計量国語学第二十七巻第二号〔Mathematical Linguistics Vol.27 No.2〕2009年

タグ付き日本語学習者コーパスの開発

李　在鎬（(独) 国際交流基金　日本語試験センター）

計量国語学第二十七巻二号〔Mathematical Linguistics, Vol.27 No.2〕2009年

ノート

タグ付き日本語学習者コーパスの開発

李 在鎬（(独）国際交流基金 日本語試験センター）

キーワード：KYコーパス，学習者コーパス，形態素解析，KWICツール

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(error analysis）研究から中間言語(interlanguage)研究へシフトしたことで，誤用のみならず，正用をも含めた包括的観点から学習者言語の実態を捉えることの重要性が強調されるようになった(長友1993，長友 1999，迫田 2001，鎌田 2006)．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付けを導いた．これを受け，90年代後半，日本語学習者の書きことば・話しことばのデータベース化が進められ，積極的に利用されつつある(鎌田 1999，大曽 1999，宇佐美 2005)．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や手作業で事例が収集されていた点など，(文法研究への利用という面から見た場合)限定的なものであり，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．こうしたことから，学習者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習得研究において利用されてきた「KY コーパス(Ver. 1.1)」（鎌田・山内 1999）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 (Ver. 2.3.3)」や「分類語彙表 —増補改訂版— (Ver. 1.0)」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，様々な習得研究へ利用できるデータを作成した．さらに，加工データに対する専用の検索ツールも作成した．このツールは Excel の VBA(Visual Basic for Applications)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，これまでコーパスツールを使ったことが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ことで，日本語教育および言語習得の研究者，さらには計量国語学の研究者にも広く利用していただくことを目的とする．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，自作データに対しても計量的な調査分析が可能になることを指摘し，関連コミュニティにおけるコーパス分析および計量言語学的研究の活性化を図りたい．

LEE Jae-Ho (The Japan Foundation, Center for Japanese-Language Testing)—Development of Annotated Japanese-learner Corpus—

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，90 人分の OPI(oral proficiency interview)テープを文字化した言語資料である[1]．90 人の被験者を母語別に見ると，中国語，英語，韓国語がそれぞれ 30 人ずつであり，さらに，その 30 人の OPI の判定結果別の内訳は，それぞれ，初級 5 人，中級 10 人，上級 10 人，超級 5 人ずつとなっている．データのサンプルを示す．

(1) T：あーそうですか，いつ日本にはいらっしゃいましたか
S：ええと，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，あの［会社名２］の，なんていう，研究社員っていうか〈ええ〉として三ヶ月，〈はい〉

KY コーパスには，日本語学習者と OPI テスターの対話が(1)に示すフォーマットで格納されている．T で始まる文字列がテスターの発話，S で始まる文字列が学習者の発話となっている．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目的と問題点

高精度の情報検索を実現するため，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ある．

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．

1として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．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も大変な手間がかかる．こうした問題を解消するためには，単語の区切りを認定する作業が必要となる．次に2として，特定の文字列に還元できない文法項目を抽出するために，品詞などの言語情報などが必要である．こうした情報が付与されることで，様々な調査研究のニーズに柔軟に対応できる．

上述の目的を果たす上で，自然言語処理の形態素解析技術が有効である．しかし，

[1] 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（牧野(他)2001)．OPI の詳細は http://www.opi.jp/を参照してほしい．

KY コーパスの形態素解析においては，大きなものとして二つの技術的問題がある．

1. 会話データの解析精度の問題
2. 誤用例や言い直しの問題

まず，1 の問題として近年高精度の形態素解析器として多用されてきている「茶筌」や「MeCab」などが示す 100%に近い解析精度は多くの場合，新聞データのように規範的な表記と固定された文体で書かれたテキストデータを元に算出した値である．対話や談話データのような省略が多い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，

(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)

(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表現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．(5)b の場合，「てる」を一つの形態素として出力し，その原形を「てる」であると解析する．(6)b の場合，((5)b と違って「てく」を一語としているのではなく)「て」と「く」を分け，「く」の原形を「く」として解析している．(5)b や(6)b の出力の正誤性は別として使う側の立場から考えた場合，検索から漏れてしまう可能性がある．というのも，学習者の「て」形の実態を調べたいと考えている人が「てる」で検索をするとは考えにくいからである．同じ理由で，「ていく」形を調べたいと考えている人が「く」で検索することも考えにくい．次の問題として(4)のような終助詞表現に誤りが目立つ．(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[2]．

[2] 本稿の解析では，「茶筌(Ver. 2.3.3)」の辞書として「ipadic(Ver. 2.6.3)」を利用した．上記の事実は形態素解析技術そのものを否定するものではなく，話し言葉の解析においては，改良が必要であることを示唆するものである．この結果の背景として，(5)b「てる」の場合，動詞の活用をそのまま保っているので，この「て」を助詞-接続助詞として解析してしまうと，次の形態素との接続が処理できなくなってしまう．よって，茶筌としては「てる」を一語(動詞)と分析せざるを得ないということがあり，全体の処

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．(8)から考えてみたい．

(8) 早いのほうがいいじゃない（CA01）

(8)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，(9)のような解析結果を出力する．

(9) 早い(形容詞)／の(名詞)／ほう(名詞)／が(助詞)／いい(形容詞)／じゃ(助詞)／ない(助動詞)

(9)の結果に関しては，解析結果の誤りは入力データが誤っている故，やむを得ないことだとしても，学習者コーパスである以上は，誤用例であるという情報を何らかの形で付与する必要があると考えられる．誤用と同様の問題として，(10)のような言い直しなども頻繁に見られる．

(10) うれしいじゃなくて，たの，楽しい，（KIM01）

(10)における言い直しの元部分を解析した場合，「た(助動詞)／の(名詞-非自立)」となる．この種の問題は形態素解析器の問題ではなく，元データに内在する問題として位置づけるべきである．

以上に示した 2 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[3]．特に KY コーパスのような小規模のコーパスで，かつ言語習得のような正確性が要求される分野の基礎資料に関してはなおさら重要な問題である．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，その誤りを修正した．同時に，(8)や(10)に関しては独自のタグを導入し，データを加工した．

理としては筋が通っているものと考えられる．またデータ作成後に分かったことであるが，「UniDic(Ver. 1.3)」を使った場合，「てる」と「てく」は一語の助動詞として解析される．

[3] 参考情報として茶筌(Ver. 2.3.3)と ipadic(Ver. 2.6.3)を使った形態素解析における正答率の平均を示す．単語区切りでは，初級 81.2%，中級 82.5%，上級 90.1%，超級 94.6%となった．また，品詞の認定では初級 79.5%，中級 77.3%，上級 89.3%，超級 92.7%となった．正答率を下げた主な原因として，初級と中級では主に学習者誤用や言い直しなどが関係していた．それに対して，上級と超級においては会話体に見られる省略表現や終助詞のところで解析エラーが目立った．

2.3. データ加工の手順

以下の手順でKYコーパスに対する言語情報を付与した．

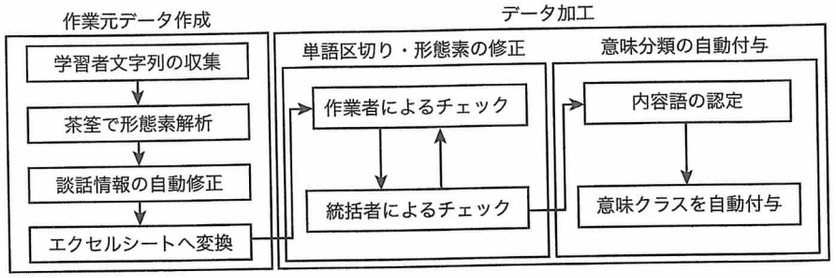


図1　データ加工の手順

加工の最初の段階としてKYコーパスのオリジナルデータから，Sで始まる文字列のみを抽出し，学習者発話データを生成した．次に，「茶筌(Ver. 2.3.3)」を用いて形態素情報を付与した．次に，談話情報の自動修正を行った．談話情報とは，テスターの相づちやポーズ，笑いなどの非言語情報であり，Perlのスクリプト処理で修正した．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Excel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．

単語区切り・形態素情報の修正は，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Excelの変更履歴として残すようにし，統括者が確認しながら，作業した．修正の際には，前節の(7)で示した「茶筌」による誤解析のみならず，(5)や(6)に示した省略形に関しても原形を手入力で修正した．さらに，(8)のような誤用例に対しては，以下の誤用タグを導入した．

(11) 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．

(12) 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（動詞-誤用）
　　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（助詞-誤用）
　　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（動詞-誤用）

(12)aのタイプには問題箇所に誤用タグをつけた．(12)bのタイプには不要な形態素のところに誤用タグをつけた．(12)cに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つけた[4]．

次に，(10)に対しては，「言い直し」というタグをつけた．また，英語母語話者において多く見られる現象であるが，母国語で話した場合，「原語発話」というタグを付与した．

作業者による第一チェックが済んだ段階で著者が，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，「内容語」に対して，「分類語彙表」（Ver.1.0）に基づいて意味情報を付与した．作業はPythonを用いて行った．現在，意味情報の付与には，一つの語が複数の意味分類に対応する場合に，次のような問題がある．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，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．その理由としては，「分類語彙表」の並びが「関係」から始まって「自然」で終わるため，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．

2.4. データ加工の結果

データ加工の結果，延べ232605語のデータベースが完成した．

表.1　レベル別集計

		全体の語数	意味分類付きの語数
初級 (15)	延べ語数	10350	2094
	異なり語数	1413	510
中級 (30)	延べ語数	67363	14282
	異なり語数	4126	1876
上級 (30)	延べ語数	99839	21725
	異なり語数	5090	2771
超級 (15)	延べ語数	55053	11878
	異なり語数	3665	2112

*( )は学習者の数

[4] 誤用の定義に関する問題として次のことを補足したい．本研究において誤用タグとは母語話者の直観に照らし合わせて捉えた場合「何らかの理由で規範的な表現から逸脱したもの」を（検索ソフトを介して）特定するための手掛かりである．ここでいう「何らかの理由」とは，文法的・形態的レベルのものだけでなく，発音や文体レベルの誤用も含まれる．さらには（確認はできないが）明らかに文字起こしの際の誤入力と考えられる例もある．しかし，これらの誤用をめぐる詳細な分類をしていない理由として次の3点がある．1)発音や文字起こしの間違いによるものは現時点では確認できない．2) 誤用の分類は寺村(1990)や市川(1997)の先行研究からも示唆されるように，分析者によって異なる可能性が高く，判断を一貫させるには膨大なコストがかかる．3) 本データベースは，日本語学習者への支援を目的とするものではなく，教師への支援ツールであるため，誤用例を発見するための検索環境を提供することが重要である．こうした理由から誤用と思われる事例に対して何らかのタグを付けておくこと，それを効率よく発見できる検索システムを作ることに重点を置いた．

表１で「意味分類付きの語数」とは，意味分類が付与された語のことである．初級であれば，全体の延べ語数は，10350 語であるが，そのうち 2094 語に対して意味分類が付与されている．一方，異なり語数の場合，初級では 1413 語が使われ，そのうち 510 語に対して意味分類が付与されている．なお，意味分類は，動詞，形容詞，名詞のみに付与されているため，実質的には（機能語に対する）内容語の数に相当する．

既述の通り，意味分類については方法論的・技術的問題から語が持つ多義性については考慮していない．このことによって生じる問題点を確認すべく，コーパス全体における多義語の割合を調査してみた．その結果，図２の分布が確認された．

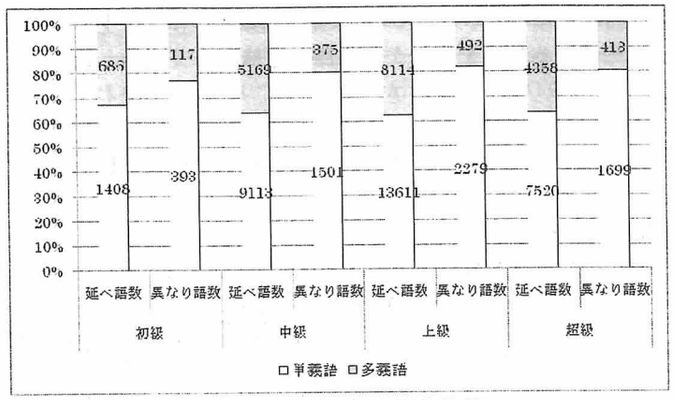


図２　意味分類の語義集計

図２において，単義語とは「分類語彙表」において一つの語義として登録されているものを指す，一方の多義語とは複数の語義として登録されているものを指す[5]．単義語の例として，イ形容詞は「安い，偉い，賢い，楽しい，嬉しい」などで，ナ形容詞は「有名だ，利口だ，大好きだ，正確だ，明らかだ」，名詞は「ラーメン，庭，楽器，食べ物」，動詞は「支払う，くれる，借りる，申す，訳す，話す，競う，行う」などである．一方の多義語の例として，イ形容詞は「優しい，遠い，おかしい，怖い，明るい」，ナ形容詞は「適当だ，嫌いだ，豊かだ，無理だ」，名詞は「平和，暴力，得意，今度，朝」，動詞は「払う，入る，受ける，持つ，聞こえる，通じる」などである．

さて，図２から読み取れる事実として，初級においては多少のずれはあるものの，全体としては同じ比率で分布している．延べ語数では，６対４で，異なり語数では８対２の比率で分布している．数値のみを見るなら，10 のうちに，４には情報の欠損があるということになるため，どこまで有効か疑問に感じるであろう．これに関しては，

5 ここで言う単義語と多義語はあくまで「分類語彙表」を使った処理上の単位であり，それ自体が日本語の中で本当に単義語か多義語かの問題には関わらない．正確に言うなら単義語は「単一の語義登録項目語」で，多義語は「多数の語義登録項目語」である．

分類語彙表の構造を見極めた上で，いくつかの活用方法が考えられる．まず，分類語彙表は「類／部門／中項目／分類項目」の４つの階層のもとですべての語彙が分類されている．例えば，「ラーメン」という単語は，「分類項目」としては「料理」に分類されるが，その上位ノードの「中項目」では「食料」になる．さらに，上位ノードの「部門」では「生産物」となり，そのさらに上位ノードの「類」では「体」に分類される．それでは，「類」や「部門」などには，どのようなものが入っているのであろうか．「類」は「体」「相」「用」「他」という非常に抽象化されたレベルで構成されている．部門は「関係」「活動」「主体」「自然」「生産物」で構成されており，日本語教育で重要視される抽象名詞や活動名詞，具体名詞といったカテゴリーを特定する上で重要なレベルである．部門に継ぐ「中項目」や「分類項目」には多種多様なものがあり，語の多義で問題になるのは主にこのレベルである．例えば，「格好」という名詞は，「中項目／分類項目」としては「様相／風・観・姿 」，「形／形・型・姿・構え」，「様相／良不良・適不適」と分類される．しかし，これらが帰属する「部門」のレベルは，そのいずれも「関係」であり，「格好」という単語が関係名詞であることは，どの語義を利用した場合も一貫している．ということは，ある程度抽象化されたレベルで，意味分類を利用した場合，多義性による実質的な影響は少なくなる[6]．よって，図２に示した問題点を認識しながら，「部門」のような抽象化されたレベルで意味分類を利用すれば，(語義と習得の関係性に対する)一定の研究成果が期待できる[7]．

3.　検索システムの構築

3.1.　開発背景

本研究では，２節で作成したデータを検索するための専用の検索ツールを作った．ツールを自作した一番の理由は，２節の手順で加工したデータを簡単に検索可能なソフトウェアが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，元データには「係り受け解析情報」も付与していた．しかし，最終的には３点の理由から「茶器」を断念し，自作ツールを作成することにした．

1. 人文系の研究者にとってはインストールが難しい．
2. 単語区切りやタグを修正した場合，動作が不安定になる．
3. コーパス格納後に，データ修正が容易ではない．

「茶器」は，奈良先端科学技術大学院大学で開発された高機能のコーパスツールで，データ抽出機能のみならず，様々な集計機能までも実装している優れたコーパスツールである．データ抽出の結果を Excel で出力する機能も備えているなど，人文系の研究者にも配慮した優れたコーパスツールである．しかし，本研究が「茶器」の利用を

6 本研究では意味分類を一つに絞るという人為的操作を行ったが，別の考え方として（漏れなく意味分類を拾うという意味では）すべての意味分類の項目を書き込んでおく方法も考えられる．

7 例えば，李・佐野・秋澤(2007)では KY コーパスにおける学習者の名詞の使い方を意味分類の観点で調査した．その結果によれば，レベルの上昇に従って関係名詞と活動名詞も増加傾向にあることが確認された．

断念した１つ目の理由として，「茶器」が利用する「MySQL」というデータベースソフトは多くの人文系の研究者にとってはインストールそのものが容易ではない．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を持た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である．２つ目として，「茶器」のデータは係り受け解析と形態素解析済みデータを利用して，検索する仕組みになっているが，２節で行ったような単語区切りやタグを変更した場合，係り受け解析の情報は事実上，機能しなくなり，全体の動作が不安定になる．さらに，タグの定義に関しても本研究のタグセットをすべて格納することはできない．３として，「茶器」の場合，一度格納したデータに関しては，事後修正が(不可能ではないが)難しく，最終的には，格納作業をやり直すことが一番早いという場合が少なくない．

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．以下の方針を立てた．

4. 可能な限り外部ソフト(データベースソフト)のインストールなしで使えるツール
5. 自作データにも簡単に応用できるツール

まず，４として「MySQL」のような外部ソフトのインストールなしで，使えるツールが必要と考えた．５として日本語の習得研究者の多くが持っているとされる，自作データにも簡単に利用可能なツールが必要と考えた．そこで着目したのが，Microsoft 社による Office の Excel である．Excel は高機能の表計算ソフトであるだけでなく，a) 多くの家庭用・業務用計算機において，インストールされている．b) 多くの人文系の研究者が日常的に利用しており，操作に慣れている．この a), b)から Excel は４，５を満足するものと考えられる．

3.2. 検索ソフト「E-KWIC」

前節の背景および方針に従って，Excel の VBA を利用し，検索ツール「E-KWIC」を製作した．このツールは，Excel のマクロを実行するだけで，図３の検索画面が表示され，data フォルダー内の Excel シートを縦方向で検索する．

図３は（形式名詞としての「の」を含まず）助詞「の」の用例を検索した場合の画面である．検索結果を KWIC (KeyWord In Context)フォーマットで表示する．また「ファイル名」には学習者 ID，「行」には Excel 上の行番号，「正誤」には正用(空白)か誤用(「誤」)かがそれぞれ表示され，学習者の OPI レベルや母語に関する情報，用法の正誤判断が行単位で把握できる．

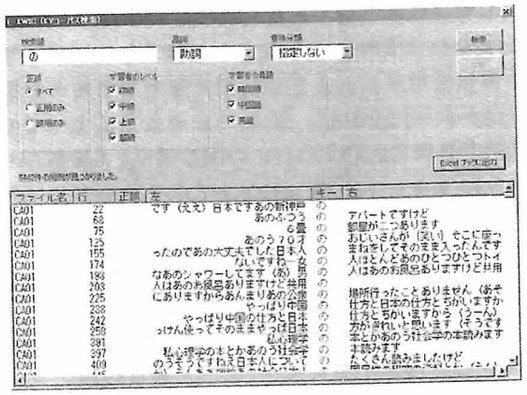
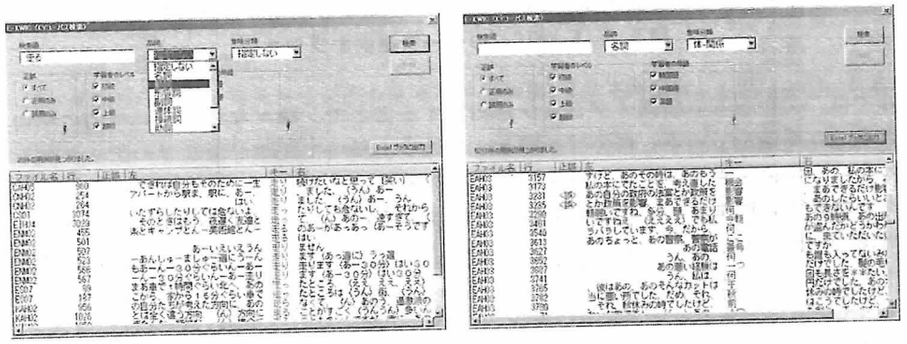


図３　「E-KWIC」の検索画面

E-KWIC は，その特徴として Excel シート内の文字列間のマッチングを取るだけの非常に原始的なツールであり，機能に関しても非常に限定的なものである．その分，計算機環境によるエラーが少なく，動作が安定する．また，VBA をベースにしているので，(自作データに合わせ)ツールの変更や加工も簡単に行える．例えば，品詞の部分を変更したり，絞り込みオプションを変更することも簡単にできる．

さて図３のツールで data フォルダ内の学習者データに対して以下の検索ができる．

① 特定の検索語からの検索
② 特定の検索語と品詞の組み合わせからの検索
③ 特定の検索語と品詞と意味分類の組み合わせからの検索
④ 不特定の検索語に対して品詞のみからの検索．
⑤ 不特定の検索語に対して意味分類のみからの検索．
⑥ 不特定の検索語に対して品詞と意味分類を組み合わせての検索．



a. 動詞「走る」の KWIC 画面　　b. 語を指定しない関係名詞の KWIC 画面

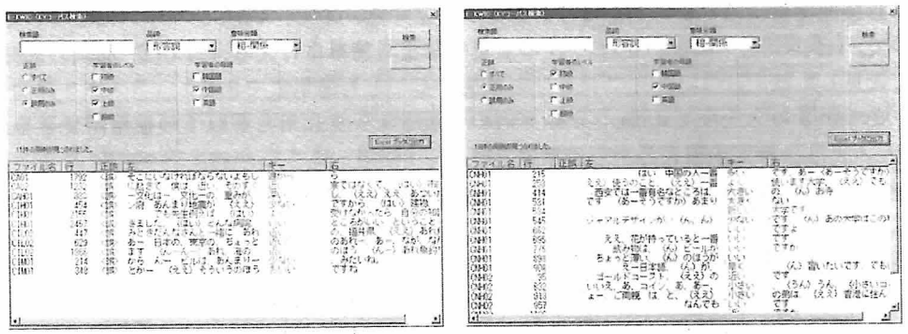
図４　「E-KWIC」の検索語オプション

検索語に関するオプションとして大きく分けた場合，検索語を指定して検索する方法

（①～③）と語を指定せず検索する方法（④～⑥）がある．語を指定して検索する方法では，①のように語のみで用例を取り出すこともできれば，②や③のように品詞や意味分類などの言語情報を指定し，よりピンポイントでデータを取り出すこともできる．仮に，その検索語が動詞であれば，図４の a のようにすべての活用形が一括で収集できる．なお，品詞の選択においては，a の図の通り，「茶筌」の品詞体系を知らなくても検索できるように工夫している．マウスでボタンをクリックすれば，コンボボックスで品詞リストが表示され，ユーザーはそれを選択するだけで，検索が実行できる．語を指定しない方法では，学習者データを網羅的・横断的にみる研究，例えば特定の品詞に関わる習得の実態を調べる研究や語義に基づく習得研究などで利用することができる．図４b では関係名詞のみを網羅的に抽出した結果である．

さて，以下のオプションを選択することで，データ選択に関する絞り込みができる．

⑦　正誤の選択: すべて，正用のみ，誤用のみから選択する
⑧　学習者レベル: 初級，中級，上級，超級から選択する
⑨　学習者の母語: 韓国語，中国語，英語から選択する



a. 中国語母語話者の
中上級学習者の関係形容詞の誤用例

b. 中国語母語話者の
初級の関係形容詞の正用例

図５　「E-KWIC」のデータ選択オプション

図５では，⑦から⑨の検索オプションを指定することで，様々な用例を効率的に検索できることを示している．例えば，図５a では中上級の中国語母語話者のデータのみを対象に，関係を表す形容詞のみを取り出した場合の画面である．b では初級の中国語母語話者の正用例としての関係形容詞を取り出した画面である．

最後に，検索の結果を Excel ファイルに出力することができるので，別の分析に再利用することもできる．

4.　終わりに

本稿では，日本語の第二言語習得研究において広く利用されてきた「KY コーパス」への言語情報付与の詳細について述べた．同時に，専用の検索ツールである「E-KWIC」の紹介を行った．最後に，本研究のデータとツールにおける問題点を述べる．４点が

考えられる．1) 現時点のデータでは「分類語彙表」を利用しているため，個々の文脈による多様な意味の問題までは記述しきれておらず，静的な意味分類のみの記述に留まっている．2) 誤用の認定が形態素単位に限定されているため，形態素を超える誤用表現に対しては充分な記述ができていない[8]．3) テスターの発話を削除しているせいで，発話の真意を把握するためには，元のテキストコーパスを参照せねばならず，会話分析的な研究ではやや使いづらい側面がある．4) 「茶筌」の形態素規定に従っている故に，日本語教育における実情を必ずしも充分に反映したものにはなっていない．例えば，「茶筌」では「ナ形容詞」は名詞の一種として扱われており，「名詞-形容詞語幹」になっている[9]．また，「指示詞」のような日本語教育においてニーズが高い品詞に関しても，「茶筌」のタグセットには含まれていないので，品詞による一括検索はできない．

現在，ウェブ経由で，利用申請をした人にデータとツールを配布している[10]．ツールはフリーライセンスで提供しており，プログラムの書き換えの制限なども設けていない．そのため，自作データに合わせてプログラムを書き換えることもできる．また，データに関しても KY コーパスの配布元と正式な利用誓約書を交わしていることを条件に無料で配布している．そして，データ配布後は，メーリングリストを経由し，サポートを行うと同時に，データの誤りを報告してもらっている．また，本データベースに基づいて作成した語彙表や N-gram 検索ツールも同じサイト内で，同様に公開している．

〈参考文献〉

[1] 市川保子(1997)『日本語誤用例文小辞典』，凡人社.
[2] 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研究成果報告書）.
[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研究成果報告書）.
[4] 鎌田修(1999)「KY コーパスと第二言語としての日本語の習得研究」『第二言語としての日本語の習得に関する総合研究』(平成８年度～平成 10 年度科学研究費補助金研究成果報告書), 227-237.
[5] 鎌田修(2006)「ＫＹコーパスと日本語教育研究」『日本語教育』130, 42-51, 日本語教育学会.
[6] 寺村秀夫(1990)「外国人学習者の日本語誤用例集」(平成２年度科学研究費特別推進研

[8] 構文レベルの間違いでも，特定の語，例えば助詞類に関わるものは取り出すことができる．例えば，「最近は，あの流行っているの小説」のような「の」の過剰使用による連体修飾構文の間違いや「全然ないことがない」のような「は」と「が」の錯誤による主題構文の間違い，さらには「人間を死んだり」のような自動詞・他動詞の構文としての誤用例は検索ツールを介して簡単に取り出すことができる．

[9] 完全ではないが，品詞と意味分類をともに指定し，検索することで，ナ形容詞のみを取り出すこともできる．たとえば，品詞を「名詞」に指定し，意味分類を「相」に指定することで，ナ形容詞を取り出せる．

[10] 配布や関連ツール類に関する詳細は，http://www30.atwiki.jp/corpus-ling/を参照してほしい．

究資料).
[7] 長友和彦(1993)「日本語の中間言語研究」『日本語教育』81, 1-18, 日本語教育学会.
[8] 長友和彦(1999)「第二言語としての日本語の習得研究—概観, 展望, 本研究の位置づけ」『第二言語としての日本語の習得に関する総合研究』(平成８年度～平成 10 年度科学研究費補助金研究成果報告書).
[9] 迫田久美子(2001)『日本語教育に生かす第二言語習得研究』, アルク.
[10] 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.
[11] 李在鎬・佐野香織・秋澤委太郎(2007)「初級学習者の名詞使用と 文生成の問題 」, OPI 国際シンポジウム（京都）(『予稿集』, pp.91-97)

〈言語資源とコーパスツール〉

[1] 奈良先端科学技術大学院大学「茶筌」(Ver. 2.1) : http://chasen.naist.jp/hiki/ChaSen/
[2] 京都大学・NTT「Mecab」: http://mecab.sourceforge.net/
[3] 奈良先端大学院大学「茶器」(Ver. 1.0) : http://chasen.naist.jp/hiki/ChaKi/
[4] 国立国語研究所「分類語彙表」(Ver.1.0) : http://www.kokken.go.jp/kanko/goihyo/syokai/
[5] 鎌田修・山内博之「KY コーパス」(Ver. 1.1) : http://opi.jp/shiryo/ky_corp.html
[6] 伝康晴・山田篤・小椋秀樹・小磯花絵・小木曽智信「UniDic」(Ver. 1.3) : http://www.tokuteicorpus.jp/dist/index.php

（2009年５月30日受付）