C-JAS（Corpus of Japanese as a second language）データ概要

2021.6.1.

１．データの概要

(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じ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』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）
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)データ収集時期

データの収集時期は 1991 年 7 月～1994 年 3 月である。

(3)データ数の内訳

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 回目（*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)

(4)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

２．文字化の概要

２－１．文字化の方針

音声データは、以下の方針に従って文字化した。

＜文字化における基本方針＞

(1)　発話番号

各発話には行頭に下記のような番号をつけ、下記のことを表す。

例）C1-1-I-00010-N ：　日本語はどのぐらい勉強しましたか

C1-1　　→学習者 C1　の第 1 期

I　　→対話データ（C-JAS ではすべて「I」の記号が付いている）

00010　→1 番目の発話（以下、00020　00030・・と続く）

N　　→調査者（日本語母語話者）

(2)　発話者の記号

発話番号の末尾に発話者を示す以下の記号をつける。

調査者（日本語母語話者）→N

学習者（日本語学習者）　→L

(3)　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。

(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【　】にその分類名とローマ数字を入れ、言いかけている固有名詞も全て置き換える。同じ固有名詞の場合は、同じ分類名、ローマ数字を使用する。

置き換える具体的な固有名詞は、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、【人名１姓】【人名１名】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が、発話者の記号は非母語話者の場合は、「L2」、日本語母語話者の場合は「N2」とし、複数以上出てくる場合はL、Nの後につける番号を適宜増やしている。

＜表記の方針＞

(1)　文字の表記方法

表記は発話内容の把握のしやすさを考慮し、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。

(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「？」を付ける。

(5)　非言語行動等

笑い声や発話に関係しそうな非言語行動は｛　｝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｝
｛テープ一旦停止｝

(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、英語で発音されたと判断された場合は英文表記も可とする。外国語が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

発話内に出てきた数字は基本的に漢数字とする。また、アルファベット表記については、その通り発音するものは全角アルファベットとする。

例) DVD USB など

(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、その意味と思われる語を( ) で補足する。

例) はいません (入りません)
しゅーじん (主人)
クラピック (グラフィック)
かい、て (買って)

また、その漢字に対し 2 つ以上の読みが考えられる場合は漢字で表記し( ) で発音された音を平仮名で補足する。

例) 入れる (はいれる)

(9) 引用発話

直接引用の発話は「 」で示す。

(10) 書籍名などのタイトル

書籍名などののタイトルは『 』で示す。

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、その部分の談話を削除している。削除した部分については、非言語情報を表わす{ } を使用し、削除した発話数の説明を表記する。

例) {続きは個人情報保護のため 4 発話分削除}

３．注意事項・その他

使用して頂く中でミスなどを発見された場合、またコーパスに関する御意見等頂ける場合は、是非御連絡頂けると幸いである。問い合わせ先は以下の通りである。

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 日本語教育研究領域

lsaj@ninjal.ac.jp

参考文献

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.