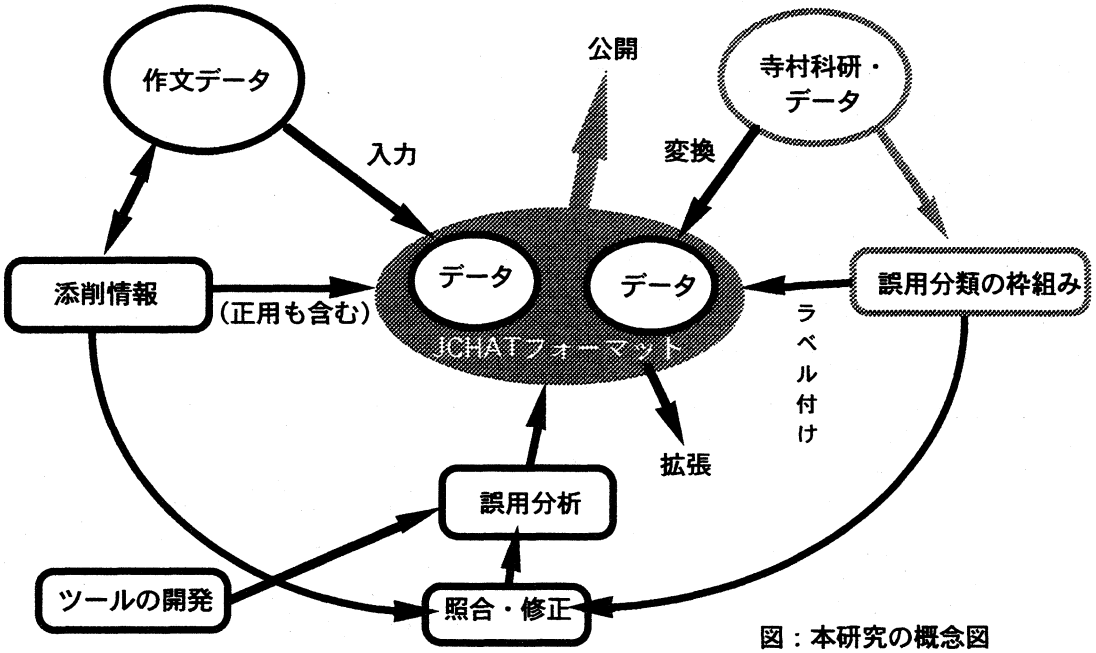
日本語学習者の作文コーパス：電子化による共有資源化

杉浦 正利，大曽 美恵子(名古屋大)，市川 保子(九州大)，奥村 学(北陸先端大)
小森 早江子(中部大)，白井 英俊(中京大)，滝沢 直宏，外池 俊幸(名古屋大)
goyoo@lang.nagoya-u.ac.jp



図：本研究の概念図

1．はじめに

本研究は、第2言語としての日本語の習得研究のため、日本語学習者による作文コーパスの構築と共有資源化およびそれを利用した誤用研究の推進を目指している。[1]

コーパス構築には、CHILDESのCHATフォーマットを採用した。CHILDESは、第1言語としての英語の話し言葉の研究のために考察されたものであるが、第2言語としての日本語の書き言葉の研究を行うため、フォーマットにいくつか独自の修正を加える必要がある。どのような修正が必要であるかということと、誤用の分類方法とそのタグ付けはどのように行うかについての議論を寺村科研と比較しながら行う。

2．日本語学習者の作文コーパス[2]

本研究では現在、独自に以下の4種類の作文データを電子化している。

1）関西外国語大学留学生別科
　特徴　：短期留学
　母語　：ほとんど英語母語話者
　レベル：初級から上級
　種類　：1授業中での課題作文
　　2学期末試験での小作文
　　3学期末試験での質問に対する応答
　量　　：298ファイル
2）名古屋大学留学生センター
　特徴　：集中日本語講座
　母語　：さまざま
　レベル：初歩・中級
　種類　：授業での課題作文
　量　　：441ファイル
3）電子パル
　特徴　：アメリカの学生の日本人宛メール
　母語　：英語
　レベル：中級
　種類　：電子メール
　量　　：15ファイル

4）名古屋大学言語文化部研究生
特徴　：日本語教授法受講生のレポート
母語　：中国語
レベル：上級
種類　：授業での課題レポート
量　　：2ファイル

3．JCHATによるデータ入力フォーマット

本研究ではデータ入力のフォーマットとして、言語習得研究用として実績のあるMacWhinny(1995)に基づくことにした。基本的な特徴としては以下の3点である：

1）1文1行
2）分析タグは各行の下の行で記述
3）データの属性をヘッダーで記述

誤用の箇所を示すために [*] という記号をデータ中に挿入しその誤用の範囲を< >というスコープ記号を使って示してある。また、それぞれの誤用に関する分析は、本文の下の行に「ディペンデントティア」と呼ばれる行をいれ、そこに誤用と正用および分類記号をかくことにした。

口頭発話を分析対象にした研究のために開発されたフォーマットであるため、表記上（文字）の間違いに関しては以下のような観点から新たに方針を立てる必要があった。

1）存在する文字かどうか
2）正用が推測できるか
3）漢字の分解と合併
4）漢字と仮名の混在

例）

```
@Begin
@Participant: GAK KOS03001
@Language of GAK:    English
@Coder:Kuno Shiho
@Date: 1990
@Location:    Kansai Gaidai, Hirakata, Osaka,
Japan
@Writing:     H, hand
@Type: exam
*GAK:  私の自分の意見は女より男の方がえらいという
       考え方 <じょ [*] > xxありませんがそう思っ
       ている人が多いと思います。
*GAK:  年とった人についてあまり<でまわ [*] ない>。
*GAK:  前は <そんなして [*] > <育だわれた [*] >
       からです。
%err:  そんなして = 風に ;
       育だわれた = 育てられた ;
%com:  添削済み
*GAK:  しかし同じ年の人の <ばい[*] > は<イライラを
       します [*] > 。
%com:  添削済み
%err:  ばい = ばあい ;
       イライラをします = イライラします ;
*GAK:  あと一週間しか <なかたら [*] > アメリカへ
       <帰りたいんです [*] > 。
%err:  なかたら = なかったら ;
       帰りたいんです = 帰りたいです ;
*GAK:  <そこで [*] > だれにも自分の
       <じょうけん [*] > を言わないで友達と家族と
       色々な事をしたいです。
%err:  じょうけん = じょうたい ;
*GAK:  <その上 [*] > <グスンドきャソヤソ [*] >
       へ行ったこと <を [*] > ないので行きxいです。
%err:  その上 = そして ;
       グスンドきャソヨソ = グランドキャニオン ;
       を= が ;
%com:  グランドキャニオンと助詞の「が」は添削済み
%com:  平仮名とカタカナが混在していた。
@End
```

4．JCHATフォーマットの現時点での問題点

JCHATフォーマットは、話し言葉のデータを想定して作られたので作文データにおける書き言葉としての性質による表記上の誤りに関してはフォーマットが定められていない。また、話し言葉における一発話を一つの単位として一行で書き分析するように作られているので、書き言葉においては一文一行ということになる。しかし、文と文との関係もしくは談話のレベルでの分析を行うためには、その分析の単位の定式化がなされていない。

また、JCHATでは誤用の分析に関して、一発話中に一カ所のエラーしか想定されていないため、複数のエラーが含まれる場合に、そのエラー間の関係およびその範囲、そして本文中のエラーと分析コード間の関係を明示する方法が定められていない。この問題に対しては、エラーコードに添え数字を付けることとスコープ記号を付けることで解決する方法を検討中である。[3]

5．誤用の分析と誤用分類案[4]

実際に誤用を分析していくと、誤用の範囲をどう判断するかということや、正用（表現意図）が不明だったばあいにどうするかとか、一つの誤用現象に複数の要因がかかわっている場合はどうするか、などの問題が起きてくる。本研究では、誤用の範囲の特定に関しては、誤用の原因が特定できるように広めに範囲を取ることとした。また、複数の要因がかかわっている場合には、それを列挙することとした。[5]

以下に、寺村科研を参考にして作成した誤用の分類案をあげる：

5.1　助詞（格助詞・取り立て詞・接続助詞・終助詞）
「私は買った本」助詞、格助詞、連体修飾

5.2 複合動詞（動詞連用形＋動詞）
「持て始めました」複V、接続
5.3 動詞（なる・する・活用形（形態的誤用）
lexical item）
「まていってください」V，活用、発音
「食べください」V、接続
「シャワーをとる」V、lexical item
5.4 形容詞
「おもしろいだ」A、ダ、品詞の混同
「おもしろかた」A、テンス・アスペクト、発音
「おもしろいなる」A、接続、なる
「おもしろいの本」A、接続、連体修飾
5.5 形容動詞
「いろいろ音楽」NA 接続、連体修飾
「しずかなる」NA、接続、なる
「しずかくなる」NA、接続、なる、品詞の混同
5.6 ダ
「もっとすぎ」NA、スタイル、発音
5.7 名詞
「病気な人」N、接続、連体修飾、品詞の混同
5.8 副詞
「あまり大きい」Ad、呼応
5.9 連用修飾
「歩き行く」V、接続、連用修飾
「大きくて開ける」A、接続、連用修飾
5.10 連体詞
「大き本」PN、A、接続、連体修飾、発音
5.11 句
「人気になる」（人気が出る）Ph
5.12 発音
「いてください」（いってください）
V、活用、発音

6．寺村科研の成果の利用

6.1 寺村科研
文部省科学研究費による特別推進研究『日本語の普遍性と個別性に関する理論的および実証的研究』（1985-1989年度代表者井上和子）の一環である「外国人学習者の日本語誤用例の収集・整理と分析」（分担者寺村秀夫）

6.2 寺村科研データ
4つの日本語教育機関より外国人学習者の作文、短文を収集・整理した。総数は4601文。データは、電子化されていると同時に、『外国人学習者の日本語誤用例集』として冊子にまとめられている。
現在利用可能なものは、元データはデータベース(CSV)形式で五つのファイルに分かれており、ほとんどが1986年度に作成された、約20カ国の国籍の、のべ339人の学生のデータである(約420KB)。台湾、中国、韓国、香港、インドなど、アジアの学生が多い。
八つの形式(自由作文、単文、穴埋め作文、聴解要約、文章要約、会話作文、パタン作文、絵からの作文)のいずれかで書かれており、一つが数行程度から二十行程度のものである。

6.2 寺村科研での誤用の分類・分析
寺村科研では、誤用と判断された部分に、どのような文法項目としてのラベルを付けるかという研究がなされた。しかし、ラベル付けの不統一さが見られる。また、誤用に対する訂正文がほとんどなく、ラベル付けだけにとどまっているため、どうしてその誤用と判断されたのか理解困難な場合もある。

6.3 電子化データの形式
各行の先頭8バイトに、データの種類(学生の番号、国籍、採取した時期と場所など)のID、その後に行番号、カンマの後、データとタグ情報がついている。
誤用の部分は |~ と ~| で括られているが、一行に複数ある場合には、番号が付けられている。
誤用は以下のように記されている。LはN、コソア、Vなどのラベルを表わす。ただし、*L2/-L1という形式で書かれている場合は、L1, L2 は具体的な助詞であることが多い。

L	Lについての使い方が誤っている
*L	Lを使うべきでないのに使っている
-L	Lを使うべきなのに使っていない
*L2/-L1	L1を使うべきなのにL2を使っている

6.4 寺村科研データのJCHAT形式への変換
●元データ
G6AJ0204001,日本に来て受けた最も強い印象.,なし
G6AJ0204002, 私は今年一月二十五日に引越ししました。.,なし
G6AJ0204003,家を出てバス停まで一分ほど|~かかります~|。.,V
G6AJ0204004,でも、そのバスは|1~限定時間~|な|2~人~|ですが、一歩|3~おくれて~|、
二十分以上も待たされたこともあります。.,1複N 2N 3*テ形／ータメニ
G6AJ0204005,|~その~|わけで自転車を買いました。.,コソア

●JCHAT形式化（Perl使用）
@Begin
@Participants: GAK G6AJ0204.cha
@Language: Japanese
@Nationality of GAK: China
@Coder: Teramura Kaken
@Date: 1986 Autumn
@Location: G
@Type: Free Format
@Warning: 白井のプログラムにより
寺村科研のデータを自動変換
*GAK: 日本に来て受けた最も強い印象
*GAK: 私は今年一月二十五日に引越ししました。
*GAK: 家を出てバス停まで一分ほど

〈かかります〉[*] 。
%com: V
*GAK: でも、そのバスは 〈限定時間〉[*1] な
〈人〉[*2] ですが、一歩〈おくれて〉[*3] 、
二十分以上も待たされたこともあります。
%com: 1複N 2N 3*テ形／ータメニ
*GAK: 〈その〉[*] わけで自転車を買いました。
%com: コソア
(中途略)
@End

6.5 JCHATフォーマット化の問題
寺村科研の元データに付けられていた誤用のタグは com (コメント)ティアとして記入した。寺村科研のデータには正用情報がないため、本研究での正用を含む分析タグ形式への自動変換は不可能であった。

7. 今後の課題

7.1 分類タグの整備
7.2 データの追加・整備・公開
電子メールのデータ化の検討も
7.3 分析ツールの開発
CHILDESの分析ツールCLANの日本語化
ローマ字変換ツールの開発
形態素解析（Juman3.1の応用）ツール
7.4 誤用分析
誤用の傾向
項目間の相関
学習歴と各項目との相関
連語に関しての否定的な情報の収集と分析
母語話者の言い間違いと学習者の誤用比較

注

1 本研究発表は、大曽美恵子を研究代表者とする科学研究費基盤研究(A)(1)（平成8-10年度、課題番号：08558020）の中間報告である。

2 データ量は1997年2月26日現在、1ファイル平均2KB。

3 しかし、スコープ間に階層関係がある場合、また、誤用以外の他の分析コードのスコープ記号との関係をどうするかが検討課題である。

4 具体的なタグ名は検討中。

5 複数の要因間の関係をどうするかは、まさにそれが誤用研究の対象となる問題である。

参考文献

市川保子(1997)『日本語誤用例文小辞典』凡人社
大嶋百合子, ブライアン・マックウィニー編 (1995)『日本語のためのCHILDESマニュアル』マッギル大学
杉浦正利, 中則夫, 宮田Susanne, 大嶋百合子(1997)「CHILDESの日本語化」『言語』第26巻第3号 大修館書店 80-87
益岡隆志, 田窪行則(1992)『基礎日本語文法』くろしお出版
松本裕治, 黒橋禎夫, 山地 治, 妙木 裕, 長尾 真 (1996)「日本語形態素解析システム (JUMAN version 3.1) 使用説明書, manuscript」 京都大学・奈良先端科学技術大学院大学
水谷信子(1985)『日英比較 話しことばの文法』くろしお出版
森田良行(1988)『誤用文の分析と研究 -日本語学への提言-』明治書院
寺尾康(1989)「発話資料のデータベース化」『言語』第18巻第6号 大修館書店 122-124
寺村秀夫(1982)『日本語のシンタクスと意味I』くろしお出版
寺村秀夫(1984)『日本語のシンタクスと意味II』くろしお出版

Cruse, D. A. (1986) *Lexical Semantics*. Cambridge University Press.
MacWhinney, B. (1995) *The CHILDES Project: Tools for Analyzing Talk, second edition*. Lawrence Erlbaum Associates.
Pustejovsky, J. & B. Boguraev eds. (1996) *Lexical Semantics*. Clarendon Press.
Saint-Dizier, P. & E. Viegas eds. (1995) *Computational Lexical Semantics*. Cambridge University Press.
Sokolov, J. & C. Snow eds. (1994) *Handbook for Research in Language Development Using CHILDES*. Lawrence Erlbaum Associates.