タグ付き日本語学習者コーパスの開発

李 在鎬[1]・淺尾 仁彦[2]・濱野 寛子[2]・佐野 香織[3]・井佐原 均[1]
1 情報通信研究機構、2 京都大学、3 お茶の水女子大学

1. 背景と目的

第二言語習得の研究パラダイムが誤用分析から中間言語へシフトしたことで、学習者言語の実態を知ることの重要性が様々な研究で強調されるようになった。こうした流れは、必然的に大量言語データに対する網羅的・探索的調査分析の必要性を強調するような方向づけを行った。このような背景のもとで 90 年代後半、日本語学習者の書き言葉・話し言葉の資料が進められ、積極的に利用されつつある([1], [3], [4])。こうした努力によって、学習者言語の実態に関する多くの研究成果が世に示された。ただ、これらの研究成果は、主として文字列を基盤にして行った調査である点で、限定的な調査に留まっていた。また、手作業によって事例収集がなされたことで、計量的調査としての信頼性に関して必ずしも十分ではない部分もあった。こうした現状から、日本語学習者のコーパスの有用性やコーパス分析が持つ潜在的な可能性に関しては必ずしも十分に示せていないのが現状であると言える。

以上の背景を踏まえ、本研究では 2000 年以降もっとも多くの研究において利用されてきた「KY コーパス」を電子的な方法で調査することを前提にした、データ加工を行った。具体的には形態素解析ソフト「茶筌」や「分類語彙表」を使用し、言語情報をタグとして付与した。そして、タグ付与の結果を人手で修正することで、より信頼性の高いデータを作成した。また、学習者の誤用や言い直しなどをタグとして付与することで、様々な習得研究へ応用できるデータを作成した。さらに、加工データに対する検索ツールも作成した。このツールは Excel の VBA を使って作成されたものであり、Excel 以外のソフトはインストールする必要がなく、テキスト処理の知識を持たない人でも簡単に使える。以下では、本研究が行ったデータ加工の詳細とその結果を報告する。同時に、専用の検索ツールを紹介する。さらに、本研究と同様の手順および分析ツールを用いることで、多くの個人研究者が有する、学習者データに対しても計量的な調査分析が可能になる、ということがあり、関連コミュニティにおけるコーパス分析の活性化を図る。

2. KY コーパスと形態素修正作業

2.1. KY コーパス

本節では、元データとなる KY コーパスとはどのようなものか簡単に紹介する。KY コーパスとは、OPI(oral proficiency interview)に基づく日本語学習者の発話データを文字化した言語資料である。なお、OPI とは外国語学習者の会話のタスク達成能力を一般的な能力基準を参照しながら対面のインタビュー方式で判定するテストであり、ACTFL(The American Council on the Teaching of Foreign Languages)によって開発されたものである([2])。その規模は、初級から超級の韓国語・中国語・英語母語話者 90 人分のデータをプレーンテキストで格納している。(1)にデータサンプルを示す。

(1) T:いつ日本にはいらっしゃいましたか
S:そうですね 1990ー年の3月末ごろ〈ああそうですか〉いや、もうそろそろ三年になるんですよね
T:ああ、そうですか、3年間、あの、ずっとこちらの［会社名1］で仕事をなさっているんですか
S:そーではないんです、〈あーそうですか〉あのーそうですね、最初来たとき、ですね

T で始まる文字列がテスターのデータ、S で始まる文字列が学習者のデータとなっている。なお、テスターの相づちや学習者のポーズなどが別表記で記されてはいるが、形態素情報といった言語情報は一切付与されていない。

2.2. データの加工における問題点

高精度の情報検索のため、KY コーパスに対する加工を行った。加工における具体的な目的としては以下の二点が挙げられる。

1 単語区切りを認定することで、より的確な検索を可能にする。
2 言語情報を付与することで、文法研究・意味研究への積極的な利用を可能にする。

日本語のテキストデータのコーパス化のためのもっとも基本的な手順として単語区切りを認定する必要がある。というのは、日本語の場合、分かち書きがないことから、単純な文字列検索では、分析者が意図しないゴミが大量に混入するという問題が発生し、データ抽出後の分析において大変な手間がかかる。また、コーパス分析的観点から見た場合、この作業によってコーパス全体の規模が明らかになり、統計的な指標を用いた共起分析なども可能になる。次に、言語情報を付与する背景として、文法研究のための利用を想定しており、文字列に還元できない類の調査に利用することを目論んでいる。

以上の目的から、学習者データを形態素解析した。しかし、学習者の会話データを形態素解析することには

いくつか本質的な問題点がある。とりわけ以下の三点が考えられる。

1. 会話文における解析精度の問題
2. 誤用例や言い直しの問題
3. 談話的情報の介入問題

まず、1 の問題として近年高精度の形態素解析器として利用されてきている「茶筌」や「Mecab」などが示す100%に近い精度は、多くの場合、新聞データのような規範的な表記と文体が固定されたテキストデータを元に産出したものである。会話や談話データのような省略や断絶的なデータに対しては十分に妥当な結果を出すまでには至っていないのが現状である。具体例を示す。

(2) a. ご飯を食べている
　　b. ご飯食べてる
(3) a. 君の元へ走っていく
　　b. 君の元へ走ってく
(4) a. それがね、とてもおいしいの
　　b. それがですね、
　　c. それをですね、

(2)から(4)のデータを形態素解析した場合、(5)から(7)になる(／は単語区切り、( )は品詞)。

(5) a. ご飯(名詞)／を(助詞)／食べ(動詞)／て(助詞-接続助詞)／いる(動詞-非自立)
　　b. ご飯(名詞)／食べ(動詞)／てる(動詞-非自立)。
(6) a. 君(名詞)／の(助詞)／元(名詞)／へ(助詞)／走っ(動詞)／て(助詞-接続助詞)／いく(動詞-非自立)
　　b. 君(名詞)／の(助詞)／元(名詞)／へ(助詞)／走っ(動詞)／て(助詞-接続助詞)／く(動詞-非自立)
(7) a. それ(名詞)／が(助詞)／ね(動詞)／、(記号)／とても(副詞)／おいしい(形容詞)／の(名詞)
　　b. それ(名詞)／が(接続詞)／です(助動詞)／ね(助詞)／、(記号)
　　c. それ(名詞)／を(助詞)／で(動詞)／すね(名詞)／、(記号)

(2)a や(3)a の標準的表現に対して、(2)b や(3)b のような会話体特有の省略体を形態素解析した場合、(5)b や(6)b が示すような解析結果を出力する。解析の一貫性の問題はさておいたとしても、ユーザーの立場から「て」の用法を調査しようとした場合、検索から漏れてしまう可能性がある。類似の問題として、(4)の終助詞表現に関しては、(7)a が示すように、終助詞の「ね」を「ねる」を基本形とする動詞であると解析する。また、(7)b では、格助詞の「が」を「接続詞」と解析するという誤りがみられる。(7)c では、格助詞の「で」を「でる」を基本形とする動詞に誤解析し、「すね」を名詞であると解析する。

次に、2の問題は、学習者データを扱う上で必然的に生じてくる問題で、多くの誤用例が含まれているという問題が挙げられる。具体例として、(8)から考えてみたい。

(8) a. 早いのほうがいいじゃない(CA01)

(8)では、中国語学習者に多いとされる、助詞「の」の過剰使用による誤用例である。この種のデータをそのまま、形態素解析した場合、表.1 のような誤った解析をしてしまう。

表.1　(8)の解析例

文字列	読み	原型	品詞の種類
早い	ハヤイ	早い	形容詞-自立
の	ノ	の	名詞-非自立-一般
ほう	ホウ	ほう	名詞-一般
が	ガ	が	助詞-格助詞-一般
いい	イイ	いい	形容詞-自立
じゃ	ジャ	じゃ	助詞-副助詞
ない	ナイ	ない	助動詞

誤用と同様の問題として、(9)のような言い直しなども頻繁に行われる。

(9) a. うれしいじゃなくて、たの、楽しい、、(KIM01)

この種の問題は形態素解析器の問題ではなく、データそのものに見られる問題として位置づけることができる。

最後に、3 の問題は、KY コーパスの場合、語間あるいは語の内部で相槌やポーズが入っており、形態素解析の精度を下げる要因になる。具体的には、(10)のような事例である。

(10)　～です〈んーそうですね〉で、
(11)　～です(助動詞)／で(助動詞)／、(記号)

(10)を形態素解析した場合、(11)になるが、テスターの相づちが介入することで、接続詞としての「で」が助動詞として解析されてしまう。

以上の示した三点の問題から、形態素解析技術の有用性は認められるものの、現状としてその結果を鵜呑みにし、調査に利用することは難しいと考えられる。そこで、本研究では形態素解析の結果をすべて人手でチェックし、修正が必要な箇所は人手で修正した。

2.3. データ加工の手順

前節の問題点を踏まえ、KY コーパスに対する言語情報を付与した。作業は次のの手順で行った。まず、KY コーパスのオリジナルデータから S で始まる文字列

を抽出し、学習者発話データを生成した。次に「茶筌」を用いて形態素情報を付与した。次に、人手によるチェック作業の利便性を考慮し、形態素解析済みデータをエクセルブック形式に変換した。以上の手順によって、作業の元データが完成した。次に、データ加工の最初の段階として、単語区切りと形態素情報の修正を行った。まず、(12)を KY コーパスの文字化ルールに従って、Perl のスクリプト処理で一括加工した。

(12) a テスター発話:"〈…〉"で記された文字列
b ポーズ:"、"で記された文字列
c 個人情報:"[…]"で記された文字列
d 非言語情報:"{…}"で記された文字列

(12)a は、学習者の発話の途中に入るテスターの相づちなどを指しており、KY コーパスでは、山括弧で記されている。(12)b は被験者によるポーズ、(12)c は被験者の個人情報に関わるものである。(12)d の非言語情報は、発話の途中に入っている「笑い」などのことであり、KY コーパスでは中括弧で記述されている。

次に、作業者によるチェックでは、5 名の作業者が第一段階のチェックを行い、誤りがある場合は、手入力で修正した。修正箇所はエクセルの変更履歴として残すようにし、統括者によるチェック時に利用した。修正の際には、「茶筌」による誤解析のみならず、以下の誤用タグを導入した。いずれも「茶筌」の品詞情報を元にしており、第一階層の後に誤用タグを付与した。

(13) a 名詞−誤用:どれくらい返額できますか
b 動詞−誤用:他の学部の人たち、分からない、知る人がないから
c 形容詞−誤用:文化の、重みが深いし、
d 助動詞−誤用:あのう機械でするです
e 助詞−誤用:10 月の日本へまいりました
f 副詞−誤用:学生たちの判断はよく尊敬します
g 連体詞−誤用:その以外のはたとえば、いろいろ、

誤用例の処理においては、本研究の限界でもあるが、形態素レベルで記述した。以下の 3 パターンに対して各々の処理を行った。

(14) a 形態素の用い方に対する誤用:三人の主人公がありま す
b 不要な形態素を用いたことによる誤用:先生からもらったの資料
c あるべき要素を省略したことによる誤用:上海いえば、川カニね

(14)a のタイプには、問題箇所に誤用タグを振った。(14)b のタイプには不要な形態素のところに誤用タグを振った。(14)c に関しては誤用ともっとも直接的な共起を持つ形態素に誤用タグを振った。また、形態素修正とともに言い直しや母語の混入による発話もタグ付与した。

作業者による第一チェックが済んだ段階で本稿の第一著者が統括者として最終確認をし、漏れや判断のゆれに対して、最終的な判定を行った。次に単語区切りや形態素の修正が終わった段階で、「分類語彙表」(Ver.1.0)による意味情報を付与した。作業は python を用いて行った。現在、意味情報の付与は、多義性に対する問題には対処できていない。というのは、本研究では、「分類語彙表」にターゲット語が複数出てきた場合、もっとも最後に出現する項目をターゲット語の意味として採用した。その理由として、「分類語彙表」の並びが「関係」から始まって「自然」で終わるので、関係のような抽象的な意味より、自然のような基本的な意味で、意味情報を付与したほうが良いと考えたからである。ただし、このことはあくまで便宜上の理由以上のものではなく、明確な根拠はない。今後の課題としたい。

2.4. データ加工の結果

データ加工の結果、総計 173,198 形態素を得ることができた。(12)など取り除いた結果、表 2 が得られた。

表.2 レベル別形態素の度数分布表

品詞および正誤		初級	中級	上級	超級
名詞	正用	2219	13519	19813	11699
	誤用	106	602	439	42
動詞	正用	594	4525	8264	5247
	誤用	59	339	265	51
形容詞	正用	169	1078	1782	845
	誤用	12	63	37	4
副詞	正用	203	1708	3826	2207
	誤用	5	57	64	29
連体詞	正用	26	505	1016	768
	誤用	4	32	28	9
助動詞	正用	943	5765	8459	5229
	誤用	29	172	146	28
助詞	正用	1163	10493	21147	12934
	誤用	111	559	535	59
接続詞	正用	80	717	1099	782
	誤用	7	14	13	3
感動詞	正用	966	2853	2787	1294
	誤用	2	7	0	1
フィラー	正用	528	3262	3775	1764
その他	正用	286	991	1306	570

表 2 のデータに関しては、母集団の個数が異なるため、平等な比較が難しい。そこで、個々の形態素別学習者の一人当たりの平均値を求め、その正誤の比率から正答率を算出した。その結果、以下の分布が得られた。

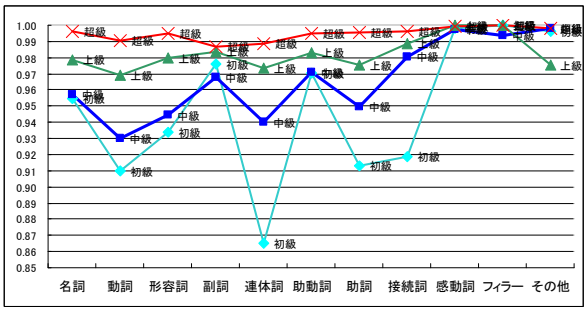

図.1　　正答率の推移

3.　検索ソフト

3.1.　開発背景

本研究では、KY コーパスのタグ付け結果を検索するための専用のツールを製作した。開発における背景としては、2 節の手順で加工したデータを自由かつ簡単に検索可能なツールは、事実として存在しないからである。というのは、本研究の開始当初は「茶器」によるデータ検索を想定しており、事実、修正の元データは「南瓜」で係り受け情報も付与していた。しかし、「茶器」が利用する「MySQL」は多くの人文系の研究にとってはインストールが難しく、簡単に利用可能な環境を保証しない。さらに、コーパスの格納段階において、細かなエラーが出ることがあるが、データベースソフトの仕組みについての理解がない利用者にとって、自らそのエラーに対策を講じることはほぼ不可能に近い。

以上の背景から、機能性より簡単さを優先したツールを製作した。その方針として可能な限り外部ソフトのインストールなしで使えるツールを心がけた。そこで着目したのが MS Office のエクセルである。エクセルは a)多くの家庭用・業務用計算機において、すでにインストールされている場合が多い。b)多くの人文系の研究者が日常的に使用しており、操作に慣れている。以上の理由からエクセル(の VBA)ベースの検索ツールは本研究の目的に合致する。

3.2.　検索ツール「E-KWIC」

前節の背景および方針に従って、エクセルの VBA を利用し、検索ツール「E-KWIC」を製作した。このツールは、エクセルのマクロを実行するだけで、図 2 の検索ウィンドウが表示される。このツールは、data フォルダ内のすべてのファイルに対して以下の検索を実行する。

1.　語彙の原型による一括検索
2.　語彙と品詞の組み合わせによる検索
3.　品詞のみの検索
4.　意味クラスのみの検索

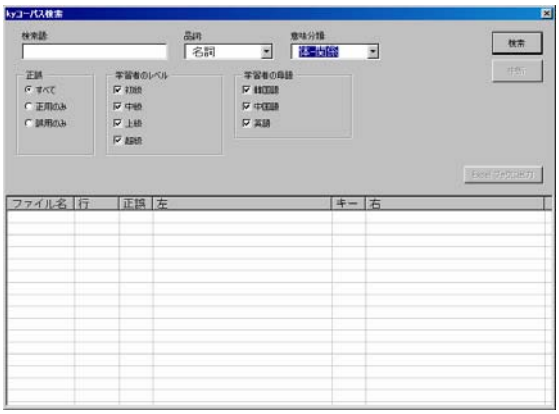

図.2　「E-KWIC」の画面

さらに、オプションを選択することで、検索対象を限定することができる。

5.　正誤の選択: すべて、正用のみ、誤用のみ
6.　学習者レベル: 初級、中級、上級、超級
7.　学習者の母語: 韓国語、中国語、英語

検索結果はエクセルファイルに出力することができる。

4.　最後に

本研究のデータとツールは、日本語教育の関係者に広く利用してもらうことを目的としている。そのため、ツールはフリー、データは KY コーパスを所有していることを前提にフリーで提供する予定である。同時にデータ加工に用いたスクリプトなども同様に公開することで他の分析者が自作データを加工する際にも再利用できる形にすることも検討している。なお、ツールやデータ配布後は、誤りを報告してもらい、著者らによって引き続き修正やメンテナンスを継続していくことを考えている。

*謝辞: 本研究は博報堂「ことばと文化・教育」研究助成(代表: 李在鎬, 06-B-0039)および科学研究費補助金(若手(B), 課題番号: 19720111)の援助を受け行った。感謝申し上げる。

〈参考文献〉

[1] 鎌田修(2006)「KYコーパスと日本語教育研究」『日本語教育』130 号 pp.42-51.

[2] 牧野成一(他)(2001)『ACTFL-OPI 入門―日本語学習者の「話す力」を客観的に測る』アルク

[3] 大曽美恵子(1999)「日本語学習者の作文コーパス:電子化による共有資源化」(平成 8 年度-平成 10 年度科学研究費補助金基盤研究(A)(1)研究成果報告書)

[4] 宇佐美洋(2005)「日本語学習者による日本語発話と母語発話との対照データベース」(平成 17 年度科学研究費補助金基盤研究(B)(2)研究成果報告書)