様式Ｃ－１９

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 25 年 5 月 31 日現在

機関番号：１２１０２
研究種目：基盤研究(C)
研究期間：2010～2012
課題番号：２２５２０５３７
研究課題名（和文）　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

研究課題名（英文）　Development of the Japanese Learner’s Written Composition Corpus based on the Natural Language Processing

研究代表者
李　在鎬（LEE JAEHO）
筑波大学・人文社会系・准教授
研究者番号：20450695

研究成果の概要（和文）：
本研究の目的は自然言語処理の技術を利用し、タグ付き日本語学習者作文コーパスを構築することである。2013 年 3 月に日本語学習者作文コーパスの開発を完了した。現在，「http://sakubun.jpn.org/」で一般公開を行っている。今後、日本語教育分野における研究資源として共有できると期待される。システム公開のほか，研究成果の公開として，著書１件，論文２件を公開した。

研究成果の概要（英文）：
The purpose of this research is to use the technology of natural language processing and to build a annotation learner composition corpus. Development of the “Japanese Learner’s Written Composition Corpus” was completed in March, 2013. A system can be used on “http://sakubun.jpn.org/”. The JSL researcher will use as a research resource. We published one Book and two papers as the result of research.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１０年度	1,100,000	330,000	1,430,000
２０１１年度	1,200,000	360,000	1,560,000
２０１２年度	800,000	240,000	1,040,000
年度			
年度			
総　計	3,100,000	930,000	4,030,000

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：学習者コーパス，自然言語処理，日本語作文

１．研究開始当初の背景

欧米諸国や韓国、中国などの諸外国では競って自国語のコーパス化を進め、言語政策などの基礎データベースとして活用している。日本の場合、法的制限からコーパス化がなかなか進まない現状にあるが、近年、母語話者の産出データに関しては国語研が中心となって急速に整備作業を進めている。一方、学習者言語のコーパス化については依然として遅れており、教育的課題に対して十分に対応できていない。

さて、学習者データをコーパス化すること

は、３つの意味を持つ。1)「外国語」が持つ生きた姿を観察することができる。James(1992)が指摘することであるが、外国語における真生性は母語話者のデータではなく、学習者の言語運用の中から見出すべきものである。2) データ同士の相互比較ができる。コーパスは設計から収集、解析、格納のすべてのプロセスにおいて一貫した方針のもとで構築されており、構造化されたデータベースである。そのため、データ間の比較が容易である。例えば、学習者同士の個別的比較やサブグループ同士の比較、さらには学習者データと母語話者のデータを比較することもできる。こうした考察を通し、外国語としての日本語が持つ様々な姿を体系的に考察することができる。3) 学習者の言語運用の実態を踏まえながら、教授法や辞書、教科書などの教育コンテンツを開発することができる。

２．研究の目的

本研究の目的は、言語処理の技術を利用し、タグ付き学習者作文コーパスを構築することである。本コーパスには多様な属性を持つ学習者の作文データが格納される。構築したコーパスは、ウェブインターフェイスを介して段階的に公開し、ユーザの評価を受けながら、漸次的に共有資源化する。このコーパスを使うことで、簡単に学習者の産出実態を、母語や習熟度や学習環境別に検索することができる。本コーパスは中間言語の研究データとして活用できるほか、習熟度に応じた教材や辞書などの教育コンテンツ開発、さらには試験問題開発などの基礎資料としても使うこともできるので、その波及効果は大きいと言える。

３．研究の方法

三つの中核的な方針のもとで研究開発を行う。1)学習環境の多様性に対応すべく、複数のサブグループを作る。すべてのグループで単一の課題による作文を行う。これにより、学習者の中間言語に影響する要因を特定することができる。2)データ解析や添削に用いるタグセットは既存の言語資源との互換性を重視して設定する。従って、本研究独自のタグセットは可能な限り設定しない。とりわけ形態素タグは国立国語研究所が進めている「日本語コーパス」の規格に概ね準拠する。また、誤用タグについても KC コーパスのタグセットを使用する。3)「コーパス＝データ公開」を大原則とする。そのため、データ収集時に著作権処理をし、二次配布以外の利用制限は設けない。また、公開の基本方針としてコーパス全体の完成を待つのではなく、モニター版を段階的に公開していく方式をとっており、ユーザからのフィードバックを受けながら精度向上を図る。

データ設計においては三つの属性を使う。一つ目は国内の学習者か国外の学習者なのか、二つ目は母語は何か、三つ目は、日本語レベルである。日本語レベルの判定は、学習者コーパスにおいてもっとも重要な部分であるため、宮岡 他(2009)が開発した語彙テストと文法テストを実施する。

データ解析においては、UniDic に基づく短単位を基本にする。誤用例については、人手による修正と誤用タグを挿入する。具体的には次の 3 種類の誤用タグが挿入する。1) 文字語彙に関連する誤用(例:私自身はそういう情況が時々発生します), 2) 文法に関連する誤用(例:習うの後は必ず復習がある), 3) 文体に関連する誤用(例:韓国も漢字を使うけど，漢字をそのまま使わないで音だけ使う)である。

データ公開においては、作文提供時に公開に関する承諾書をとっており、必要な著作権処理を行った。

４．研究成果

当初の申請通り，2013 年 3 月に日本語学習者作文コーパスの開発を完了し，検索環境も含めてウェブで利用できるシステムを開発した。収集データの詳細は以下の通りである。

表 1. 収集した作文数

	韓国語 母語話者	中国語 母語話者	総計
初級学習者	11	20	31
中級学習者	50	99	149
上級学習者	83	41	124
総計	144	160	304

単位：編

表 1 では、コーパスに登録された作文数を示した。表 2 では、語数を示す。

表 2. 収集した作文の語数

	韓国語 母語話者	中国語 母語話者	総計
初級学習者	4055	8303	12358
中級学習者	17448	38335	55783
上級学習者	29615	15798	45413
総計	51118	62436	113554

単位：語

表 2 の語数が示すように、合計 113,554 語の

データに対してアノテーションを完了した。そして、人手で作成した誤用例も文字語彙に関する誤用が 1575 件、文法に関する誤用が 8877 件、文体に関する誤用が 951 件含まれている。母語別の一人あたりの平均値を見ると以下のような分布になる。

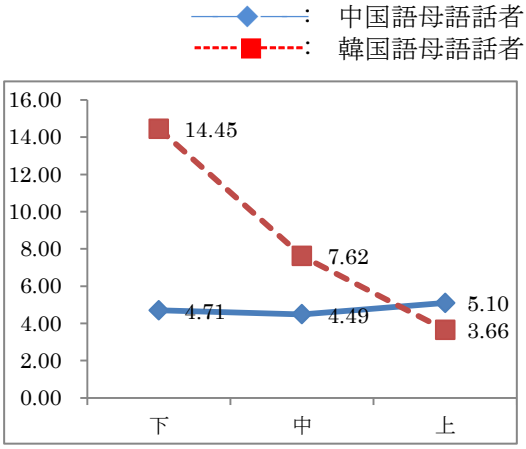

図 1a. 語彙的誤用

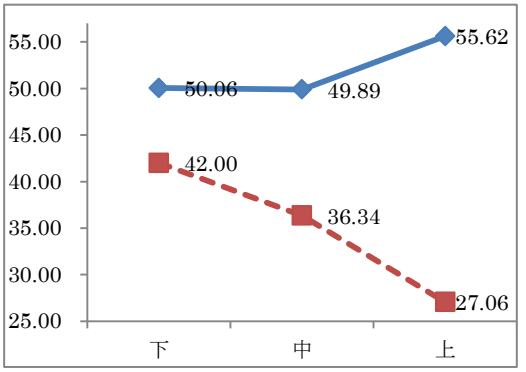

図 1b. 文法的誤用

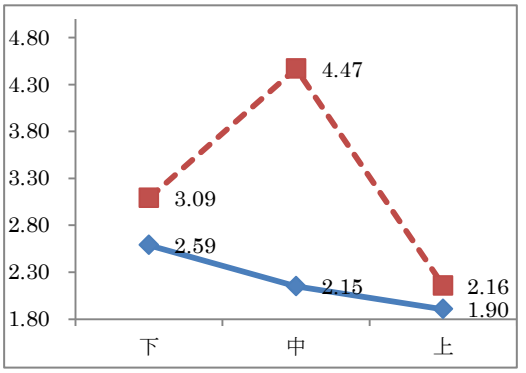

図 1c. 文体的誤用

図 1 の a)は，語彙に関する誤用，b)は文法に関する誤用，c)は文体に関する誤用である。いずれも，一作文あたりの平均値である。ここで，まず注目すべきこととして，学習が進むにつれ誤用が減ることで日本語能力が上がったことを教師としては実感すると考えられるが，実際のデータは必ずしもそうなっていないということである。語彙の誤用に関して言えば，韓国語母語話者の場合，日本語能力があがるにつれ，減少する傾向が見て取れるが，中国語母語話者の場合ほとんど変化がない。次に，文法の誤用に関して言えば，韓国語母語話者の場合，日本語能力があがるにつれ，減少する傾向が見て取れるが，中国語母語話者の場合，下グループ（初級学習者）から中グループ（中級学習者）にいくにつれわずかに減ったものの，上グループ（上級学習者）では上昇している。さらに，文体の誤用については，中国語母語話者の場合，緩やかではあるが，レベルが上がるにつれ減少する傾向があるが，韓国語母語話者の場合，増えてから減るという逆Ｖ字型の分布を見せている。これは，母語によって表れる誤用の頻出パターンが異なることを示すものであり、大規模なコーパス分析によって明らかになる事実と言える。

アノテーション済みのコーパスは、現在，「http://sakubun.jpn.org/」ドメインで一般公開している。

図 2. 検索画面

図 2 の検索画面から、キーワード検索を行うことで、作文データの中身を確認することができる。図 2 の検索オプションとしては、以下のものが用意されている。

- 誤用と正用の選択：キーワードに対して正用だけ、誤用だけ、あるいは両方のいずれかを選択することで、表示する用例の絞り込みができる。
- 学習者の選択：学習者のレベルや性別，データ収集場所が国内か国外かを選択することで、検索対象を絞り込みができる。
- 誤用タイプの選択：誤用のタイプが文字、文法、文体のどれなのかを選択し、表示

する用例の絞り込みができる。
- 学習者の母語，学習歴の選択：学習者の母語や学習歴を選択することで、検索対象を絞り込みができる。
- テスト成績の選択：テスト成績別に検索対象を選択できる。値として、A は上位グループ(得点率 80%～100%)，B は中位グループ（得点率 79%～60&)，C は下位グループ（得点率 0%～59%)，X は受験していない学習者を意味する。

そして、検索の結果として、図 3 が表示される。

図 3. 検索結果

図 3 では、KWIC 形式により、検索結果を表示する。さらに、以下の機能を持つ。

- クロステーブルの表示：レベル×母語でクロス集計表を生成する。
- 学習者のレジスタの表示：センテンス単位で学習者のレジスタが確認できる。
- 全文表示：センテンスを含む全文データが確認できる。
- 検索結果をダウンロードできる。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計２件）

① 李在鎬・林炫情・宮岡弥生、学習者コーパスと言語テスト：言語テストの得点と作文のテキスト情報量の関連性、言語教育評価研究、査読有、３巻、2013 年(印刷中)

② 林炫情、李在鎬、宮岡弥生、柴崎秀子、趙堰熙、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、日本文化学報、査読有、No.56、2012、129-142.

〔学会発表〕（計０件）

〔図書〕（計１件）

① 李 在鎬、石川 慎一郎、砂川 有里子、くろしお出版、日本語教育のためのコーパス調査入門 、2012、250

〔その他〕

ホームページ（作文コーパスシステム）
http://sakubun.jpn.org

６．研究組織

(1)研究代表者
李在鎬（LEE JAEHO）
筑波大学・人文社会系・准教授
研究者番号：20450695

(2)研究分担者
宮岡弥生（MIYAOKA YAYOI）
広島経済大学・経済学部・教授
研究者番号：10351975
林炫情（LIM HYUNJUN）
山口県立大学・国際文化学部・准教授
研究者番号：30412290
柴崎秀子（SHIBASAKI HIDEKO）
長岡技術科学大学・工学部・教授
研究者番号：00376815