立命館大学における 2 つの日本語会話コーパスのデザインと編纂

田中良[*]・川本真佐美[**]

要旨

立命館大学において 2009 年度と 2010 年度に計 2 つの日本語会話コーパスの作成を行った。1 つは母語話者データで、主に会話分析や談話分析を目的としたコーパスである。もう 1 つは母語話者と学習者の接触場面データで日本語教育に役立てることを目的としたコーパスである。これらのコーパスは今まで扱うことが難しかった様々なデータを自在に扱うことができるものである。本稿は、このコーパス構築の仕組みと各作業工程での実際の判断基準、更にコーパスチーム運営のコツを提示することで各研究者が独自のコーパスを構築することができるようになることを目的とする。

[*] 2008 年度　立命館大学　大学院　言語教育情報研究科修了生、関西学院大学　大学院　言語コミュニケーション文化研究科　院生
[**] 立命館大学　大学院　言語教育情報研究科　院生

1. 2つのコーパスと仕組み

本稿は、立命館大学において、2009 年度、2010 年度に編纂した 2 つの日本語会話コーパスの編集作業方法と方針、各種判断基準を示し、読者が各自で同じ仕組みのコーパスを作成できるようになることを目指すものである。

2つのコーパスは同じ仕組みで作られている。設計は「多種情報記述による再現性の高い自然会話コーパス構築システム」(田中,2011)を基にしている。それぞれ日本語母語話者同士による会話（「立命館日本語会話コーパス」以下「立命会話コーパス」）と、日本語母語話者と学習者の接触場面会話（「立命館日本語学習者会話コーパス」、以下「立命学習者コーパス」）で、仕組みは同じであり多くの情報タグは同一のものが付与されているが、一部の情報タグには違うものがある。共通するタグは、「表記形」「基本形」「品詞」「品詞下位分類」「活用形」「活用型」「読み」「母音配列」「モーラ数」「プロソディ」「発話の重なり」である。それぞれに個別のタグは、立命会話コーパスでは主に談話分析、会話分析の視点でデータを選別し、「発話権」「笑い」「視線」「頭」「上体」を付与している。対して立命学習者コーパスの方は、日本語教育に特化して情報タグを選別し、「正表記形」「誤用」「語彙レベル」「文型レベル」を付与している。

タグには機械処理で自動で付与されるものと人手での作業により付与されるものがある。全体の編集は、まず録音した音声の文字化を行いタグ付けを自動で行い、更にそれぞれの工程での編集をしていくという流れになる。

具体的な編集は専用エディタを作りそれで行った。その編集ソフトのメイン機能はほとんどを「多言語対応コンコーダンサーHASHI(田中,2010)」のバージョン0.8.10.0以降に移植済みである。それらの機能はこのソフトによって一般に公開する予定なので、そちらのマニュアルを参照いただきたい。つまり、HAHSIがあれば本コーパスとほぼ同等のものを各自が作れるということである。ただし、日本語教育レベルでの語彙、文型タグの自動付与は行えない。専用に作成した文型辞書が、参考文献で提示された文型をリスト化した上で様々な文法情報を付与したものであり、著作権対応に不安があるためである。

本コーパスの編纂工程は、データ協力者やスタッフの募集、音声データの録音、文字化方法と方針、専用エディタでの編集、の順になる。本稿では、具体的な作業方法の説明は最小限に控え、主に、各項目を編集する際に出てくる様々な状況に対する判断根拠を示す。

作業方法であるが、一般には本専用エディタではなく、HASHIによってその機能を公開するので、ここで提示するものとは完全には一致しないがほぼ同一の作業で各自が行える。

コーパス作成の内容を分けると、機械技術的なもの、日本語文法などを基にしたコーパスの内容判断、作業スタッフの運営管理になる。本稿では主にコーパス判断を記し、必要に応じて運営管理上の注意点を示す。

2つのコーパスは「立命会話コーパス」をはじめに、その経験を元に「立命学習者コーパス」を作成した。本稿では特記無い限り「立命学習者コーパス」についての記述とする。

2. 作業手順

編纂作業の流れは、①スタッフ選別、②データ協力者選別、③会話音声録音、4音声の文字化、⑤各種タグ修正、付与作業、⑥最終コーパス整形となる。

③で録音したファイルを「音声ファイル」、④で作成する文字化のスクリプトを「文字化ファイル」、⑤で文字化ファイルに一次の整形をし、様々なタグを編集するために作成したファイルを「一次整形ファイル」、⑥で一次整形ファイルを最終コーパス形式へ再整形し、検索や統計などの利用法を可能にしたファイルを「最終コーパスファイル」とする。「一次整形ファイル」は様々な工程を経て編集が行われるため、作業段階によって内容が大きく変化するが、これを総称して「編集ファイル」とする。

それぞれの作業の詳しい内容は、②、③を 3 章で、③を 4 章で、④を 5 章で、⑤を 6、7 章で示す。

3. 作業スタッフ、データ協力者

作業スタッフは学内メーリングリスト等で募集し、作業説明会を行った。スタッフは言語や文学を専攻する学部生と院生から構成される。その際に、本コーパスではデータ協力者の個人情報を扱うため、その秘匿に関しての同意書をもらった。この同意書は管理者も含めた全スタッフが提出した。

同様にデータ協力者からも、データ録音第1回目に別途作成の個人情報に関する同意書を提示しながら口頭での説明の上に署名をもらった。データ協力者は日本語学習者と日本語母語話者のペアである。詳しい説明は、田中(2011，pp159-161)を参照されたい。

データ協力者の数であるが、データに必要な人数よりも多く用意し、余計にデータを録音しておく方がいい。特に今回は一定期間を開けて数回に亘ってデータを録音する縦断的データであったので、データ録音開始から数回後に協力者が参加の辞退を申し出てくるなどがあった場合、時間を遡って他のペアでデータを取り直すことは不可能であるため非常に不都合となる。したがって予備のデータは2~3組は有った方がいい。予備データは特に問題が起こらなければ文字化もせずにキープしておき、使用不可能なデータが出た時に利用するようにすれば、通常時では特に作業の負担にもなることもない。

逆にスタッフは予定よりも多く集まった場合、それぞれが希望する分量や内容の作業が割当たらない可能性があり、参加意欲の低下の原因となりやすいので注意が必要である。管理者の目的はコーパスの完成や研究の推進であり、ついその視点でのみ考えがちだが、スタッフのうちいくらかは「バイト」という感覚で収入を目的に参加する者もいる。どのような目的であるにせよ、意欲的に参加し丁寧に作業をするスタッフは非常に戦力になる。互いの意思の齟齬が無いように各スタッフの参加目的を把握しておくことは大事である。

4. 音声録音

会話コーパスの編纂では最初に会話の音声データを用意するところから始まる。本コーパス

は自然会話のデータを採用したため、データ協力者の会話の録音から行った。データの録音は常に協力者2人一組で行い、個別の部屋に2人きりになるようにし、録音者も退出した。これにより、完全な二人だけの空間での会話を採取した。

録音時の状況は、机にハの字で向かうように2つの椅子を用意し、2つのICレコーダー、3つのカメラを用意した。

具体的な録画場面は以下の通りになる。

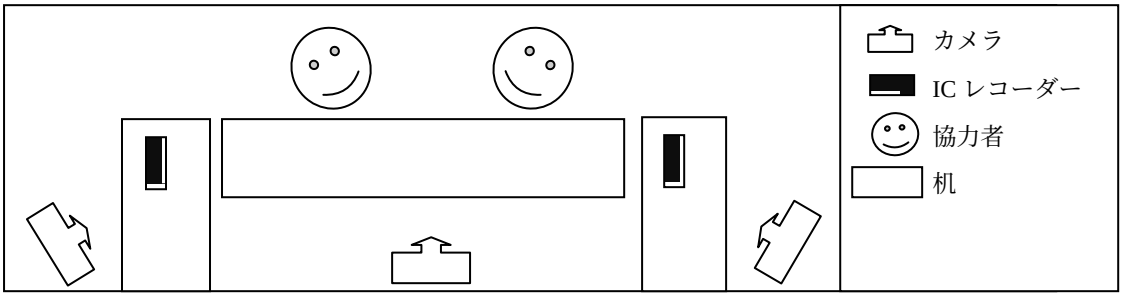

図1. 録音時の状況

カメラは3台用意し、正面に1台、斜めからお互いの顔を録画できるように2台使用した。ICレコーダーは2つを用意した。配置は、発話者による机の振動の音を拾わないようにデータ協力者の座る机とは別の机の上に置き、集音性を高めるために高機能マイクを接続した。録音の際の別の注意点として、雑音の混入には特に気を付ける必要がある。特に空調の音を拾うと文字化に大きく支障が出るくらいの雑音になる。録音時に部屋の中で耳で確認してほぼ問題無いと思われても、録音終了後文字化する際にはかなりの雑音になっていることがある。本コーパスでも秋から真冬にかけて録音したが、空調はできるだけ切って行った。また、ICレコーダーで記録した音声が聞き取りにくい場合に、本来映像確認用に撮っておいたカメラの映像の方が、付属のマイクの性能が高いため音声がクリアであり、大いに役に立った。録音された音のクオリティによってその後の作業効率が大きく変わり、場合によって作業時間や人件費に大きく影響を及ぼすので、十分気を付けたい。

5. 文字化

音声の録音後はその内容を文字に起こす作業を行う。作業時に用いた音声ファイルの再生ソフトは「Audacity」である。これは高性能な音声編集用のフリーソフトである。

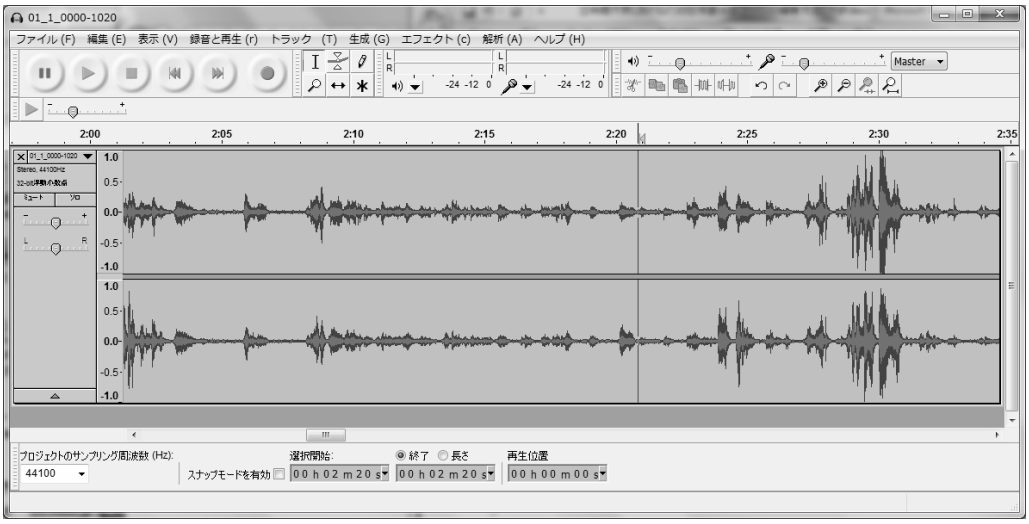

図2．Audacityの作業画面

本コーパスでは、発話の重なりの記録の際にひらがな1文字の単位で重なり幅の記録をし、ポーズの長さを表す「・」と「、」の厳密な運用のために01秒単位でポーズ箇所を見るので、音の波形や、時間をあらわすメモリが必要なためこのソフトを利用した。

文字化内容はテキストファイルに記録するので、作業はメモ帳でも秀丸でも、各自が一番作業しやすいソフトを使用して行った。

文字化の際に使用した記号とその意味合いは以下の通りとなる。

表 1．スクリプト内使用記号一覧

ー	音に震えの伴わない長音
～	音に震えの伴う長音
h	笑い声
f	強い息
X	聞き取り不可
・	1 秒以上の沈黙
、	1 秒未満のポーズ
,	語の区切りを明確にする

文字化ルールと注意点は田中(2011, pp162-166)で詳しく説明しており、そちらを参照されたい。

各スタッフへ文字化の作業を割り振る単位は1分の音声とした。作業にかかる時間はおよそ45分である。立命会話コーパスの場合には文字化作業を10分単位としたが、作業者によって進行に大きく差が出て文字化ファイルの提出が揃わなかったため、単位時間を短くした。音声の厳密な文字化は非常に根気のいる作業で、文字化10分の作業は10時間程度になることも多い。

このため、作業者の負担感が強く時間もかかるため必然的に締め切りが遠くなることもあり作業に向かい始める時間が遅くなる。そこで1分を最低単位とし、作業の進行具合によってスタッフごとに2分ぶん、4分ぶんなどに小分けにして割り振ることで作業の負担感を減らし、また各箇所の文字化作業が遅延した際のリスクも分散した。

作業時に規定した文字化ルールが徹底されにくかったため、専用エディタに「文字化確認」機能を入れ、文字化完了ファイルの記号や書式の運用に不備が無いかを作業者各自が自動でチェックできるようにした。記号運用の不備は特に全角と半角の違いが多く半角のスクウェアブラケット"[ ]" とすべきところが、キーボード上でこのキーを押した際に全角モードの場合に表示される鍵括弧"「 」" になることや、半角丸括弧"( )" が 全角丸括弧"（ ）"になることなどが多かった。これは、日本語コーパス編纂のため常時全角モードで文字化作業を行い、記号の入力の際に半角への切り替えを忘れるためと思われる。

自動での文字化チェックは非常に有効である。ただしこれは記号や書式しかチェックできないため、文字化内容が元の音声ファイルと一致しているかまでは判別できず、人手によるチェックが必要であった。この文字化確認作業を全ファイル2回、作業者を変えて行った。1分区切りで文字化したファイルを10分ぶん連結して1ファイルとし、その後の全ての作業ファイルの単位とした。二重チェックを行うことで文字化の質を統一できると考えられたが、実際にはこれでも漏れは多く残った。残った不備は一次整形後のタグ編集作業時に随時修正を行った。しかし、テキストファイル上で修正することと専用エディタ上で修正することでは、前者の方がはるかに行いやすいため、できるだけ文字化段階での精度を高めることがその後の全ての作業を早く行える元となる。

6. 一次整形後の作業

6.1 タグ編集作業項目

文字化の確認までが完了した後は、それに対しいったん自動処理で大量のタグを付与し、「一次整形ファイル」とする。以降は、これらの自動付与されたタグの修正や、機械処理では自動的に付与できないタグを手作業で付与する作業を行う。

自動で付与されるタグは「表記形」「基本形」「品詞」「品詞下位分類」「活用形」「活用型」「読み」「母音配列」「モーラ数」「発話の重なり」である。手作業で付与するタグは、「プロソディ」「誤用」である。その他として、自動付与の結果を大きく利用するが個別の判断を行わなければいけないものに「正表記形」、いくつかの手作業でのタグ修正の結果を基に自動付与できるタグに「語彙レベル」「文型レベル」がある。

また、立命会話コーパスでは主に談話分析、会話分析の視点でデータを選別し、「発話権」「笑い」「視線」「頭」「上体」がある。これらは全て手作業でのタグになる。

文法項目等のタグ付与は奈良先端科学技術大学院大学で開発されたChaSenを使用し、内部辞書には伝康晴・他(2009)により開発されたUniDic version 1.3.12を用いた。

他に、話者の属性や発話番号などの行ごとに付く情報はすべて自動でタグが付与される。こ

の中で「開始時間」「終了時間」は手作業での修正が必要である。会話コーパスの場合はこれに加え「発話番号」を手作業による修正を行った。

以後の作業は、これらの情報タグを作業領域ごとにまとめて行った。具体的な編集作業のカテゴリーは「時間付け修正」「形態素分け修正」「正表記形」「プロソディ付与」「読み修正」「重なり修正」「誤用」「各文型」となる。

これらの各作業をスタッフごとに割り振った。作業順序も基本的にこの通りになる。

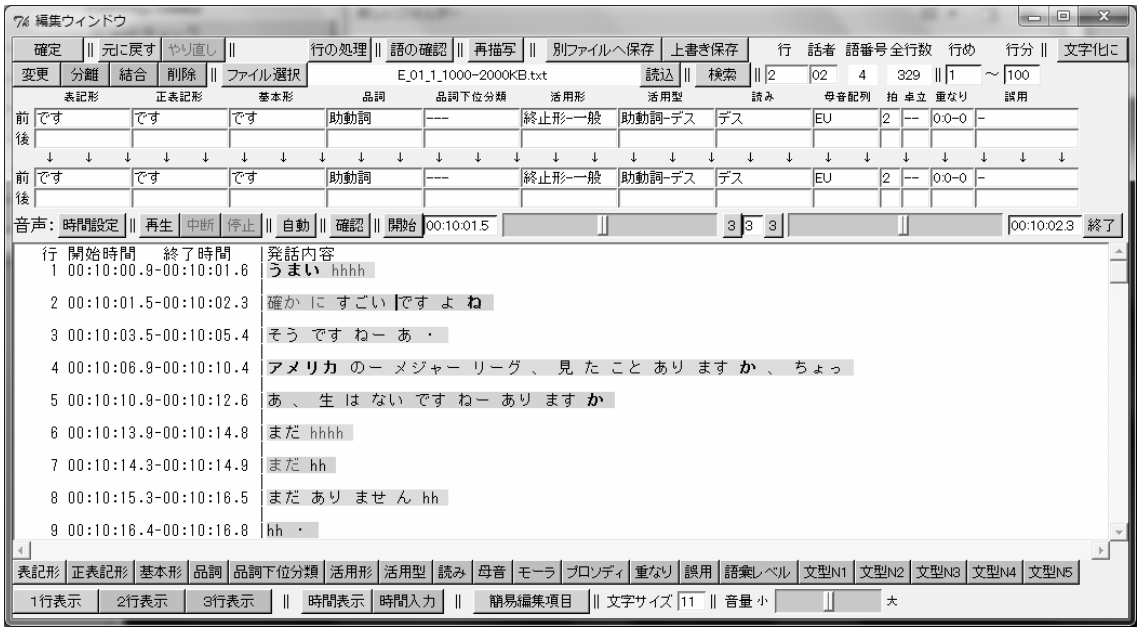

図3．専用エディタ編集ウィンドウ

以降の作業は全て専用エディタによって行う。

6.2 時間付け修正

本コーパスは会話コーパスであるので利用中に該当行の音声を確認する必要が出てくる。このため音声ファイルと対応するよう各行の開始時間と終了時間を0.1秒単位で記録した。

「開始時間」と「終了時間」はまず自動で全行に付けられる。しかしこの自動で付けられた時間はずれが大きいため修正が必要である。編集する行をクリックののち「確認」ボタンなどで音声を聞きながらスライダーで調整し、各ボタンで修正した時間を付与する。

この作業で各行の開始時間と終了時間を付けることにより、以降そのファイルの編集中にいつでも該当箇所の音声を確認できるようになる。これは、発話の重なり、プロソディ付与などの作業をするために必要な工程で、他の作業においても文字上では内容の確認が取りにくい際に音声の確認をする必要が出ることが多いため、重要な工程である。

時間付与は各行の音声が正に0.1秒ずれると切れてしまうというギリギリの幅で付与した。前の行の終了時間と次の行の開始時間は連続しなくても構わなく、沈黙が長く現れればどの行

に付与された時間にも入らない時間帯も存在する。逆に、該当行の音声が終わる前に次の話者の音声が重なって始まっていた場合、該当行の文字化された内容が終了する位置で終了時間を付与したため、行ごとに時間の重複する箇所が存在する。他の行の発話内容の最中に重なって現れた音声の行の場合、前の行の開始時間と終了時間の間にある時間帯にその行の開始時間、終了時間が付与されることも有りうる。つまり、どのような場合でもその行に文字化された内容のギリギリの開始位置から終了位置までの時間を付与するという作業になる。

6.3　形態素分け修正

一次整形により専用エディタで自動で分けられた語の区切りや付与された文法などのタグには解析ミスの箇所も多く含まれる。その訂正作業として形態素分け修正が必要となる。修正する箇所は、判断の微妙なものではなく、明らかにおかしい箇所で、「動詞」を「名詞」としている場合や、語の区切りがおかしいなどといったもののみを修正する。

ここでは、起こりやすい形態素解析ミスと、判断が付きづらい分析結果を提示する。本コーパスでは、形態素解析をChaSen＋UniDicの組み合わせにより行っているので、以降の提示する内容もその結果である。

文字化の際に長音を2文字以上で伸ばして記述された箇所や、通常ではあまり使わない記録方式で長音を記述された箇所はまず解析が意図通りされない。「あの～」とした箇所は自動解析では「あの」「～」に分割される。修正でこの2つを結合し、「あの～」という一形態素にする。基本形は長音記号を抜かした「あの」にし、読みは「アノ～」にする。母音配列は「AO-」にする。このように「～」などの記号は、1つで独立した形態素と解析されることが多い。同様に、「えぇぇ」などは「えぇ」「ぇ」と2形態素に分かれるため、「えぇぇ」と1形態素に結合する。このように「ぁぃぅぇぉ」が入る場合も単独で1形態素とされることがある。「電車で40分くらいかなあ」の「なあ」が「な (助詞　終助詞)」「あ (感動詞 一般)」と解析されるため、これを「なあ (助詞　終助詞)」と結合する。

「その人がいます」の「その」は「連体詞」と解析されるが、「あの人がいます」の「あの」は「感動詞」とされる。このように感動詞と分析される語は多く、「その」「あの」などは気を付ける必要がある。

「いきたがる」が、「た (助動詞-タ)」「が (格助詞)」「る (記号)」になるなど、明らかにおかしい解析がされることもある。本コーパスでは文字が記号と解釈される場面はあり得ないので、「記号」という品詞は注意する必要性が高い。

「そうか」、「そっか」、「そか」のように、縮約形で発話されやすく、様々な形で現れる語のとき、この場合では「そうか」が「そう (副詞)」「か (副助詞)」と解析されるので、他の形もその解析結果と同様に、「そっ」、「そ」を副詞とし、基本形はすべて「そう」とした。その他の類似の語も同様にする。

「なんか」は大きく3つの形で解析される。

①「そうなんかな」　→「な (助動詞-ダ)」「ん (準体助詞)」「か (助詞 終助詞)」

②「なんかいます」　→「なん (代名詞)」「か (副助詞)」
③「旅なんかしない」→「なんか (副助詞)」

となるのが基本であるが、「私なんか無理です」は③になるはずだが、①と解析され、「そこになんかいます」は②のはずだが、③になる。「なんか」は1つ1つ注意して確認しないと解析ミスを見落としやすい。

否定の「ない」も「助動詞」と「形容詞」の2つに分かれる。

①「綺麗でない」→「助動詞-ナイ」
「ない」は活用語尾として「綺麗な」という語の一部となるため助動詞
②「綺麗ではない」→「形容詞」
「ない」と「綺麗な」の間に「は」という係助詞があり、「ない」は「綺麗な」とは独立した自立語なので形容詞
③「綺麗じゃない」→「形容詞」
「じゃ」は「では」の融合形なので「綺麗ではない」と同じルールで「ない」は形容詞

縮約形では、「じゃ」は「助動詞-ダ　連用形-融合」、「じゃん」は「終助詞」

「っす」は、「オレです」→「オレっす」のように本来が「です」の縮約形であるが、「行くっす」のように本来は「ます」が来る位置に代わりに使われることもある。本コーパスでは、語幹との活用がおかしくはなるが「っす」の基本形は一律「です」とした。

関西方言の終助詞、助動詞は以下の通りとした。

「そうやってん」「てん」　→「て (助動詞-テ)」「ん (準体助詞)」
「無理やん」　「やん」　→「や (助動詞-ヤ)」「ん (助動詞-ヌ)」
「そうやねん」　「ねん」　→「ねん (終助詞)」
「行かへん」　「へん」　→「へん (助動詞-ヘン)」
「できひん」　「ひん」　→「ひん (助動詞-ヒン)」
「してはる」　「はる」　→「はる (助動詞-ハル)」
「そらあかん」　「あかん」→「あか (動詞 あく)」「ん (助動詞-ヌ)」
「おまえや」　「や」　→「や (助動詞-ヤ)」

また、解析ミスではないが修正を加える要素として、一つの語の中に1秒未満のポーズがあった場合、文字化の際にはそのポーズの箇所を「、」によって区切って記述し「た、い変」の形で表記されているため、自動での形態素解析結果では「た」「、」「い」「変」と解析される箇所がある。これを表記形で「た、い変」という一語にし、基本形を「大変」に、読みを「タイヘン」に修正することで、1形態素として扱うこととした。また、同様なことが活用する語に起こった場合、「な、って」となっている場合は、表記形は「な、って」で一語にし、基本形は「なる」に修正した。

この作業では形態素の分けられる位置を修正するので、それに付随して6.1で提示した、自動で付与されるタグのうち「表記形」から「モーラ数」の修正も兼ねる。以降のタグ編集作業の基礎となる。

作業は「変更」「結合」「分割」「削除」機能を使い分けて行う。基本的に各項目を直接入力して変更するが、入力した例文を自動で解析する確認ウィンドウがあり、修正後の各形態素の文法項目はその結果をクリックすることで簡単に指定することができる。

ここまでの工程は提示した順番通り行う必要があるが、ここまでが完了後は、以降の「正表記形」「プロソディ付与」「読み修正」「重なり修正」の作業が同時並行でできる。ただし、同時並行で以降の作業をする際でも、現実的にはこれまでの工程で完了しているはずの文字化のミス、形態素分け修正漏れ、時間のずれなどが必ず見つかることになる。これらは結局見付け次第後から修正する必要があり、1つのファイルでその箇所を修正したとしても、同時並行で作業している全ての同じファイルでデータを合わせる必要があり、そこから更なるミスにつながりやすい。編纂スケジュールの余裕などで、可能であれば1ファイルの作業工程は直列式に行う方が安全である。しかし、やはり同時並行で作業を行った方が効率は高いため、並列で行うとすれば、問題が出た時の効率的な解決法を決めておく必要がある。

6.4　正表記形

学習者コーパスの場合に発話に誤用が含まれる場合がある。その際に、形態素レベルでの誤用が有った場合、発話した内容をそのまま記録した「表記形」が、本来は存在しない語の形になることがある。その語の本来の形を「正表記形」に記録することで、本来の形で検索をかけた時にその語の誤用形を含めて全て検出することができるようになる。

他にも本コーパスでは1秒未満のポーズを「、」で記録し、それが１つの形態素の中で起こったとしても記述するというルールがあるため「全、部」などの形で１形態素で記述されることがある。これも「正表記形」に本来の「全部」として記録する。

本コーパスでは、データ対象とした学習者のレベルが高かったためか、形態素レベルの誤用はほぼ起こらず、この作業はそれほど時間のかかるものではなかった。ただ、後の工程である「誤用」「語彙レベル」「文型レベル」を自動で付与するために必要な情報なので、工程のできるだけ早い段階でやっておくと便利である。

6.5　プロソディ付与

発話内の各形態素の音声的高低強弱を、高い「↑」、低い「↓」、強い「！」で記録する。自動での付与はできないため、各行の音声を聞いて1つずつ判断して付与する作業になる。作業は簡易リストに「↑」「↓」「！」が既に指定してあり編集画面内の発話内容の該当する形態素をクリックするだけで付与できる。そのため作業方法は極めて簡単であるが、微妙な違いを聞き分ける作業が続くため、長時間行うと音の高低強弱の感覚が混乱してくる。長く連続して作業せず、適度に区切って行う方が効率を持続できる。

作業は、編集する行の音声を確認しながら行う必要があるため、時間付け修正の作業が完了している必要がある。

6.6　読み修正

発話された音声のそのままを「読み」のタグに記録している。実際には文字化の作業は漢字仮名交じりで行われるため、漢字で文字化された形態素が解析された際に、発話音声とは違う読みで解析されることがある。これの修正作業となる。

作業は直接「読み」のタグに書き込んで修正して行く作業になる。その際に「母音配列」と「モーラ数」も同時に変更する必要がある。

各行の音声を1つずつ確かめながら行うため、時間のかかる作業にはなるが、基本的に他の作業に比べ、高度な判断の伴いづらいものである。しかし、「私」が実際には「わたし」や「あたし」や「ったし」という風に違って発音されるようなことも非常に多いため、どの程度精密にすべきか、こだわりだすと切りが無くなる。どこまで踏み込むかを事前に決めておく必要がある。本コーパスでは、「発音」ではなく「読み」のため、漢字の読みとしての基礎的な形の記述に留めた。

作業は、編集する行の音声を確認しながら行う必要があるため、時間付け修正の作業が完了している必要がある。

6.7　発話の重なり修正

2発話者の音声が一時的に重なって現れた場合の記録をする。発話の重なりの記録方式は以下の通りとなる。

0:0-0 重なっていない（エンピツ）

1:0-0 全てが重なっている（[エンピツ]）

2:1-2 語の初めから数えて1文字目から2文字分重なっている([エン]ピツ)

2:2-3 語の初めから数えて2文字目から3文字分重なっている(エ[ンピツ])

これは「発話の重なり」のタグに記録するが、文字化の際に［ ］の記号で囲った箇所が重なりであるとしたので、専用エディタで一次整形する際に、その範囲内の形態素にはすべて「1:0-0」と自動で付くようになっている。本作業はこれの細かい修正になる。

ひらがな1文字の単位で音声の重なっている幅を付けていく。

作業は、編集する行の音声を確認しながら行う必要があるため、時間付け修正の作業が完了している必要がある。

6.8　誤用

日本語文法として誤用がある箇所にその種類を付ける。自動では不可能なため、完全に一からの判断、作業となる。詳しくは7章で記す。

6.9　各文型

発話内容で使われたそれぞれの形態素が、日本語教育でのどの文型に当たるものかの情報を付ける。自動で付与されたものを一つずつ修正する作業である。詳しくは 8 章で記す。

6.10　笑い

以降、立命会話コーパスのみのタグについて記す。

発話の中で笑っている箇所がある。この形態素に笑いのタグ「h」を付与した。

笑いは、完全に笑い声のみの音声で構成され、形態素を構成する音素にならずに発声されている箇所と、形態素が発声されているが同時に笑いが含まれる箇所がある。前者は表記形自体も「h」で記録するが、後者は表記形にはその発話内容の形態素を記録する。そのため、どこで笑っているかを記録するために本タグを用意した。

笑い声のみの場合も、笑いながらの場合も両方「笑い」タグへの記録をする。笑いは、音声化されたもの、笑顔の状態、声にならない笑い、笑い終わった後の声を吸っている状態など極めて複雑であるが、本コーパスでは基本的に音声データを元にしたコーパスであるため、音声として聞こえるもののみを対象とした。しかしこれは判断が極めて難しいため、笑いの開始と終了の境界を決めるのは難しく、今後のコーパス改良の課題である。

6.11　発話番号、発話権

2者の発話では、それぞれの箇所で基本的に会話参加者のどちらかに発話権がある。主に発話権の有る方が「話し手」、発話権の無い方が「聞き手」となる。音声を発していたとしても、あいづちは、相手の発話を受け入れたり促したりするために行われるものであるので、発話権の無い音声となる。

また、発話権のある発話ごとにまとめて順に付与した番号が「発話番号」である。この２項目は非常に密接にかかわるため、作業工程は違うが説明は本節で同時に行う。

本コーパスでは、全ての発話に対して、発話権の有り、無し、を付与した。付与の単位は形態素ごとになる。1行の発話であったとしても、あいづちから始まり、途中から発話を奪い取り発話権が発生する場合は、あいづちの位置までを発話権「無」、それ以降を発話権「有」とする。

まず、「発話権」の付与基準を示す。

当話者が主体的に発話を行っていて、発話を進めているものを発話権有りとする。短い内容で、発話権が微妙なものでは、直前の話し手である相手の発話が「YES」「NO」を求めるもので、それに対し「はい」「うん」などで「YES」か「NO」という返答をしたものは、積極的な意見の表明として発話権の有りと扱う。あいづちとの判断が微妙な場合、相手が更にそれに対して明確に反応したものは発話権有りとする。それが完全に重なりの中で行われていても発話権有りとする。

対し、「YES」「NO」などの意見が含まれずに相手の話の促しのみの働きのものは発話権無しとする。話し手が発話をしている間や直後に聞き手が送り、ターンを取ることを目的としない音声をあいづちとし、発話権無しとする。

フィラーの場合は音声のつなぎを行い、そこから更に自分の発話へつなげる元となる場合には発話権有りとする。

どのような場合でも、前発話と1秒以上空いていれば独立するものとして発話権有りとする。

次に「発話番号」の付与基準を示す。

行単位で付与する。

同一発話者による発話権の有りの発話が続いて、間に相手の発話権有りの発話が挟まらなければ同一発話番号とする。発話権が移った行から次の発話番号とする。発話番号は2者の間での通し番号とする。発話権の無い形態素のみの行の場合、その前の行である相手の発話の発話番号と同じ番号になる。複数行にまたがって同一発話番号になることもある。

相手の発話に重なって発話権有りの発話が起きた場合、次の発話番号とする。

発話番号が変わった最初の発話者がその発話のメインの発話者である。

あいづちで返され、それには発話権が無かったとしても、それに対し明確に返答をしたら次の発話から次番号とする。あいづちのみを挟んで前発話と後ろ発話に間が無ければ同じ発話番号とする。同じ状況でも前発話と後ろ発話に1秒以上間があれば別の発話番号とする。

時間的に1秒空いていなくても、聞き手の「YES」「NO」を表す発話が現れれば、それが重なりだったとしても、その次から別発話番号とする。全く重なりがなくても、発話者が話し終えて1秒以内にしたあいづち的な形式のものはあいづちと扱い、その発話権は前発話の話者とし、発話番号も前発話の番号をつける。

以上、発話権、発話番号の判断基準を示したが、具体的な作業時では、特に「感動詞」に注意する必要がある。「ああ」、「ええ」などの判断が微妙なものの場合、該当形態素の品詞が「感動詞、フィラー」の場合は、発話権を維持し、次の発話をするための音声的なつなぎであることが多く、「発話権 有」である。「感動詞、一般」となっている場合は、相手の発話へ対する反応であり、その際は相手に発話権があることを認めていると考えられることが多く、「発話権 無し」とするのが基本である。しかし、これも個別の条件によって判断すべきである。また、「感動詞 フィラー」と「感動詞 一般」の自動付与はこちらの意図通り行われないことが多いため、形態素修正の際にはほぼすべてを判断し直す必要があり、この作業は特に大変である。今回、立命会話コーパスでは「発話権」と「発話番号」の情報付与を行ったため、感動詞の下位分類の判断が重要になったが、立命学習者コーパスでは「発話権」「発話番号」の付与を行わなかったため、出現数の多さや個別の判断を要するために作業にかかる時間と効果の対比を考え、特に強く区別せずに自動付与の結果をほぼそのまま採用している。

6.12 非言語

自然発話の記録であるので、発話には様々な非言語情報が含まれる。本コーパスでは「視線」、「上体」、「頭」を記録した。

「視線」は、各形態素が発話されたときに視線がどの位置を向いているかの情報である。完全に「位置」の情報であるため、視線の移動や意味合いの情報は記録しない。各時点での位置情報であるため、全ての形態素に必ず付く。

「上体」は、各時点での上半身の情報である。「まっすぐ」、「腕組み」、「ほうづえをつ

く」、「もたれる」、「動く」、「立つ」になる。位置の情報と動きの情報の2つが混在する。このタグは全ての形態素に付く。

「頭」は、基本的に移動があった場合に付けるが、特殊な状態を維持している間も付与する。「うなずく」、「横ふり」、「傾げる」「うつむく」になる。うなずくの場合に、1形態素の中で複数回うなずくこともあるため、「うなずく：2回」や「うなずく：3回」などと、回数も記録した。複数形態素にまたがって1度うなずく場合は、そのまたがる全形態素にそれぞれ「うなずく」と付けた。「頭」の情報は特殊な状態の記録となるため、全形態素には付与していない。

6.13 その他

他に、「手」の項目を用意したが、これは非常に判断が難しいため長い作業の結果断念した。手は、左右が別々に動き、場合によってそれぞれが意味を持つ場合、まったく意味を持たない場合、2つ合わせて一つの意味を持つ場合など複雑である。手をただの位置として記せば、「机の上」、「頭」などの「その形態素を発している最中にある手の位置」を記せるが、動きを記すことができなく、またそれによってその動きがジェスチャーである際の「意味」を記録できなくなる。「円を描く」、「指し示す」などのジェスチャーが発話中に大量に発生するが、手の動きが意味をもったジェスチャーである場合と、ただの手の移動である場合、またその境界があいまいな場合など、様々な要因が複雑に混ざって出現する。左右が一体となって1つの意味を示す場合、左だけが意味を持ち、右はただの動作の場合などがあり、記録方法、ルール化、判断が非常に難しいため、この要因は本コーパスでは断念した。何度もデータを見ることで記述ルールを作成し、今後のコーパスで記録可能とすることが今後の課題である。

7. 誤用

7.1 誤用判定の基準

誤用判定の基準は、一般の native speaker を想定し、日本語として許容できるかどうかで判断した。

この点、誤りの判定は、一般の人に比べ日本語教師の方が厳しい。「一般の人は，日本語として許容可能かどうかという基準を自分の語感」や直観によって判定するのに対し、専門家は、「文法に照らし合わせて」判定する。しかし、言語の機能を、「社会の中でのコミュニケーションとしてとらえる」ならば、「一般人の言語的直観によって判定される能力こそ，communicative competence を反映している」と言えるだろう（遠藤・大井，1992）。

そこで、日本語教師が日本語教育における文法に照らし合わせ、誤用と判断するものでも、一般の native speaker が直観的に許容できると考えられるものは、誤用無しとすることにした。

確かに、一般の native speaker という基準は曖昧であり、許容範囲も人によって異なる。タグ付け作業前の段階では、一般の native speaker の直観で判断する場合と、日本語教師が文法に照らし合わせて判断する場合に分け、誤用のレベルを二つにするという案もあった。しかし、日

本語教師の基準といえども、考え方によって判定に幅があると思われ、両者の基準の区分は難しく、却って作業が煩雑となる。また、実際に母語話者が話す言葉は、すべてがいわゆる文法ルールに則しているわけではなく、文法も絶対的な規範とはいえない。そこで、今回は、作文ではなく会話コーパスであることから、コミュニケーションを重視し、一般の native speaker の基準のみで判断した。

7.2 タグの種類

誤用のタグには、誤用無しという意味の「-」と、誤用有りと判断した場合の「脱落」「付加」「誤形成」「混同」「位置」「その他」「発音」「活用」がある。誤用の分類は、市川（2010）に準拠し、それに加え「発音」と「活用」の二つを新設した。「活用」のタグを新設した理由は、「誤形成」に分類される誤用には、用言の活用の間違いが多く含まれるためである。

さらに、「脱落」と「混同」の場合、他の誤用と異なり、何が脱落したか、あるいは何と混同したかを示す必要がある。そこで、脱落したと思われる語句、混同したと思われる語句を、（ ）内に表示した。例えば、「ビタミン C をふく、含める」（会話 05 第 2 回 148 行）の場合、「含める」に対して「混同（含む）」のタグを付与した。

なお、「脱落」のタグは、脱落している箇所の前の形態素に付与した。例えば、「私も一回生にはそうだった」（会話 02 第 5 回 60 行）という発話で、「一回生」と「に」の間の「の時」が脱落している場合、「生」に対し「脱落（の時）」というタグ付けを行った。

誤用タグの中には「脱落」「付加」「誤形成」等の誤用の性質を表す部分に加え、「付属」という誤用の重要度を表す部分がある。これは、誤用が複数形態素にまたがり、「誤用数＝形態素数」でない場合、表現の核となる形態素を「本体」、それ以外を「付属」とし、「付属」の形態素に付与したものである。そもそも、一形態素の誤用の場合、「誤用数＝形態素数」となる。しかし、複数形態素にまたがる場合、検索の際その複数形態素すべてを別々に抽出するため、誤用の数と一致しない。そこで、誤用の重要度を分けることにより、誤用の数を特定できるようにした。すなわち、誤用の「本体」で検索すると、「付属」と付けられた形態素以外の形態素が抽出され、誤用の数を特定できる。

例えば、「入試試験は」（会話 03 第 1 回 7 行）という場合、「入試」が入学試験の略語のため、「試験」に「付加」のタグが付き、誤用数と誤用の形態素数は等しくなる。他方、「でもここで、関西の友達の多いですか」（会話 05 第 1 回 87 行）の場合、前の話の逆接ではないため、「で」と「も」に「付加」のタグが付与される。その結果、誤用を検索すると、二つの形態素が抽出される。しかし、誤用の数としては、「でも」で一つの誤用と捉えるべきであるから、表現の核となる「で」を「本体」とし、「も」に対し「付加＜付属＞」というタグを付けた。これにより、誤用を形態素単位ですべて抽出できるとともに、誤用の「本体」で検索すると、誤用の数がわかるようになった。また、「お姉さんと一緒に出か、けるときは、なんかお姉さん、らしくない、とよく言われている、なんか、俺の、ことのほうが、なんかお兄さんと」（会話 03 第 2 回 256 行）という場合、「こと」と「の」に「付加」のタグが付与される。その結果、

誤用を検索すると、二つの形態素が抽出される。しかし、「こと」を付加しなければ、「の」の付加という誤用も起きなかっただろうから、誤用の数としては全体で一つの誤用と捉えるべきと思われる。そこで、核となる形態素の「こと」に付属する「の」に対し「付加＜付属＞」というタグを付与した。

7.3　タグ付けの具体例

タグ付けの判断では、できるだけ学習者の発話意図を汲むようにし、訂正する場合は、原文を最大限残すよう心掛けた。

誤りの判断レベルは、形態素・統語レベルの誤用に止め、学習者の発話意図を確認できないことから、文脈レベルの判断には立ち入らないこととした。また、語用論レベルの判断も、待遇表現に関する判断も行っていない。

全体として、学習者の日本語レベルが総じて高く、誤用が少なかった。それに加え、誤用の場合も、発話意図を比較的推測しやすかったと言える。しかし、中には判断に困った場合もあり、そのような例は以下の三つに分けられる。

①発話が聞き取りづらく、正しいかどうか分からないもの
②誤りであることは確かだが、何と言いたかったのか分からないもの
③文として正しいが、発話意図の推測次第で誤りの可能性があるもの

①の例には、「留学生みんな金曜日になるとデンカーデス friday」（会話 01 第 2 回 27 行）や、「大学シテイは大学チョウ」（会話 03 第 1 回 121 行）がある。これらは、文字化の作業者や文字化確認の作業者、そして筆者も聞き取りづらかったものである。前者は、カタカナ部分が英語として正しい可能性もあるが、日本語としては正しくない。しかし、本当は何と言いたかったのかが分からないため、「脱落」「付加」「誤形成」等のタグは付与できない。また、後者は、各音素の発生に誤りがあった場合ではないので、「発音」のタグを付けることはできない。このような場合、「その他」のタグを付与した。

②の例には、最近 Facebook を始めたという母語話者に対する、以下のような学習者の発話がある。なお、例文中の「<03>」、「<04>」は、話者番号を表す。

会話 02 第 4 回（115 行～117 行）
<03>
あ、教えて
<04>
あ、いいですよ、最近[いつで]

<03>

[あでも、]私の方が安いかも

この場合、「安い」は、明らかに誤りである。おそらく何かと混同したと考えられるが、真意がわからなかったので、「その他」に分類した。

また、日本語能力試験に関する話題での、以下のような学習者の発話がある。なお、例文中の「<07>」、「<08>」は、話者番号を表す。

会話 04 第 2 回（121 行・122 行）
<08>
どうゆう風に勉強してるんですか、そういう教材みたいなの使ってー
<07>
はい教材ー、韓国からもらった教材でー

上記の例も、下線部が誤りであることは確かである。しかし、学習者が本当は何と言いたかったのか分からない。そこで、「その他」に分類した。

このように、誤りであることは確かだが、何と言いたかったのか分からないものには、「その他」のタグを付与した。

③には、学習者も母語話者も朝に弱いという話から、一限の授業について「ひとつも取れなかったです」（会話 05 第 3 回 118 行）と言った例がある。この場合、一限の授業の受講登録をしたが、朝が弱くて単位を落としたという意味であれば、誤りではない。しかし、朝が弱いので、そもそも受講を諦めたというのであれば、「取らなかった」が正しい。このように文として間違っていないが、発話意図によっては誤りの可能性があるものは、「-」のタグを付けた。

以上のように、判断に困った場合の多くは、学習者の発話意図を確認できないことが理由であった。

他には、エディタの仕様に起因する場合がある。例えば、「働き忙しくって結婚しなかった」（会話 04 第 3 回 125 行）という場合、学習者の発話意図を汲み、原文を最大限残すなら、「働くのが忙しくって結婚しなかった」と訂正すべきであろう。しかし、一つの形態素に対し、二つの誤用のタグを付けられない、すなわち、「働き」という形態素に対し、「働く」の「活用」の誤りを示すタグと、「のが」の「脱落」を示すタグを重ねて付けられないため、「働き」を「仕事が」との「混同」であるとした。同様に、「いつから始めてわからないです」（会話 05 第 1 回 158 行）の場合、訂正文は「いつから始めたかわからないです」となる。分析すると、助詞「て」と助動詞「た」の「混同」、「て」の後の助詞「か」の「脱落」が起きたと考えられる。しかし、「て」という形態素に対し、二つの誤用のタグを付けられないため、「て」には「脱落（か）」のタグを付け、「始める」にテ形ではなくタ形という意味で「活用」のタグを付けた。

7.4　タグ付け作業からみえた課題

筆者は、誤用のタグ付け作業を通して、次の二点を今後の課題と考えた。

一つは、誤用判定の基準であり、これは未だ迷うところである。日本語教師として、学習者の誤用の原因を探り、どのように指導するかを考える場合、文法に照らし合わせて間違っているものをすべて抜き出す方が、資料を豊富に得ることができる。一方、実際の会話では、許容範囲であれば多少の文法の間違いは問題にならない。個々の表現が誤りかどうかということと、その誤りが許されるかどうかは次元を異にし、どちらが正しいという問題ではないとすると、やはり二つの基準を併存した方が、幅広い研究の資料となり得るだろう。また、今回は一般の native speaker を基準としたが、そうであるならば、実際に日本語教育とは関係ない一般の人を何人か集め、日本語として許容できるかどうか判断してもらう方が、合理的と思われる。この場合、許容できない理由について、日本語教師が文法面から理論づけるとすると、誤りかどうかということと、その誤りが許されるかどうかということが、比較的容易に併存できるのではないだろうか。

もう一つは、学習者の発話意図をどのように確認するかである。作文の場合、同じ内容を母語で書いてもらったり、フォローアップインタビューをしたりするなどの方法で、意図を尋ねることができる。しかし、作文に比べて会話は、意図を尋ねるのがより困難と言える。誤用の要因を明らかにするためには、学習者が本当は何と言いたかったのか、確認できる方法を考える必要があると思われる。

8.　文型レベル

8.1　作業手順

本節では、文型レベルのタグ付けに関し、自動でタグを付与する仕組みと、その結果を確認・修正する作業手順を説明する。

はじめに、正表記形までの作業が終わった編集ファイルを、「日本語教育語彙、文型レベル編集ウィンドウ」で指定する。すると、友松・他（2010）の新日本語能力試験各レベルの文型リストを基に、各文型で使用される形態素に自動でタグが付与される（田中，2011）。その結果、発話内容に各レベルの文型を表す色(1)が付けられる（図 4）。例えば、図 4 の「そうかー私も留学してみたい・」の場合、「て」と「み」それぞれに N4 レベルの文型「てみる」のタグが付与され、N4 レベルの文型を表す黄緑色が表示される。また、「たい」には、N5 レベルの文型「たい」のタグが付与され、N5 レベルの文型を表す水色が表示される。

図 4. 日本語教育語彙、文型レベル編集ウィンドウ

このようなタグが付与される形態素の範囲は、友松・他（2010）に記載された文型の範囲であり、基本的に動詞や「です・ます」等コピュラの部分は含まない。具体的には、「メジャーリーグ、見たことありますか」（会話 01 第 1 回 4 行）の場合、N4 レベルの文型「たことがある＜経験＞」が使われているので、動詞の「見」とコピュラの「ます」の部分を除く「た」「こと」「あり」に対し、それぞれ N4

レベルの文型「たことがある＜経験＞」のタグが付与される。ただし、N2 と N4 レベルの文型「のだ」や、N4 レベルの文型「のですか」のように文型にコピュラが含まれる場合、コピュラに対してもタグが付与される。例えば、「僕海外いっぱい行ってみたいんですけども」（会話 01 第 1 回 10 行）の場合、「ん」「です」に対しそれぞれ N4 レベルの文型「のだ＜説明＞」のタグが付与される。また、タグ付けの範囲に動詞は含まないのが基本だが、例外として、N4 レベルの文型「よう＜意志＞」や「ようとおもう」の場合が挙げられる。すなわち、「ゆっくりしようかなーと思ってるんです」（会話 05 第 3 回 157 行）の場合、友松・他（2010）に記載された文型の範囲で、動詞部分は含まないとすると、「よう」の部分にのみタグが付けられるべきである。しかし、タグは形態素単位で付与され、この場合、「しよう」で一形態素となるため、動詞の部分も含むことになる。

続いて、自動でタグ付けされた結果について、人手による確認・修正を行う。手順は以下の通りである。

図 4 の発話内容の下、中央にある N1、N2、N3、といったボタンをクリックすると、そのレベルを示す色だけが発話内容に表示される。つまり、図 4 で N2 のボタンをクリックすると、発話内容の「じゃない」と「か」の部分にピンク色だけが表示され、他のレベルの色は消える。同様に、N3 のボタンをクリックすると、「と言うより」、「ぐらい」、「とか」の部分のオレンジ色だけが残る。他のレベルも同様である。

一方、N1、N2、N3 のボタンをクリックした時、ボタンの下の文型リストには、使用されていると判断された文型に対し、文型レベルの色が表示される。さらに、使用されている可能性のある文型に対し、赤色の表示がされる（図 5）。そこで、使用されていると判断され、そのレベルの色が表示されたものと、使用されている可能性のあるため赤色の表示がされた文型について、個別に確認を行う。言い換えると、例えば、N2 レベルの場合、文型リスト中ピンク色の文型と赤色の文型について、正しいかどうか確認する。

ウィンドウ下部の文型リストにある個々の文型をクリックすると、発話内容での該当箇所が色付けされる。図 5 は、N2 レベルの文型リストにある「ではないか＜判断＞」をクリックした場合である。N2 レベルの文型「ではないか＜判断＞」の場合、友松・他（2010）に基づき、「ではないか」、「ではありませんか」、「じゃありませんか」、「じゃないか」の形と一致する発話内容は、N2 レベルの文型「ではないか＜判断＞」とされ、ピンク色で表示される。それに加え、「ではないか」、「ではありませんか」、「じゃありませんか」、「じゃないか」の形と完全に一致するものではないが、一部の形態素を含む発話内容は、赤色で表示される。これが、図 5 の発話内容の「じゃないか」と「じゃない」に、それぞれピンク色と赤色が付されていることの意味である。そこで、前者が N2 レベルの文型「ではないか＜判断＞」であるという判断が正しいか、後者が N2 レベルの文型「ではないか＜判断＞」かどうかを確認し、必要があれば修正する。この場合、修正の必要はないが、仮に、自動でタグ付けされた結果が間違っていて、それを修正する場合には、ピンク色の箇所にカーソルを合わせ、Ctrl キーを押しながらクリックすると、赤色に変えることができる。反対に、赤色の箇所にカーソルを合わせ、Ctrl キーを押しながらクリックすると、ピンク色に修正できる。すなわち、仮に、図 5 で、N2 レベルの文型を示すピンク色で表示された「じゃないか」が、「ではないか＜判断＞」の文型に当たらないと判断した場合、「じゃないか」にカーソルを合わせ、Ctrl キーを押しながらクリックすると、ピンク色が赤色に変わり、一部の形態素を含むが N2 レベルの文型でないことを表示できる。反対に、赤色で表示された「じゃない」が、「ではないか＜判断＞」の文型に当たると判断した場合、「じゃない」にカーソルを合わせ、Ctrl キーを押しながらクリックすると、赤色がピンク色に変わり、N2 レベルの文型であることを表示できる。

このように、全ファイルの全レベルで、そのレベルの色が表示された文型と赤色の表示がなされた文型について確認し、必要があれば修正を行う。

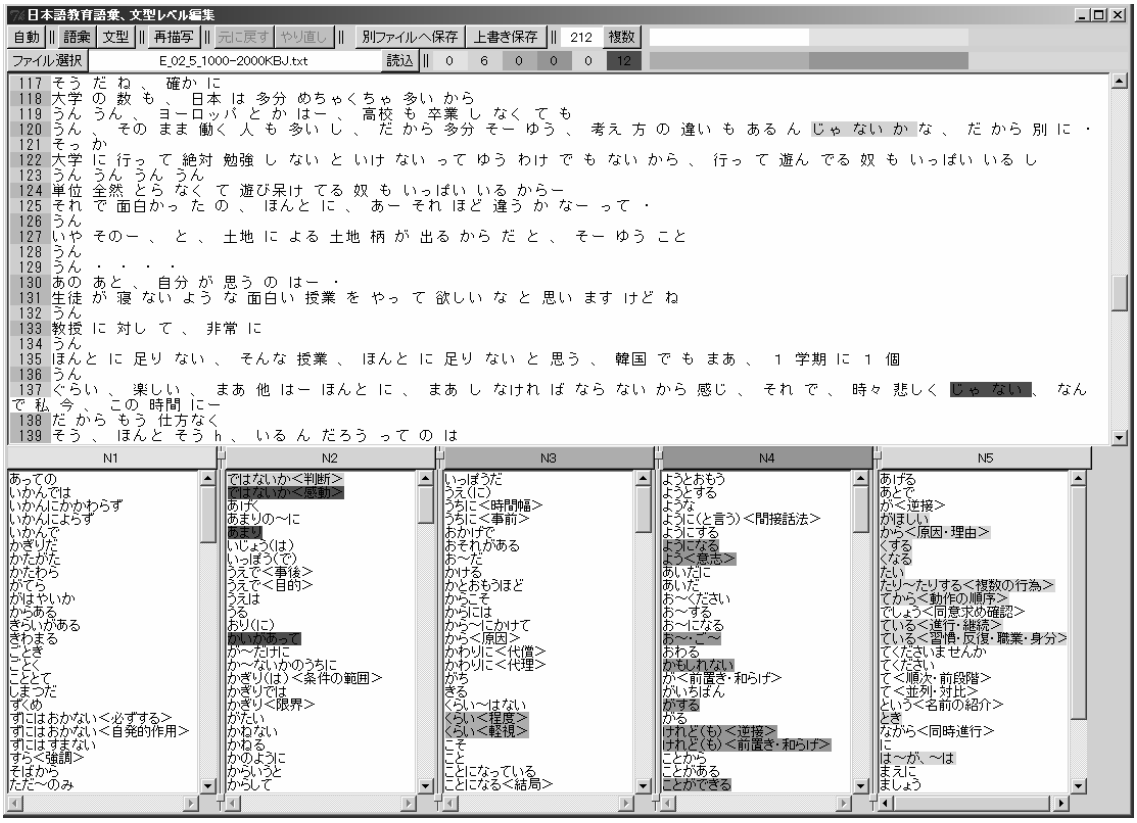

図 5. N2 レベルの編集ウィンドウ

8.2 確認作業の具体例

自動でタグ付けされた結果を確認・修正する際も、友松・他（2010）を基準とする。個々の文型について、発話内容を同書の例文や解説に照らし合わせ、タグ付けされた結果が正しいかどうか確認した。同書は、豊富な例文によって文型の使い方が示され、くだけた話し言葉の例もあるので、当該文型かどうかの判断がし易かった。また、各文型の見出し語への接続の形が示されている点も、判断に役立った。ただし、実際の会話には、同書にない用法や接続のものもあり、判断に困るものもいくつかあった。そのような例や特徴について、レベルごとに以下に述べる。

N1 レベルは、自動でタグが付与された結果、N1 レベルの文型として該当する形態素にタグが付与され、このレベルの文型を示す黄色が表示された箇所はなかった。よって、確認作業をしたのは、N1 レベルの文型とする条件と完全には一致しないが、一部の形態素を含むため、このレベルの文型の可能性があるとして赤く色付けされた箇所のみだった。その確認が必要な文型の数も、発話内容において赤色で表示される形態素数も、他のレベルに比べて少なかった。確認作業の結果、全ファイルにおいて N1 レベルの文型は認められなかった。

N2 レベルでは、準拠する文型辞典にない使われ方をした文型に、「ではないか」がある。例えば、「あれ食堂にあるじゃないですか」（会話 01 第 1 回 313 行）や、「短期、留学あるじゃないですか」（会話 04 第 1 回 146 行）のように、確認の意味で使っている場合である。これらは、

形としては N2 レベルの文型「ではないか」に当たるが、話者の判断や感動を表すものではないので、N2 レベルの文型には当たらない。他方、発話意図からすると、N5 レベルの文型「でしょう＜同意求め　確認＞」に当たると思われるが、この文型で使用される形態素に一致しないため、N5 レベルの文型にも当たらない。その結果、どちらの文型にも当てはまらないものとした。これは、日本語母語話者の発話に多く見られた。

その他、確認に音声が関係するものとして、「そうなんですか」（会話 05 第 3 回 53 行）がある。これは、形として N4 レベルの「のですか」の文型だが、話者が納得した場合にも使われ、その場合 N2 レベルの「のだ＜納得＞」の文型になる。そのため、会話を聞き「んです」の部分が、N2 レベルの文型に当たると判断した。

N3 レベルでは、準拠する文型辞典にない接続の例に、「くらい」が挙げられる。友松・他（2010）は、「くらい」を程度を表す場合と、軽視の意味で使う場合の二つに分け、「くらい＜程度＞」の場合、普通形（主にイ形容詞と動詞の現在形）に接続するとし、それに沿った例文が挙げられている。しかし、実際には「めっちゃ顔くらい大きい」（会話 03 第 1 回 282 行）や、「かかる料金もそのぐらい」（会話 02 第 3 回 189 行）など、名詞・代名詞や連体詞に接続するものが多かった。他にも、「朝、6 時ぐらいに」（会話 05 第 3 回 114 行）や、「4 泊で、7 千円ぐらいです」（会話 01 第 5 回 42 行）など、「頃」や「約」で言い換えられるものも多く見られた。前者は「程度」ではないが、後者はある程度の幅を持つため判断が難しく、N3 レベルの文型「くらい＜程度＞」のタグが付けられた形態素と、そうでないものとが混在する。

N4 レベルは、確認が必要な文型の数も、発話内容において赤色で表示される形態素数も、他のレベルに比べて多く、確認に時間がかかった。このレベルの文型で、判断に困った例としては、「て＜緩い連結＞」が挙げられる。そもそも、友松・他（2010）は、「イ形容詞、ナ形容詞、名詞の『て』の形、『で』の形によって、緩やかにつながれた文である」と解説している。しかし、接続助詞の「て」が前後をつなぐのは当然で、「緩い連結」という説明自体曖昧と考えられる。そして、同書の例文も、「昨夜は暑くて、寝られなかった」は、「意味の弱い原因」であり「て＜緩い連結＞」とする一方、「て＜理由・原因＞」の例文に、「母のことが心配で眠れなかった」とあり、両者の区別が難しい。これは、発話内容が N4 レベルの文型に当たるかどうかの判断に困った例ではなく、基礎となる判断基準の理解に困った例と言える。

また、実際の会話では、「入学式、七千人来て」（会話 01 第 3 回 37 行）や「出席はちゃんとしてー」（会話 02 第 5 回 100 行）のように、動詞の「て形」で終わる場合がよく見られた。この点、準拠する文型辞典は、動詞に接続する文型として「て＜並列・対比＞」、「て＜順次・前段階＞」、「て＜方法・状態＞」、「て＜理由・原因＞」を挙げている。しかし、「て形」の後ろに文がない場合、上記のどれに当てはまるか判断が難しい。このようなものについて、消去法で「て＜緩い連結＞」に振り分けているところがある。

他には、同じレベルの文型が重なる場合が問題となった。例えば、「何言おうと思ったんだっけ、」（会話 03 第 3 回 214 行）の場合、N4 レベルの文型「ようとおもう」に当たるが、同じレベルの文型「よう＜意志＞」と「と＜間接話法＞」を含んでいる。この場合、本来なら「言

おう」に N4 レベルの文型「よう＜意志＞」と「ようとおもう」のタグを、「と」に「と＜間接話法＞」と「ようとおもう」のタグを、「思っ」に「ようとおもう」のタグを付けるべきである。しかし、エディタの仕様上、一形態素に対し同じ文型レベルのタグを付与できない。そこで、小さい文型（「よう＜意志＞」と「と＜間接話法＞」）を含む大きい文型（「ようとおもう」）のタグを付与した。すなわち、前述の「何言おうと思ったんだっけ、」の場合、「言おう」「と」「思っ」に対し、それぞれ N4 レベルの文型「ようとおもう」のタグを付与した。なお、一形態素に対し異なる文型レベルのタグを付けることはできる。例えば、「濃く、にしようと思いましたが、」（会話 04 第 5 回 121 行）の場合、「しよう」に対し、N5 レベルの文型「くする」のタグと、N4 レベルの文型「ようとおもう」のタグを付けることはできる。

N5 レベルでは、「は～が、～は」に関し、一行の発話で使用されていないものを、どう判断するか基準を設ける必要があった。つまり、「味噌は大丈夫だけど、納豆はほんとに、遠慮」（会話 01 第 5 回 104 行）の場合、一行の発話で使用されているので、「は～が、～は」の文型であると容易に判断できる。しかし、例えば以下のように、複数行にまたがる表示の場合、一まとめに「は～が、～は」の文型とできるか問題となる。なお、例文中の「<03>」、「<04>」、「<07>」、「<08>」は、話者番号を表す。

会話 02 第 1 回（91 行～95 行）
<03>
プザンは、韓国と、あいさつ、おはようとか、これ、韓国に語[にしたら]同じですけど、うん
<04>
[あ、は]
<03>
ヴェジュウ島は
<04>
ん
<03>
ひどい

会話 04 第 3 回（100 行～102 行）
<07>
多分こっちには、この曲にはピアノは弾かないと思うんですけど
<08>
うんうん
<07>
他の曲、によっては、なんかそれで、弾く、ん、あります

このように発話の表示が近接し、相手の発話が相槌である場合、ほぼ一つの内容と捉えられるので、一まとめに「は～が、～は」の文型と判断した。

8.3 確認作業からみえた課題

筆者は、この確認作業に参加して、次の二点を今後の課題と考えた。

一つは、緻密な文法ルールと用例の蓄積である。準拠する文型辞典に記載のない表現について、どのように判断するか、また、一度判断したものがぶれないためにも、緻密な文法ルールと用例の蓄積が必要と思われる。

もう一つは、多人数での作業において、情報をどのように共有するかである。一人の強力なリーダーが全てを把握し指示を出す場合を除き、関係者の情報共有が不可欠となるのは、この作業に限ったことではない。今回の作業では、自動でタグが付与されたものを、複数人で確認・修正したが、各ファイル・各レベルの確認回数は、2 回とやや心許ない。さらに、1 回目の修正に関する申し送りや、1 回目と 2 回目で判断が異なった時の話し合いもなかったことは、作業に携わった者として反省すべき点である。限られた時間で多くのファイルを確認しなければならず、それぞれ忙しくて時間が合わないといった理由もあるが、判断に迷ったものや人によって作業結果の異なるものについては、作業メモを残し関係者が見られるようにするなどすれば、より精度を高めることができるのではないだろうか。

9. 作業管理全体

作業の割り振りは1週間ごとに行い、全ての作業を週の初めに割り振り、締め切りを週末とした。週末に全スタッフに翌週の作業可能時間を提示してもらい、それに合わせた作業分量を割り振った。翌週になり、担当の作業が完了したスタッフには順に提出をしてもらい、早いうちに割り振られた作業が完了したスタッフには引き続き別の作業を割り振る方式を取った。このように、できるだけ多くの仕事を抱えさせないようにし、各スタッフが少量の仕事を一つずつこなし、達成感を維持しながら作業を進められることを心がけた。

コーパス作成には様々な工程があるが、作業者によって、作業の向き不向きや好みがある。今回はできる限りそれぞれの好みを優先した。これは、作業効率と作業からの撤退を考えたためで、とにかくはまず完成させることが大前提であるためである。それぞれが研究、授業、サークル、バイトなど個別のスケジュールを抱えている学部生、院生に編纂作業を依存する環境では、作業者の負担感や作業への参加意欲などが作業進行に大きく関わってくる。苦手意識の有る作業や望みの作業でない場合、作業速度や量に大きく差が出るだけでなく、作業からの撤退も考えられる。もちろんスケジュールが変わり、各自の予定が付かなくなるなど様々な要因があり把握は難しいが、各スタッフの参加意欲の維持は総じて重要な要因である。また、作業からの撤退の際はほとんど例外なく突然連絡が付かなくなる。事前連絡があることはごくまれであった。通常でも、多くのスタッフが 2～3 週間程度は忙しく作業に参加できないことはままある。したがって、離脱したスタッフの見分けが付きづらく、現状での参加スタッフの全数

を把握することが極めて困難である。この点は連絡方法、状況を徹底しておくべきであった。

いずれにせよ、作業進行の見通しは低めに見積もり、作成予定のデータを増やしすぎたり内容を詰めすぎたりせず、リスクを分散するなどの対策を取っておいた方がいい。

10. まとめ

以上、本稿で提示したようにコーパス編纂には実に様々な要因が関わってくる。データの構造をどうするか、何を記録するか、どのようなデータを記録するかといったコーパス設計から、各項目を実際に作業するに当たって、細かい基準をどうするべきか。更にどうチームを運営して作業を進めるか。すべてを踏まえて作成に臨む必要がある。

編集内容では、言語を扱う以上、あらゆる項目で完全なルールに作り、付与するタグの内実を精密に決定をすることはできない。つまり、判断が微妙になり、完全にルールに合致するタグが付けられない要素も多く出てくる。しかし、コーパスというものは全ての形態素に対し必ず何らかのタグを決めて付ける必要がある。この矛盾する状況がコーパス編纂を難しくする。完璧を求めると必ず抜けられなくなってしまう。そこで、本コーパスでは、できるだけ作業を先へ進め、全体の完成をさせることを優先して編纂を行った。様々な実験的試みの含まれたコーパスであるので、何年も寝かして完璧を求めて大幅に完成が遅れるよりも、まずはこの仕組みを作ることによってデータがどう動くか、どれほどの効果を発揮するのかを試すことが大切である。実際に完成品を動かしてみること、そして世の中の目にさらすことの方が大事であると考えられる。編纂されたコーパスがある程度課題の残るものであったとしても、早い段階でそれを提示し、作成や使用のノウハウを蓄積し、様々な意見を受けルールを改良し、２度目、３度目のコーパスをより良く作ることで、より早く質の高いものにたどり着けると考えられる。

本コーパスではこのような基本方針で作業を行った。本稿に触れた各研究者にも、それぞれ独自の視点でデータを構築したコーパスを編纂してもらいたいと考える。必ずしも本コーパスで提示したタグを採用する必要はない。今回は、あくまでも筆者やそのチームにとって必要性のある情報を記録したに過ぎず、実際に言語が使用される場面ではまだまだ様々な情報が溢れている。本コーパスによって、どのような情報でも記録し、利用することができることを提示したつもりである。新しい視点の様々なアイディア溢れるコーパスが大量にできることで、言語研究が益々発展していくはずである。本コーパスや本稿がその一助となることを真に望む。

注

(1) 文型レベル N1 は黄色、文型レベル N2 はピンク色、文型レベル N3 はオレンジ色、文型レベル N4 は黄緑色、文型レベル N5 は水色とする。

参考文献

市川保子(2010)『日本語誤用辞典―外国人学習者の誤用から学ぶ日本語の意味用法と指導のポ

イント』スリーエーネットワーク.
遠藤裕子・大井恭子（1992）「誤答分析の最近の傾向と母国語話者による外国人の作文の誤用判定」『東洋女子短期大学紀要』24、1-16.
岡本牧子・氏原傭子・山本修 著、真田信治監修(1998)『聞いておぼえる関西(大阪)弁入門』アルク.
佐々木仁子・松本紀子(2010) 『「日本語能力試験」対策 日本語総まとめN1 語彙』 アスク出版.
佐々木仁子・松本紀子(2010) 『「日本語能力試験」対策 日本語総まとめN2 語彙』 アスク出版.
佐々木仁子・松本紀子(2010) 『「日本語能力試験」対策 日本語総まとめN3 語彙』 アスク出版.
田中良（2010)「コンコーダンサー「HASHI」による日本語検索と処理」『社会言語科学会第25回大会発表論文集』. 198-201.
田中良・波多江優子・加藤理恵(2010)「立命館日本語会話コーパス :話者, 文法, 音声, 非言語情報付きコーパス –使用一例としての自然会話に現れた終助詞・視線・性別による分析-」『社会言語科学会第26回大会発表論文集』. 190-193.
田中良（2011)「『多種情報記述による再現性の高い自然会話コーパス構築システム』とその実装としての『立命館日本語学習者会話コーパス』」『Studies in Language Science 言語科学研究』第 1 巻、147-176.
友松悦子・宮本淳・和栗雅子（2010）『新装版 どんなときどう使う 日本語表現文型辞典』アルク.
山崎由紀子(2010)『「日本語能力試験」対策 にほんごチャレンジ N4 ことば 』 アスク出版.
Audacity ver 1.3.14 http://audacity.sourceforge.net/?lang=ja
伝康晴・山田篤・小椋秀樹・小磯花絵・小木曽智信(2009)UniDic version 1.3.12. http://www.tokuteicorpus.jp/dist/
奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座(2007)形態素解析システ『茶筌』 version 2.4.0. http://chasen.naist.jp/hiki/ChaSen/