日本語学習者会話データベースの利用手引き

（平成 22 年 5 月　国立国語研究所）

目　　次

１．データの概要

このデータベースの目的は，日本語教育研究，言語習得研究，会話分析など，研究活動のための基礎資料として「外国語母語話者と日本語母語話者の生の会話データ」を提供することにある。ここで公開するデータは，文字化データ 339 件とそのうちの 215 件の音声データである。音声データの録音時間は 1 件あたり 30 分程度である。

１.１．データ項目

日本語教育ネットワークの Web ページ[1]では，以下のようにデータ番号とインフォーマントである日本語学習者の日本語レベルや属性など 10 項目を示し，検索ができるようになっている。

【検索画面】

(1) データ番号
(2) OPI レベル
(3) 年齢
(4) 性別

[1] http://dbms.ninjal.ac.jp/nknet/ndata/opi/　このデータの利用には「日本語教育ネットワーク」への会員登録が必要である。

（5）出身国
（6）母語
（7）職業等
（8）日本滞在期間
（9）日本語学習期間（参考）
（10）日本語能力試験合格情報（参考）

データ項目中のインフォーマント属性情報は原則として本人の記述のままであるが，選択肢の部分はカテゴリーごとに分けるなどの整理をした。インフォーマントカードの記載に不確実な箇所がある場合には，想定される情報に「?」を付け加えている。未記入などで不明な場合は空白とした。

なお，（9）の日本語学習期間については，正確に学習の期間や頻度を測ることが難しいことに加え，設問を「日本での日本語学習期間」と誤解して申告したと思われる場合があることから参考項目とした。（10）日本語能力試験合格情報も記載があった場合のみ示し，また合格年の情報を得ていないためこちらも参考項目とした。

１.２．評価（OPI レベル）

評価の方法としては，American Council on the Teaching of Foreign Languages（ACTFL:全米外国語教育協会）[2]の Oral Proficiency Interview（OPI）という面接式口頭能力テストの方式を用いた。OPI とは，面接者（テスター）が学習者に対して，定められた方法により 30 分以内のインタビューを行い，そのレベルを判定するものである。

このプロジェクトでは，試験時の面接者（テスター）が最初の評価（ファーストレイティング）を行い，その後すべてのデータについて別のテスター認定者による二度目の評価（セカンドレイティング）を経ている。両者が一致しない場合はさらに三度目の評価（サードレイティング）を行った。結果として，2 つのレベルが付いたデータに関しては，下の方のレベルを示した[3]。レベルが 3 つに分れたデータについては，判定不能（アンレイタブル）と考え，公開の対象としなかった。

OPI では，判定尺度として，超級（Superior），上級（Advanced），中級（Intermediate），初級（Novice）を設けている。超級以外の級では，それぞれの級をさらに上中下に下位分類している。したがって，全体としては「超級」から「初級の下」までの 10 段階の評価尺度があることになる。

各級の判定の基準[4]を以下に示しておく。

---

[2] http://www.actfl.org/
[3] OPI 研究会の判断による。
[4] 牧野成一他（2001）『ACTFL-OPI 入門』アルク　P18，19 より転記の上整理。

【概略】I

超級　：（機能・タスク）裏付けのある意見が述べられる。仮説が立てられる。言語的に不慣れな状況に対応できる。（場面／話題）フォーマル／インフォーマルな状況で，抽象的な話題，専門的な話題を幅広くこなせる。（テキストの型）複段落

上級　：（機能・タスク）詳しい説明・叙述ができる。予期していなかった複雑な状況に対応できる。（場面／話題）インフォーマルな状況で具体的な話題がこなせる。フォーマルな状況で話せることもある。（テキストの型）段落

中級　：（機能・タスク）意味のある陳述・質問内容を，模倣ではなく想像できる。サバイバルのタスクを遂行できるが，会話の主導権を取ることはできない。（場面／話題）日常的な場面で身近な日常的な話題が話せる。（テキストの型）文

初級　：（機能・タスク）機能的な能力がない。暗記した語句を使って，最低の伝達などの極めて限られた内容が話せる。（場面／話題）非常に身近な場面において挨拶を行う。（テキストの型）語，句

【正確さ】

超級　：（文法）基本構文に間違いがまずない。低頻度構文には間違いがあるが伝達には支障がない。（語彙）語彙が豊富。特に漢語系の抽象語彙が駆使できる。（発音）だれが聞いてもわかる。母語の痕跡がほとんどない。（社会言語的能力）くだけた表現もかしこまった表現もできる。（語用論的能力）ターンテイキング，重要な情報のハイライトの仕方，間の取り方，相づちなどが巧みにできる。（流暢さ）会話全体が滑らか。

上級　：（文法）談話文法を使って統括された段落が作れる。（語彙）漢語系の抽象語彙の部分的コントロールができる。（発音）外国人の発音に慣れていない人にもわかるが，母語の影響が残っている。（社会言語的能力）主なスピーチレベルが使える。敬語は部分的コントロールだけ。（語用論的能力）相づち，言い換えができる。（流暢さ）ときどきつかえることはあるが，一人でどんどん話せる。

中級　：（文法）高頻度構文がかなりコントロールされている。（語彙）具体的で身近な基礎語彙が使える。（発音）外国人の日本語に慣れている人にはわかる。（社会言語的能力）常体か敬体のどちらかが駆使できる。（語用論的能力）相づち，言い換えなどに成功するのはまれ。（流暢さ）つかえることが多いし，一人で話しつづけることは難しい。

初級　：（文法）語・句のレベルだから文法は事実上ないに等しい。（語彙）わずかの丸暗記した基礎語彙や挨拶言葉が使える。（発音）母語の影響が強く，外国人の日本語に慣れている人にもわかりにくい。（社会言語的能力）暗記した待遇表現だけができる。（語用論的能力）語用論的能力はゼロ。（流暢さ）流暢さはない。

その他　下位レベルを決めるときは，すぐ上の主要レベルの特徴をどれだけ維持できるかで判断[5]。

１.３．音声データ

音声は録音時の MD から変換した mp3 形式のデータである。音声データと文字化データの対応関係を示すため，文字化テキストには「★02★」「★04★」...のように 2 分ごとに時間情報を挿入した。ただし，再生ソフトによっては数秒の時間差が生じることがある。また，インタビューの前後に，直接関係のない不要な会話がわずかに入っているケース等もあるので，ご注意願いたい。

１.４．文字化データ

文字化にあたっては，当初，さまざま研究を行う際の最も基本的なデータを提供するという意味から，一般的な日本語母語話者が聞いた音声そのままをできうる限り文字にするという漠然とした方針を立てて聞きとり作業を開始した。

しかし，予想されたこととはいえ，ポーズの長さやイントネーション，発話の重なり等，もともとプレーンな形では表現しづらい要素に加え，中間音的な音声など，日本語の限定された文字では表しがたいものも多く出てきた。外国語母語話者による日本語での会話という特性もあり，結局どのように文字化を行っても聞き手による表現の差は存在し，また同一の聞き手であっても毎回同じように聞き取るとは限らないことが作業が進行するにつれてはっきりしてきた。

このプロジェクトの目的が，汎用的な基礎データを提供するというところにあったため，どのような研究目的に利用されるかという想定が弱く，このことがひいては文字化方法の決定に多くの時間を要する原因となった。どの要素に注目して何に関する研究を行うかを十分に検討した上で文字化を行うことの重要性を再認識した次第である。

最終的には，（1）録音された音声データ自体も提示する（2）文字化は音声の一つのサンプルとして提示する（3）日本語母語話者が聞こえたままを文字化するという当初の方針は維持する（4）文字化は一般的で内容が把握しやすいものとする（5）文字化担当者がその語の意味を主観的に判断した可能性がある場合はできる限りそれを示す，という方針のもと，文字化の規則を作りなおした。これらは，作業の進行に伴い修正したものもあるため，表記等に関しては，データごとにある程度のばらつきがある可能性がある。この場合でも，同一データ内ではできる限り統一するようにしている。文字化の規則の詳細に関しては４．をご覧いただきたい。

２．調査方法

２.１．インフォーマント（データ提供者）の構成

[5] 牧野成一他（2001）『ACTFL-OPI 入門』アルク P. 26 より。

収集された 390 件のデータのインフォーマント属性[6]については，以下に示す通りである。

■年齢

年齢別では 20 歳～24 歳が 43%，25 歳～29 歳が 35%であり，20 代がおよそ 3/4 を占めている。次いで 30 代（13%）10 代（7%）の順となった。性別では，男性が 31%女性が 66%で男女比は 1：2 であった。

■母語

日本語学習者の母語別でみると，韓国語が 53%，中国語 17%，英語 8%で，韓国が約半数を占めている。

■所属・職業等

所属・職業等では，日本語学校生 53%，大学・大学院生 27%，会社員 4%と，大部分が日本語学校や大学の学生であった。

■滞日期間

日本滞在期間では，「3 か月以上 6 か月未満」（23%）と「1 年以上 2 年未満」（23%）及び「6 か月以上 1 年未満」（22%）がほぼ同数であり，この 3 つを合わせると，日本滞在が 3 か月以上 2 年未満が 68%となる。次いで 3 か月未満の滞在者が 17%あった。

■日本語学習期間

日本語学習期間については，前述したように参考項目ではあるが，その範囲では 3 年以上（26%），2 年以上 3 年未満（20%），1 年以上 1 年 6 か月未満（17%），6 か月以上 1 年未満（16%）となっている。

２.２．データ収集の方法

データ収集の際に利用したのは Oral Proficiency Interview（OPI）という面接式口頭能力テストの方式である。

■OPI とは

OPI とは，30 分以内という時間制限の中で面接者（テスター）と外国語学習者が，学習者に身近な話題に基づいた会話を行い，その会話能力を判定する口頭試験である。会話展開の方法やレベル判定について厳密な基準が定められており，学習者のレベル判定者であるテスターに関しても ACTFL による認定が必要であること，またその質を維持するための方策が取られているなど，より均一なデータと安定したレベル判定が得られると考え，この方式を採用することとした。

なお，同じ OPI の方式を用いて収集したデータとしては，すでに「KY コーパス」[7]というデータがある。

---

[6] 詳細は「資料」を参照のこと。なお，「資料」では公開データではなく，収集されたデータ 390 件すべてのインフォーマントが母数となっている。

[7] 90 件の文字化とインフォーマントの日本語レベルが示されている。作成者である鎌田修氏（南山大学教授），山内博之氏（実践女子大学教授）の頭文字をとって「KY コーパス」と称される。当プロジェクトのデータ収集依頼先である「日本語学習者会話コーパス研究会」のメンバーである。

■データ収集の時期と場所

データの収集は，OPI の専門家による研究会である「日本語学習者会話コーパス研究会」[8]に依頼して行った。データ提供者である日本語学習者（インフォーマント）及び面接者（テスター）の選定，調査場所の選定もすべて同研究会による。

390 件のデータの収集時期は以下の通りである。調査場所は，東京 256 件，京都 46 件，大阪 28 件，名古屋 26 件，高知 20 件，神戸 14 件であった。データは MD で録音され納品された。

2007 年 2 月～4 月（データ番号 001～090）
2007 年 7 月～10 月（データ番号 091～186）
2007 年 11 月～2008 年 1 月（データ番号 187～390）

３．著作権について

収集された 390 件のうち，文字化・音声とも公開の許諾を得られたものは 295 件，文字化のみの許諾は 95 件[9]であった。

収集に際してはインフォーマント及びテスターに対し，データの利用許諾を得ている。承諾書の文面は，Web 上での公開を念頭に法律の専門家と相談の上作成した。承諾書は，韓国語，中国語，英語，ブラジルポルトガル語の翻訳版も用意したが，その文面については結果として高度に法律的なものとなり，インフォーマントである日本語学習者が理解しにくいものであったことは否めない。本来であれば文面をよりわかりやすい日本語にした上で提示すべきであった。この点は大きな反省材料である。文面のわかりにくさに対応するため，データ収集にあたっては，テスターから口頭で十分な説明を行ってもらうことにしたが，限られた時間内で調査を終えなくてはならないというデータ収集時の状況を考えると，必ずしも十分な説明が行えたかどうかは定かではない。

こうしたことから，このプロジェクトでは，承諾は得てはいても，(1) データを使用できる者を限定する（会員登録制）(2) データ提供者が特定できないような措置を講ずる (3) 音声の公開にあたっては，その旨をデータ提供者に事前に通知する (4) 以上の結果として，データ提供者からデータ取り下げの申し出があった場合はただちに取り下げるというさらなる段階を設けた。

４．データ整備方法

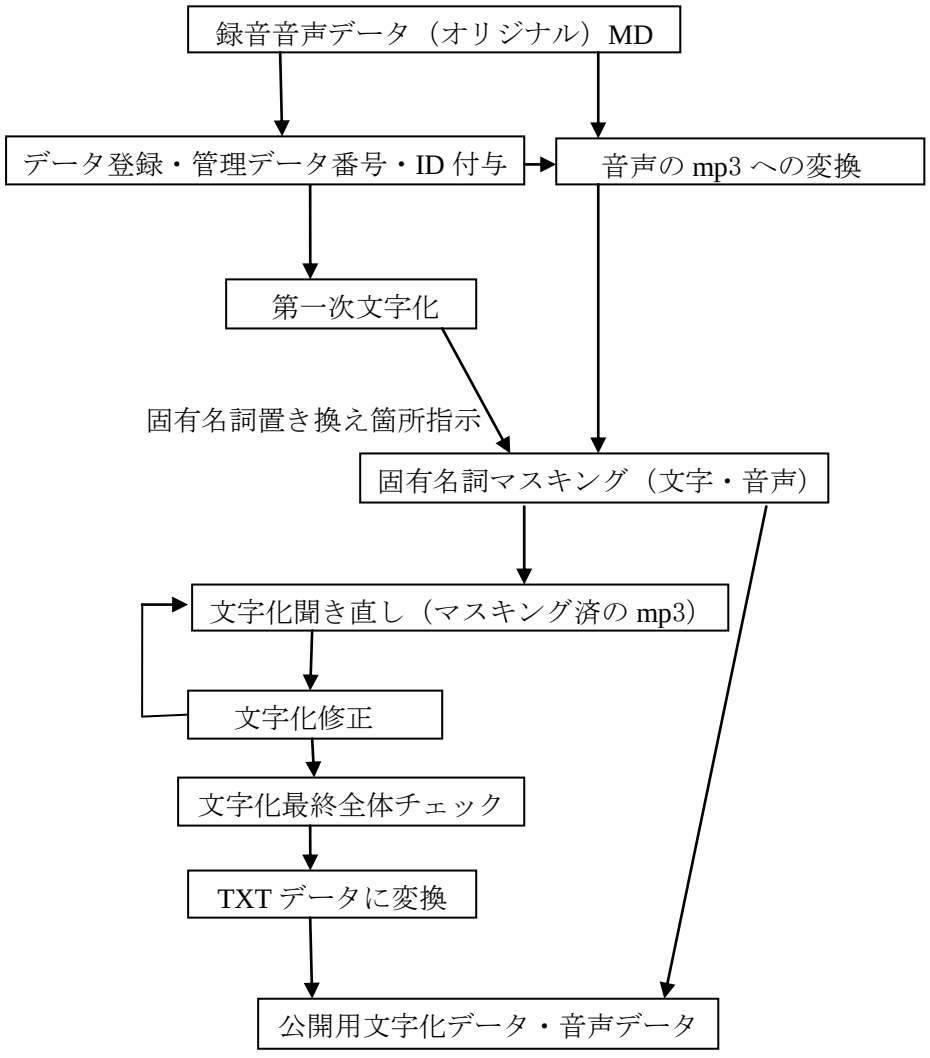
４.１．データ整理の流れ

[8] 代表：鎌田修，事務局長：嶋田和子（イーストウエスト日本語学校副校長），分析委員：山内博之

[9] 2007 年 2 月～4 月に収集した 90 件に関しては，そもそも承諾書に音声公開についての許諾項目を設けていなかったため文字化のみの許諾となった。他の 5 件に関しては，インフォーマントが「文字化のみの公開」を選択したものである。

文字化は以下の手順で行われた。まず，「一般的な日本語母語話者が聞いた通り」という方針のもと，文字化専門業者に委託し，でき上がったものを第一次文字化とした。このデータを，文字化に対して専門的知識・経験を持った者が聞き直しを行って確認したが，この時点で第一次文字化の修正箇所が予想以上に多いことがわかったため，同じデータについて，別の熟練者による表記等形式的な修正も兼ねた二度目の聞き直しを行うことにした。結果的に聞き取りは，外注業者も含め 3 人の作業者が合計 3 度行ったことになる。その後担当者が全体確認と調整を行い，必要に応じて 4 度目の聞き取りを行った後，Web での公開用にテキストデータ化した。データ整理の流れは以下の図に示す通りである。

また，mp3 に変換した音声データについては，データ提供者のプライバシー保護の面から，個人が特定できる可能性のある箇所にホワイトノイズ（ピー音）をかぶせてマスキングした。さらに聞き直しの際にその箇所が正しくマスキングされているかどうかを確認している。



４.２. データの整備

納品された音声データ（MD）については，データ番号や ID 番号の付与等必要な整備を行い，その情報を管理用ファイルに登録した。これらの情報は文字化の際のテンプレートにもヘッダーとして記載した。音声データ及び文字化データは，作業段階での間違いを防ぐため，データ番号及びインフォーマント名をファイル名とし，作業の最終段階まで残したが，公開の時点ではデータ番号のみを残してすべて削除している。ただし，作業の段階では，外注業者及び整理補助者がこの個人情報を目にすることになるので，文字化を担当した会社との契約事項に必要項目を盛り込み，また個々の作業者に対しては個人情報に関する遵守事項についての念書を取った。

４.３. 固有名詞の置き換え

一般的に考えて，それによりインフォーマントやテスター個人が特定できる可能性が高い情報は，以下の通り文字化データの該当箇所をアルファベットに置き換え，音声はマスキングした。

４.３.１. 置き換えをする場合

置き換えを行うのは以下の場合である。

(1) 個人名，個人電話番号，個人住所，最寄りの駅名等
(2) 個人が所属している学校名，会社名，店名
(3) 個人の出身地などで，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。ただし国名は置き換えないこととする。
(4) ロールプレイで，テスターや友人等，実在する人物の個人名，電話番号，住所等を使用していると思われる場合
(5) 聞き取り不能箇所（「＊＊＊」で示される）が，上記に該当しそうな場合は，聞きとれないものでも念のために置き換え対象とする（この場合カテゴリーの通しアルファベットは付けず【不明】とする）。

以上に該当しない固有名詞等は，原則としてそのままとした。映画名，俳優名，本のタイトル，著者名等もそのままとしたが，ビル名，会社名，バス停名，駅名，都市名等で，状況により個人が特定できそうな場合には例外的に置き換えている場合がある。

なお，聞き直しの作業段階で新たにこれらに該当する固有名詞が見つかった場合は，個人が特定される可能性が低ければ文字化のみを訂正し音声はそのままにしていることがある。

４.３.２. 置き換えの法則

アルファベットの置き換え法則は以下の通りである。

(1) テスター（T：） → Aで統一

姓名 → 【姓名A】

姓のみ → 【姓A】

名前のみ → 【名A】

ニックネームは【名A】扱い。特に必要があれば【名Aの別称】【姓Aの別姓】などとしている場合もある。どちらかわからないときは【姓？名？A】，ほぼどちらかと分かる場合は【名？A】とする。(2) も同様。

(2) インフォーマント（I：） → Bで統一

姓名 → 【姓名B】

姓のみ → 【姓B】

名前のみ → 【名B】

(3) その他

テスター，インフォーマントには必ず「A」と「B」を用いるが，それ以外の置き換えは，種類にかかわらず「C」から順に通しのアルファベット（全角）を用いる。

大学名 → 【大学名C】

日本語学校名 → 【日本語学校名D】

会社名 → 【会社名E】

店名 → 【店名F】

料理店名 → 【料理店名G】

→ 【レストランH】

地名 → 【地名I】

駅名 → 【駅名J】

以上を原則としたが，詳細なカテゴリーを必要とする場合は，状況により【旅行代理店名】，【韓国料理店名】【ファストフード店名】【支店名】などの下位カテゴリーで示した場合もある。当該固有名詞が略称で用いられた時も，確実な場合には併せて置き換え，必要に応じて【○○の略称】のように示した。OPI のロールプレイ中に出てきた会社名や人名等で，確実に架空のものと考えられる場合は置き換えていない場合もある。

また，一括して置き換えても会話内容に影響しない場合はまとめていることがある。

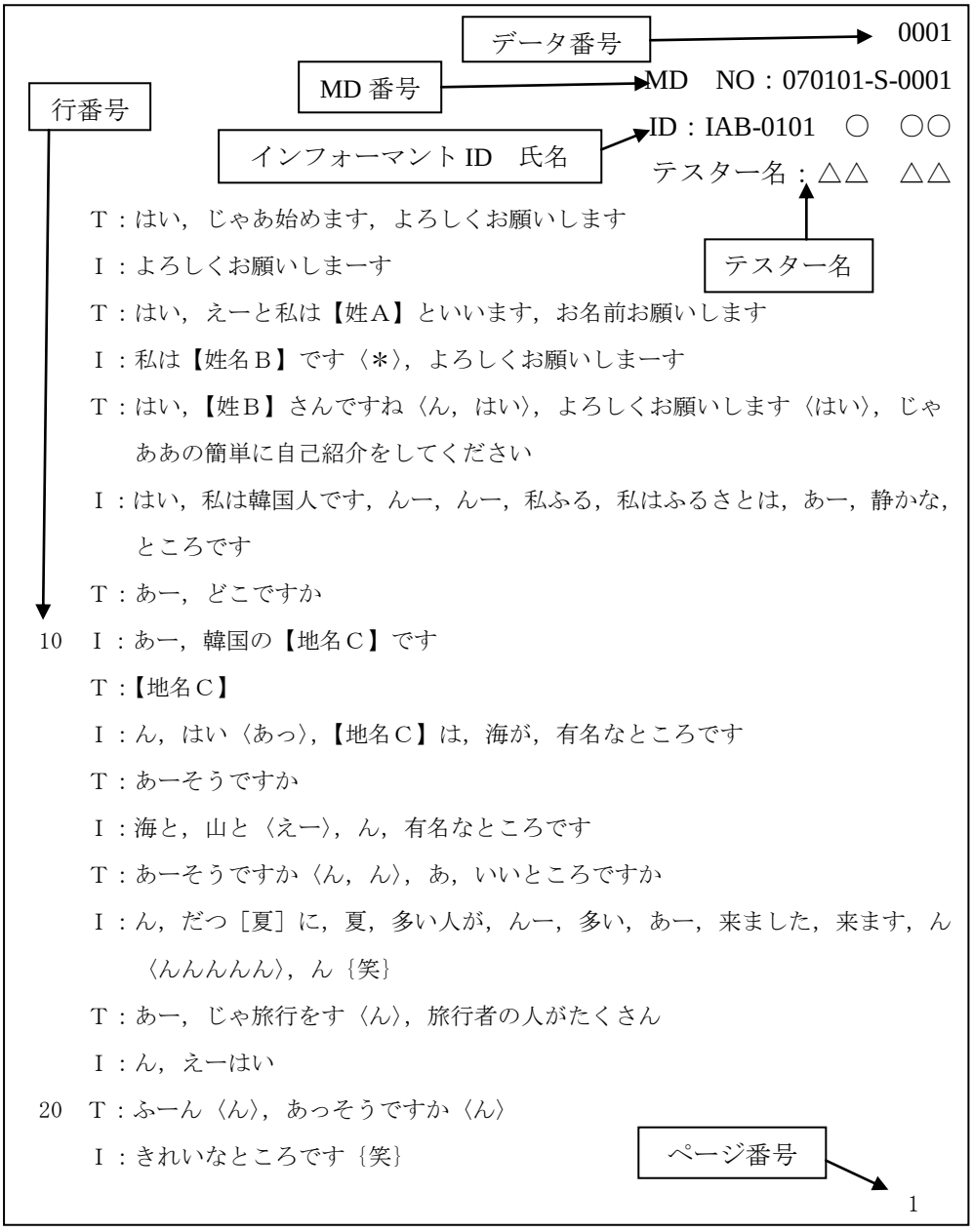
例：「03 の****の****」（電話番号）→【電話番号E】

言い間違いや一部のみを言っている場合は，【姓Aの言い間違い】【姓 B の一部】のように示した。

４.４. テンプレート

文字化を行うにあたっては，Word のテンプレートを用意した。テンプレートにはデータ番号等をヘッダーとして右肩に入れた。また各ページにはページ番号，各ページには 10 行ごとに行数を入れている。その他，字数・行数の設定等はすべて規定通りとした。なお，ID や氏名等の個人情報は，最終的に Word からテキストデータへの変換する段階ですべて消去した。

【文字化用テンプレート】



４.５. 文字化の基本方針

文字化の基本方針は以下の通りである。ひらがな・カタカナの使い分け等については,４.６.を参照していただきたい。

【形　式】

(1) 行番号：表示しない。

(2) 発話文の定義：文の単位は考慮しない。文末を示すことになる「。」や疑問符「？」は用いない。

(3) 改行：話者交替で改行。発話の主導権をどちらが持っているかをおおよその基準とするが厳密には定めない。

(4) 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。いずれも全角大文字で示す。

テスター（面接者・日本語母語話者）→　Ｔ：

インフォーマント（日本語学習者・外国語母語話者）→　Ｉ：

(5) イントネーション：表記しない。そのため，相当に癖のあるイントネーションであっても文字上では表われないことになる。

(6) 間（ポーズ）：長短に関わらず「,」1つで示す。

(7) 発話の重なり：示すことができないので，別々の発話として立てるか，もしくは相手の発話中に分割して割り込ませる。

(8) 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。あいづち等の語形は，４.６.４ に示すように統一する。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，別々の発話として立てることも可とする。

(9) 非言語行動等：笑いや発話に関係しそうな非言語行動は｛　｝で示す。

(10) 長音：長さにかかわらず「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ない。

(11) 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[10]。

(12) 聞きとり不能箇所：聞きとり不能箇所に関しては，おおよその音節数を＊の数で示す。

【表　記】

(1) 全体：表記は，発話内容の把握のしやすさに考慮し，一般的な漢字かなまじり文とした。表記することが難しい音についても，同一ファイル内ではできるだけ統一するようにする[11]。

[10] この結果, 全角アルファベットがテキスト中で用いられるのは, インフォーマント, テスターを表す「Ｉ」「Ｔ」の記号と「４.３.２」で示した固有名詞を置き換えた場合のみとなる。

(2) 補足記号［　］の使用：通常と異なる発音がなされた場合やポーズ等が挿入されて漢字で表せない場合はひらがなで表記し，その意味と思われる語を［　］で補足する。また，その漢字に対し 2 つ以上の読みが考えられる場合もひらがな表記し［　］で漢字を補足する。この場合，前後関係から読みが特定できる場合は補足しないこともある。

例：おと，さん［おとうさん］
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］
いちがつにじゅうよんにち［1 月 24 日］
かよって［通って］

(4) 短縮形：「大丈夫（だいじょうぶ)」に対して「だいじょぶ」などの短縮形がある場合は少なくも短縮形に関しては漢字表記しない。

(5) インフォーマントが質問の意味がわからず，テスターの言葉を単に繰り返していると思われる場合はひらがなで示した。

例：T：あなたは，落し物をして交番に行きました
I：こ，こうばん，こうばんってなんですか
T：警察，警察に行きました

(6) 助詞かどうか判断できない「wa」「o」「e」の音は，「わ」「お」「え」と表記する。

(7) 引用発話：「　」で示す

(8) 書籍等タイトル：『　』で示す

4.6. ひらがな・カタカナ等の表記規則

ひらがな・カタカナの使い分けもできる限り統一するようにした。また，送り仮名は用字用例辞典に従った。記載がない場合はひらがなが多い方を採用し，同音異義語も用字用例辞典に従っている。基本的に本動詞と補助動詞があるものは，本動詞は漢字，補助動詞をひらがなとした。

4.6.1. ひらがなで表記するもの

ひらがなで表記するものの例を以下に示す。

(1) 代名詞，連体詞

これ，それ，ここ，その，この，その，だれ，あらゆる，いかなる

[11]例えば促音は母語の影響でかなり入る学習者もいるが，厳密にその音が促音か促音でないかを区別することは難しい。長音，ポーズ，「しゅばらしい（すばらしい)」などの発音をその語の表記としてどこまで許容するかも同様である。これらは聞き取り者が自らの基準で，できる限りそのデータ内では統一することとする。

(2) 形式名詞

こと，もの，ところ，はず，ため，つもり，ふう，ほう，まま，やつ，よう，わけ，へん，ほか（ほかに）

(3) 接続詞

しかし，でも，ただし，したがって

(4) 助詞，助詞に準じる語

まで，くらい，ほど，だけ，ばかり，について，とともに，

例外：に関して，に対して

(5) 助動詞など

ようだ，そうだ，たい，てあげる，ていただく，てみる，てください，てもいい，という人

(6) 接頭語・接尾語

お皿，ご両親，1 日ごと，私たち，～とおり，文字どおり，大きめ，僕ら，私ども，～ごろ，～くらい

例外：多目に，3 年目，～か月

(7) 当て字，熟字訓

いとこ，ふさわしい

例外：派手，地味，立派

(8) 部首

しんにょう，にんべん

(9) えと

さる年，ひのえうま

(10) 擬態語

ぐずぐず，そわそわ

(11) 多くの副詞

こう，このように，よく，いちおう，いまいち，いろいろ，だいたい，たくさん，せっかく，たぶん，なんとなく，なかなか，ぜひ

例外：急に，突然

(12) 応答詞

えー，はー

(13) 終助詞

ね，よ，わ

(14) あいさつ語，慣用語

おはよう，こんにちは，ありがとう，しょうがない

(15) 疑問詞の一部

なぜ，どうして，なんで（「なんでそうなったの」などの場合），いつ，

例外：何（「なに」と読む場合），誰

（16）動詞

わかる

4.6.2. カタカナで表記するもの

カタカナで表記するものは以下の通りである。

（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，英語の場合のみ確実と思われる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

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的に漢字とする。「台北」は「たいほく」と読む場合のみ漢字とし，「タイペイ」の場合はカタカナ。ウーロン茶，ギョーザはカタカナ。なお，「ヴ」は原則として用いないが，特に明示することが必要な場合には使用可。

（3）動植物名

シマウマ，ニホンザル，ニンジン，バラ，ユリ，スミレ，チューリップなど。ただし，犬，猫，稲，麻等，常用漢字 1 字で書くことができ，かつ一般的なものは漢字。

（3）擬声語

カチカチ，ワンワン等。

4.6.3. 表記例

文字化の中で形を統一した表記の例を以下に示しておく。

【表記例】

語	使用する表記	例	使用しない表記等
あさごはん	朝ご飯		同様の例：昼ご飯，晩ご飯，夕ご飯
あと	あと		×後（「のち」とも読めるため）
ありがたい	ありがたい		×有り難い
ありがとう	ありがとう		×有り難う
あまり	あまり	あまり多くない，喜びのあまり	「残り」の意味のときは「余り」
いい	いい	成績がいい	×良い。ただし「よい」と発音している場合はひらがなで「よい」と表記
いちおう	いちおう		

うつくしい	美しい		
うれしい	うれしい		
えび	エビ		
おもしろい	おもしろい		×面白い
おれい	お礼		
おわび	お詫び		
かき	柿	干し柿	
かき	カキ	カキの養殖	
かげつ	か月	3 か月	×3 ヶ月　×3 カ月　×箇月
かた	かた	読みかた，こちらのかたが	×方（「ほう」とも読めるため）
がんばる	がんばる		×頑張る
きゅうに	急に		多くの副詞はひらがなが原則だが，例外的に漢字
きょう	きょう		×今日（「こんにち」とも読めるため）
きれい	きれい		
くらい	くらい，ぐらい		×位（「くらい」「ぐらい」の区別ができないため
けっこう	けっこう	けっこうです	×結構
ご	ご		×後（「あと」とも読めるため。ただし午後，数日後等は熟語なので漢字
こども	子供		
ごめん	ごめん		
ころ，ごろ	ころ，ごろ		×頃（「ころ」ごろ」の区別ができないため）
こんにち	こんにち		×今日（「きょう」とも読めるため）
ずいぶん	ずいぶん		
すでに	すでに		×既に
すみません	すみません		聞こえたまま。「すいません」と言っている場合は「すいません」
ぜひ	ぜひ	ぜひ来てください	「是非を問う」は漢字
た	た	そのた	×その他（「そのほか」とも読

			めるため
たいてい	たいてい		×大抵
だいぶ	だいぶ		×大分
たくさん	たくさん		×沢山
ただし	ただし		×但し
たぶん	たぶん		×多分
ちょうど	ちょうど		×丁度
とう	等		「など」と読む場合はひらがな
とき	とき	日本にきたとき，そのとき	「時を刻む」は漢字
とき	時	時を刻む	「日本にきたとき」はひらがな
ときどき	ときどき		×時々
とし	年	年が離れている，その年に	×歳　熟語以外で「ねん」と読む場合は，ひらがな
とつぜん	突然		ただし，多くの副詞はひらがなが原則
ない	ない	お金がない	×無い
など	など		×等
なに	何		「なん」と発音する場合はひらがな
なん	なん		×何。「なんじ（何時）」「なんだい（何台）なども全体をひらがなで示す。必要な場合は［　］で漢字を補足
のち	のち		×後
はたち	はたち		わかりづらい場合は「はたち［20 歳］と補足
ほう	ほう	このほうが	×方（「かた」とも読めるため）
ほかに	ほかに	勉強のほかに	×他に（「た」とも読めるため）
ほしい	ほしい		×欲しい
ほんと	ほんと		×本当。「ほんと」「ほんとう」を区別するため。ただし，「ほんとう」と発音している場合は「本当」も可
また	また		
まち	町	きれいな町	×街

もちろん	もちろん		
わかる	わかる		×分かる
わたし	私		「わたくし」「あたし」はひらがな

４.６.４. あいづち等の表記

あいづち等は，原則として以下のように表記する。Yes/No を答える応答詞の場合は「ん」でなく「うん」を用いる。ただし，その区別は厳密なものではない。

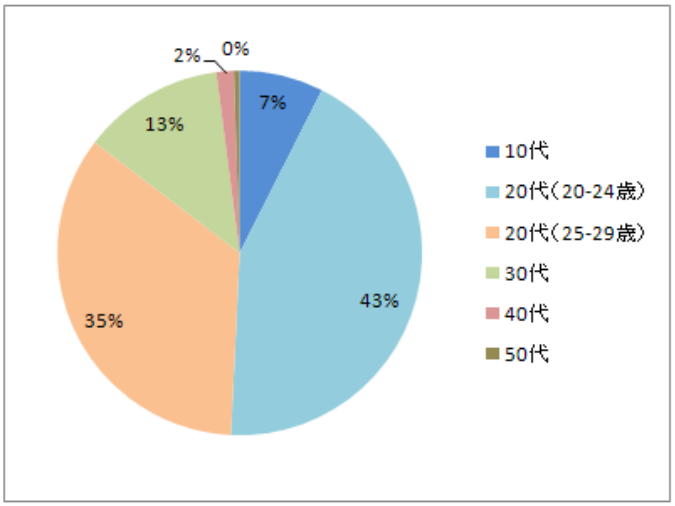
- Yes/No　：はい（はー），うん，えー，そう／いいえ（いえ），ううん
- あいづち　：ん，んー（んーんー），はい，はー（はーはー），えー（えーえー），あー（あーあー），そう，ほー（ほーほー）
- 言いよどみ：んー，あー，えー
- 呼びかけ　：ね，ねー
- フィラー　：あのー，そのー，えーと，えっと

資料　インフォーマントの属性

1.インフォーマント属性：年齢別

（単位：人）

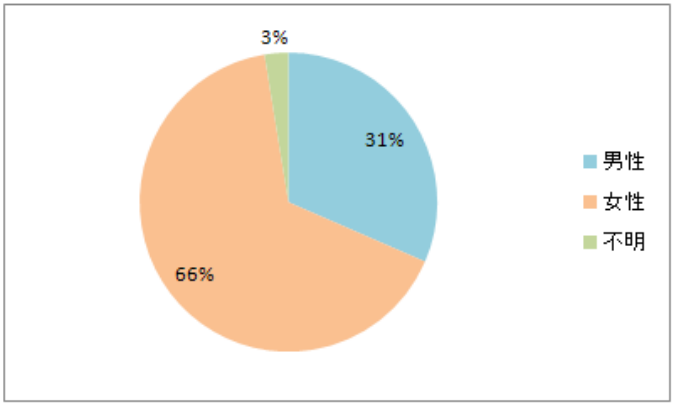
10代	29
20代(20-24歳)	169
20代(25-29歳)	135
30代	49
40代	6
50代	2
合計	390



2.インフォーマント属性：性別

（単位：人）

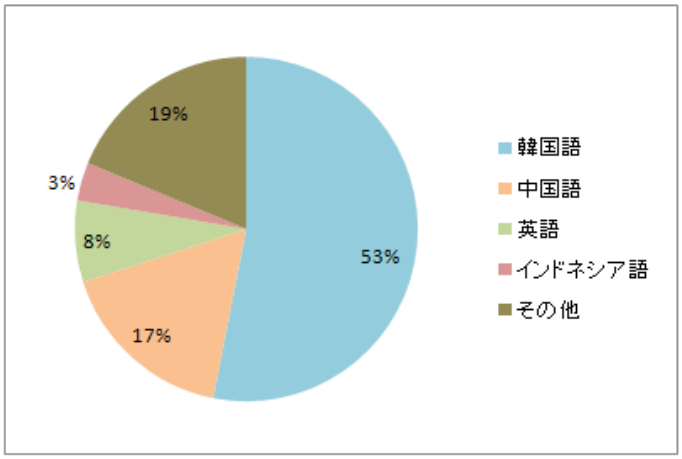
男性	123
女性	257
不明	10
合計	390



3.インフォーマント属性：母語別（複数回答の場合は最初に記された言語）

（単位：人）

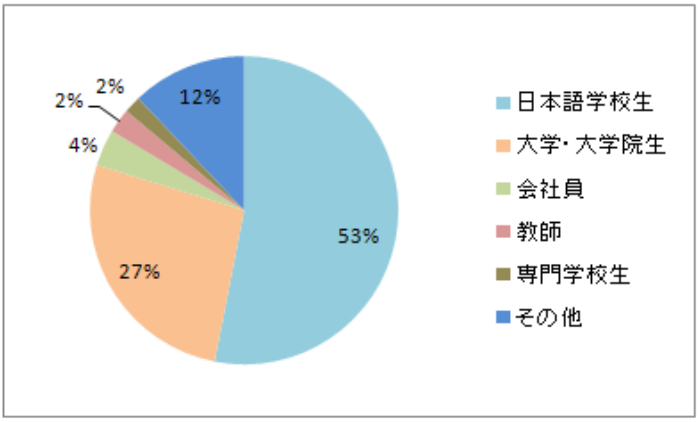
韓国語	207
中国語	66
英語	30
インドネシア語	14
その他	73
合計	390



４.インフォーマント属性：職業

（単位：人）

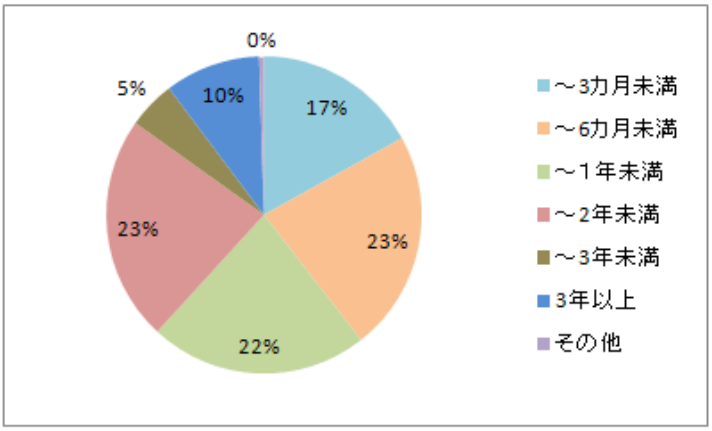
日本語学校生	207
大学･大学院生	104
会社員	15
教師	10
専門学校生	7
その他	47
合計	390



５.インフォーマント属性：日本滞在期間

（単位：人）

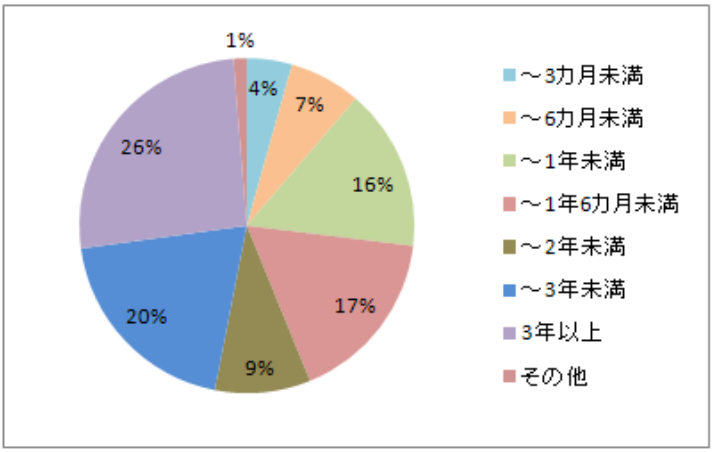
～3カ月未満	66
～6カ月未満	88
～1年未満	87
～2年未満	90
～3年未満	19
3年以上	38
その他	2
合計	390



６.インフォーマント属性：日本語学習期間

（単位：人）

～3カ月未満	17
～6カ月未満	27
～1年未満	61
～1年6カ月未満	66
～2年未満	36
～3年未満	77
3年以上	101
その他	5
合計	390



あとがき

この「日本語学習者会話データベース」は，独立行政法人国立国語研究所日本語教育基盤情報センターのプロジェクト「日本語教育データベースの構築」による成果である。データベース作成の目的は，日本語教育研究，言語習得研究，会話分析など，研究活動のための基礎資料として「外国語母語話者と日本語母語話者（テスター）の生の会話データ」を提供することにあった。

プロジェクトは，2006（平成 18）年度から 2009（平成 21）年度上半期までの 3 年 6 か月の間行われた。2010（平成 12）年度までの 5 年計画で開始されたが，2009 年 10 月の国立国語研究所の大学共同利用機関法人移管により期間が 1 年半短縮されている。担当したのは，旧国立国語研究所日本語教育基盤情報センター整備普及グループで，グループ内の分担は以下の通りである。

企画・統括：野山広（整備普及グループグループ長・現大学共同利用機関法人国立国語研究所日本語教育研究・情報センター研究員）

データベース作成・整備・事務処理担当：早田美智子（整備普及グループ研究員・現大学共同利用機関法人国立国語研究所研究情報資料センター専門職員），高橋悦子（旧国立国語研究所日本語教育基盤情報センター研究補佐員），塩谷由美子（旧国立国語研究所日本語教育基盤情報センター非常勤研究員）

グループ長である野山は，このプロジェクトの企画と対外的調整及び全体的な統括を行った。早田はデータベース作成・整備及び発信に関しての全体的な調整管理とこの報告の作成を担当した。高橋はデータ番号，ID の付与方法等の決定，テンプレートの作成をはじめとする収集データについての管理及び事務処理を担当した。塩谷は文字化作業開始時の文字化の基本的な方針の策定と表記規則の決定を担当した。文字化データの確認，音声マスキング箇所の特定等データベースに関わる作業は作成担当の 3 名が分担して行った。

なお，表記規則の決定については，旧国立国語研究所日本語教育基盤情報センター評価基準グループが「日本語学習者による日本語／母語発話の対照言語データベース（発話対照 DB）」で作成したマニュアルを参考にしている。

謝辞

まずはじめに，データ収集の趣旨をご理解くださり，研究・教育の発展のためにご自分の発話データを提供してくださったすべてのインフォーマント及びテスターの方々に心からの感謝を申し上げます。

「日本語学習者会話コーパス研究会」のメンバーである鎌田修，嶋田和子，山内博之の各氏には，データ収集という最も中心となる業務を研究会として請け負っていただいたほか，プロジェクトの企画計画の時点での検討にも加わっていただき，プロジェクトの方向

性を決定していただきました。また節目節目では貴重なご意見をお寄せくださり，議論の中心となってくださいました。ここに改めて御礼を申し上げます。

熊倉加代子，肥後玉衣，伊藤啓子，犬飼芳子，李明熙，王鴿，伊藤直美，宮武かおりはじめ，文字化作業にあたってくださった方々には，格別の御礼を申し上げます。

この方々の試行錯誤とフィードバックのおかげで，何とか文字化データとしての全体的な統一を目指すことができました。この方々の長期にわたる粘り強い努力がなければデータとしての完成はありませんでした。もちろん最終的にすべてのデータ確認を行ったのは担当者であり，誤りや不統一等があったとしても，それは担当者の責任であることはいうまでもありません。

このほか，さまざまな形で御協力くださったすべての方々にプロジェクト担当者一同，深く感謝いたします。