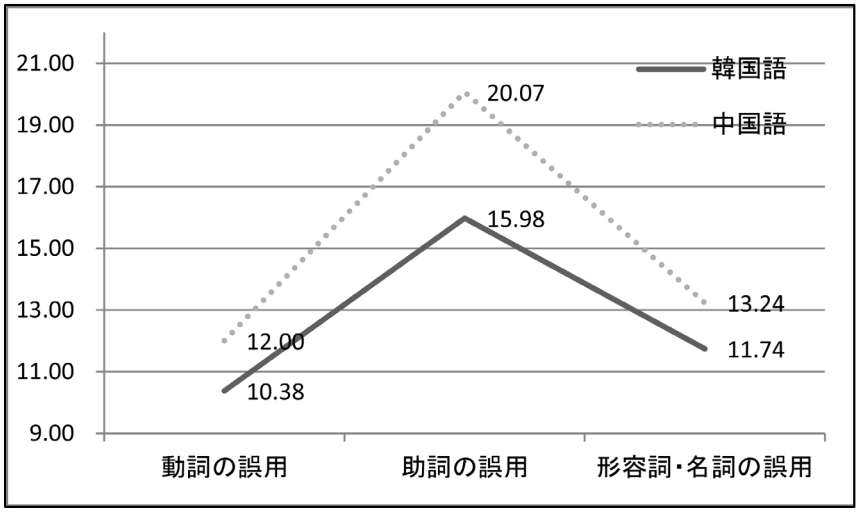

図３．主要な文法的誤用の平均値

一方、母語の相違に注目した場合、動詞や形容詞、名詞の誤用に見られる差に比べ、助詞の誤用に見られる差が相対的に大きく、中国語母語話者においては、助詞の学習が大きな壁になっていることが示唆された。

3.3　レベル判定について

学習者コーパスにとって、データの日本語レベルを明示する作業は不可欠である。多くの学習者コーパスでは、学習歴などの関連情報をもとに、レベルを推定する手法が取られている。しかし、日本語学習歴は必ずしも日本語能力を反映する情報とは言えない。学習者の日本語能力に関しては、主観を排除し、客観的な評価基準をもとにレベル判定を行う必要がある。このことを踏まえ本研究では、宮岡・玉岡・酒井(2011)の日本語テストを修正して新たに作成した日本語テストの得点にもとづいてレベル判定を行った。まず、語彙テストと文法テストの得点をもとに、三つのグループに分けた。80%以上をAグループ、79%から60%をBグループ、59%未満をCグループに設定した。そして、語彙テストと文法テストのいずれにおいても、Aグループに入る学習者を上級にした。語彙テストがAグループであるが、文法テストがBないしはCグループに入る学習者と文法テストがAグループであるが、語彙テストがBないしはCグループに入る学習者を中級にした。そして、語彙と文法テストのいずれもCグループに入る学習者を初級にした。それぞれの集計は、以下の通りである。

表 6．レベル分け

母　語	初級	中級	上級	合　計
韓国語	13	47	32	92
中国語	18	47	21	86
合　計	31	94	53	178

表6のレベル分けで、JCコーパス内の全データは、ひも付けされており、キーワードとレベルを組み合わせて柔軟な検索ができる仕組みを開発した。

4．検索ツール開発の概要

JCコーパスは、主たるユーザーとして一般の日本語教師を想定しており、パソコンに関する専門的スキルを必要としない利用環境を用意する必要があった。また、世界中の日本語教師に使ってもらうことを考えているため、OSや言語に左右されることのない、利用環境を構築する必要があった。この 2 点の要請を踏まえ、ウェブインターフェイス上で、データにアクセスし、検索およびダウンロードができる環境を作ることにした。開発環境としては、スクリプト言語の「python」(http://python.jp/)とウェブCGI (Common Gateway Interface) 5)の技術を利用し、ウェブ・アプリケーションとしてのコーパス検索ツールを開発した (図4) 。図4のウェブ・アプリケーションを利用した場合、以下の検索が可能になる。

①形態素単位で検索する：語基単位で検索を実行し、KWIC6)データが取得できる。
②文字列単位で検索する:表層の文字列単位で検索ができる。
③検索オプションを指定する：学習者のレジスタに応じた条件指定ができる。また、検索したいキーワードの誤用が文法的な誤用か、文字・語彙的な誤用か、文体の誤用か選択できたり、作文課題および学習者の母語、日本語学習歴、作

5)CGIとはWebサーバが、Webブラウザからの要求に応じて、プログラムを起動するための仕組みのこと。荻野・田野村 (編) (2012) 参照。

6)KWIC(keyword in context)とは、検索キーワードを中心に、その前後の文脈を同時に表示する索引手法であるが、調べる語や語句がどのような前後関係で使われるのかを調べるのに大変有効である。

文と同時に行ったテスト成績情報を使った絞り込みができる。

図４．JCコーパス検索画面

以上の検索を行ったあと、検索の結果は、図５のKWIC列によって確認することができる。KWIC列を見ることで、キーワードの前後にどのような語彙や表現が使用されているかを確認できるとともに、誤用や添削に関する具体的な情報を見ることができる。

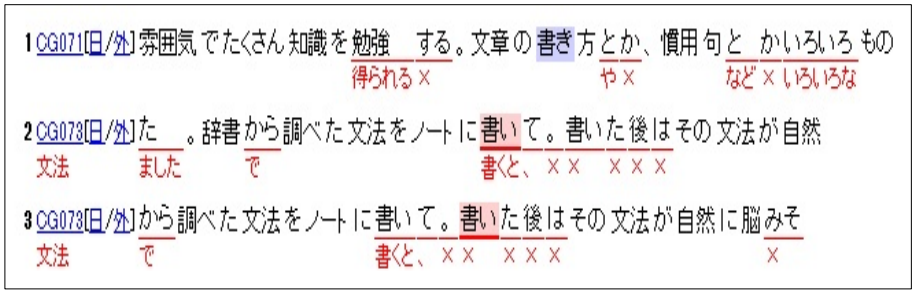

図５．KWIC表示

図５の下線の下に表示されている情報は誤用例に関する添削情報である。また、誤用の分類も表示され、文法的な誤用か、語彙・文字表記に関わる誤用か、文体的な誤用かが表示される。また、学習者のIDをクリックすることで、図6の学習者プロフィールや作文の全文閲覧ができる。そして、［日/外］をクリックすると、元の手書き作文データをPDFファイルで見ることもできる。なお、「日」は日本語、「外」は外国語のデータを示す。

さらに、検索の結果は、エクセルファイルないしはテキストファイルでダウンロードすることができるため、ローカル環境でさらに分析を進めることが可能である。

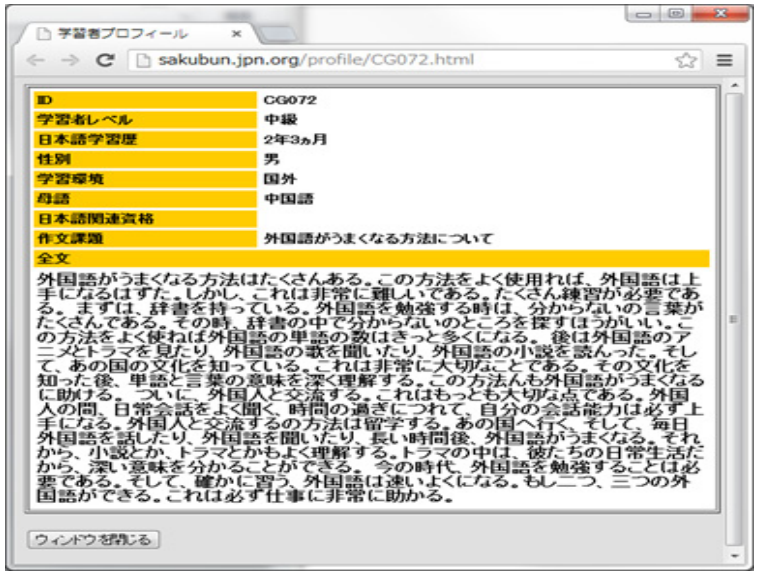

図6. 学習者プロフィール画面

5. おわりに

以上、JCコーパスの概要について述べてみた。現在、中国語母語話者と韓国語母語話者に関しては収集が終わり、合計178名の作文が利用可能な形になっている。今度の課題としては、①英語母語話者も含めた幅広いデータを収集すること、②タグの誤りを詳細にチェックすることがあげられる。今後は、データの拡充を図りながら、複数の添削者による添削、タグの誤りの検出と修正、タグの細分類、そしてレベル判定の精緻化を実施していきたいと考えている。また、JCコーパスのより効果的な日本語教育への応用を視野に入れ、その活用可能性を模索していきたい。

【参考文献】

石川慎一郎（2008）『英語コーパスと言語教育』大修館書店
李在鎬(2009)「タグ付き日本語学習者コーパスの開発」『計量国語学』27(2). 60-72
李在鎬(2011)「大規模テストの読解問題作成過程へのコーパス利用の可能性」『日本語教育』148. 84-98
李在鎬・石川慎一郎・砂川有里子(2012)『日本語教育のためのコーパス調査入門』くろしお出版
李在鎬・林炫情・宮岡弥生・柴崎秀子(2012)「言語処理の技術を利用したタグ付き日本語学習者コーパスの構築」日本語教育学会2012年春季大会デモンストレーション
荻野綱男・田野村忠温(編)(2012)『講座　ITと日本語研究8　質問調査法と統計処理』明治書院
近藤安月子・小森和子(編)(2012)『研究社 日本語教育事典』研究社.
迫田久美子・木下藍子・小西円・李在鎬（2012）「日本語学習者の縦断的会話コーパスの構築と習得研究―3年間のデータから文法習得の過程を探る―」『2012 年度日本語教育国際研究大会予稿集』. 206
寺嶋弘道(2011)「日本語教育におけるコーパスの応用　－データ駆動型学習とその実践方法の考察－」『POLYGLOSSIA』20. 91-102
寺村秀夫(1990)『外国人学習者日の日本語誤用例集』(大阪大学;データベース版、国立国語研究所 2011年)
宮岡弥生・玉岡賀津雄・酒井弘（2011）「日本語語彙テストの開発と信頼性: 中国語を母語とする日本語学習者のデータによるテスト評価」『広島経済大学研究論集』34(1). 1-18

【日本語学習者データベース】

■話し言葉を収録した学習者コーパス
「KYコーパス」http://opi.jp/shiryo/ky_corp.html

■書き言葉を収録した学習者コーパス
「作文対訳DB」http://jpforlife.jp/taiyakudb.html
「日本語学習者言語コーパス」 http://cblle.tufs.ac.jp/llc/ja/index.php?menulang=ja
「寺村誤用例集」http://www.ninjal.ac.jp/teramuragoyoureishu/
「日本・韓国・台湾の大学生による日本語意見文データベース」
http://www.tufs.ac.jp/ts/personal/ijuin/koukai_data1.html

謝辞：本研究はJSPS科研費22520537「自然言語処理の技術を利用したタグ付き学習者作文コーパスの開発」（代表者：李在鎬）の助成を受けたものである。

要 旨

本稿では、自然言語処理の技術を利用した日本語学習者作文コーパス（JCコーパス）の開発プロセスおよび現状について紹介した。学習者コーパスは、学習環境や母語などの要因が学習者の産出にどのように影響を与えるのかを調べることができるため、学習者のレベルや学習環境に即した授業法や教材の開発、テスト作成などに活用できる点で、言語教育への貢献の余地と利用価値は高い。しかし、開発に多くのコストがかかることから、研究ニーズに答えるだけの十分な量・質のデータが存在するわけではない。また、すでに開発されたコーパスであっても、1）言語情報のアノテーションがされていないため、文字列検索以上のことはできず、多様なニーズに答えるには不十分である。2）誤用例に対する添削やタグ情報が標準化されていない。3）学習者の日本語レベルに関する情報が入っていない。といった問題点が挙げられる。

JCコーパスの開発にあたっては、1）「誤用分析に影響を及ぼしうる外的要因をできる限り排除すること」、2）「使いやすい検索ツールの開発」、3）「コーパス構築＝データ公開」の三つの点を基本方針とした。現在、中国語母語話者と韓国語母語話者に関しては収集が終わり、合計178名分の作文が利用可能な形になっている。今度の課題としては、①英語母語話者も含めた幅広いデータを収集すること、②タグの誤りを詳細にチェックすることがあげられる。今後は、データの拡充を図りながら、複数の添削者による添削、タグの誤りの検出と修正、タグの細分類、そしてレベル判定の精緻化を実施していきたいと考えている。また、JCコーパスのより効果的な日本語教育への応用を視野に入れ、その活用可能性を模索していきたい。

キーワード：学習者コーパス、日本語学習者作文コーパス、言語処理技術、
学習環境、学習者レベル、誤用分析、検索ツール

투　　고：2012. 11. 30
1차 심사：2012. 12. 15
2차 심사：2013. 1. 5