「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」

田中　良*

要旨

本稿では、新たなコーパスの可能性を提示するために設計した「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、それを元に今回作成した、「立命館日本語学習者会話コーパス」を紹介する。日本語学習者と日本語母語話者 2 人 1 組の会話を、5 組、5 回記録し、縦断的データのコーパスとした。発話内容を文字化し、全てを形態素単位に分け、発話に含まれる様々な情報を付与した。情報は日本語教育で有用なものに特化した。語形、文法情報、音声情報、発話の重なり、誤用、使われた文型の日本語教育でのレベル、発話者の属性、などの情報を付与することにより、これらの情報を任意に組み合わせて検索できるようになり、コーパス使用の自由度が格段に高まった。

* 立命館大学　言語教育情報研究科　研修生

1．システム概要

本稿では、新たなコーパスの可能性を提示するために設計した「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、それを元に今回作成した「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」）を紹介する。「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」とは、発話内に存在する語の品詞や、音声の高低、発話の重なり、など多種類の情報をコーパスに記録することで、使用時に、発話時の状況や含まれる情報を再現して見られるコーパスの作成システムである。

このシステムを元に作成した1つ目のコーパスは、「立命館日本語会話コーパス(田中 波多江 加藤 2010)」（以下、「立命会話コーパス」）である。今回の「立命学習者コーパス」はそのシステムを実装した2つ目のコーパスになる。立命会話コーパスは談話分析を目的とし、非言語や発話権など、それに特化したタグを付与した母語話者会話コーパスであるが、本コーパスは日本語学習者の発話を記録し、日本語教育に役立つタグを付与した学習者会話コーパスである。

本システムにおける用語の定義は以下の通りとする。

情報　発話中に含まれる語の基本形や品詞や発話の重なりなどの様々な事柄
項目　情報を具体的に分けたタグの種類
　　「基本形」「品詞」「読み」「プロソディ」「性別」「母語」など
要素　項目の内容
　　「話す」「動詞」「オオキイ」「↑」「男」「日本語」など

現在、コーパスを使った研究手法がかなり進んでおり、高度な数学的手法を用い、目を見張るような多くの処理や分析ができる。しかし、それとは別の方向に目を向ければ今までとは違った更に多くのことができるはずである。

本システムでは、現在のコーパス利用の主流と言える統計による分析を行うことを主眼とせず、データ全体や個別のポイントを自在に見比べ分析できるコーパス、を設計思想とした。それを実現するために、実際の発話に含まれる膨大な情報を可能な限り多く扱えるようにし、それによって使用者が興味を持つことを自由自在に探し出せるようになった。

今回作成したコーパスで、多くは今までコーパスにあまり縁が無かったような、現場で教えている日本語教師、日本語学習者、談話分析や会話分析を研究する研究者など、より多くの使用者によって幅広い用途におけるコーパスの利用が可能となるはずである。

ただ、本システムを用いる上で注意しなければならない点として、データ量を多くできないということがある。これにより、データの種類も限定され、使用上得られた結果の一般性に留意する必要がある。しかし、これを踏まえた上で使用すれば大いに活用できるコーパスになり、コーパスの新しい利用価値を提示出来るものである。

2. 画面の表示例と検索例

本章では、2.1 節で、コーパス付属ツールで行える処理、2.2 節で、会話内容の表示の形式、2.3 節で、検索項目一覧、2.4 節で KWIC という処理による検索結果の表示、2.5 節で検索例から得られる学習者の傾向を提示する。

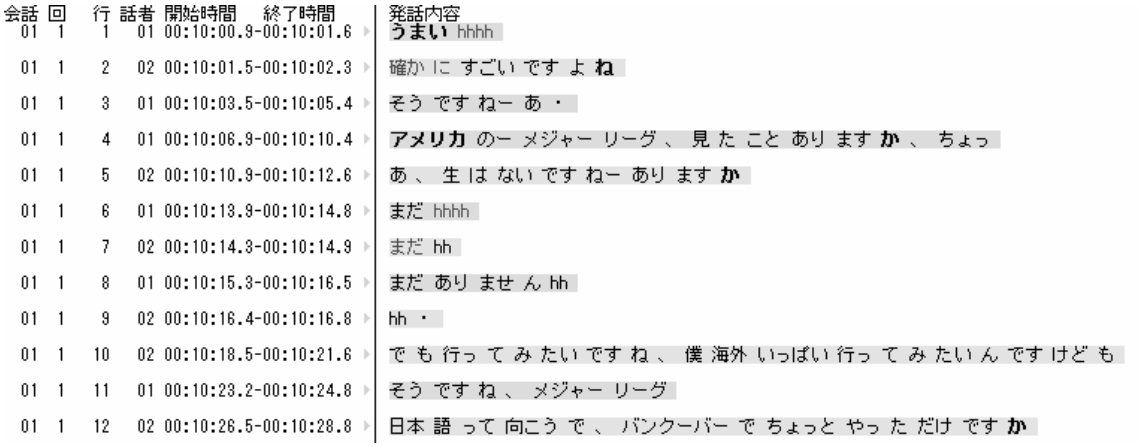
2.1 処理内容概要

ここでは、立命学習者コーパスを実際に使用する視点からツールの概要を提示する。なお、ここで使うツールは本コーパスに付属のものだが、立命館大学内でのみ使用可能なバージョンであり、予定している Web 公開の一般バージョンとはインターフェイスや一部機能に違いがある。ここではバージョンの違いに関わらない処理の内容を提示する。また、ここで使用するコーパスデータは 2011 年 3 月 15 日現在でまだ編纂作業中のもので、公開にあたり修正されるものである。

図 1. 処理一覧

図 1 は、本付属ツールで扱える処理の一覧である。処理は大きく以下の 2 つに分かれる。
①読むことを前提として発話内容を表示する。また、それを前提とし検索結果を抽出する。
②コーパス中の総語数や検索語とその共起語の数などを元とした各種統計値を算出する。
ここでは、主に①の、データを読む、検索する処理を使い説明する。

2.2 会話表示(Conversation)

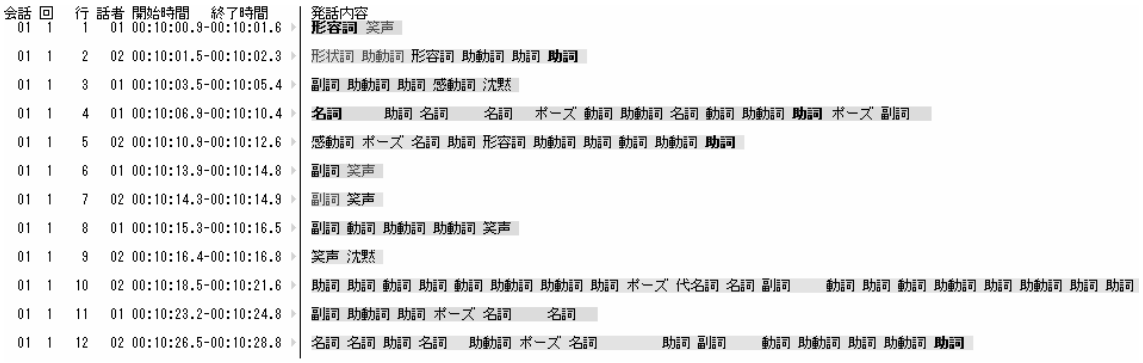


会話	回	行	話者	開始時間	終了時間	発話内容
01	1	1	01	00:10:00.9	00:10:01.6	**うまい** hhhh
01	1	2	02	00:10:01.5	00:10:02.3	確か に すごい です よ **ね**
01	1	3	01	00:10:03.5	00:10:05.4	そう です ねー あ ・
01	1	4	01	00:10:06.9	00:10:10.4	**アメリカ** のー メジャー リーグ 、 見 た こと あり ます **か** 、 ちょっ
01	1	5	02	00:10:10.9	00:10:12.6	あ 、 生 は ない です ねー あり ます **か**
01	1	6	01	00:10:13.9	00:10:14.8	まだ hhhh
01	1	7	02	00:10:14.3	00:10:14.9	まだ hh
01	1	8	01	00:10:15.3	00:10:16.5	まだ あり ませ ん hh
01	1	9	02	00:10:16.4	00:10:16.8	hh ・
01	1	10	02	00:10:18.5	00:10:21.6	で も 行っ て み たい です ね 、 僕 海外 いっぱい 行っ て み たい ん です けど も
01	1	11	01	00:10:23.2	00:10:24.8	そう です ね 、 メジャー リーグ
01	1	12	02	00:10:26.5	00:10:28.8	日本 語 って 向こう で 、 バンクーバー で ちょっと やっ た だけ です **か**

図 2．会話表示、1 行表示

図 2 は、コーパス内の発話の全体を表示させる画面である。各行の左側にその発話の番号や時間など詳細が記される。右側には発話の内容が全て表示される。この画面ではコーパス内の発話を全て読むことができ、会話の流れ、詳しい内容の確認に使うことができる。

発話者ごとに行の色が分かれる。2 者の発話が重なる箇所は赤い文字で表示され、音に高低や強さのある箇所は太字で表示される。何も選択しない状態では、発話に含まれる様々な情報は表示されず、発話の内容を読むことできる形式になる。



会話	回	行	話者	開始時間	終了時間	発話内容
01	1	1	01	00:10:00.9	00:10:01.6	**形容詞** 笑声
01	1	2	02	00:10:01.5	00:10:02.3	形状詞 助動詞 形容詞 助動詞 助詞 **助詞**
01	1	3	01	00:10:03.5	00:10:05.4	副詞 助動詞 助詞 感動詞 沈黙
01	1	4	01	00:10:06.9	00:10:10.4	**名詞** 助詞 名詞 名詞 ポーズ 動詞 助動詞 名詞 動詞 助動詞 **助詞** ポーズ 副詞
01	1	5	02	00:10:10.9	00:10:12.6	感動詞 ポーズ 名詞 助詞 形容詞 助動詞 助詞 動詞 助動詞 **助詞**
01	1	6	01	00:10:13.9	00:10:14.8	副詞 笑声
01	1	7	02	00:10:14.3	00:10:14.9	副詞 笑声
01	1	8	01	00:10:15.3	00:10:16.5	副詞 動詞 助動詞 助動詞 笑声
01	1	9	02	00:10:16.4	00:10:16.8	笑声 沈黙
01	1	10	02	00:10:18.5	00:10:21.6	助詞 助詞 動詞 助詞 動詞 助動詞 助動詞 助詞 ポーズ 代名詞 名詞 副詞 動詞 助詞 動詞 助動詞 助詞 助動詞 助詞 助詞
01	1	11	01	00:10:23.2	00:10:24.8	副詞 助動詞 助詞 ポーズ 名詞 名詞
01	1	12	02	00:10:26.5	00:10:28.8	名詞 名詞 助詞 名詞 助動詞 ポーズ 名詞 助詞 副詞 動詞 助動詞 助詞 助動詞 **助詞**

図 3．会話表示、1 行、品詞表示

図 3 は、発話内容を品詞で表示したものである。このように、指定の項目に切り替えて表示できる。切り替えられる項目は、コーパス内の各形態素に付与されている情報の全てで、文字化した文字列そのものである表記形や、活用されない形の基本形、品詞、活用形、音の強弱、誤用の有無、日本語教育レベルごとの文型、などの様々な種類がある。使用者が任意の項目の表示に切り替えられ、発話に含まれる様々な情報を確認できる。

図 4．会話表示、2 行、品詞表示

図 4 は、発話を 2 行で表示したものである。表示の切り替えでは、発話内容をそのまま表示しながら同時に各形態素に付く項目を表示することもできる。その際、2 行で表示され、上の行には発話内容、下の行には指定した項目となる。これにより、発話を読みながら同時に各形態素に付く情報を確認することができる

表示は 3 行まで同時にすることができる。その際、上の行には発話内容、中の行には直前に指定した項目、下の行には 2 つ前に指定した項目が表示される。これにより、発話内容に加え、品詞を見ながら活用形を確認する、文型を見ながら誤用を見る、などの使い方が可能である。

2.3 検索

図 5．検索条件一覧

図 5 は、検索で指定できる項目の一覧である。検索条件は、発話自体の持つ情報からの項目、発話内容の各形態素に付けられている項目、の全てを任意に指定でき、複数の項目を組み合わせることもできる。

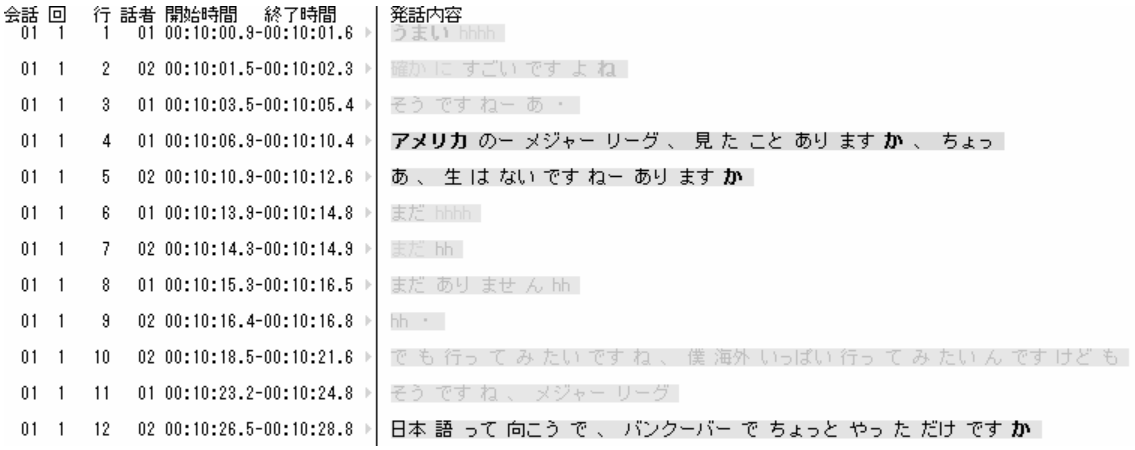


図 6．会話表示、1 行、「か」検索

図 6 は、表記形を「か」と指定し検索をした結果で、該当する形態素を含む行はそのまま、含まない行は薄く表示される。これにより、発話内容を読みながら、コンテキスト内での言語使用や機能を分析対象とする談話分析でも効果的に使用できる。

2.4 検索語と使われている文脈(KWIC)

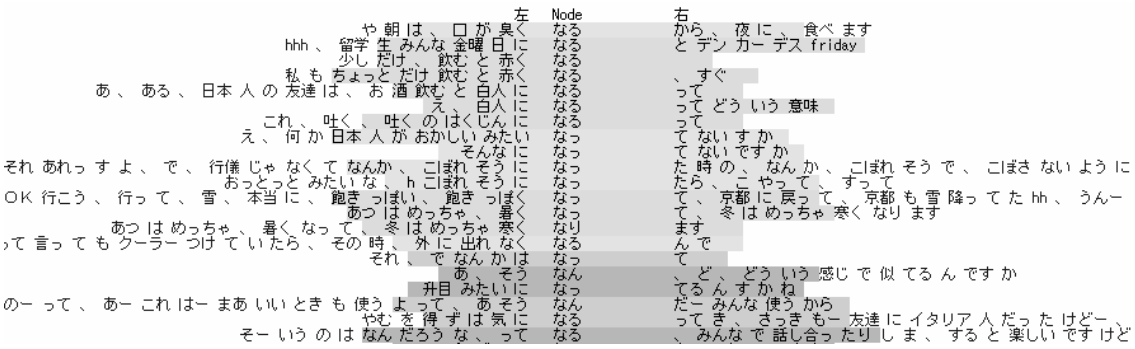


図 7．KWIC、基本形「なる」検索

図 7 は、KWIC という方法で表示される検索結果の画面である。ここでは「基本形」の項目を「なる」と指定して検索している。KWIC では検索した形態素を中心に並べて表示し、その左右に、前後の内容が表示される。会話表示同様に話者ごとに色が分かれる。表示の切り替えも会話表示の処理と同様に指定項目で行える。

2.5 検索例と学習者の傾向

検索例として、日本語母語話者と日本語学習者の使用する接続助詞＋終助詞を比較する。

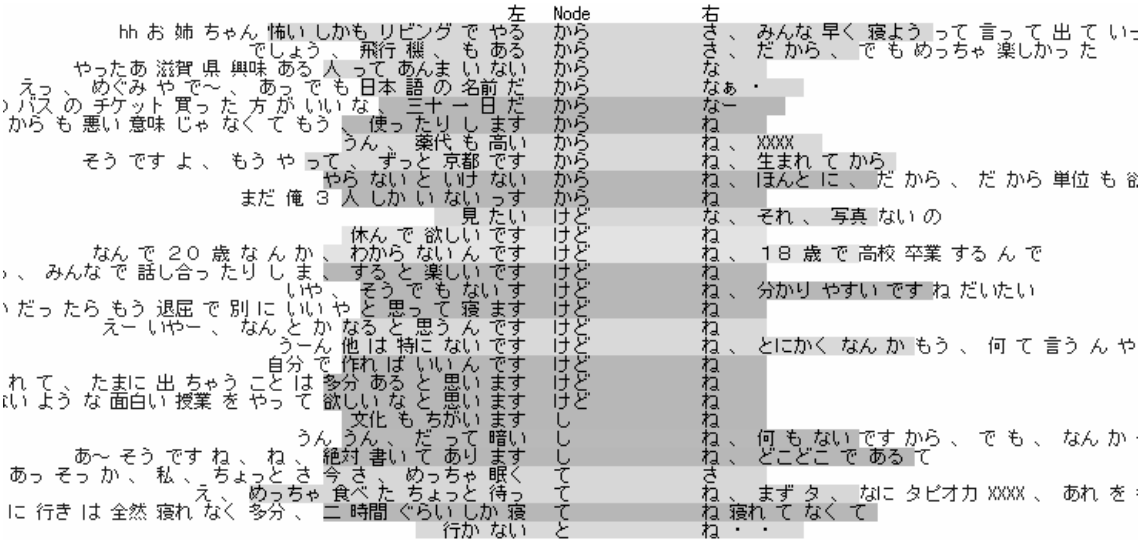


図 8．KWIC、日本語母語話者、接続助詞＋終助詞

図 8 は、「母語」の項目を「日本語」、「品詞下位分類」の項目を「接続助詞」と指定して検索し、更に検索語の右 1 つ隣に、「品詞下位分類」の項目で「終助詞」がある結果のみを抽出し、それを、検索語の位置を第 1 条件、右 1 を第 2 条件でソートしたものである。つまり、日本語母語話者が使った「接続助詞」＋「終助詞」の例を、使用された接続助詞と終助詞でまとめて並べたものである。検索結果は 28 で、内訳として「けど＋ね 10」、「から＋ね 5」、「し＋ね 3」、「と＋ね 3」などが多いことが分かる。次に学習者の使用例と比べる。

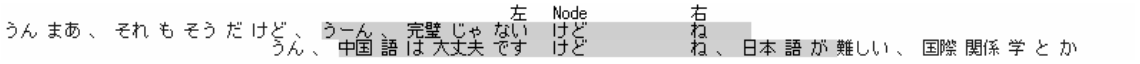


図 9．KWIC、日本語学習者、接続助詞＋終助詞

図 9、条件のうち「母語」の項目を「中国語|韓国語」に変更したもの、つまり、図 8 での例を日本語学習者に変えたものである。検索結果は 2 で、共に「けど＋ね」である。

母語話者に比べ、学習者は「接続助詞」＋「終助詞」を圧倒的に使用していないことが分かる。また、母語話者も学習者も一番多く使用される表現は「けど＋ね」で、学習者はこれ以外を全く使用しておらず、このデータだけで言うと使用傾向は偏っている。この「接続助詞」＋「終助詞」という形の習得が難しいからか、言語習慣的に用いない表現であるからか理由の確定は今後の研究が必要だが、この使用数の差は明確な特徴であると言える。

本コーパスでは、簡単な検索やソートでこれらの例を見付けることができる。

3．理念、設計思想

本来コーパスには大きく 2 つの利点がある。1 つ目は、その規模を根拠とした統計を行える点、2 つ目は、必要な例をすばやく検索して抽出することができる点である。

特に、1 つ目の「規模を根拠とした統計処理」というのが一般的にコーパスの一番のメリットとされ、コレスポンデンス分析や、カイ二乗検定など様々な統計手法が使われていて大きな成果を果たしている。データが大規模であることは、その量によって、他の方法ではできない言語の様々な側面の発見ができるという強みを持っている。その効果や有意義さ、これまでの成果は疑う余地がない。しかし、コーパスは必ずしもこの様な高度な統計処理を前提として使用する必要はない。実際に使用された言語資源の記録である以上、記述文法や談話分析の研究などで使用するデータと本質的に変わることはない。

現在、R や SPSS など、不慣れな研究者や現場教育者にも扱いやすい、優れた統計ソフトがいくつもあるが、それらを利用するには、まず目的のデータを確実に抽出することが必要である。日本語コーパスの整備も著しく、多くのコーパスが作られていて、また、学習者コーパスもそれなりに揃ってきているが、膨大なデータの中から必要な部分のみをどう抽出するかが問題である。テキスト内から目的の部分を抽出するためには、目印となる情報が付与されている必要がある。文字化したままの状態であるプレーンテキストに、様々な文法情報を付与する形態素解析ソフトは高性能なものが作られている。しかし、現状形態素解析ソフトで付与できる多彩な情報を余すところなく扱える分析ツールはほぼ無い。

そこで、筆者はまず、コーパス利用の根本となる「検索する」ことに特化して、「HASHI(田中 2010)」というコンコーダンサーを開発した。「コンコーダンサー」とは、ここではコーパス中のデータの検索や統計を行うためのソフトウェアのこととする。その HASHI で、形態素解析ソフトによって付与できる情報を出来るだけ多く扱い、自由で複雑な組み合わせで検索することを可能とした。しかし、「検索」ということに限って言えば、それを高度に突き詰めると、向かう方向は「“ 文字列 ” で検索するために、正規表現によりアプローチの自由度を高める」と、「アクセスできる条件の種類の増すために、テキスト中に付与する情報を増やす」の 2 点になる。そのうちの後者を選ぶ場合、テキストに「情報」を付与するには形態素解析ソフトに頼らざるを得ないが、それよって自動で付けられるタグには限りがある。時としてそれらの情報だけでは研究に必要な項目が扱えないこともある。基本形、品詞、活用形、アクセント型、などの辞書を基にした情報を画一的に付けるしかできない。もちろん、それができるようになったというのは革新的で、言語研究に多くの可能性をもたらした偉大な業績であることは疑う余地が無い。しかし、その素晴らしさゆえにそれらのソフトが一気に普及し、既に形態素ごとの各情報の付与が当たり前となっている。その上で次の段階としてその他の、談話分析や日本語教育の用途でコーパスを使う場合や、それらに関係する多くの情報で検索を行う場合に、それを叶えるタグ付与のソフトが現在無いのである。それらの情報を扱いたいときは、元々のデータ自体に自分の使いたいタグを

付与するしかなく、そのためには、コーパス自体を作成するしか方法がない。その目的を叶えるコーパスを作るために「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」を構築した。これは「できうる限りのあらゆる情報を付与し、多次元的な複合検索を行えるコーパス」という方針で設計されている。本稿で扱う立命学習者コーパスは、このシステムで作られたコーパスである。

規模を拠り所とする統計処理を主目的としたコーパスとしては設計せず、今まででは到底扱うことのできなかった、様々なニーズでの多様な情報を扱い、きめの細かい条件で、その時すぐにほしい結果を、簡単に入手できるよう、強力な検索機能を持つという方向を目指した。つまり本システムで作られるコーパスは、統計よりむしろ直接データを読みながら使う、談話分析のような研究に資するものとして設計した。そして、必要な結果を自由に抽出できるように、コーパス中に膨大な種類の情報タグを付与することを前提とした。これにより理論上、会話中に含まれる非言語や話者の性質や言語習得レベルなどを含むあらゆる情報の付与、取り扱いが可能となる。

このように、膨大な種類の情報をコーパス中に付与することを前提とするのだが、それを効率良く扱うためにはどうしても専用コンコーダンサーが必要になる。そのコンコーダンサーはコーパスの価値を最大限に引き出すために、そのコーパスデザインに特化したものでなくてはならない。また逆に、技術的にコンコーダンサーで扱えない情報は残念ながらコーパスに織り込んでも利用することができない。データであるコーパスを最大限生かせる機能のツール必要があり、また、ツールであるコンコーダンサーの機能を最大限生かせるデータ形式が必要である。ツールの機能強化に伴いデータ形式の修正を行い、また、データの種類の追加に伴いツールを改良する。それを交互に何度も行うことで、コーパスとコンコーダンサーが有機的に結びつき、一体となるし、ならねばならない。その点でコーパスデザインとコンコーダンサー設計は同義であるといえる。本システムではこのようにして、コーパスデザインに特化した専用コンコーダンサーの付属を前提とする。

本システムで作成されるコーパスでは、膨大な情報を簡便に扱えるようなるのだが、それゆえにコーパス編纂の際、単純な文字化コーパスに比べ何倍もの人的労働力を必要とする。単純に、人手も時間も予算もより多く必要になるのである。自動では付けられない様々な情報をテキストに付与する関係から、多くの作業を、基本的に人的資源に負わねばならない。よって、編纂作業には膨大な時間と手間がかかり、そのため、多くの予算が必要になるという性質がある。余程の予算や開発期間が無いと規模の拡充は望めないという構造上のウィークポイントがあるのである。更に発話コーパスの場合、最初にデータを文字化しスクリプトにすること自体から始めなければならず、また、それは音声データと限りなく一致するものである必要がある。このため本システムは、その採用を決定した段階において既に大規模コーパスの構築という方向はほぼ望めず、現実的に小規模コーパスという選択肢しかないものと言える。ある意味でコーパスの一番の利点と言える「データの分量」に頼ることができないため、あらかじめその使用やデータの選定には工夫が必要になる。

4. コーパス構築システム

コーパスは、編纂されたデータが「物質」だとすると、設計は「法則」である。本章では本コーパスを構成する法則を示す。以下、4.1 節で、本システムの概要、4.2 節で、情報をコーパスに記録する方式、4.3 節で、情報タグ付与の単位、4.4 節で、実装を提示する。

4.1 構築システム

本稿で提示する「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」は、3 章の設計思想の元構築したものである。同システムで初めて実装したコーパスは「立命会話コーパス」であり、2 つ目が今回の「立命学習者コーパス」である。

4.2 データ記録法概要

まず、3 章で提示した設計思想でコーパスを設計する際、具体的にどうやって実現するかが問題となる。特に、発話内に含まれる情報をコーパス内にどう記録するかが問題である。以下、情報タグのコーパス内への記録形式を 3 つ提示する。

昨日、しゅっ＜言い淀み＞、受験を受けた＜誤=重複＞けど[N5:けれど（も）＜逆接＞]、あのかっこう＜正=学校＞に私（わたし）（うんうん他者(女)）行けるかな↑

図 10．タグ内在形式の図

昨日、しゅっ、受験を受けたけど、あのかっこうに私行けるかな

図 11．タグ無し形式の図

発話者A	昨日、しゅっ、受験を受けたけど、あのかっこうに私行けるかな
種別	言い淀み
誤用	重複
文型	N5:けれど（も）＜逆接＞
正しい形	学校
読み	わたし
音声	↑

発話者B	うんうん

図 12． 階層別タグ記述形式の模式図

図 10 から図 12 は、全て同じ発話内容をコーパス化したものである。発話内容と発話内情報があり、3 つの図は全て、発話内容は同じだが、情報タグの記録形式に違いがある。

発話内容と、発話内に記録された情報を分けて表示すると以下になる。

発話内容

昨日、しゅっ、受験を受けたけど、あのかっこうに私行けるかな

発話内に記録された情報

しゅっ	言い淀み
受験を受けた	誤用：重複
けど	日本語能力試験 N5 相当文型：けれど（も）＜逆接＞
かっこう	正しくは「学校」
私	読みは「わたし」
うんうん	別の発話者による重なり
行けるかな	上昇イントネーション

図 10 は、発話のスクリプト内に様々な情報が付与される形式である。この形式は情報過多で、どこが発話内容で、どこが発話に含まれる情報かが分かりにくい。それでもなんとか人間には読むことができるが、機械で検索するには不都合が生じる。発話内容も発話内に付随する情報も、同じ「文字列」であるため、機械では区別しにくいからである。例えば「受験」という語を検索しその周囲 10 文字を抽出すると、「しゅっ＜言い淀み＞、受験を受けた＜誤=重複＞」となる。この中で発話内容は「しゅっ、受験を受けた」だけで、「＜言い淀み＞」、「＜誤=重複＞」は発話内の情報である。発話内容のみを目的に検索をしても、不要な情報まで抽出される。ただ、「＜言い淀み＞」と検索することで言い淀みの箇所すべてを抽出できるような、発話内の情報での検索ができるメリットはある。ただし、その際も抽出範囲に発話内容と発話情報が混在することは同様で、この結果を元に語数を数えるなどの次段階の処理が行えない。人間には読めるが機械には扱えないコーパスとなる。

図 11 は、反対に、内容の読みやすさと機械での扱いやすさを重視し、発話のスクリプト以外の一切の情報を記録しない形式である。この形式は読みやすく、また、検索の際、発話内容のみが抽出される。図 10 の場合と同様「受験」と検索すると、「昨日、しゅっ、受験を受けたけど、あのか」となり、この結果を元に検索語の周囲の語を数えるなどの処理も行える。しかし、これでは発話に含まれる豊富な情報や、付属の情報をデータに取り込めず、検索では発話内容の「文字列」にしかアクセスできない。機械処理には向くがデータの豊富さを捨てているのである。

多くのコーパスはこれら図 10 と図 11 のように、「機械では扱いにくいが豊富な情報が付与され人間には読めるコーパス」か、「発話に含まれる情報はほぼ扱えないが機械処理に向いているコーパス」のどちらかになる。

図 12 は、その両方のメリットが両立する形式で、本システムの形式である。この形式で

は、まず発話者ごとに発話の記録を分ける。その上で発話の記録と発話情報の記録を分ける。更に、発話情報の種類ごとに記録場所を分ける。記録は、発話内容の層、発話情報Ａの層、発話情報Ｂの層・・・と階層式にする。一番上の層に発話内容を記述する。その下に何層にも分かれて様々な情報を種類ごとに記述する。また、発話情報の記述は、それが起こった発話内容の真下に記述することとする。これにより、発話内容の検索を行う際にも各種情報が検索結果に入り込むことなく発話内容のみを抽出できるようになる。また、発話情報を指定して検索をすれば、その条件のある場所の真上の発話内容の語を抜き出すことができる。例えば、「文型」の層で「N5:けれど（も）＜逆接＞」と指定すれば、それが見つかった場所の一番上の層にある「けど」を検索でき、その周囲の語を抽出する際には「受験を受けたけど、あのかっこうに私」と、発話内容のみを抜き出すことができる。また、階層を分けることで階層別に検索条件を指定でき、情報の組み合わせによる検索が可能となる。可能性としては、「文型」が「N4 のどれか」＋「誤用」が「有り」や、「種別」が「言い淀み」＋「音声」が「↑」などの、複数に指定した条件が全て揃う結果のみを検索できるようになる、など複雑な条件での検索が扱えるようになる。

4.3 タグ付けの単位

このシステムでは、文字化した発話内容を全て、決まったある基準の単位で分ける。その他の、発話中に含まれるあらゆる情報は、発話内容を、基準の単位で区切った位置と時間軸上で全く同じ位置で区切り、全て、その区切られたユニットごとにタグとして付与し、保持、運用する。発話を一つの同じ単位で区切れれば、その単位ごとに何層でも多重に情報を付与できる。今回の立命学習者コーパスではその単位を形態素とする。本来は、発話中に含まれる情報の全てが「形態素」の基準で区切れるものではないが、敢えて１つの単位を基準に区切ることで、あらゆる情報を、統一の仕組みで各ユニットに付けることができる。情報の付け方は、各ユニットに複数の階層を設け、１つの階層に１つの情報タグを付与するという方法にする。つまり、発話を、時間軸上で横に区切った上で、各ユニットを階層として縦に区切り、そこに様々な情報タグを付けるという記録方法になる。

表記形	京都	に	ずん	で	京都	弁	を	使い	ます
基本形	京都	に	住む	で	京都	弁	を	使う	ます
活用形	---	---	連用形	---	---	---	---	連用形	終止形
読み	キョウト	ニ	ズン	デ	キョウト	ベン	ヲ	ツカイ	マス
プロソディ	--	--	--	--	!	!	--	--	--
重なり	有	有	無	無	無	無	無	有	有
誤用	-	-	発音	-	-	-	-	-	-
文型 N4	---	---	---	て(緩い連結)	---	---	---	---	---

図 13.　形態素単位での情報付与の模式図

図 13 は、あくまでも模式図であり、情報タグ付与のイメージ図であるが、この形式で発話を記録することにより、発話中に多面的に含まれる様々な情報を、漏らすことなく記録できる。その結果、全ての種類の情報を一律の仕組み検索できるようになり、扱い方においても情報間の差異が無くなる。また、階層を分けて情報を付与することで、複数の情報をいくつでも組み合わせて検索できるようになる。理論上、ここに付与できる情報は、語形、文法事項、音声情報、発話権、言語能力情報、非言語情報など、多岐にわたり、ほぼどんな情報でも付与することができる。

また、形態素ごとに分けて付与する情報以外に、性別や母語や出身地などの発話者の情報や、発話番号などの、行ごとに一定している情報がある。これらはいちいち形態素ごとに付与すると情報が無駄に多くなり、処理に大きく負荷がかかるため、行自体を多階層に分けて記録した。つまり、「形態素単位」と「行単位」の二段構えになる。行ごとに発話者の属性、発話者同士の関係性などの項目を付与することで、発話内容に含まれる情報と、発話者の属性情報を組み合わせて検索することが可能となる。これらの情報も、実際のコーパス使用の際は、形態素ごとに付与した情報と全く同じ操作で扱え、検索項目の組み合わせに加えることができる。この仕組みにより、発話内容、発話内の情報、発話者の情報など、発話が行われる際に存在するあらゆる情報をコーパス内に付与できる。それによって、コーパスを使用する際に、発話された場面や話者の状況など、その発話に含まれるあらゆる内容を再現することができるようになる。

4.4 実装

この「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」を元に、これまで 2 つのコーパスを作成した。1 つ目は「立命会話コーパス」で、2 つ目は、今回の「立命学習者コーパス」である。ともに収録データの言語は日本語である。

立命会話コーパスは、談話分析で使用することを目的に作成した、母語話者会話コーパスである。付与した情報タグもそれに特化したものとして選別した。本コーパスは、日本語学習者の発話を記録した学習者会話コーパスである。また、日本語学習者と日本語母語話者の会話を記録した接触場面コーパスでもある。日本語学習者の発話を扱うため、情報タグは日本語教育での研究に有用なものに特化し、選別、追加した。本コーパスは、多次元多項目で日本語教育に特化した情報の付与されるもの、かつ、それらの情報を簡単な操作でいくらでも複雑に組み合わせて検索できるコーパス。という方針に基づいて編纂した。

5. データ

コーパスのデータには、日本語学習者と日本語母語話者の接触場面を記録した。来日したばかりの日本語学習者と日本語母語話者のペアの会話を、一定期間継続的に取り、学習

者の日本語の変遷や、同一ペアの間に起こる会話の変化を記録した。この、非常に珍しく貴重な縦断的データにより、学習者の数カ月にわたる会話の、一定の期間おきの時点を扱え、各時点で起こった発話の様々な状況を再現して分析することができる。データ収録の方針として、高レベルの学習者のみを対象としてデータを集めた。

発話者の属性を以下の表に記す。

表 1 .話者属性表

話者	年齢	性別	出身国（出身地）	最長居留地	母語	日本語レベル	来日時	日本語の勉強法
01	21	女	台湾	台湾	中国語	上級	2010 09/	大学の授業
02	19	男	東京都	大阪府	日本語			
03	22	女	韓国	韓国	韓国語	上級，2 級	2010 09/17	独学
04	20	男	愛知県	愛知県	日本語			
05	23	男	台湾	台湾	中国語	上級，N1	2010 09/15	大学の授業
06	20	女	滋賀県	日本	日本語			
07	20	女	韓国	韓国	韓国語	上級	2010 09/	大学の授業
08	22	女	愛知県	愛知県	日本語			
09	23	女	中国	中国	中国語	上級，1 級	2010 09/14	大学の授業
10(1)	21	女	愛知県	愛知県	日本語			
10(2)	20	男	京都府	京都府	日本語			

会話録音日時を以下に示す。

表 2．会話録音日時

	第 1 回	第 2 回	第 3 回	第 4 回	第 5 回
会話 01	2010 10/28	2010 11/18	2010 12/9	2010 12/23	2011 1/20
会話 02	2010 10/28	2010 11/18	2010 12/9	2010 12/23	2011 1/14
会話 03	2010 10/29	2010 11/19	2010 12/10	2010 12/24	2011 1/20
会話 04	2010 11/5	2010 11/24	2010 12/16	2011 1/7	2011 1/22
会話 05	2010 11/16	2010 12/8	2011 2/8		

会話ペアの数え方は、それぞれのペアを 2 桁の通し番号で示し、会話 01、会話 02・・・とする。話者番号も同様とし、各会話中の、学習者を奇数、日本語母語話者を偶数とする。

学習者の日本語レベルに関しては、日本のレベルと、保持していれば日本語能力試験の結果を併記している。

出身地と最長居留地に関しては、学習者は国、母語話者は都道府県レベルとする。属性調査の際に学習者には「国」で記入するよう提示したが、回答は上記の通りであった。こ

こでは記入者の意思を尊重し、このまま扱うこととする。

会話 05 に関しては、日本語母語話者である話者 10 が第 2 回で参加停止したため、3 回目以降は別の発話者に変更している。便宜上、最初の協力者を話者 10(1)、交代した協力者を話者 10(2)とする。

発話協力者は、2010 年 9 月に来日した、立命館大学独自の短期（半年～1 年）留学生受け入れプログラムである、SKP（Study in Kyoto Program）の留学生 5 名、うち中国語母語話者 3 名、韓国語母語話者 2 名、と立命館大学の学部生日本語母語話者である。これら、外国人留学生と日本語母語話者の接触場面を、2010 年 10 月 28 日～2011 年 02 月 8 日にかけて、約 3 週間おきに記録した。具体的には、学習者と母語話者 2 名 1 組のペアを 5 組設定し、同じペアの会話をそれぞれ 5 回収録すると計画している。ただし、うち、4 組は全 5 回の会話の収録を終了したが、1 組のみ、本稿執筆時点（2011 年 3 月 15 日）では、発話協力者の都合により、第 3 回目で収録が停止しているため、全 23 会話の収録が済んでいる。今後、残り 2 会話を加え、合計 25 会話を記録する予定である。1 会話を各 60 分録音し、そのうちの 10 分目から 40 分目の 30 分ずつを文字化してコーパス化し（以下、「文字コーパス」）、更に、そのうちの 10 分目から 20 分目のデータには情報タグを付与し、精密なコーパス（以下、「タグ付きコーパス」）とした。

2 種類のコーパスのうち、本稿で主に扱うのはタグ付コーパスである。そのデータの分量は、本稿執筆時点（2011 年 3 月 15 日）で 4 組が全 5 回収録済み、残り 1 組が 3 回目まで収録済みで、合計 23 会話、230 分を収録している。収録の完了している会話は、会話 01～会話 04 で、未完の会話は会話 05 である。総語数である TOKEN 数は 34140 である。本コーパスは現時点で編纂過程にあり、今後データの追加や修正により時間数や語数などの変更を予定している。また、一般公開後、文字化の不備やタグの付与の相違などの指摘や意見を受けた場合、それを精査しコーパスデータ修正に反映させる可能性もあるため、この TOKEN 数は漸次的なものである。

6. コーパス設計

本章では、コーパスの具体的な設計内容を示す。6.1 節で、本システムで使用する用語の定義、6.2 節で、文字化ルール、6.3 節で、形態素単位に付与する情報タグの種類、6.4 節で、行単位に付与する情報タグの種類、6.5 節で、コーパスに情報タグを付与する単位の規定とその理由を提示する。

6.1 用語の定義

本システムにおける用語の定義を再度提示する。

情報　発話中に含まれる語の基本形や品詞や発話の重なりなどの様々な事柄

項目　情報を具体的に分けたタグの種類
　　「基本形」「品詞」「読み」「プロソディ」「性別」「母語」など
要素　項目の内容
　　「話す」「動詞」「オオキイ」「↑」「男」「日本語」など

6.2 文字化ルール

以下本コーパスの文字化のルールを記す。

◎文字化規則

発話内容を文字化したスクリプトは「表記形」という項目の層に記録されるが、ここには、実際に発した音声以外の情報は、基本的に一切記述しない。本コーパスでは、発話内容以外のあらゆる情報は、別項目として階層を分けて記述することができるためで、発話に含まれる様々な情報は、文字化スクリプト内には記述しない。

発話内容は、フィラー、言いよどみ、言い直し、言い間違いなど、全てひらがなとカタカナの単位で再現できる程度まで発声通り記述する。

◎1 行の単位

1 つのまとまった発話を連続して記述する単位を「行」とする。1 行は、同一話者による連続した音声とする。行を区切る条件は、「同一話者が 1 秒以上音声を発しなくなる箇所」とし、原則としてそれ以外では行は区切らない。発話中、相手話者が発話を開始したり、2 者の発話が一時的に重ったり、2 者がともに沈黙したり、様々な状況が発生するが、いずれにせよその行に記述する話者の発話が 1 秒以上止まる場所でのみ区切ることとする。1 つの行には、1 人の話者の音声が記録され、その発話自体が 1 秒以上止まるまで記述される。

行には、通常日本語で扱う「文」を複数含むものもある。逆に 1 文が複数の行に分かれるものもある。現実に発せられる発話は、多くの場合「文」と明確に区切ることが難しいものが多く、また、その区切りは意味に頼らなければいけない場面が多い。発話記述の区切りに、発話内容の意味を持ち込み、ひとつひとつを識別するのは現実的には非常に難しいため、今回、それは行わないとした。文の区切りの根拠を各発話全てに提示できない以上、意味を伴う「文」の採用は回避すべきで、その理由から今回は「行」を単位とした。「文」という単位を採用しないので、「。」「！」「？」のような文末を表す記号は用いない。単純に音声が連続しているかどうかのみを、行を区切る基準とする。

では、この「行」を区切る「1 秒」という時間に付いてであるが、これは実際に扱いやすい長さということで規定したのであり、学術的な根拠は無い。0.2 秒などの短いポーズで区切ることも考えられるが、その場合は区切りが細かく、1 行が短くなり過ぎるため、語を検索し周囲の語の傾向を見ようとする際に、抽出範囲内にほとんど語が無いという状況になる。このため、運用と見やすさを考え、やや長めの 1 秒とすることとした。

◎記号

本コーパスの文字化に使用する記号の一覧を示す。

表 3．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**

○長音記号

長音の記述は「ー」と「～」を併用する。長音化された音声部分に、音の高低や強弱での振えが無ければ「ー」を使用し、震えが有れば「～」を使用する。

音の長さは「ー」を連続で記す個数で表現する。「うーーん」「うーーーん」など、「ー」の長さは周囲の語を参照し、1 モーラに「ー」1 つを前提とする。音に震えが全く無い 3 モーラ程度の長音は「ーーー」と記す。同様に震えの有る 3 モーラ程度の長音の場合は「～～～」とする。「ー」と「～」の区別には音の長さは関係無いものとする。

○笑い声

発話を伴わない笑い声は「h」で記述する。長い笑いは「hhh」のように連続させ「f」の個数で再現し、長さは周囲の音声部分の文字列を参照する。発話しながらの笑いは発話のみを記述する。

○息

発話中、強い息が有り目立つ所は「f」で記述する。長い息は「fff」のように連続させ「f」の個数で再現し、長さは周囲の音声部分の文字列を参照する。

○聞き取り不可

聞き取り不可部分は「X」で記述する。長い聞き取り不可部分は「XXX」のように連続させ「X」の個数で再現し、長さは周囲の音声部分の文字列を参照する。

○沈黙

話者 2 人がともに 1 秒以上沈黙し、音声が無い個所には「・」を打つ。1 秒につき「・」

を１つ打ち、5 秒沈黙があれば「・・・・・」と 5 つ連続して記入する。「・」は、前の発話の続きの行末に打つ。この箇所だけ「秒」という単位がコーパス中に混在する。

〇短いポーズ

１人の発話内での、１秒未満のポーズの個所には「、」を打つ。

１つの形態素の中であったとしても、完全に音声が連続していなければ形態素の途中でも「、」を打つ。「オリン、ピック」のようになる。

〇連続する語の区切り

「４、５回」「中国、韓国」など、音声では完全に連続するが、連続して記述すると意味が分からなくなる箇所には「,」を挿入する。本来、通常日本語で表記する際、見やすさで打つ「、」が入る箇所だが、「、」は本コーパスでは１秒未満のポーズという意味を持つため使用しない。しかし、「４、５回」を「４５回」と記述すると意味が変わるので、「、」の代わりに「,」を用いる。ただし、この箇所に実際に音声として少しの区切りがあれば「、」を挿入する。あくまでも音声に区切れ目が有るか無いかのみが基準となる。

〇発話の重なり

音の重なっている所は [ ]で囲う。ただし、これは文字コーパスでの記述方式であり、タグ付きコーパスでは重なりの箇所は別の記録方式になり、表示の際には赤い文字で記される。以下、例文を示す。例文中の「<05>」、「<06>」はそれぞれ話者番号を表す。

例文１

<05>

改めて、はい話してくださ[いって言われるとぉー]、なかなか、話せないものですよね・・

<06>

[って言われるとね]

上記のように、話者 05 の「いって言われるとぉー」と、話者 06 の「って言われるとね」が重なっている場合、それぞれの箇所を[ ] で囲い、後から発声された話者 06 の発話を次の行に記す。重なりは、ひらがな１文字の単位までの厳密さで記録する。そのため、漢字で表記される形態素の中に音声の重なりの区切れ目がある場合、例えば「学校」という形態素のうち、最後の「う」の部分だけが重なっている場合、つまり「がっこ[う]」のように記されるような場合は、「学こ[う]」と、漢字で表記できる部分は漢字のままにし、重なりの区切りを含む漢字だけをひらがなで表記し記す。

本来は、発話の重なりを付ける箇所は、厳密に音声同士が重なる場所とするのだが、ひらがな１文字単位で細かく記録するというルールと、１秒以上の連続した同一話者の音

声は 1 行に記録する、という 2 つのルールを現実的に運用するために、発話の音声と直接重ならなくても、「、」の部分に相づちなどが入る場合重なりと同等に扱うとした。

例文 2
<08>
あっそゆこと[かぁ、]大学名かと思った[、]うーん、はいっ
<07>
[はい hhh]
<07>
[はい]

例文 2 の中の話者 08 の 2 つ目の[ ]の箇所は音声の無いポーズの箇所だが、ここを、重なりが有るとして扱い、話者 07 の「[はい]」と対応させている。

◎表記法

原則的に発話の文字化は、実際に聞こえる発話された音声そのままを記録し、日本語であればひらがなで再現できる限りの精度で正確に文字化する。

〇漢字のルール

基本的に、日本語で一般に漢字で記述する程度の語は漢字にする。「わかる」「分かる」など、漢字やひらがなのどちらで記述されても自然なものは漢字を使う。別階層に各形態素の読みを付与できるので、「わたし」「あたし」「わたくし」など、同一漢字で読みに揺れがある語は、その代表形として「私」と漢字で記述する。形式名詞の「こと」「もの」「よう」などはひらがなで記述する。また、主に「動詞の連用形＋接続助詞のテ」に後節する補助動詞の「いく」「みる」「いる」「おく」「くださる」などはひらがなで記述する。「ている」が実際の発声では「てる」や「てく」となっている場合など、音声通り記述する。記述に幅がある表記を統一することで、コーパス使用時に検索の効率が上がる。

〇別言語、数字、アルファベット

日本語で発音された部分の数字や、アルファベットでの語は、全角で記述する。
英語で発音された部分は、半角アルファベットで、数字は半角で記述する。
中国語で発音された部分は、中国語の漢字、数字は全角で記述する。
韓国語で発音された部分は、韓国語のハングル、数字は全角で記述する。

〇微妙に違いのある音声の表記の分け方

「そっか」「そうか」、「うん」「うーん」、「よね」「よねぇ」など、非常に近い音声も、

出来る限り実際の音の通りに記述する。「よねぇ」「よねえ」の違いは、語尾の語気が強いかどうかとし、音の高低は関係ないものとする。

6.3 語情報 - 形態素単位のタグの種類

形態素単位に付与する項目別のタグの種類をを示す。タグの詳細は、別に作成するマニュアルに記述する。形態素分けは奈良先端科学技術大学院大学で開発された ChaSen version 2.4.1 で行い、内部辞書に伝康晴・他(2009)により開発された UniDic version 1.3.12 を用いた。ChaSen で自動付与したタグは、その後、全て人手で修正した。語彙レベルと文型レベルは、独自開発のプログラムにより自動付与し、その後文型レベルは全て人手で修正した。それ以外は全て初めから人手で付与した。

タグ情報は、「表記形」「正表記形」「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」「重なり」「誤用」「語彙レベル」「文型レベル N1」「文型レベル N2」「文型レベル N3」「文型レベル N4」「文型レベル N5」である。

以下、個別に説明する。

◎表記形

形態素が実際に発話された形で、活用語であれば活用された形である。具体的には文字化した文字列そのままを、形態素ごとに区切ったものである。会話コーパスであるので、出来る限り発声に正確に記述するため、「歩くぅ」や「高校ー」のように、通常の形態素の形が変形されたものも、付随する長音記号なども含め、表記形とする。

◎正表記形

誤用された表記形が正しく使われた場合の形のこと。活用された形で表記する。誤用以外では、母語話者であっても、表記形の「なるっ」に対し「なる」とするなど、一般的な表記形の変化形に設定する。特に誤用や特殊な表記形でなければ表記形をそのまま用いる。

◎基本形

形態素の基本形のこと。実際に発話中に出現した形態素の、活用がされていない形。活用語の活用をまとめたもの以外で、名詞など、活用の無い語の場合も表記形と基本形が違う場合がある。会話コーパスであるので、出来る限り発声に正確に記述するため、「机～」や「手紙ぃ」という表記になることがあるが、その際は「机」「手紙」を基本形とする。同様に感動詞など、例えば「ああ」「あ～」「あぁ」などを「ああ」を基本形とし統一する。

◎品詞

形態素の品詞のこと。UniDic での品詞大分類を採用しているため、国語文法や日本語文法の品詞分類とは異なる場合がある。本コーパスで使用する特殊な記号や形態素分けに対

して、いくつかオリジナルの品詞を設定した。それぞれの記号に対し各品詞を

X　　聞き取り不可
h　　笑い声
f　　息
・　　沈黙
、　　ポーズ

とし、他に、言い淀みの箇所を「言い淀み」と設定、日本語以外の言語で発話された箇所に関しては、「別言語」という品詞を設定し、下位分類をそれらの個別言語の名前とした。

◎品詞下位分類

形態素の品詞の下位区分のこと。具体的にはUniDicでの品詞中分類以下である。品詞同様、国語文法や日本語文法の品詞分類とは違う場合がある。UniDicで品詞下位分類の割り当てが無い場合、品詞名をそのまま下位分類にも適応する。

◎活用形

形態素の活用形のこと。「連用形-イ音便」や「命令形」など。活用形に下位分類が有る場合、「連用形-一般」「連用形-イ音便」「連用形-撥音便」と、続けて表記される。非活用語の場合は、--- となる。

◎活用型

形態素の活用型のこと。「五段-カ行-イク」や「下一段-マ行」など。活用型に下位分類が有る場合、「上一段-ア行」「上一段-カ行」「上一段-マ行」と、続けて表記される。非活用語の場合は、--- となる。

◎読み

形態素の読みをカタカナで表記したもの。「私」などのように複数の読みがある場合は、「アタシ」「ワタシ」「ワタクシ」のように正確に発声通り記述する。カタカナの単位で表現できる限り精密に記述する。

◎母音配列

形態素に付与される、Unidicでの「発音」タグのカタカナを、その中に含まれる母音をAIUEO に、「ン」を N に、「ッ」を Q に、「ー」を - に置き換えたもの。
「開始」は、読みでは「カイシ」になり、母音配列は「カイシ」を元に「AII」になる。
「交差」は、読みでは「コウサ」になり、母音配列は「コーサ」を元に「O-A」になる。
「調査」は、読みでは「チョウサ」になり、母音配列は「チョーサ」を元に「O-A」になる。

◎モーラ数

形態素を日本語として扱った際のモーラ数のこと。具体的には「母音配列」の文字数。「X(聞き取り不可)」や「h(笑い声)」などは 0 モーラとする。

◎プロソディ

音声に高低や強さがあるかどうか。プロミネンスに近い概念であるが、特別強い発声で強調された形態素以外にも、いわゆる文末に音の上昇を伴って発声される形態素にも付く。本コーパスでは、文末の上昇イントネーションの際によく用いられる「？」などの記号を使わないためである。このため、完全にプロミネンスとイコールではない。1 語の中での上昇下降などを表すアクセントの概念は含まない。具体的には以下の要素となる。

！	音声が周囲に比べ明らかに強い
↑	音声が周囲に比べ明らかに高い
↓	音声が周囲に比べ明らかに低い
--	音声が周囲と同等の高さや強さ

◎重なり

2 者の音声が重なっている箇所のこと。重なり幅の単位は、「読み」の文字とする。表記形では漢字が混じるため、表記される文字幅と発音された音の数に誤差が出る。表音文字であるカタカナの場合、記述された文字数を、発声されたモーラ数の近似値と見ることができ、また「母音配列」よりも人間には見やすいため、精密さと可読性の双方で妥協できるレベルであるため、読みのカタカナを、重なりを記録する単位とした。

重なりのタグの中身は大きく 3 つに分かれる。書式は「0 or 1 or 2」＋「:」+「数字-数字」の形を取り、最初の「0 or 1 or 2」は、重なりの大まかな種類を表す。その形態素に、他者の音声が一切重なっていない場合は「0」、1 形態素が丸々重なっている場合は「1」、形態素の一部だけが重なる場合は「2」とする。「:」以下は重なりの種別が「2」の場合のみ使い、「何文字目」から「何文字分」重なるというルールとする。重なりの種別が「0」か「1」の場合は「:」以下は「0-0」とする。以下に、ルールと具体的な例を示す。

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])

◎誤用

日本語教育の基準での正誤のこと。「正」の場合は誤用無しという意味で「-」、「誤」の場合は誤用の下位分類で誤用のタグを付ける。市川(2010)を参考に「脱落」「付加」「誤形成」「混同」「位置」「その他」の他、別途設定した「発音」「-」の 8 つを用いる。同書を基準

とした理由は、既に誤用の分類をコーパスに持ち込むという先行研究を行っている、東京外国語大学で開発されている「オンライン日本語学習者作文コーパス(小柳・他 2010)」に準拠したためである。独自設定の「発音」の誤用は、「学校」を「かっこう」と発音するなど、誤形成のレベルではないが、各音素の発生に誤りが有った場合とする。あくまでも母語話者では生成しないレベルの誤用を扱うため、学習者の発話にしか付与していない。

◎語彙レベル

形態素が、新日本語能力試験の基準での語彙に該当するかどうか。該当する語彙であれば、レベルの番号とその語彙をタグとして付与する。例えば、N2 相当の語彙で「うまい」であれば、「2:うまい」とする。

基準は佐々木・他(2010)ほかとする。同書を基準とした理由は、現時点で、各語彙を日本語能力試験、新試験のレベル別に分けた、使用するに足る量のリストを提供する唯一のシリーズであるためである。同書で提示されている語彙リストを元に、コーパス内の全形態素をチェックし、該当する形態素に自動でタグ付けする独自開発のソフトにより付与する。リストが膨大なため、人手による修正は行っておらず、タグの精度は低い。

「語彙」であるが、リストにある「自転車置き場」や「削除する」は、「自転」「車」「置き場」や「削除」「する」など複数形態素になるため、連続する複数形態素の全てに同じ語彙タグが付く。

◎文型レベル N1
◎文型レベル N2
◎文型レベル N3
◎文型レベル N4
◎文型レベル N5

以上 5 つのタグは、新日本語能力試験の基準で各レベルとされる文型で使われている各形態素のことで、その文型見出しをタグとして付与する。基準は友松・他(2010)とする。同書を基準とした理由は、現時点で、各種文型を日本語能力試験の新試験に対応しレベル分けした唯一の基準であるためである。

同書で提示されている各レベルの文型リストを元に、コーパス内の全形態素をチェックして該当する形態素に自動でタグ付けする独自開発のソフトにより付与。その後全てを人手により修正。必ずしも見出し形そのままでなくても、友松・他で各見出しに提示されている例文に該当するものは、その文型であるとする。

6.4 コンテキスト - 行単位のタグの種類

行単位に付与する項目別のタグの種類を以下に示す。これらは行ごとに一定である。

タグ情報は、「発話者」「年齢」「性別」「出身国」「最長居留地」「母語」「性差」「来日時

期」「日本語レベル」「日本語学習法」「会話回数」「時間(以降)」「時間(以前)」である。

タグの一部は 5 章の表 1 に詳しい。以下、個別に説明する。

◎発話者

発話者の番号のこと。二桁の数字になり、「01」～「10」となる。

◎年齢

発話者の年齢のこと。「19」～「23」となる。

◎性別

発話者の性別のこと。「男性」か「女性」となる。

◎出身国

発話者の出身国のこと。学習者は国名、母語話者は都道府県名となる。

◎最長居留地

発話者が 5 歳～14 歳で一番長く住んでいた場所。学習者は国名、母語話者は都道府県名。

◎母語

発話者の母語のこと。「日本語」「中国語」「韓国語」となる。

◎性差

発話者と相手話者の性差のこと。「同性」か「異性」となる。

◎来日時期

学習者が日本に来た時期のこと。学習者のみに設定している。

◎日本語レベル

学習者の日本語のレベル。具体的な日本語能力試験の結果を保持している者はその級も記す。

◎日本語学習法

来日前までの日本語の学習法のこと。「大学の授業」か「独学」となる。

◎会話回数

会話を収録した回数の番号のこと。「1」～「5」となる。

◎時間(以降)

会話開始から何分目以降の発話かの指定のこと。「00:00:00.0」の形式で指定し、「時間:分:秒」となる。秒は 0.1 秒単位で指定できる。

◎時間(以前)

会話開始から何分目以前の発話かの指定のこと。「00:00:00.0」の形式で指定し、「時間:分:秒」となる。秒は 0.1 秒単位で指定できる。

6.5 単位の規定

前述の通り、形態素分けは基本的に UniDic に基づく。つまり、立命学習者コーパスでの形態素の単位は、国立国語研究所の提唱している「短単位」に近似したものとなる。全体を統一した基準とするため、単位として形態素を採用したが、本コーパス内で扱う項目は、本来は形態素を基準として分けられないものもいくつかある。

「プロソディ」は、それぞれの形態素が発音される際に「高い」「低い」「強い」かで、発せられた発話部分に付けているが、主に「プロミネンス」として強調されるものと、「イントネーション」として上昇などにより疑問の意味が加わる、などの効果の有るものが対象となっている。ただ、1 語の中で、何音目が高く、どこから下がるというものではないので、「アクセント」ではない。この場合、イントネーションであれば、「文末にかけて徐々に変化が加わる（猪塚・他 2003)」ものなどがあり、これは、当然形態素の単位で扱えない。また、プロミネンスも基本的に語の単位で扱うが、例えば「おばあさん」は、短単位では「お（接頭辞)」「ばあ（名詞)」「さん（接尾辞)」の 3 形態素に分かれるので、「おばあさん」が強調された場合、この 3 つの形態素にまたがってタグを付与しなければならない。声の高低強なので、決して完全に形態素の単位とは一致せず、場合によって形態素より大き範囲であったり、逆に形態素よりも小さい範囲で出現したりする。

また、各文型レベルであるが、例えば N4 レベルの文型の、「たことがある＜経験＞」の場合、発話中に「それやってみたことがあるよ」という表現が有った場合、「た」「こと」「が」「ある」の 4 形態素全てにタグが付けられる。

本来は形態素よりも小さい範囲で現れる事柄の場合、検索や統計では「それを含む形態素」という、それよりも大きな範囲でしか扱えず、逆に形態素よりも大きい範囲で現れる事柄の場合、検索や統計をした際には「その範囲にある全ての形態素」でしか扱えない。例えば、1 つの文型である「というのは」が、スクリプト内に表れる場合、「と」「いう」「の」「は」と、4 形態素として扱われる。文型としては 1 つだけ存在するに過ぎないが、形態素では 4 つに分かれるため、形態素を基準とした検索では 4 つの形態素がそれぞれヒットし、4 例検索されたように扱われる。

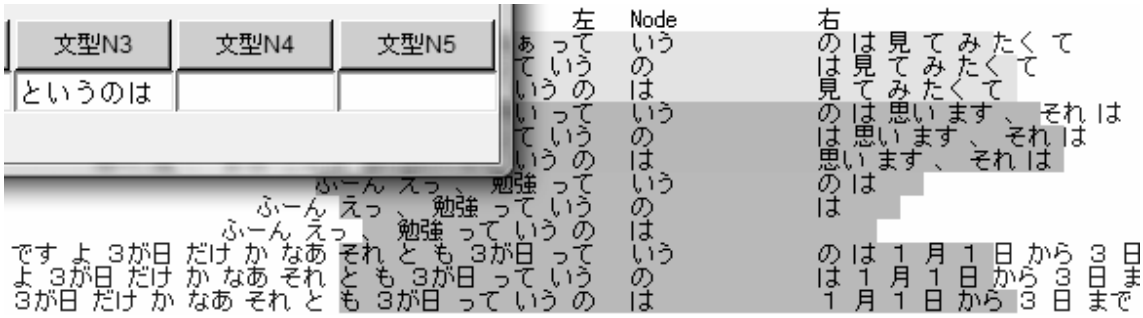


図 14.KWIC、文型 N3「というのは」検索

このように、本来は複数の単位が混在するはずのデータであるが、データの扱いに統一性を持たせるため、1 つの単位で統一的に扱わざるを得ない。それを前提とする場合、比較的細かい単位の方が融通が効きやすい。しかし「単音」や「音素」単位まで細かくなると、文法その他の項目が扱いにくくなる。以上の理由から、あくまで「言語」を扱うコーパスの基準として、文法項目などが扱いやすい「形態素」を単位として採用した。

7. ツール

コーパスデータを詳細に扱うにはコンコーダンサーが欠かせないが、6 章の方式で記録したデータを誰にでも簡易に扱えるように、専用のコンコーダンサーを付属した。使用したコンコーダンサーの原型は「HASHI」で、この上で実装させたため、多項目でありながら極めて簡単に扱うことができるというコーパスとなった。

統計や数学を基準に使うコーパスではなく、生データを目で見ることをメインに置いて設計しているので、検索を細かく使い、必要な情報を効率良く抽出し、一工程一工程自分で行い、目的のデータの核にアプローチするように使うことができる。そのため、データで利用できる項目の全てを使用者が直接選択、選別でき、想定できるあらゆるニーズに対応できる機能を用意した。そして、これらは全て使用者が自分で選択し、検索でき、結果を自分の考えた通りに絞り込め、表示を見たい通りに切り替えることができる。これによって、データのあらゆる要素をあらゆる視点から見ることができ、自在に操ることができる。高度なソフトが何か素晴らしいことをしてくれるのではなく、使用者自らが、直接データを自分の目で見、自分の手で触れ、目的とする箇所を抜き出してくる。その瞬間に思いついた興味やニーズを、瞬時に確認できる。そして、その結果を見て、より詳細に焦点を絞ることもでき、また、見るべきポイントの違いに気づけば、すぐに、より核心に近づくように修正できる、正に使用者の目となり、繊細な指先となるようなソフトとして設計した。このコンコ―ダンサーの付属により、初めて本コーパスの真価を発揮できるようになる。

8. データについて

本章では、8.1 節で、本コーパスに収めたデータの使用上の注意点、8.2 節で、データの選択理由、8.3 節で、データの持つ利点を提示する。

8.1 注意点

1 章で提示した注意点であるが、立命学習者コーパスではデータの量が少なく、また、種類が少ないという点に注意する必要がある。具体的には、サンプルの絶対数が少ない、学習者の母語の種類が少なく偏っている、学習者の日本語レベルも比較的高レベルの学生に偏っている、という 3 点が挙げられる。

サンプル数の少なさは、データに現れた現象がどの程度一般性を持つものかの判断に大きく影響する。個別の事例が、調査対象の言語使用者全般にわたるものか、単に個人的傾向に過ぎないものなのか、判断しづらいということである。

学習者の母語の種類の少なさは、サンプル数の少なさと同じ問題も含むが、それに加え母語のバランスの問題がある。今回の協力者は母語別に、中国語 3 人、韓国語 2 人を収録している。日本語学習者は一般に、特にその 2 言語の話者が圧倒的に多いという状況を反映してはいるが、それ以外の言語の母語話者のデータは収録されていないので、本コーパスを使用することでそれらのデータを見ることはできない。コーパスの設計上扱えることであるが、データの数によりそれが制限されているといえる。

学習者の日本語レベルも偏りがある。具体的には旧日本語能力試験 2 級から新試験 N1 である。本コーパスでは、話者の使用する文型や誤用なども扱えるため、検索によって、どの文型にどの誤用が多くあるかなどを抽出できるものの、データの学習レベルが偏っているため、学習段階ごとの誤用の傾向や、使用文型の分布は見ることができない。

本コーパスを使用する際には以上の点に注意する必要がある。

8.2 データ配分の理由

上記した点は、そもそもがデータのサンプル数自体の少なさのために起こるものである。これは、1 つのデータに多くの作業時間をかけて、様々な情報を付与するという設計思想からくるものであり、この方向でのコーパスを志向した以上、必ず起こることであり表裏の特徴であるとも言える。この際、もしサンプル数が少ないのにも関わらず、母語、学習レベル、性別など、様々な要素をそれぞれ均等に分配した場合、全ての話者のあらゆる属性がバラバラになり、共通点が無くなってしまう。データに表れた特徴が起因する要因を探ろうとしても、識別する材料が散らばり過ぎ判断できなくなる。逆に、話者のほとんどの属性を合わせた上で、1 つだけの属性が違うのなら、それを、発見した特徴の根拠とすることが可能であるが、複数の属性が全て違う場合、どの属性が起因して起こったのか分からなくなる。この点から、データの各属性にあまり幅を持たせないということは、現実的に、

本コーパスを使用しうるに足るものにしていると言え、次善の策であると言える。また、これらの要因を抱えていたとしても、このコーパスの設計自体は革新的であり、それを補って余りある利点をもたらすものと確信している。

8.3 利点、縦断的データ

上記の点を解決するには、データのサンプル数を増やすことが根本的な策であり、それ以外に方法はないと言える。しかし、今回は敢えてその方針は取らなかった。サンプル数を増やすことよりも、決まった組み合わせのペアを定期的に複数回追いかけ、縦断的にデータを取ることとした。非常に手間と時間がかかる今回のコーパス設計の場合、いたずらにデータの種類を拡散させるより、有意義な一点を追求する方が、効果が高いと思われるためである。

現状有る学習者コーパスの中で、この縦断的なデータというのは非常に珍しいものであり、高いオリジナリティと、データの価値であると言える。これにより、今までなかなか見ることのできなかった日本語学習者の発達段階や、ペアの関係性の変化などを見ることができるようになった。本コーパスにおけるデータの特殊性、優位性はまさにこの点にあると言える。コーパス構築の設計と並び、この縦断的データという点も、今後の様々な研究に非常に大きく貢献できる点であると言える。

9. コーパス編纂作業からみたコーパスの姿

筆者は短い研究生活の中で、幸運にも 4 つのコーパスの編纂に関わることができた。更にそのうち 2 つは、自分でコーパスデザイン、作成、作業スタッフの運営管理を行うことができた。そこで得た知見は、コーパスというのは一見とても機械的であり、設計思想や概念をそのまま実体化させたもののように見えるかもしれないが、実はとても人間的なものであるということである。

コーパスを設計するということは、実際に発話される中にある豊富な情報を、どう記録するかということである。発話には語彙のほかに、それらの文法情報、音声情報、非言語情報、更にその発話が内包する様々な社会言語学的情報や談話情報などのあらゆる情報を含んでいる。そのうちどれを収録しどれを除くか、その選択が必要である。これはコーパスを使用する目的に沿って取捨選択するものであるが、その根本には、コーパスの作者が言語のどの側面を見ているか、言語をどう捉えているかがある。日本語教育の視点からで言語を見る者は第二言語習得や言語喪失などの側面を重視するかもしれないし、社会言語学の視点から言語を見る者はスピーチアコモデーションやスタイルのシフトを重視するかもしれない。つまり、コーパスというものは、その作者の言語観というフィルターを通した言語の姿とも言える。

また、編纂作業の面からいうと、文字化から各種タグを付与する作業まで、実際の作業は多くを人手に頼らなければならないということがある。特に機械処理による一斉タグ付与で起こる様々なミスの修正や、機械ではできないきめ細かい編集に必要な、極めて大切な判断は人間のスタッフが行うしかない。重要な作業をするスタッフはみな人間である。それぞれに知識レベルや信条、作業環境などの様々な違いがある。それにより、作業スタッフによって同じデータを扱うとしても、判別基準や作業内容に若干のずれが生じる。それに加え当然うっかりしたミスもするし、思い違いによるミスもある。運営者はそれの調整をし、基準を設け、スタッフ間で見解の統一を促さねばらない。スタッフの性格的な違い、研究の専門などの背景的な知識レベルの違い、作業にかけられる時間の違いなど、様々な要素を把握した上で適材適所での作業分配をしなければならない。運営者にとってコーパスの作成をするというのは、ほとんどの仕事が、スタッフをどうケアするかということに尽きる。それによってコーパスを完全に１つの統一した基準によって構築することを目指す。しかし、どんなに基準の統一を図ろうとしても、自然言語の完全な文法構築が不可能なように、それを完全に行うことはできない。誰が作業をしたとしても判断が非常に難しい要素は、データ内に散在している。全員が完全に納得する作業基準の根拠の提示は難しく、作業スタッフにより編集結果に微妙なずれが生じる。各スタッフが作業を繰り返すことによって、自分の中での基準を少しずつ明確にし、それをコーパスプロジェクト自体に還元する。それにより判断の難しかった要素を１つずつクリアし減らしていく。この繰り返しでコーパスの質も「成長」していく。最初から完璧な基準ですべての作業を通すことは現実的には非常に難しい。このことからもコーパスというものは極めて人間的であるといえる。理想的な言語使用者というものは存在しえず、実際には個別の言語使用者しかいないように、ラングは概念の中にしか存在しえなく、全ての言語は実体化する際にはパロールにならざるをえないように、コーパスもデザイン段階では概念的であり理念であるが、編纂され実装される段階では「実体」にならざるをえない。

あくまで人間がコーパスを作り、作成されたものは理念とはイコールになりえない「実体」である以上、その使用者も、そこから得られるデータを、絶対のものや理想化されたものと見なすことはできない。コーパスの使用には必ず注意が必要であるが、それはコーパスのデータがどういう種類のものであるか、コーパスデザインは何か、というマニュアル的なものだけでなく、そのコーパスに付けられている様々なタグがどのような信条、方針で付けられているのかという部分にまで及ぶ。実際に使用された発話（文）の集合がコーパスであるので、その中に収められている発話データの内容は、発話した人間を反映したものであるが、「実体」として作成されたコーパス自身もまた、作成した人間を反映したものである。多くのスタッフの様々な労力や考えが込められたものであり、そこに魅力があると言えるし、使用者にとっては注意しなければならない点であるとも言える。コーパスを使用する際には必ずそれらを前提し、留意する必要がある。

10. まとめ

今回、「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」に基づく立命学習者コーパスの作成で、言語研究におけるコーパスの新しい使用法を提案した。本稿は、その仕組み、定義、設計思想を示したに過ぎず、本格的な使用や研究はこれからである。本コーパスで行える研究は多様で、学習者発話の品詞の傾向、誤用しやすい文型、母語話者と学習者の発話時間の差、学習者のプロソディ、など様々である。これらは全て本コーパスに予め付与されたタグで対応出来、簡単な検索で誰でも調べられる。他にも工夫次第で、はるかに高度で複雑なことが行える。使用者の多彩な興味をカバーする程、本コーパスの潜在的能力は高い。研究者、教育者の別、コーパスへの習熟度などを問わず、むしろ全くコーパス経験の無い多くの人にこそ本コーパスを大いに使用されることを心より願う。

【参考文献】

庵功雄・高梨信乃・中西久美子・山田敏弘(2000). 『初級を教える人のための日本語文法ハンドブック』 スリーエーネットワーク.

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.

猪塚元・猪塚恵美子(2003)『日本語教育トレーニングマニュアル① 日本語の音声入門 解説と演習＜全面改定版＞』バベルプレス.

小柳昇・テレンス・シャア(2010)「オンライン日本語学習者作文コーパス・誤用コーパスの作成と日本語教育への活用」『2010 年度 日本語教育学会秋季大会予稿集』. p9.

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』. pp198-201.

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』. pp190-193.

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.

中村純作(2010)「中村純作教授退職記念講義 何故、コーパスか？」立命館大学.

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.

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)形態素解析システム『茶筌』 version 2.4.0. http://chasen.naist.jp/hiki/ChaSen/