大学共同利用機関法人 人間文化研究機構

国立国語研究所

基幹型研究プロジェクト「多文化共生社会における日本語教育研究」
学習者の言語環境と日本語の習得過程に関する研究

C-JAS(Corpus of Japanese as a second language)構築

に関する報告書

2014 年 3 月

研究代表者　迫田 久美子
（国立国語研究所 日本語教育研究・情報センター　センター長）

目次

はじめに

学習者の誤用は、誤用ではない。学習者の誤用は、学習者が自らの言語習得の仮説を検証しようとして試行錯誤している証である。間違っていると思って使っているのではなく、『これ、使えるかな』と新しい語彙や言語環境に適用し、うまく適用できなかった結果として誤用になる。

長年、日本語教育に携わってきて、日本語学習者の誤用が面白いと感じていた。「おもしろいだった（→おもしろかった)」「会館で（→に）住んでいます」「先生は結婚しましたか（→結婚していますか)」など、母語の異なる学習者から、同種の誤用が産出される。しかし、多くの研究では、学習者の誤用が母語の影響が原因だと結論づけている。果たして、そうなのであろうか。この疑問が学習者の言語研究をスタートさせるきっかけとなった。1980 年代後半である。大学の外国人留学生の日記や作文データに基づく 3 年間の誤用分析の基礎研究を行い、話し言葉に対象を移して 1991 年、日本語の指示詞の習得研究を開始した。指示詞を取り挙げたのは、同期の大学院生に台湾からの留学生が 2 名在籍し、日本語が非常に流暢にも関わらず、初対面の自己紹介で 2 名とも指示詞で同種の誤用（ソを使うべき場面でアを使った）を犯した（例「留学した優秀な先輩がいて、私もあの人（→その）のようになりたいと思っています」）ことに因る。

この C-JAS (Corpus of Japanese As a Second language)は、その研究のために 1991 年から 1993 年まで、3 年間、実施した縦断調査のデータをまとめたものである。最初の 1 年間は同じ民間の日本語学校の教室指導を受けた学習者が、その後、国立大学、私立大学、別の民間日本語学校などの異なった進路に進んでも定期的に収集した対話データをコーパス化したものである。

本データは、韓国語母語話者 3 名、中国語母語話者 3 名の 6 名の 48.5 時間、約 87 万語のコーパスである。一般公開は、データ収集から 20 年を経て、2013 年 1 月となったが、その背景には、当初は「サコダコーパス」として、一部の研究者の間のみで使用されていたが、3 年を経た縦断調査のデータの希少性を考え、文字化作業および形検索システムを付与して公開することにした。

この 6 名の学習者は、研究の面白さと深さを気付かせてくれ、筆者の研究の基盤を支えてくれた学習者たちである。そして、この学習者コーパスの生みの親である。彼らの協力がなかったら、世に出ることはできなかった。また、筑波大学の李在鎬先生には、コーパス構築に関して、多くのご支援を賜った。この報告書を完成させてくれた佐々木(木下)藍子氏、小西円氏は、育ての親である。二人がいなかったら、一般公開にも報告書にも成長していなかった。ここに記して、深く感謝したい。

2014 年 3 月 4 日　　迫田久美子

1.　プロジェクトの概要

1－1.　研究目的

本研究「学習者の言語環境と日本語の習得過程に関する研究」は、国立国語研究所の日本語教育研究・情報センターの基幹型共同研究プロジェクト「多文化共生社会における日本語教育研究」のサブプロジェクトの一環としてスタートした。

本研究は、第二言語習得研究の枠組みを基盤としつつ、言語心理学、対照言語学等の関連諸領域との協働により、日本語学習者の言語環境と日本語の習得過程との関係を実証的に解明しようとするものである。具体的には、(1)「母語環境と第二言語環境」「教室指導環境と自然習得環境」などの学習者外部の言語環境の違いが日本語習得に及ぼす影響に関する研究、(2)学習者内部の言語環境である学習者の母語が日本語習得に及ぼす影響（言語転移）に関する研究、そして、(3)そのための基礎資料として有用な日本語学習者の発話や作文のコーパスの内容と構造に関する研究を行う。これらの研究は、学習者のソトとウチの両面から第二言語習得を総合的に分析する研究の開拓、ならびに第二言語習得研究のための基礎データの整備につながる。

1－2.　研究組織

【統括リーダー】　迫田　久美子

【基幹型共同研究プロジェクト名称】

多文化共生社会における日本語教育研究

－学習者の言語環境と日本語の習得過程に関する研究－

本研究においては、「研究目的」に記した３つの研究について、それぞれ研究班を設けた。以下は、各研究班の関係と各班の主要メンバーを記載したものである。

（１）「言語環境と日本語習得」班

既存のあるいは新規に収集した日本語学習者の発話や作文のデータを資料として、外部の言語環境の異なる日本語学習者の習得過程の比較を行い、その類似点と相違点を明らかにする。

共同研究者：白井恭弘，岩立志津夫，渋谷勝己，南雅彦，小柳かおる　他

（２）「言語転移と日本語習得」班

既存のあるいは新規に収集した日本語学習者の発話や作文のデータを資料として、母語の異なる日本語学習者の日本語習得過程の比較を行い、その類似点と相違点を明らかにする。

共同研究者：奥野由紀子，田中真理，タサニー・メーターピスィット　他

（3）「学習者コーパス研究」班

上記２班の研究方法を参考にしながら、日本語学習者の発話や作文のコーパスの内容と構造に関する研究を行い、既存の日本語学習者の発話や作文のデータの活用について検討する。

共同研究者：山内博之，野山広，金田智子　他

【共同研究者】（50 音順，敬称略）（平成 26 年 3 月 3 日現在）

井上　優　(麗澤大学)
岩立　志津夫　(日本女子大学)
大関　浩美　(麗澤大学)
奥野　由紀子　(首都大学東京)
金田　智子　(学習院大学)
家村　伸子　(広島修道大学)
川崎　千枝見　(広島国際学院大学)
小柳　かおる　(上智大学)
渋谷　勝己　(大阪大学大学院)
白井　恭弘　(ピッツバーグ大学)
砂川　有里子　(筑波大学)
タサニー･メーターピスィット　(タマサート大学)
田中　真理　(名古屋外国語大学)
中石　ゆうこ　(広島大学大学院)
仁科　喜久子　(東京工業大学名誉教授)
野山　広　(国立国語研究所)
橋本　ゆかり (横浜国立大学)
福永　由佳　(国立国語研究所)
南　雅彦　(サンフランシスコ州立大学)
峯　布由紀　(東洋学園大学)
山内　博之　(実践女子大学)
横山　詔一　(国立国語研究所)

【C-JAS 担当 プロジェクト非常勤研究員】

佐々木（木下）藍子　(国立国語研究所)
小西　円 (国立国語研究所)

（迫田久美子）

2. C-JAS の概要（データ）

C-JAS とは、Corpus of Japanese As a Second language の略で、日本で日本語を第二言語として学んでいる学習者の約 3 年間の縦断的発話コーパスである。このコーパスは、外国人の日本語習得に興味を持ち、研究する方々や日本語教師の方々に利用して頂きたいと考え、作成した。

本コーパスには、以下の４つの特徴がある。

（1）中国語、韓国語を母語とする特定の学習者を約３年間調査して収集した発話データである

（2）文法習得の研究を目的として収集された自然な会話データである

（3）コーパス付属の検索システムを備え、オンラインで使用できる

（4）統語的、文法的、発音などの誤用タグが付与されている

第二言語習得研究は、母語とは別に学ぶ外国語・第二言語の学習・習得にかかわるさまざまな現象を研究する分野であり、データが不可欠である。本コーパスがその分野の研究や日本語指導の資料として少しでも貢献できれば、本コーパスのデータ収集に協力してくださった学習者や作成者たちの喜びであると考える。

2−1. 学習者の概要

学習者の性別、母語、調査期間の年齢、学習者の環境を表１にまとめた。詳細は以下の通りである。下記６名の学習者は全員、日本における教室環境学習者であり、最初の 1 年間は同じ日本語学校で同時期に初級から日本語を学んだ。その際使用していた教科書は『日本語初歩』[1]である。

表１．学習者の概要

	性別	母語	調査期間の年齢	学習者の環境
C1	女	中国語	25 歳～28 歳	1 期：日本語学校 3～4 期：大学 1 年生（看護系） 5～8 期：大学 2 年生
C2	女	中国語	20 歳～23 歳	1 期：日本語学校 2～5 期：短大 1 年生（国文系） 6～8 期：短大 2 年生
C3	女	中国語	22 歳～25 歳	1～2 期：日本語学校 3～5 期：大学研究生（商学系） 6～8 期：大学 1 年生（他大学商学系）

[1] 国際交流基金日本語国際センター（1985）『日本語初歩』 凡人社

K1	男	韓国語	21 歳～24 歳	1～2 期：日本語学校 3～4 期：別の日本語学校 5～8 期：専門学校 1 年生
K2	男	韓国語	18 歳～21 歳	1～2 期：日本語学校 3～4 期：大学 1 年生（工学系） 5～8 期：大学 2 年生
K3	女	韓国語	21 歳～24 歳	1～3 期：日本語学校(3 期後やめる) 4～5 期：主婦兼アルバイト 6～8 期：大学 1 年生（商学系）

2－2. データの収集時期とデータ数の内訳

データの収集時期は 1991 年 7 月～1994 年 3 月である。学習者 1 人につき 8 回の調査が行われた。一回の調査は、約 60 分の対話形式である。データの名称として、1 回目から 8 回目までの調査時期ごとに 1 期から 8 期と呼ぶこととする。C1 のみ 2 期目（*****1）のデータが欠けているため、データの総数は 47 本である。また、K1 の 2 期目（*****2）のデータは 30 分である。データそれぞれの内訳と調査日は以下の表２の通りである。

表２．データの内訳と調査日

中国語母語話者			韓国語母語話者		
C1	C2	C3	K1	K2	K3
C1 - 1 期 ('91/7/24)	C2 - 1 期 ('91/6/27)	C3 - 1 期 ('91/8/22)	K1 - 1 期 ('91/9/9)	K2 - 1 期 ('91/7/10)	K3 - 1 期 ('91/9/12)
*****1	C2 - 2 期 ('92/5/1)	C3 - 2 期 ('92/3/15)	*****2 K1 - 2 期 ('92/2/24)	K2 - 2 期 ('91/12/4)	K3 - 2 期 ('92/3/13)
C1 - 3 期 ('92/8/5)	C2 - 3 期 ('92/7/19)	C3 - 3 期 ('92/7/16)	K1 - 3 期 ('92/7/22)	K2 - 3 期 ('92/7/17)	K3 - 3 期 ('92/7/5)
C1 - 4 期 ('92/12/20)	C2 - 4 期 ('92/11/30)	C3 - 4 期 ('92/11/23)	K1 - 4 期 ('92/12/21)	K2 - 4 期 ('92/12/5)	K3 - 4 期 ('92/11/29)
C1 - 5 期 ('93/4/26)	C2 - 5 期 ('93/3/2)	C3 - 5 期 ('93/3/21)	K1 - 5 期 ('93/4/20)	K2 - 5 期 ('93/4/2)	K3 - 5 期 ('93/3/18)
C1 - 6 期 ('93/7/27)	C2 - 6 期 ('93/7/16)	C3 - 6 期 ('93/8/2)	K1 - 6 期 ('93/7/27)	K2 - 6 期 ('93/8/31)	K3 - 6 期 ('93/8/22)
C1 - 7 期 ('93/12/12)	C2 - 7 期 ('93/12/16)	C3 - 7 期 ('93/12/29)	K1 - 7 期 ('93/11/27)	K2 - 7 期 ('93/12/27)	K3 - 7 期 ('93/11/11)
C1 - 8 期 ('94/3/9)	C2 - 8 期 ('94/3/8)	C3 - 8 期 ('94/3/8)	K1 - 8 期 ('94/3/10)	K2 - 8 期 ('94/3/4)	K3 - 8 期 ('94/3/12)

2−3. インタビューのテーマ

8 回の調査はそれぞれ共通の話題が設定されており、それを含めた母語話者との自由会話となっている。8 回の共通の話題は以下の通りである。

1 期：小・中学校の先生の思い出
2 期：留学 1 年を振り返って
3 期：私の日本人の友達
4 期：私の学校生活
5 期：日本人について
6 期：休日の過ごし方
7 期：日本の衣食住について
8 期：日本での 3 年間を振り返って

（佐々木（木下）藍子）

3.　コーパス作成について

3−1.　コーパス作成作業の経緯

コーパス作成の大まかな手順は以下図１の通りである。今回の作業では、「文字化」の途中部分より作業を行った。本コーパスは、検索システムを備えるため、文字化データを形態素解析する必要があった。そのため、まず元データである文字化データを形態素解析に適した形となるよう修正を行い、形態素解析を行うという流れで作業を行った。

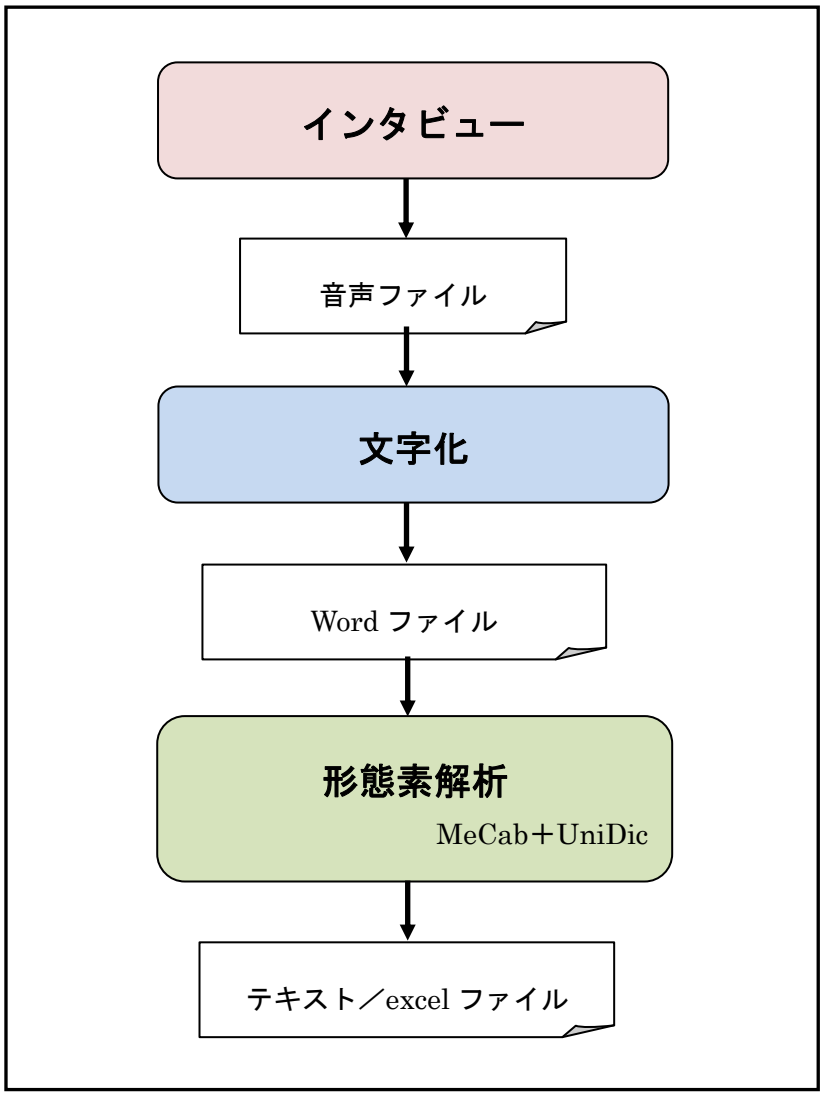

図１．コーパス構築の工程

3−2.　文字化作業について

以下、文字化の方針について詳細を述べる。本プロジェクトでは、すでに文字化されている元データの記号や表記の統一を図るため、再度音声データを確認し、本コーパスの文字化の方針に従って修正するところから始まった。

3-2-1. 文字化の方針

＜文字化における基本方針＞

（1）発話者の記号

発話の行頭に発話者を示す以下の記号をつける。いずれも半角大文字で示す。

調査者（日本語母語話者）→NS

学習者（日本語学習者）→C1、C2、C3、K1、K2、K3

（2）文の単位・改行

本データでは、文の単位は考慮しないため、文字化資料には句点「。」は使用しない。改行は発話の主導権が交替したと思われる際に入れるが、厳密には定めない。

（3）発話番号

各発話には行頭に 4 ケタの番号をつける。

例）0001NS ： 日本語はどのぐらい勉強しましたか

（4）あいづち

一般的にあいづちとみなされる発話は〈 〉で相手の発話の中のおおよその位置に挿入する。また、相手の発話と完全に重なるあいづちは、その発話の区切りにまとめて示すか、別の発話として立てる。

（5）発話の重なり

発話が重なっている場合は表記が困難なため、別の発話として扱うか、もしくはあいづち同様〈 〉を使用して相手の発話中に挿入する。基本的に短いものであれば挿入し、長いものは次の発話として扱う。

（6）固有名詞

音声データに表れる固有名詞のうち、以下に相当するものは【 】にその分類名とアルファベット 1 文字を入れ、言いかけている固有名詞も全て置き換える。固有名詞が多く出現するデータの場合はアルファベッドが 2 文字にわたる場合もある。 1 データ内で同じ固有名詞が使用された場合は、同じ分類名およびアルファベットを使用する。使用される分類名およびアルファベットは、 1 学習者の 1 データごとの通しで付与され、異なるデータで同様の固有名詞が出現した場合でも、関連しないこととする。

例）0098NS：【人名 C】先生はどうして,【人名 C】先生を知ってたの?

置き換える固有名詞は、以下の通りである。

・個人名

・個人が所属している学校名、会社名、店名（アルバイト先等）

・個人の出身地（大都市の場合は除く場合もある）、個人に関係のある駅名、個人が特定される可能性の高い地名、あるいは個人に深く関係のある者の出身地等で、当該データのみでは個人は特定できないが、他のデータとの関係で特定される可能性が

高い場合

・実在する人物の個人名、会社名、大学(学校)名、店名、施設名等

・学習者の母国と日本以外の第 3 国

・宗教名

・上記以外のもので個人の情報を特定する可能性がある場合

以上を原則とするが、状況により置き換えが必要な場合は、適当な分類名を使用し、置き換える。確実に架空のものと考えられる場合は置き換えしていない場合もある。また、人名で特に姓と名を区別する必要がある場合は、【姓 A】【名 B】とし、固有名詞が略称で用いられた時も、正式名と同様の置き換えで表記する。

(7) 第 3 者の発話

第 3 者（調査者・学習者以外の人物）の発話も文字化する。発話者の記号は非母語話者の場合、「NNS1」、日本語母語話者の場合「NS2」とし、複数以上出てくる場合は NNS、NS の後につける番号を適宜増やし表記する。

＜表記の方針＞

(1) 文字の表記方法

表記は、一般的な漢字仮名交じり文を用いる。表記することが困難な音についても、同一データ内ではできる限り統一する。

促音、長音、拗音（「しゅばらしい（すばらしい)」などの発音、およびポーズなどの表記をコーパス全体で厳密に統一することは困難であるが、同一データ内ではできる限り統一する。

(2) 長音

前の音節が長く伸ばされていることを表す。長さに関わらず「ー」1 つで示す。ただし、ひらがなで表記されることが一般的な長音はこの限りではない。

例）0001C2：ちょーとねー

(3) 間・ポーズ

発話が途切れることを表し、長短に関わらず「,」 1 つで示す。

(4) 上昇イントネーション

発話末のイントネーションが上昇調である場合、疑問符「?」を付与する。

(5) 非言語行動等

笑い声や発話に関係のある非言語行動は{　}で示す。また、聞き取りが困難な場合もこの記号を使用し、状況を説明する。

例）{笑}　{音声不良のため、聞き取り不可}　{一時停止}

(6) カタカナ表記

①外国語

外国語は、意味を考慮しつつ聞こえたようにカタカナで表記することを原則とし、英語で発音したと判断された場合は英文表記も可とする。外国語かどうか不明な場合はひらがなで表す。

②一般的な外来語・外国地名・外国人名等

基本的にはカタカナで表記する。(ブルガリア、ニューヨークなど) ただし、北京、釜山等、漢字表記が普通のものは例外とする。

③動植物名

一般的に常用漢字で書くことができる犬、猫等は漢字で表記し、シマウマ、バラ、ユリ、等のような動植物はカタカナで表記する。

(7) 数字・アルファベット

すべての数字はアラビア半角、またすべてのアルファベットも半角とする。

(8) 補足情報記号[ ]について

通常と異なる発音や縮約形、ポーズ等が挿入され、漢字で表せない場合はひらがなで表記し、文脈から意図される語を[ ]で補足する。

例) おと，さん[お父さん]

じぇーったい[絶対]

クラピック[グラフィック]

かいて[買って]

また、漢字に対し 2 つ以上の読みが考えられる場合もひらがなで表記し[ ]で漢字表記を補足する。

例) はいれる[入れる]

(9) 引用発話

直接引用の発話は「 」で示す。

(10) 書籍名タイトル

書籍名のタイトルは『 』で示す。

(11) 誤用マーク Φ について

学習者の発話には、誤用と思われる箇所のうち、本来あるべきなんらかの語が脱落していると判断した箇所に、文字数に関わらず「Φ」を付与している。

3-2-2. 個人情報保護について

本データは調査期ごとに話題が設定されているが、基本的には自由会話であり、また継続的に収集したデータであるため、固有名詞等を伏せても個人を特定するような内容、あるいは個人のプライバシーに関わる内容が読み取れる可能性がある。本データでは、そのような箇所の談話を削除し、非言語情報を表わす{　}に、削除した発話数の説明を表記した。

例）{続きは個人情報保護のため 4 発話分削除}

また、学習者の誕生日、来日した日付、個人の電話番号、個人の住所等は＊で消している。

例）0020NS：はいはいすいません，えーと，生年月日教えてくれますか?

0021K1：あー，《ろくじゅきゅねん》[69 年]?〈うん〉＊月＊日

3−3. 形態素解析作業ついて

3-3-1. 形態素解析

本コーパスは検索の利便性を上げるため、一般的な文字列検索だけでなく、形態素情報を用いた検索が行えるよう、検索システムを備えることとした。そこで、本コーパスの文字化データに対し、形態素解析を行った。

形態素解析とは、コンピュータを用いて文を形態素単位に分割し、それぞれの品詞を同定する作業のことである。形態素解析には、文を形態素に区切る形態素解析エンジンと、それに品詞を振る辞書が必要であるが、本コーパスでは前者に MeCab、後者に UniDic を使用している。

UniDic では、表記が異なっても、同じ語であれば一つの見出しにまとめるという方針をとり、語を階層化した形で辞書登録している。この階層の最上位を語彙素と呼び、その下に、語形、書字形、発音形という階層がある。語彙素とは、国語辞典の見出し語に相当するレベルで、元来同一と見なしうる語をまとめ上げたものである。語形は、異語形を区別するレベル、書字形は異表記を区別するレベルである。発音形は発音などの情報が記載される（小木曽・中村 2009:7）。

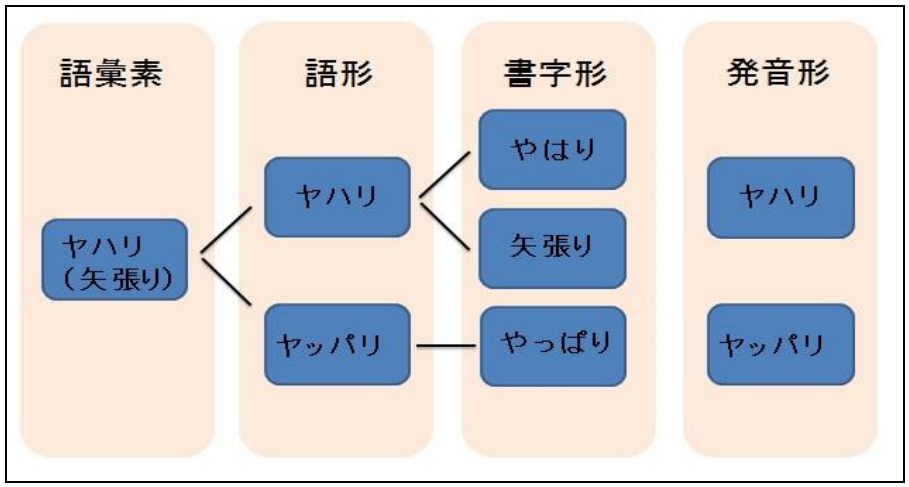

図２．UniDic の見出し構造の例

本コーパスの形態素解析結果も、UniDic の階層に準じている。例えば、本コーパスに収録されている「入る時はね、私すごく楽しみにしてた、けどね」という文は、以下のように形態素解析される。

表１．形態素解析の例

書字形	発音形	語彙素読み	語彙素	品詞	活用型	活用形	語形	語種
入る	はいる	はいる	入る	動詞-一般	五段-ラ行-一般	連体形-一般	ハイル	和
時	とき	とき	時	名詞-普通名詞-副詞可能			トキ	和
は	わ	は	は	助詞-係助詞			ハ	和
ね	ね	ね	ね	助詞-終助詞			ネ	和
,			,	補助記号-読点				記号
私	わたし	わたし	私	代名詞			ワタクシ	和
すごく	すごく	すごい	凄い	形容詞-一般	形容詞	連用形-一般	スゴイ	和
楽しみ	たのしみ	たのしみ	楽しみ	名詞-普通名詞-一般			タノシミ	和
に	に	に	に	助詞-格助詞			ニ	和
し	し	する	為る	動詞-非自立可能	サ行変格	連用形-一般	スル	和
て	て	てる	てる	助動詞	下一段-タ行	連用形-一般	テル	和
た	た	た	た	助動詞	助動詞-タ	終止形-一般	タ	和
,			,	補助記号-読点				記号
けど	けど	けれど	けれど	助詞-接続助詞			ケド	和
ね	ね	ね	ね	助詞-終助詞			ネ	和

ここで示されている書字形は、発話を文字化したテキストデータの文字列そのものを指す。発音形は書字形の発音を示す。各書字形は、辞書的見出し語である語彙素を付与され、語彙素読みでその読みが示されている。活用のある品詞の場合は、活用に関する情報も付与される。検索に使用できるのは、「書字形」「発音形」「語彙素」「品詞」であり、活用に関する情報は利用できない。

形態素解析ツールの解析精度は、いわゆる書き言葉であれば 98%である[2]が、ブログのような話し言葉の特徴を含んだ文字言語や、音声言語を文字化したものの場合、やや精度が落ちる。形態素解析されたデータは、検索の利便性が向上するが、一方で誤解析を完全には排除できないという問題点もある。誤解析とは、発話の意図するものとは異なって解析されてしまったもので、検索システムでの検索精度を向上させるためには、このような誤解析をできるだけ排除したデータを作成することが好ましい。

---

2 以下を参照されたい。
http://www.ninjal.ac.jp/corpus_center/cmj/doc/05ogiso.pdf#search='%E5%BD%A2%E6%85%8B%E7%B4%A0%E8%A7%A3%E6%9E%90%E5%99%A8+%E8%A7%A3%E6%9E%90%E7%B2%BE%E5%BA%A6'

3-3-2. 形態素解析に備えた前処理と注意点について

本コーパスの文字化作業では、可能な限り発話に忠実に文字化することを心掛けた。しかし、形態素解析するにあたっては、上記でも述べた通り、文字化された発話データをそのまま形態素解析すると、誤解析が多く出現するのではないかと考えられた。そこで、誤解析ができるだけ起こらないよう処理を施すこととした。誤解析となる可能性の高い箇所として考えられたのは、学習者の発話内に出現する誤用部分であった。そのため、形態素解析を行う前に、学習者の誤用部分は、自動の形態素解析から排除されるよう、タグ《　》(二重山かっこ) を付与することとした。この作業は文字化データの表記修正の段階で、同時に行った。

そして、学習者の誤用と考えられる《　》を付与した箇所は、形態素解析を行った後、人手で形態素解析を行った。この人手での作業の詳細は、次の節で詳しく述べる。

学習者の誤用部分に注意を払い形態素解析を行ったが、本コーパスは発話データであったため、学習者の誤用部分以外にもいくつかの誤解析が出現してしまった。以下のような箇所である。

（1）1形態素中に長音が含まれる場合

本コーパスでは発話をできる限り忠実に文字化するという方針を取ったが、1 形態素中に長音を含むものも見られた。形態素解析後に確認されたが、それらは発話の意図したものとは異なって解析されていた。

以下の表２は、「ほーんとに」と発話された「ほーんと」を形態素解析したものであるが、本来であれば、1形態素で解析されるべきところが、「ほー｜んと」と解析されてしまっていた。

また、このような誤解析は名詞のみならず、その他の品詞でも同様であった。表３は「お願いしまーす」と発話された「しまーす」を形態素解析した例であるが、本来なら「し｜ます」と分割されるべきところが、「しま｜ー｜す」と解析されてしまっている。

表２．発話例「しまーす」の形態素解析例

書字形	発音形	語彙素読み	語彙素	品詞	活用型	活用形	語形	語種
ほー	ほー	ほう	ほう	感動詞-一般			ホー	和
んと	んと	うんと	うんと	感動詞-フィラー			ント	和

表３．発話例「しまーす」の形態素解析例

書字形	発音形	語彙素読み	語彙素	品詞	活用型	活用形	語形	語種
しま	しま	しま	縞	名詞-普通名詞-一般			シマ	和
ー			ー	補助記号-一般				記号
す	す	です	です	助動詞	助動詞-デス	終止形-一般	ッス	和

（2）1形態素中にポーズが含まれる場合

この場合も、（1）の長音の場合と同様に自動で形態素解析した場合、1形態素中にポーズが含まれるため、発話の意図したものとは異なって解析されてしまった。

表4の例では、本来なら「探し｜たい」と解析されるべきものが、「探し｜た｜，｜い」となり、「たい」の部分がうまく解析されていなかった。

表4．発話例「探した，い」の形態素解析例

書字形	発音形	語彙素読み	語彙素	品詞	活用型	活用形	語形	語種
探し	さがし	さがす	探す	動詞-一般	五段-サ行	連用形-一般	サガス	和
た	た	た	た	助動詞	助動詞-タ	連体形-一般	タ	和
,			,	補助記号-読点				記号
い	い	い	イ	記号-一般			イ	記号

（3）音が省略されている場合

この例も、上記（1）（2）と同様に、発話を忠実に文字化した結果、日本語の発音の怠けや癖によって音が省略されたものに誤解析が見られた。その代表例としては、「だから」の「ら」が省略された「だか」であった。発音上は、「だか」であっても、「だから」と解析されるような処理が必要であった。

（4）未知語・不明語の場合

本コーパスは日本語学習開始、半年からのデータを収集しているため、初級段階ではうまく発話できない部分（未知語）や語彙が不明瞭な部分、母語と混同してしまっている部分なども見られた。形態素解析ツールは前後の品詞情報を参考に解析を行っているため、その部分やそれらに隣接する箇所で上手く解析できていなかった箇所が見られた。以下のようなものがその例の一部である。

（例1）みまやー，難しいのピアノないでしょ

（例2）みよいするん，じゃないんですか

（5）擬音語・擬態語を通常より多く繰り返した場合

上記（4）同様に、日本語学習者の発話には、擬音語や擬態語を正確に産出できていない箇所も見られた。例としては、「私はぼろぼろぼろずっと泣いて」や「動いたらぼーんぼんぼんするんですけど」などのように、通常より多く繰り返されたものである。通常通り発話されていれば、うまく解析されるが、繰り返しの部分が通常より多い場合、以下の表5のように他の品詞として、解析されてしまう可能性が高い。

表５．発話例「ぼろぼろぼろずっと泣いて」の形態素解析例

書字形	発音形	語彙素読み	語彙素	品詞	活用型	活用形	語形	語種
ぼろぼろ	ぼろぼろ	ぼろぼろ	ぼろぼろ	副詞			ボロボロ	和
ぼろ	ぼろ	ぼろ	襤褸	名詞-普通名詞-一般			ボロ	和
ずっと	ずっと	ずっと	ずっと	副詞			ズット	和
泣い	ない	なく	泣く	動詞-一般	五段-カ行	連用形-イ音便	ナク	和
て	て	て	て	助詞-接続助詞			テ	和

（６）指示詞とフィラーの判別

「あの」や「その」は指示詞として使用される場合と、フィラーとして使用される場合がある。本データは発話データであり、フィラーとして使用される場合も多く、自動の形態素解析ではなかなかその判別が難しいようであった。そのため、この部分については、誤解析のチェックを行う際に、人手で確認し修正を行ったが、文字上では判別が非常に困難なものもあった。この点については、文字化の際に音声データを参照し、処理をしておくべき箇所であったと考える。

また、フィラーについては、「あの」「その」だけでなく、その他の表現も決まった形だけでなく、様々なバリエーションで産出されるため、誤解析となりやすいものが多かった。そのため、フィラーについては、形態素解析を行う前に特に処理が必要な箇所であると考える。

（７）言いよどみ、言いかけの場合

発話には、書き言葉と異なり、言いよどみや、言いかけが多く出現する。今回の作業では、形態素解析を行う際には、特に処理を施さなかったが、これらも誤解析となる原因の一つであった。これらについても、形態素解析後、人手で修正を行うこととなってしまった。これらの箇所はかなりの量の出現するが、学習者の誤用部分だけでなく、これらにも何らかの処理を施しておけば、人手での作業の軽減となったと考える。

以上のような箇所に誤解析が出現してしまったため、今回の作業では最終的に誤用以外の箇所も人手で誤解析の確認、修正を行うこととなった。今回の作業では形態素解析ツールの仕様に対する考慮が欠けており、誤用部分以外は特に加工を施さず形態素解析してしまったことが原因であった。文字化データを形態素解析するためには、データを詳しく観察してその特徴を十分に把握し、形態素解析ツールの仕様を考慮した上で、それに対する対策を施した文字化データを作成することが好ましいと考える。

3－4.　誤用タグ付与作業について

本コーパスは日本語学習者の習得過程のうち、特に文法項目の習得過程を探ることを目的として作成されたコーパスである。そのため、学習者の誤用箇所も検索可能にするため、誤用にもタグを付与することとした。

3-4-1.　誤用タグの付与基準

誤用タグは研究目的によって付与方針がさまざまである。そのため、どのような誤用タグを付与するかは大きな課題であった。本コーパスでは誤用タグとして、以下の基準で「誤用箇所を示すマーク」と「正用例」の 2 種類のタグを付与した。付与基準は以下の通りである。

①統語的、文法的、あるいは発音が誤用だと判断される場合に、誤用箇所を示すマークを付与し、正用例を記述する。
②正用が想定されにくいものについては正用を記述せず、誤用箇所を示すマークだけを付与する。
③日本語母語話者数名で協議して判断が一致しない場合、および、イントネーション、アクセントの誤用、話し手と聞き手の関係や発話場面が影響する文体や待遇表現に関しては、誤用の判定を行わず、いずれのタグも付与しない。

本コーパスにおいてこれらの誤用タグを付与した利点は、できる限り検索対象を増やす、誤用箇所が判別しやすくなることである。具体的な情報の付与方法は以下を参照願いたい。

3-4-2.　学習者の誤用に対する対処について

ここでは、学習者発話の誤用箇所に対する対処方法および情報付与の方法について説明する。

誤用タグを付与した学習者の発話の誤用部分は、表現としては誤用であっても日本語の形態素として成立していれば、原則、形態素解析の解析結果をそのまま付与することとした。これは、学習者の発話に忠実に語彙素と品詞を振ることにより、正確に検索されるようにするためである。

以下の表は学習者の発話の誤用部分を形態素解析した箇所である。表 6 の例では、「30 分も待ってる」と言うべきところを、「30 分が待ってる」と発話しているが、人手での形態素解析では正しい形の「も」に訂正し情報を付与するのではなく、「が」のまま解析し、情報を付与した。

表６．誤用の基本的な処理例

書字形	発音形	語彙素読み	語彙素	品詞	活用型	活用形	語形	語種
3	さん	さん	三	名詞-数詞			サン	漢
0	ぜろ	ぜろ	ゼロ	名詞-数詞			ゼロ	外
分	ふん	ふん	分	接尾辞-名詞的-助数詞			フン	漢
《が》	が	が	が	助詞-格助詞			ガ	和
待っ	まっ	まつ	待つ	動詞-一般	五段-タ行	連用形-促	マツ	和
てる	てる	てる	てる	助動詞	下一段-タ	終止形-一	テル	和

しかし、学習者の誤用のうち、発音・活用の間違い、文脈から想定される発話の意味と異なって解析されたもの、「～じゃないくて」「こわいくない」「おもしろいかった」のような学習者に特有の誤用は、日本語の文法体系から逸脱しているため、形態素解析ツールではうまく解析されない。そのため、これらには特別な対処が必要であった。そこで、本コーパスでは、以下の通りに対処した。

（１）発音・活用の間違い

発音と活用の間違いは、そのまま形態素解析を行うと全く異なる語として登録されたり、誤解析となったりするため、検索システムで検索した場合その語が検索されなくなってしまう。また、学習者が特定の語をどのように間違えるかということは想定しきれないこともある。そこで、このような誤用には書字形・発音形に発話通りの形を残し、語彙素と語彙素読みは文脈から想定される正しい語の形態素解析情報を付与する。形態素の分割が必要な場合は分割した上で情報を付与する。このように対処することで、形態素単位（語彙素）で検索する際、発音や活用の誤用個所も正用とともに検索することが可能となる。その上、書字形は学習者が発話した通りに登録してあるため、語彙素で検索した場合、書字形を確認すれば、学習者がどのように間違っているのかを確認することができ、文字列検索では、誤用の形でも検索することが可能である。

以下の表７は「経済」を「けいさん」と言い間違えている例であるが、書字形・発音形には発話通りの「けいさん」を残し（発音形では長音部分は「ー」で表示されるため「けーさん」となる）、語彙素読みより右側には正しい形の「経済」の形態素情報を付与する。以上の処理により、「経済」で検索する際には、「けいさん」という誤用も取り出すことが可能である。

表７．発音の間違い

書字形	発音形	語彙素読み	語彙素	品詞	活用型	活用形	語形	語種
《けいさん》	けーさん	けいざい	経済	名詞-普通名詞-一般			ケイザイ	漢

発話の通り

「経済」という正しい形の情報を付与

また、表８は「聞いて」とすべき活用を「聞きて」と間違えているが、この場合も発音の間違いと同様に、書字形・発音形には発話通りの「聞きて」を残し、語彙素読みより右側には文脈より想定される正しい形「聞いて」の形態素情報を付与する。ただ、この「聞いて」の場合は２つの形態素に分割されるため、２行に分割し修正している。

表８．活用の間違い

書字形	発音形	語彙素読み	語彙素	品詞	活用型	活用形	語形	語種
聞き	きき	きく	聞く	動詞-一般五段-カ行-一般連用			キク	和
て	て	て	て	助詞-接続助詞			テ	和

発話の通り

「聞いて」という正しい形の情報を付与

（２）日本語の文法体系から逸脱するような誤用

学習者の誤用のうち、「～じゃないくて」のような誤用は形態素解析にかけると、表９のように「食べ物｜じゃ｜な｜行く｜て」と分割され、「な」と「行く」のように誤解析となってしまう。このような誤用に対しては特別な処理が必要である。以下表 10 のように書字形、発音形では形容詞の「ない」に「く」までを１形態素として登録し、語彙素読みより右側には文脈より想定される正しい形で形態素情報を登録した。以上のように処理することで、「ない」は発話意図通り、否定の「ない」と登録することができる。

表９．「食べ物じゃないくて」をそのまま形態素解析にかけた場合

書字形	発音形	語彙素読み	語彙素	品詞	活用型	活用形	語形	語種
食べ物	たべもの	たべもの	食べ物	名詞			タベモノ	和
じゃ	じゃ	では	では	接続詞			ジャ	和
な	な	だ	だ	助動詞	助動詞-ダ	連体形-一般	ダ	和
いく	いく	いく	行く	動詞-非自立	五段-カ行	終止形-一般	イク	和
て	て	て	て	助詞-接続助詞			テ	和

誤解析となっている

表 10. 日本語の文法体系から逸脱するような誤用の対処法①

書字形	発音形	語彙素読み	語彙素	品詞	活用型	活用形	語形	語種
食べ物	たべもの	たべもの	食べ物	名詞			タベモノ	和
じゃ	じゃ	では	だ	助動詞	助動詞-ダ	連用形-融合	ダ	和
《ないく	ないく	ない	無い	形容詞-非自立	形容詞	連用形-一般	ナイ	和
て》	て	て	て	助詞-接続助詞			テ	和

発話の通り

正しい形の形態素情報

また、「おもしろいかった」の場合もそのまま形態素解析にかけると、以下表 11 のように「駆る」と解析され誤解析が起こってしまう。そのため、表 12 のように「おもしろいかっ」で１形態素とし、語彙素読みより右側には文脈により想定される正しい形で形態素情報を登録した。

表 11.「おもしろいかった」をそのまま形態素解析にかけた場合

書字形	発音形	語彙素読み	語彙素	品詞	活用型			種
おもしろい	おもしろい	おもしろい	面白い	形容詞-一	形容詞			
かっ	かっ	かる	駆る	動詞-一般	五段-ラ行	連用形-促	カル	和
た	た	た	た	助動詞	助動詞-タ	終止形-一	タ	和

誤解析となっている

表 12. 日本語の文法体系から逸脱するような誤用の対処法②

書字形	発音形	語彙素読み	語彙素	品詞	活用型	活用形	語形	語種
おもしろいかっ	おもしろいかっ	おもしろい	面白い	形容詞-一	形容詞	連体形-一	オモシロイ	和
た	た	た	た	助動詞	助動詞-タ	終止形-一	タ	和

発話の通り

正しい形の形態素情報

（3）言いよどみ・語断片・意味の分からない語

学習者の発話には言いよどみや語を言いかけて途中でやめたもの（以下、語断片と呼ぶ）、意味の分からないものも多々見受けられる。以下の表にそれぞれの対処法を示した。人手で処理した部分は枠線で囲った部分である。表 13 は語断片「とも」、表 14 は意味の分からない語「みまやー」の例である。言いよどみと語断片は、品詞欄に「未知語-語断片」という情報を付与し、意味の分からない語には「未知語-不明語」という情報を付与した。また、母語や外国語の発音で発話している場合は、表 15 のように「外国語」という品詞で登録した。

表 13. いいよどみ、語断片の対処法

書字形	発音形	語彙素読み	語彙素	品詞	活用型	活用形	語形	語種
あだし	アダシ	ワタシ	私	代名詞			ワタシ	和
も	モ	モ	も	助詞-係助詞			モ	和
お	オ	オー	おー	感動詞-フィラー			オ	和
とも	トモ	トモ	とも	未知語-語断片				
,			,	補助記号-読点				記号
わだし	ワダシ	ワタシ	私	代名詞			ワタシ	和
の	ノ	ノ	の	助詞-格助詞			ノ	和
友達	トモダチ	トモダチ	友達	名詞-普通名詞-一般			トモダチ	和
も	モ	モ	も	助詞-係助詞			モ	和

表 14. 意味の分からない語の対処法

書字形	発音形	語彙素読み	語彙素	品詞	活用型	活用形	語形	語種
《みまやー》	みまやー	みまやー	みまやー	未知語-不明語				
,								
難しい	むずかしー	難しい	難しい	形容詞-一般	形容詞	終止形-一	ムズカシイ	和
の	の	の	の	助詞-格助詞			ノ	和
ピアノ	ピアノ	ピアノ	ピアノ	名詞-普通名詞-一般			ピアノ	外
ない	ない	ない	無い	形容詞-非自立可	形容詞	終止形-一	ナイ	和
でしょ	でしょ	です	です	助動詞	助動詞-デ	意志推量形	デス	和

表 15. 外国語を使用している箇所

書字形		語彙素読み	語彙素	品詞	活用型	活用形	語形	語種
あー		アー	あー	感動詞-フィラー			アー	和
カンド	カンド	カンド	カンド	外国語			カンド	外
{韓国語で「強盗」}				非言語情報				
と	ト	ト	と	助詞-格助詞			ト	和
かー	カー	カ	か	助詞-終助詞			カア	和

外国語発話

引用文献

小木曽智信・中村壮範(2009)『特定領域研究「日本語コーパス」平成 20 年度研究成果報告書『現代日本語書き言葉均衡コーパス』形態論情報データベースの設計と実装』文部科学省科学研究費特定領域研究「日本語コーパス」データ班.

（佐々木（木下）藍子）

4. 検索システムについて

4－1. 検索システムの構築

C-JAS は発話の全文テキストデータに加え、付属の検索システムを備えることとした。検索システムの構築は、李在鎬氏（筑波大学）が行った。検索システムは、プロジェクトメンバー内での内部公開を経て、いくつかの改良を行い、一般公開となった。

本検索システムでは、検索の利便性を向上させるため、一般的な文字列検索だけでなく、形態素情報を用いた検索が行えるように設計した。その他、話者や調査期の指定、品詞での絞り込みや、意味分類からの検索も可能である。

4－2. 検索画面の説明

図 1 は、2014 年 2 月現在の C-JAS のサイトにログイン後に現れる画面である。ここからコーパスの検索を行うことができる。以下、画面上の各項目について説明する。

図 1．検索画面

① 形態素単位で検索

形態素による検索を行う場合、こちらを選択する。検索キーワードに対して、⑧～⑫で示す検索オプションを加えることで、より高度な検索ができる。

② 文字列で検索

表層の文字列に対して検索する場合、こちらを選択する。ただし、「文字列で検索」を選択した場合、形態素情報を利用しないため、⑧～⑫の検索オプションは指定できない。

③ 検索ボックス

検索したいキーワードを、形態素または文字列で入力する。形態素単位での検索を行う場合、入力する文字種に注意が必要である。詳細は以下⑧を参照願いたい。

④ 話者

日本語学習者を選択して検索を行うことができる。チェックをいれると、検索対象となる。

⑤ 収集時期

データの収集時期単位で検索を行うことができる。チェックをいれると、検索対象となる。

⑥ 発話の種類

学習者または調査者の単位で検索を行うことができる。チェックをいれると、検索対象となる。

⑦ 文脈表示幅

検索結果に表示されるキーワードの前後の文脈に表示される形態素数を選択することが出来る。５、10、30、50、100 語の５種から選択する。30 語以上を選択した場合、⑯の検索結果に表示される文脈が見づらくなるが、データをエクセルにダウンロードした場合（⑮参照）は、一列に表示されるため問題ない。そのため、web 画面上でのみデータを閲覧する場合は、５または 10 語での検索を推奨したい。

⑧ 形態素解析情報を用いた検索オプション 「語彙素」「発音形」「書字形」

検索キーワードが「語彙素」「発音形」「書字形」のどれにあたるかを指定することができる。それぞれの入力に関する注意事項は以下のとおりである。

- 「語彙素」を指定する場合、本コーパス作成時に形態素解析で利用した辞書、UniDic の登録に従い、漢字仮名交じり表記で入力する必要がある。登録に使用されている漢字は、常用漢字以外の漢字も使用されているため、その点に注意が必要である。詳細は 5-3 を参照のこと。

  例）×はしる、〇走る／×ある、〇有る／×この、〇此の／×する、〇為る

- 「発音形」を選択した場合は、すべてカタカナ表記で入力する必要がある。また、長音は「ー」と表記する。

  例）×走る、〇ハシル／×計算、×ケイサン、〇ケーサン

- 「書字形」の場合は、文字化されたテキストデータと一致した表記で入力する必要があ

る。

例）〇走る／〇てんわ（電話）

⑨形態素解析情報を用いた検索オプション 「完全一致」「部分一致」

検索キーワードと部分一致で用例を収集するか、完全一致で収集するかを指定することができる。

⑩形態素解析情報を用いた検索オプション キーワードの「品詞」

検索キーワードの品詞を指定することができる。検索キーワードが空欄でも検索可能である。品詞・詳細はプルダウンメニューから選択することができる。品詞分類は、UniDic に従っている。

⑪形態素解析情報を用いた検索オプション キーワードの「意味分類」

検索キーワードの意味分類を指定することができる。意味分類は『分類語彙表』によっているため、このオプションが利用できるのは、名詞、形容詞、動詞のみである。意味分類の詳細は、『分類語彙表』を参照されたい。具体的な検索キーワードを指定せず、意味分類から形態素を抽出することも可能である。

⑫形態素解析情報を用いた検索オプション「誤用」「正用」の指定

学習者の誤用と正用、あるいはその両方を指定することができる。

⑬全文会話のダウンロード

コーパスに収録されている全会話の文字化資料をダウンロードすることができる。

4－3. 検索方法および検索結果

4-3-1. 語彙素の入力

本コーパスの特徴である形態素単位による検索について、詳細を述べる。形態素単位での検索を選択し、さらに語彙素で検索をする場合、語の入力に注意が必要である。4-2 ですでに述べたが、本コーパスは形態素解析の際に UniDic を使用しているため、語彙素の登録は UniDic の表記に準じている。そのため、検索語を入力する前に、語彙素が UniDic にどのように登録されているかを知る必要がある。登録は、おおむね一般的に使用する表記と同じであるが、中には表１のように使用し慣れていない表記が用いられている場合がある。

表１．注意すべき語彙素の例

検索したい語	語彙素	検索したい語	語彙素	検索したい語	語彙素
あげる	上げる	くれる	呉れる	もらう	貰う
する	為る	やる	遣る	いつ	何時
これ	此れ	それ	其れ	あれ	彼れ
ちょっと	一寸	ほとんど	殆ど	もし	若し

また、「てある」「なければならない」など、文型と呼ばれるような表現を検索する際にも注意が必要である。「てある」は、語彙素に区切ると「て|ある」という 2 つに区切られ、それぞれの語彙素は「て」「有る」である。「なければならない」は、語彙素に区切ると「なけれ|ば|なら|ない」という 4 つに区切られ、それぞれの語彙素は「ない」「ば」「成る」「ない」である。そのため、「語彙素で検索」を選択した状態で検索ボックスに「てある」や「なければならない」と入力しても、何もヒットしない。一方、この場合、文字列検索では検索が可能である。

語彙素を用いて正しく使用例を検索するためには、形態素解析の特徴を理解したうえで検索をする必要がある。語彙素を知るためには、「茶まめ」を利用することができる。「茶まめ」は UniDic を使用して形態素解析を行うのを補助するためのソフトウェアで、以下からダウンロードが可能である。

http://sourceforge.jp/projects/unidic/

4-3-2. 検索結果の見方

次に、検索結果の見方について説明する。「形態素単位で検索」「語彙素で検索」を選択し、「食べる」という語を検索した場合を例としたい。

検索ボタンを押すと、ヒット件数の表示とともに、図２のような検索結果のクロス集計表が現れる。検索結果をもとに、検索キーワードの出現回数を、学習者ごと、時期ごとに自動で集計し、一覧で出力したものである。

話者	第1期	第2期	第3期	第4期	第5期	第6期	第7期	第8期
K1	2	0	1	3	21	0	5	37
K2	0	8	3	2	6	5	17	3
K3	0	10	6	12	1	16	29	3
C1	2	0	3	0	1	4	11	1
C2	1	0	1	0	3	7	24	0
C3	0	0	3	14	2	2	20	0

図２．検索結果のクロス集計表の例

ヒットした発話は、以下のように表示される。図３は、Ｋ１を検索対象にした「食べる」の検索結果である。検索キーワードをハイライトしながら、KWIC 形式（Key Word in Context）で、キーワードの前後の文脈を含め表示される。緑字で示されているのは調査者の発話、黒字で示されているのは学習者の発話である。表示される形態素数は、検索画面の「文脈表示幅」で選択することができる。１形態素ごとに区切って表示されるが、調査者の発話にはさらに１形態素ごとに下線が付されている。

行頭の「全文閲覧」（青字）をクリックすると、その用例を含む全文が参照できる。また、発話

末に記載されている「発話番号」は、「全文ダウンロード」でダウンロードできる全文テキストに記載されている発話番号である。

「全文閲覧」という文字の下に赤字で「誤用」と表記されている場合は、検索キーワードに誤用タグが振られている場合である。また、赤い下線が付されている箇所は、検索キーワード以外で誤用タグが付されている箇所であり、下に赤字で考えられる正用が表示されている。正用は、全ての誤用に表示される訳ではなく、また、表示されている場合であってもそれが唯一の正用であるとは限らない。また、誤用と思われる箇所であっても、赤線が表示されない場合もある。これは、誤用の判定そのものが大変難しい作業であり、タグ付与の際に検討を重ねたものの、判定者の判断が一致しない箇所があったためである。そのため、誤用箇所および正用の表記は、参考として扱っていただきたい。

検索結果の発話の下の赤字の[　]でくくられた内容は、補足情報（参照：3-2-1.文字化の方針<表記の方針>(8)）であり、下線の上の発話の補足情報である。たとえば、検索結果１の「わだし」の下には赤字で[私]、「ちょとー」の下には[ちょっと]と記されている。これは主に発音の間違いであり、　[　]が付されている箇所は、形態素解析結果を手作業で修正する中で、正しい語彙素に修正してある。例えば、検索結果１の「わだし」「ちょとー」の場合、書字形は「わだし」であっても語彙素は「私」、書字形が「ちょとー」であっても語彙素は「ちょっと（一寸）」としてある。一方、[　]のない赤字部分は、形態素情報の修正をしていない（参照：3-2-2.学習者の誤用に対する対処について）。

「Φ」は学習者の発話のうち、必要な語が脱落していると思われる箇所である。

図３．「食べる」の検索結果（一部）

4-3-3.　検索結果のダウンロード

検索結果は、テキストファイル（TXT）とエクセルファイル（XLS）でダウンロードすることができる。図４にエクセルファイルでダウンロードした場合の例を示す。

	A	B	C	D	E	F
1	file	error	left context	keyword	right context	utterance ID
2	K1-c		L:んですね，わだしは，ちょとーもうこはん	食べ	たら，わだしのー部屋に行きますからー，	0238
3	K1-d		L:，お風呂入って，N:〈うん〉L:ゆタご飯N:〈うん〉L:	食べ	て，したら，N:〈うん〉L:11時ぐらい	0062
4	K1-d		L:，ぐらいでー，韓国人と，昼ご飯	食べ	て，N:〈うん〉L:話をしてジュースでも	0076
5	K1-d		L:さんは，他の女の人は，	食べる	だけです，N:〈うん〉L:それでも，【人名AA】さん	0548
6	K1-e		N:どこで食べるの，いつもL:あの部屋で	食べ	ますよN:買ってきて食べる?L:はいN:あのー	0070
7	K1-e		L:と，私は好きなものはも，	食べる	ものでも，よく，なんと言います	0076
8	K1-e		L:嫌いですねN:〈あーそう〉L:，それで今は少し	食べ	ますけど，韓国にいた時はじぇんじぇん	0076
9	K1-e		L:ますけど，韓国にいた時はじぇんじぇん	食べ	なかったですねN:魚?L:魚でも，	0076
10	K1-e	誤	N:?L:魚でも，ひとちゅふたちゅだけ	食べ	たから，それ名前は分からないですけど	0078
11	K1-e		L:肉も，小学生Φどきは肉も	食べ	なかったΦですよN:K1君は一体何	0082
12	K1-e	誤	N:のーL:私もそれ，それで，何を	食べ	たかよく分からないですねN:[笑]L:ええ，	0084
13	K1-e		L:ですね，それで作ったものを，	食べ	たらおいしかったですがN:キムチはL:キムチもよ	0096
14	K1-e		L:はしょうがつ，小学生5年生までは	食べ	なかったΦですよ，韓国人でも	0098
15	K1-e	誤	L:なかったΦですよ，韓国人でも	食べ	なかて，あの，親戚の人が一緒	0098
16	K1-e	誤	L:とき，あの，親戚の人がキムチよく	食べ	たΦですね，おいしそうに食べるから	0098
17	K1-e		L:よく食べたΦですね，おいしそうに	食べる	から，私も1回2回食べて	0098
18	K1-e		L:に食べるから，私も1回2回	食べ	て，そのとき習ったΦですよN:あー	0098
19	K1-e		N:とかってほんと?韓国の人L:私，	食べ	たことがないですね，それ，見	0102
20	K1-e		L:と他のもの，肉とかいれて	食べる	ΦですよN:おいしい?L:「おいしい」と言っ	0104
21	K1-e		L:おいしい」と言ったですけど，私は	食べ	たことはないですねN:ふーん，そう，	0106
22	K1-e		N:どんなもの食べてるの?L:今はべんとうよく	食べ	たり，それと，ラーメン買ってN:インスタントラーメン	0120
23	K1-e		L:た時ですね，それで，夜ご飯	食べ	て，私は，あの，いとこさんの	0136
24	K1-e		L:ですね，それでもあの日，ご飯	食べ	て，トイレに行くΦかなーと思っ	0136
25	K1-e		L:に，私が，その時も，こはん	食べ	てトイレへ行った時ね，ですけど	0176

図４．「食べる」の検索結果のダウンロード（エクセル）

エクセルファイルの「file」列は「K1-a」のように、各学習者記号とその時期が表示される。最初の 2 文字（「K1」）は学習者を、ハイフンに続くアルファベットは時期を示している。時期は 1 期の「a」から 8 期の「h」までである。「error」列には、検索キーワードに誤用タグが付与されている場合に、「誤」と表示される。前後の文脈（left context および light context）に示される文脈は、L（学習者）と N（調査者）という文字によって、発話者が表示される。「utterance ID」列には「File」列に表示されたデータの全文テキスト中に表記されている発話番号が表示される。

4ー4.　形態素単位の検索を用いた検索例

前述の通り、本コーパスの特徴の一つとして、文字化データに形態素解析を行い、形態素単位での検索を可能にしていることがある。形態素解析結果を用いることで検索の利便性が向上する事例を紹介する。

4-4-1.　多様な活用形を語彙素でまとめて検索する場合

文字列検索で動詞「行く」の多様な活用形を検索する場合、それぞれの活用形ごとに検索をかける必要がある。しかし、形態素解析されたデータであれば、語彙素「行く」を指定することで、多様な活用形および書字形を一度に検索することができる。

図５では、ハイライトされた検索キーワード「行き」「行っ」「行か」が検索されているのが分かる。「語彙素」による検索の場合は、表層の書字形が漢字・ひらかな・カタカナのどれであっても、形態素解析結果において語彙素が「行く」になっている例が検索される。

図５．語彙素「行く」の検索結果（検索画面一部省略）

4-4-2.　品詞情報を利用して形態素を検索する場合

助詞の「は」や「が」など、ひらがな１～２文字の形態素を文字列検索しようとすると、不必要な例が大量にヒットしてしまう。しかし、C-JAS は形態素解析されたデータであるため、品詞を指定して検索することで、不要な例を除外することが可能である。

図６では、助詞の「に」の検索結果を示す。助詞「に」を検索する場合、「形態素単位で検索」を指定し、検索キーに「に」と入力する。そして、検索オプション２で「語彙素で検索」にチェックを入れ、キーワードの品詞は「助詞」を選択する。

しかし、C-JAS が話し言葉、かつ学習者の発話であることによる誤解析には注意が必要である。また、助詞の「に」には多様な意味分類があるが、そのような意味に関わる分類では絞り込みは行えない。

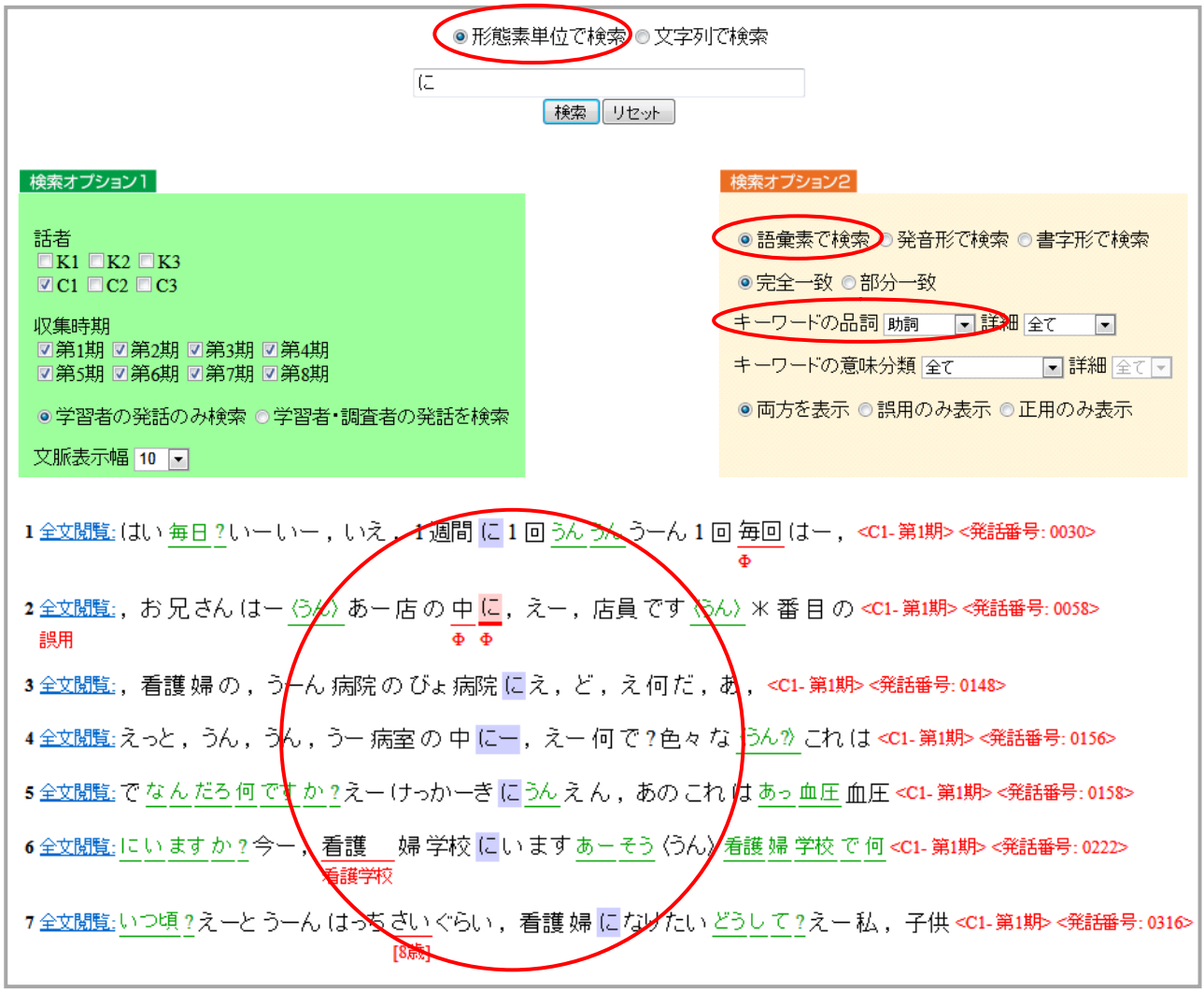

図６．語彙素「に」（助詞）の検索結果（検索画面一部省略）

（小西円・李在鎬）

5. 研究報告

C-JAS を使用した研究および発表は以下の通りである。

（1）図書

1．迫田久美子「非母語話者のコミュニケーションの工夫」野田尚史（編）『日本語教育のためのコミュニケーション研究』pp.105-124，東京：くろしお出版，2012.

（2）論文

1．Irena SRDANOVIC, Kumiko SAKODA (2013) Analysis of Learner's Production of Adjectives Using the Japanese Language Learner's Corpus C-JAS: The Case of *takai*. *Acta Linguistica Asiatica vol3*.pp.9-24.

※　次ページに論文掲載

（3）発表

1．迫田久美子・木下藍子・小西円・李在鎬「日本語学習者縦断コーパスの構築について」2011 年度国立国語研究所公開シンポジウム「多文共生社会におけるコミュニケーションとその教育」，ポスター発表，2012 年 2 月

2．迫田久美子・木下藍子・小西円・李在鎬「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－」日本語教育学会国際大会（JCJLE2012），ポスター発表，名古屋，2012 年 8 月

3．迫田久美子・木下藍子・小西円・李在鎬「日本語学習者の縦断的会話コーパス『C-JAS』の構築」2012 年日本語教育学会秋季大会，デモンストレーション，北海学園大学，2012 年 10 月

4．木下藍子・迫田久美子・小西円・李在鎬「日本語学習者のタグ付き発話コーパス『C-JAS』－C-JAS（Corpus of Japanese as a second language）開発と利用－」国立国語研究所 2012 年度「多文共生社会における日本語教育研究」研究発表会，ポスター発表

おわりに

本コーパスおよび検索システムは、日本語の第二言語習得研究の発展を願い、日本語を第二言語として学ぶ学習者の日本語に興味を持ち研究する方や日本語教師の方に活用して頂きたいという思いから作成することとなりました。

本来、コーパスを構築する場合、まずコーパスの最終的な形を設計し、それを踏まえた上で、調査およびデータ収集、そしてデータ化と作業を進めていくものだと思います。本プロジェクトでは日本語学習者の 3 年間の縦断的発話データという、大変貴重なデータが扱えることとなりました。途中段階からの作業開始でしたが、さらに本コーパスの活用の幅が広がるよう、検索システムを備えることになりました。

今回の作業では、データを検索システムで使用するため、様々な作業を行いました。その過程で、発話データを文字化する困難さやデータ全体の表記に一貫性を持たせることの困難さ、日本語学習者の産出する予想外な発話を形態素解析し、データ化することなど、多くの壁にぶつかりました。また、作業を進める過程で、データをこんな風にするともっと使いやすくなるのではないか、検索システムの仕様をこんな風に変えるともっと多くの人にも使ってもらえるのではないかなど、様々なアイデアが湧くこともありました。

途中段階からの作業であったこともあり、作業がスムーズにいかないことや、失敗もありました。しかし、失敗から学べたことも多く、我々にとっても大変勉強になったと感じています。本報告書が、今後コーパス構築を目指す方々にとって、少しでも役に立つものとなれば大変嬉しく思います。

本コーパスおよび検索システムは、データ収集に協力してくださった学習者の方々、文字化の修正や形態素解析の修正に協力してくださったアルバイトの方々、ご助言くださった皆様方のご支援により完成させることができました。また、筑波大学の李在鎬先生には、多大なるご指導に加え、多くのご支援を賜りました。ここに感謝申し上げます。

2014 年 3 月 4 日

迫田久美子（国立国語研究所）
佐々木（木下）藍子（国立国語研究所）
小西円（国立国語研究所）