

多言語母語の日本語学習者横断コーパス
「I-JAS」

書き起こし規程集

2020 年 3 月版

目次

1. 書き起こしの基本方針	1
2. タスクの種類	1
3. 基本的な表記方法	1
3-1. 漢字仮名交じり文	1
3-2. その他の書き分け	3
3-3. 音声の書き起こし特有の問題	3
4. タグセット	5
4-1. タグの概要	5
4-2. 除外指定タグ（語の断片など、解析から除外すべき箇所を指定する）（X タグ） ..	6
4-3. 品詞指定タグ（品詞解析ミスが起こりうる箇所に正しい品詞を指定する）	7
4-3-1. フィラー（F タグ）	7
4-3-2. 外国語や登場人物名（N タグ）	8
4-3-3. 連体詞「あの」「その」「この」（R タグ）	11
4-4. 表記指定タグ（解析ミスが起こりうる非標準表記を標準表記に置換する）	11
4-4-1. 一般的な規則とは違う箇所に長音やポーズがある場合（T タグ）	11
4-4-2. 発音や活用間違い（G タグ）	12
4-4-3. 表記の間違い（K タグ）【SW のみ】	13
4-5. 発音指定タグ（文字だけでは発音が不明な箇所に本来の発音を記載する）	14
4-5-1. 発音が曖昧であることを示す（H タグ）	14
4-5-2. 読みを指定する（Y タグ）	14
4-6. 補足情報タグ（その他データ利用者に有益となりうる情報を記載する）（「?」） ..	17
5. 個人情報について	19

1. 書き起こしの基本方針

多言語母語の日本語学習者横断コーパス（以下 I-JAS）は、主として日本語学習者の文法習得、談話習得などの研究に活用することを目的として収集されたデータである。そのため、テキスト化では調査協力者と調査者の発話をできるだけ発音に忠実に書き起こすこととした。一方、音声・音韻に関する研究や音声的特徴を扱う社会言語学的な研究は対象としていないため、発話の重なりやポーズの長さ、声の大きさなどの情報は付与しない。

本コーパスは書き起こされたテキストデータと音声データを公開するだけではなく、テキストデータを形態素解析し、その形態論情報を利用した検索も可能なコーパスを目指す。そのため、本コーパス専用の形態素解析システムを作成し、それによる自動の形態素解析を念頭において、書き起こしのルールおよびタグを策定した。また、本コーパスでは、原則として、誤用および正用の表示は行わない。誤用の認定は非常に困難であり、研究内容によってもその判定基準が異なるからである。

2. タスクの種類

（1）独話

・ST（ストーリーテリング）

4～5 コマのコマ割り漫画を見てストーリーを述べるタスクで、「ピクニック」と「鍵」の2種類がある。それぞれおよそ2～3分のデータである。

・D（ディスクリプション）

絵の状況説明をさせて、動詞活用（「テイル形」など）の使用状況を調査するタスクである。およそ3～8分のデータである。

（2）対話

・I（インタラクション、対話）

調査協力者と調査者が1対1で、30分程度の会話を行うタスクである。

・RP（ロールプレイ）

「依頼」と「断り」という言語機能を持った2種のロールプレイを行うタスクで、それぞれおよそ2～5分のデータである。

3. 基本的な表記方法

3-1. 漢字仮名交じり文

音声の書き起こしは、基本的に日本語の一般的な漢字仮名交じり文を用いる。ただし、形態素解析で誤解析を起こす可能性のあるものについては、I-JAS 特異の表記もある。おおよその目安を示せば、以下のようになる。

(1) 動詞や形容詞は漢字

例) 見る、食べる、掃除する、交代する、見つかる、置く、近い、暑い

例外) いる、ある、わかる、できる、ください、する、ほしい、いい、ない

(2) 補助動詞は平仮名

例) てください、ていく、てくる、てみる、ておく、～いたします

(3) 実質名詞は一般的な表記

例) 学校、誕生日、日本語、スカイプ、ピクニック、サンドイッチ

- ・外国の地名、外国語はカタカナで書く。中国語の名詞（地名や料理名）は日本語で一般的に漢字表記のものに限り漢字で書く。中国語の発音で話されているものはカタカナで表記する。

例) グアンジー省、ソウル、ウィーン、北京、上海、シーアン（西安）、
麻婆豆腐、シャオロンパオ（小籠包）

- ・ST に関しては、イラストに書かれている指示の名詞は、書かれている通りに書く。

例) ケン、マリ、犬、バスケット、地図、鍵、梯子、警官

(4) 形式名詞は平仮名

例) ほう、こと、の、ところ、はず、ため、つもり、まま、わけ

例外) 時（日本語を勉強している時に、この辞書を使います）

(5) 副詞は平仮名

例) いろいろ、だいたい、たくさん、たぶん、やっぱり、よく、いっぱい

例外) 急に、突然

(6) 接続詞は平仮名

例) でも、だから、もし、ただし、したがって、また、しかし

(7) 挨拶語などの定型句は平仮名

例) ありがとう、おはようございます、～でございます、

(8) 数字は漢数字

例) 兄は二十三歳です／アルバイトは週に三日／家族は五人です／そうしたら給料
が二万三千円も上がったんです／学校まではバスと電車で四十五分くらい

- ・ID 番号はローマ数字でよい。例) 「IDはKKD20です」
- ・「ではロールプレイ1を始めます」などの「1」はローマ数字でよい。
- ・音声の書き起こしでローマ数字を使う場合は、全角を用いる。

(9) 擬音語・擬態語はカタカナ

例) 猫がニャーニャー鳴いています／雨がザーザー降ってきて

(10) アルファベット表記でその通り発音するものは全角アルファベット

例) アルファベットで書く例 (全角) ...PC、DVD、CD、ID、USB など
(パスワードは「PW」とは書かない。)

3-2. その他の書き分け

以下は、上記以外の書き分けで、特に注意の必要なものをまとめて表にしたものである。

表1. 書き分けるもの

発音	表記方法	備考
いえ	家	漢字表記
うち	うち	ひらがな表記
えがく	描く	「かく」とは読まない
おじいさん	おじいさん/お爺さん	「お祖父さん」は使用しない
おばあさん	おばあさん/お婆さん	「お祖母さん」は使用しない
かく	書く	「絵を書く」のようにも使う
かん	缶	「カン」とは書かない
くらい/ぐらい	くらい/ぐらい	「位」は使用しない
(傘を) さす	(傘を) 差す	「傘を差す」の「差す」
そば	そば	「傍」は使用しない
ときどき	時々	
びん	瓶	「ビン」とは書かない
まち	町	「街」は使用しない
ふるさと	ふるさと/古里	「故郷」は使用しない
bag (かばんの意)	バッグ	「バック」ではない
bed (寝床の意)	ベッド	「ベット」ではない

3-3. 音声の書き起こし特有の問題

(1) 1文の単位、改行

発話音声で、1文を定めるのは大変困難である。そのため、I-JASでは、以下のように対応する。

- ・対話 (I・RP) : 話者のターン交代をもって改行する。文末に「。」(句点)は打たないが、上昇イントネーションの「？」は付与する。
- ・独話 (ST・D) : 調査協力者が独話する場面では、話者交代がなくても、1文だと思われる単位で改行する。その際も、文末に「。」は打たないが、上昇イントネーションの「？」は付与する。

(2) 間、ポーズ 「、」

発話が途切れる場合、「、」（読点）で表す。ポーズの長短に関わらず、読点1つで表記する。本コーパスではポーズの秒数などはカウントしないため、読点の付与は、書き起こし作業者の主観的な判断となってしまうが、タスク内で基準を統一する。

また、通常の日本語の統語的なルールと異なる箇所に読点を付け過ぎると、自動の形態素解析で誤解析が起こりやすくなるので、あまり気にならない程度のポーズであれば、読点は打たなくてもよい。読点を打つべきポーズがあり、それが通常の日本語表記と異なる場合は、T タグを使用して処理する（4-4-1 参照）。

(3) 長音 「ー」

前の音節が長く伸ばされている場合、「ー」で表す。ただし、平仮名で表記されることが一般的な場合（例：おねえさん）は、長音を使用しない。本コーパスでは、間や長音は厳密な秒数などをカウントしないため、長音を付与するかは作業者の主観的な判断になるが、タスク内で基準を統一する。実際の音声で、長音が一拍以上ある場合でも「ー」で表し、「ーー」のように長音記号の連続表記はしない。

(4) あいづち 「〈 〉」

あいづちとみなされる発話は、〈 〉（やまかっこ）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、別の発話として独立させる。本コーパスでは、〈 〉に入れられたあいづちは一律、品詞が「あいづち」と付与される。そのため、あいづちが長い場合や感動詞だけでなく文としてまとまりのある発話の場合は、別の発話として改行し、表記する。

(5) 発話の重なり

発話が重なっている場合は、別の発話として改行して表記するか、もしくはあいづちとして相手の発話中に挿入する。原則としては、長い発話の場合は別の発話として扱い、感動詞のみのような短い場合はあいづちとして扱う。

(6) 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、「？」（全角）を付与する。

(7) 聞き取り不能の音声 「＊」

音が相手の発話と重なることによって、全く聞き取れない場合、おおよそのモーラ数を＊の数で示す。

(8) 非言語行動 「{ }」（中かっこ）

笑いや咳などの非言語行動は、本コーパスにおいてはそれほど重要視しない。しかし、笑いそのものが一つのターンを取っている場合や、咳込に対して「すみません」と言った

り、相手が「大丈夫ですか」と言ったりする場合がある。そのような笑いは、「{笑}」「{咳}」と書き、「あははっ」や「ごほんごほん」などのようには書き起こさない。声が小さい・大きい、大きく息を吸い込むなどの非言語的な要素も書き起こさない。

(9) 個人情報 「【 】(すみつきかっこ)」

個人が特定される可能性があると思われる箇所は【 】を使用し、一般名詞を用いて伏せる。個人情報の詳細については、5章を参照のこと。

以下に上記で説明した書き起こしで使用する記号類についてまとめた表を載せる。記号は全て全角で表記する。

表2. 記号一覧

意味	記号(全角)	記号の名称
あいづち	< >	山カッコ
間・ポーズ	、	読点
長音	ー	長音
上昇イントネーション	?	疑問符
聞き取り不能	*	星印
非言語情報	{ }	中カッコ
個人情報	【 】	すみつきカッコ
直接引用	「 」	かぎカッコ
書名、映画名、ドラマ名	『 』	二重かぎカッコ
複数の読みのある場合のフリガナ (タグの中)	()	小カッコ

4. タグセット

4-1. タグの概要

タグには大きく分けて以下の5種類がある。

- (1) 除外指定タグ（語の断片など、解析から除外すべき箇所を指定する）
- (2) 品詞指定タグ（品詞解析ミスが起こりうる箇所に正しい品詞を指定する）
- (3) 表記指定タグ（解析ミスが起こりうる非標準表記を標準表記に置換する）
- (4) 発音指定タグ（文字だけでは発音が不明な箇所に本来の発音を記載する）
- (5) 補足情報タグ（その他データ利用者に有益となりうる情報を記載する）

原則として、タグ内の記号は半角とする。ただし、Yタグに使用する()は例外として、全角とする。以下の表3はI-JASで作成したタグセットの一覧である。

表 3. I-JAS のタグ一覧

タグの種類	具体的な内容	タグ表記 (α は発話通りの文字列、β は解析用の語)	タグ表記例
(1) 除外指定 タグ	意味不明語、語の断片	[α=X]	[みつー=X]
(2) 品詞指定 タグ	フィラーを感動詞に指定	[α=F]	[えー=F]
	外国語を名詞に指定	[α=N]	[シジャン=N]
	連体詞に指定	[α=R]	[あの=R]
(3) 表記指定 タグ	語中の長音、ポーズ	[α=T=β]	[マ、リ=T=マリ]
	語や活用や発音の誤り	[α=G=β]	[書きて=G=書いて]
	PC 入力時の変換ミス (SW のみ)	[α=K=β]	[持って=K=持って]
(4) 発音指定 タグ	発音不明瞭 (α1 か α2)	[α1/α2=H]	[が/は=H]
	複数の読みがある漢字語 α	[α (読み) =Y]	[何 (なん) =Y]
(5) 補足情報 タグ	補足情報を付与する	[α?補足のメモ] [α=タグ?補足のメモ] [α=タグ=β?補足のメモ]	

4-2. 除外指定タグ（語の断片など、解析から除外すべき箇所を指定する）（X タグ）

自動の形態素解析で正しく解析できない箇所は、解析から除外するため、「X タグ」を付与する。このタグを付与して形態素解析した場合、語彙素は空欄、品詞は「解析困難箇所」となる。X タグを付与するのは「語の断片」「不明語」「学習者によるオリジナルの語」の 3 種である。

（1）語の断片：正確な語の前に現れる語の断片

語の断片とは、「ちよ、ちよつと」や「たべ、たべ、たべます」のように、正確な語の発話の前に現れるものを指す。その他にも、「ちよ、えと、ちよつと」のように間にフィラーなどが入る場合もある。

また、「ちよ、ちよつと」のように、正確な語の一部である場合と、「お、ちよつと、すん、お尋ねします」のように、正確な語の一部ではない場合がある。後者の場合は、「すん」が「お尋ねします」の一部ではないが、何かの言いかけであると判断され、かつ、フィラーや感動詞でもない。このような場合は、語の断片と判断する。

例) K：マリは、[けけ=X]、警官さんに状態を説明しました

例) K：[ゆ、ゆうた=X]、[ゆうかた=G=タ方]は、パソコン、[て=X]、テスト

例) K：うちのレストランは〈うん〉ふぐのレストランですからー〈うん〉私は免許
を{笑}[と取らなかつ=X]、取らないんです、けど

（この例の場合、「と取らなかつ」の部分を X とする。「取らなかつ」は、形態

素で分割して「取ら + ない + た」のように考えれば、「取ら」の部分は正用と考えることもできるが、そのように分割してタグ付けしても誤解析を誘発するので、「と取らなかつ」全体を一つの語の断片とする。）

(2) 不明語：全く意味のわからない語や表現

学習者の発話には、意図する事柄が全く分からない語や表現も見られる。

例) K：でもちょっと、田舎のが、[ぐうぐ=X]が好き

例) K：怖かった、わたしは、〈うん〉[あいびげ=X] {笑} 〈ん?〉事故、事故が怖い、怖い、です

X タグを付与した箇所は形態素解析からは除外するが、不明語であっても前後の文脈などから意図する事柄が想像できる部分がある。訂正する候補が複数考えられる場合、X タグを付与し、候補をリストアップしておく。X タグの後に補足情報を示すタグ「?」(4-6 参照)を書き、その後に候補となる複数の語を「/」で区切って書く。

また、修正候補を複数想定することはできないが、修正候補に確信が持てない場合も、X タグと補足情報「?」で対応する。

例) 宿題が[もでません=X?もう出ません/持てません/持っていない]

(「もでません」の修正候補として「もう出ません」「持てません」「持っていない」の3種類が想定される。)

例) ぬいぐるみとか、[どーと=X?ノート]、とか

(「どーと」は「ノート」だと思われるが、文脈上絶対そうとも言い切れず、確信が持てない。)

(3) 学習者によるオリジナルの語

擬音語・擬態語には、学習者のオリジナルの語が現れやすい。原則的に、「ひとつの修正候補が思いつくかどうか」を判断基準とした。「ピャーピャー怒る」「トゥ、トゥ、トゥ降る」は、学習者によるオリジナルの擬音語・擬態語であり、一般的に使用される擬音語・擬態語の候補が1つに絞れない。そのような場合はX タグを付与し、補足情報(4-6 参照)の「?擬音語・擬態語」も付ける。

例) お父さんが[ピャーピャー=X?擬音語・擬態語]怒って

例) 雪が[トゥ、トゥ、トゥ=X?擬音語・擬態語]、降りました

4-3. 品詞指定タグ(品詞解析ミスが起こりうる箇所に正しい品詞を指定する)

4-3-1. フィラー(F タグ)

フィラーとは、言い淀みなどに出現する場つなぎ的な表現である。フィラーは誤解析を誘発することが多いので、F タグを付与することで、解析時の品詞をあらかじめ指定する。指定する品詞は「感動詞」である。

場つなぎ的なフィラーには、次のようなものがある。以下は一例である。

「あ」「あっ」「あー」「あの一」「あーん」「ああと」
「う」「うん」「うーん」「うんと」
「え」「えと」「えっと」「んと」「えーっと」
「お」「おー」「とー」「っと」「のー」(あの)

例) [あっ=F][えっと=F]七時半に、家を、出ましたか？

例) [うー=F]、[うーん=F]、[おたずん=X]、[う=F]、[お=F]、ちょっとー[すん=X]、お尋ねしたいんですが

誤解析を起こしやすい感動詞には、場つなぎ的なフィラーのほかにも、驚きや落胆を表す感動詞（感情表出系感動詞）がある。これらは、文頭（発話の最初）に現れた場合は誤解析になりにくい、文中にある場合は、誤解析になりやすい。そのため、文中にある感情表出系感動詞には、タグ F をつける。

感情表出系感動詞：「ありゃ」「うわ」「えっ」「ほー」

例) 十時間ぐらい、[ほー=F]じゃあ何日か行くんですか

「まあ」はフィラーか副詞かの区別が大変難しい。修飾する述語があり、完全に副詞であると分かる場合はタグなしでいいが、そうでない場合は、F タグを付与する。

例) まあ、いいか

例) それは[まあ=F]、えーっと、じゃあ次の質問なんですけど

学習者の発するフィラーは、日本語のフィラーとは異なるようなフィラーがみられる。その場合も日本語のフィラーと同様に F タグを付与する。

例) [やの一=F]

例) [あの一ん=F]

例) [あむー=F]

4-3-2. 外国語や登場人物名 (N タグ)

音声データには、形態素解析用の辞書に登録されておらず形態素解析ができない名詞が多数現れる。それらの名詞のほとんどは、「外国語で表現された語」か「アニメなどの架空のカタカナ語の固有名詞」である。それらには、N タグを付与し、品詞が「名詞」となるよう指定する。

N タグを付与した箇所は、形態素解析の際に用いる辞書 (UniDic) に登録されている語であれば、登録されている品詞情報が付与される。しかし、登録されていない場合は、一律に「名詞」という品詞情報がつく。N タグを付与した箇所は、1 短単位¹として見なされ

¹ 「短単位」の規程については、https://pj.ninjal.ac.jp/corpus_center/bccwj/morphology.html を参照。

るため、2 短単位で成り立っている固有名詞などは、別々に N タグを付与する必要がある。

(1) 外国語で表現された語

対話の中で外国語が現れる場合の多くは、「地名」「食べ物の名前」「観光地の名前」「本や映画のタイトル」である。聞こえた通りにカタカナで書き起こし、N タグを付与する。何を示しているか分かる場合には、補足情報のタグ「?」(4-6 参照)を使って、以下の例のように記入する。ただし、前後に「市」や「という食べ物」といった発話があり、その外国語が何を指し示しているかが明白な場合は、記載しない。

例) 学校は[グレンイーデン=N?地名]にあります

例) 生れたところは[オボネキーシン=N?地名]

例) [モウマカイン=N]という食べ物があります

例) 湖南省にありますけど、[チンシュウ=N]市という市ですが

例) C: あー[えっと=F]、もう一度名前をお願いします何て言う、タイトルですか?

K: 『[クロムシャットレジ奥斯=N?小説名『鋼殻のレギ奥斯』]』

C: 『[クロムシャット、レジ、ジ奥斯=N?小説名『鋼殻のレギ奥斯』]』?

表 4. N タグの補足情報の正しい表記例

カテゴリー	記入例
地名	[イズミツツ=N?地名] [シャンシー=N]省
料理名	[タットリタン=N?料理名] [ちょうどうふ=N?料理名「臭豆腐」]
外国語発話 (1 語の場合) ※ 1	[ヤン=N?中国語で「羊」] [エングリー=N?英語の「angry」]
アニメやドラマに関して	[ナルト=N?アニメ名] [ハウル=N?登場人物名] [ジブリ=N?映画会社名]
バンド名	[ラルク=N?バンド名]
路線名	[アストラム=N?電車路線名]
観光地	[アタクレ=N?観光地名]
店名	[イケア=N?店名]
テレビ番組のタイトル	[イッテQ=N?テレビ番組名]
会社・企業の名前	[インテル=N?企業名]
施設の名前	[カニオン=N?施設名]
川や山の名前	[ガンガ=N?川名「ガンジス川」] [アッポウ=N?山名「アポ山」]
短期プログラムや研修	[ジェット=N?プログラム名]

外国語は、動詞など名詞以外で使用されている場合も N タグで対応する。しかし、英語の場合、2 語以上や文単位になっている場合は、タグを付与せず、アルファベットで表記する。アルファベットは半角、スペースは全角とする。英語以外の言語の場合は、カタカナで対応する。また、中国語に限っては、日本語で通常漢字で表記される地名（「上海」「北京」など）や固有名詞は漢字で書く。日本語にある物の名称や映画などのタイトルなど、英語名と日本語名が同じ場合は、英語風に話している場合でもカタカナで表記する。

例) 梯子に[クライム=N]した

(「クライム」は名詞ではないが、N でよい。)

例) [あー=F]、I have so many great experiences、とても、すてきな、誕生日の、経験

例) スカイツリー、に行きました

例) [ス=X]、スターウォーズが好きです

学習者が間違って発話している箇所の正しい語が形態素解析用の辞書に登録されている場合、G タグ (4-4-2 参照) を付与し、日本語で表記されるように書く。複数の短単位に分割される場合、補足情報は最後の短単位に記入する。

例) [ルヒ=G=ルフィー?登場人物名]

例) 『[ウィッチ=G=リッチ][メン=G=マン][ポウ=G=プア][メン=G=マン?ドラマ名『リッチマンブアウーマン』]』

(2) カタカナ語の固有名詞

カタカナ語の固有名詞は、形態素解析用の辞書に登録されているかを基準とし、登録されていない場合は、N タグを付与する。その語がさらに複数の短単位に分割される場合は、それぞれにタグを付与することとする。一般的によく使う外来語、地名であれば、たいていは辞書に登録されている。

例) [キャナル=N]シティ、行きました

例) 赤坂[サカス=N]、と、レインボーブリッジ、有名です

例) ニュースキャスターもやってるし

例) 『ドラゴンボール』の[ベジータ=N?登場人物名]、[ベジッタ=N?登場人物名「ベジータ」]、強いです

例) 『[ガンツ=N?映画名『GANTZ』]』という映画を見ました

(3) 小説や映画のタイトル、登場人物の名前など

日本語の固有名詞には N タグを付与しない。小説や映画のタイトル、登場人物の名前などであっても、日本語の語から成る場合はそのまま書き起こす。

例) 嵐かっこいい。桜井君が一番好き。

例) 小説読みました、『その時彼によろしく』って

4-3-3. 連体詞「あの」「その」「この」(R タグ)

書き起こされたテキストを読む場合、「あの」「その」「この」は、連体詞かフィラーか区別が付きにくい。また、これらは自動の形態素解析の際にフィラー（感動詞）と解析されることが多い。そのため、音声からそれらの区別ができる場合、連体詞の「あの」「その」「この」は、連体詞という品詞を指定するよう「R タグ」を付与する。

また、「あん時」「そん時」というような発話にも、R タグを付与する。

例) [あの=R]、警官が、言いました

例) [あの=F]、ちょっとお願いが、あるんですが

例) [あん=R]時は、えーたぶん『デジモン N』

G タグ (4-4-2 参照) で訂正した「その」が連体詞である場合、R タグがつけられない。その場合は、補足情報の「?」をつけて対応する。

例) 科目〈うん〉があります[てど=G=けど]〈はい〉[との=G=その?連体詞]中は日本語があります

4-4. 表記指定タグ（解析ミスが起こりうる非標準表記を標準表記に置換する）

4-4-1. 一般的な規則とは違う箇所に長音やポーズがある場合 (T タグ)

一般的な統語規則や語の表記とは異なる箇所に、長音やポーズがある場合、正しく解析されない。そのため、そのような箇所に正しい語や表現を指定するよう「T タグ」を付与する。一般的な統語規則や語の表記とは異なる箇所とは、主に次のような場合である。

(1) 1 短単位の中に長音やポーズがある場合

1 短単位中に長音やポーズがある場合は、T タグを付与する。語末の長音は、自動の形態素解析にほとんど影響しないので、T タグは付与しない。ただし、長音の後に読点がない場合はタグを付与する。

例) [だーから=T=だから]、あーこの、坂本龍馬は

例) ケンと[マ、リ=T=マリ]さんが、言って

(2) 通常の日本語の統語規則と異なる箇所に長音やポーズがある場合

通常は、文や節の切れ目でポーズが入る。しかし、音声の場合は、2 つの文や節が連続して発話されたり、通常とは異なる箇所にポーズが入ったりすることがある。このような場合、誤解析になりやすいため、T タグを付与する。特に、文や節の最初にくる「接続詞」、文や節の最後に来る「助詞」「接続助詞」が対象になる。

例) 「ありますが、日本語より少ないです」と言いたいのに、以下のように言っている。

K : あります、がー、日本語より、少ないです

タグ : あります[、がー=T=が]、日本語より、少ないです

例) 「知ってません、でも後輩に聞いて」と言いたいのに、以下のように言っている。

K : [う、あー=F]知ってませんが、後輩に、聞いて、
 タグ : [う、あー=F]知ってません[でも=T=、でも]、後輩に、聞いて、
 例)「一週間 (に) 三日は、働いて」と言いたいのに、以下のように言っている。
 K : 一週間、三日、は一 (うん)、働い、てー、
 タグ : 一週間、三日[、は一=T=は] (うん)、[働い、てー=T=働いて]、

(3) 文の終わりではない箇所に上昇音がある場合

文の終わりではない箇所に聞き逃せないほどの上昇音が入る場合がある。その際には上昇を表す「？」を記入するが、「？」がある箇所は文末と判断され、誤解析を引き起こしやすい。そのため、そのような「？」は T タグを用いて除外する。

例) 発話 : 地図？を見てる間に
 タグ : 地図[？を=T=を]見てる間に

4-4-2. 発音や活用間違い (G タグ)

長音やポーズの他にも、学習者独特の間違いがあり、そのような箇所は正しく解析できない。それらのうち、発音と活用の間違いについては、訂正例が明確に 1 つに絞られるため、「G タグ」を付与し、訂正例を指定する。解析は指定した訂正例で解析される。

基本的に、発音や活用の間違いで、形態素解析がうまくできないために、タグをつけるものである。そのため、日本語として解析できるものには付与しない方針である。長音やポーズ以外の学習者の間違いで、訂正すべき候補が 1 つに絞られる場合には、以下のようなものがある。

(1) 発音が違っている場合

例) [けーじゃい=G=経済]
 例) [ちゅがつくー=G=中学]三年生
 例) [よこ=G=よく][くいす=G=クイズ]があります
 例) ニュース[カスター=G=キャスター]もやってるし
 例) [もりごろろ=G=毛利小五郎?登場人物名]っていう人
 例) 岐阜県の[ぎぶ=G=岐阜]市に行きました
 例) 雨が[ジャジャ=G=ザーザー]って降って
 例) ねこが[ニヤニヤ=G=ニャーニャー]鳴いています

(2) 語の活用が違っている場合

例) それから、作文を[かきて=G=書いて]
 例) 友達が[来って=G=来て]から
 例) 試験を受け、受け[ないでも=G=なくても]いいです
 例) [おいしじゃない=G=おいしくない]

活用や接続を誤って、必要な語が足りないような場合 (脱落系) も、G タグで修正する。

例) [難しで=G=難しくて]

例) [おいしくない=G=おいしくない]

(3) 相手の発音を正確にリピートできずに間違っている場合

例) C: 日本の俳優さんも好きですか

K: [あー=F][はいゆ=G=俳優]?

(4) 発音のミスの部分が日本語の語彙にある場合

例) K: [えー=F]バスに[えーえっとー=F][う=X][うえ=G=家]に帰って

C: 上?

K: あ、家、家

(5) 語や表現の一部が間違っている場合

例) 今[覚えてーの=ない=G=覚えてない]です

例) 誰でも[たすか=X]、助けてくれ、くれません=G=助けてくれません

例) 箱に、[飛び込みます、みました=G=飛び込みました]

例) [ちゃいもー=G=チャイム]、[な=X]、[ならっていました=G=鳴ってました]

例) [飛んでし、ました、飛んで、飛んでしま=X]、[飛んでします、しました=G=飛んでしましました]

例) [休みだった、でした=G=休みでした]

(6) 数字の間違い

数字の発音が間違っている場合も、G タグで対応する。その場合、「発音通りの書き起こし」の部分は平仮名で聞こえた通りに書き、「指定した語や品詞」の部分は漢字で書く。数字は基本的に漢数字とする。

例) [よんとう一、にん=G=四十人]、[よんじゅにか=G=四十人が]、

例) [は、え=X]、はい、[とっぷ=X]、[ろっ=X]、[ろっかげす=G=六か月]から一八、[はちがす=G=八月]、から

4-4-3. 表記の間違い (K タグ) 【SW のみ】

ストーリーライティングは、他の発話データとは異なり、基本的に学習者自身がコンピュータに入力したデータ（一部手書きもある）であるため、表記の間違いや、平仮名のみで書かれたデータもある。このようなデータを自動の形態素解析にかけると、誤解析が多発する可能性が高まる。そのため、表記の間違いや、漢字仮名交じりで書くべき箇所が平仮名のみで書かれている箇所には K タグを付与する。

例) そばに地図で所を[さがして=K=探して]います

例) マリは[置きます=K=起きます]

例) バasketを[持って=K=持って]、公園に行きました

4-5. 発音指定タグ（文字だけでは発音が不明な箇所に本来の発音を記載する）

4-5-1. 発音が曖昧であることを示す（H タグ）

発音があいまいで、一つの文字に決定できない場合は、H タグを付与する。タグは[一番可能性の高い候補/二番目に可能性の高い候補=H]と表記する。形態素解析は、一番可能性の高い候補で行うよう指定される。

H タグの付与は、助詞などの 1 音の短単位 の音が曖昧な場合が中心である。「心配（しんぱい）」か「先輩（せんぱい）」かなど、発音の曖昧さが語の判断にまで及ぶ場合は、別のタグの範疇とする。

例) パソコン[を/の=H]すごい技術を

4-5-2. 読みを指定する（Y タグ）

一つの漢字表記に複数の読みがある場合は、漢字の後に（ ）で読みを書き、Y タグを付与する。読み方を書く（ ）は例外的に全角表記とする。

例) 音 声 : いちにちだけです

書き起こし: [一日（いちにち）=Y]だけです

例) 音 声 : きょうはついたちです

書き起こし: 今日は[一日（ついたち）=Y]です

「なんですか」「なにですか」は通常の漢字仮名交じり文では漢字で書くことが一般的なので、両方漢字で表記する。しかし、漢字表記した場合、複数ある読み方のうち、どちらで、発音されたのか不明なため、読み方を書いて Y タグを付与する。「なんか」は平仮名表記が一般的であるため、平仮名で表記する。

例) 音 声 : それは、なんですか

書き起こし: それは、[何（なん）=Y]ですか

例) 音 声 : それは、なにですか

書き起こし: それは、[何（なに）=Y]ですか

例) 音 声 : なんか、ちょっと

書き起こし: なんか、ちょっと

ロールプレイでは「ににちかん」「さんにちかん」という表現が頻出する。「二日間」「三日間」と表記した場合、正しい読み方は「ふつかかん」「みっかかん」が一般的であるため、「ににちかん」「さんにちかん」は Y タグでなく、G タグを付与する。Y タグは表記の問題として複数の読み方がある場合に使用し、G タグは、発音や読み方が間違っている場合に使用する。

例) アルバイトは[ににちかん=G=二日間]が、いいです

例) 今は[さんにちかん=G=三日間]、働いて一

同一表記で漢字と仮名で書き分ける必要がある場合や、Y タグで処理する場合の一覧表を以下に示す。

表 5. 書き分け一覧

漢字表記	発音	表記例	備考
開く	あく	[開く (あく) =Y]	
	ひらく	[開く (ひらく) =Y]	
言う	いう	言う	
	ゆう	[言う (ゆう) =Y]	「そうゆうこと」「ああいうこと」「こういうこと」には平仮名書きするが、補足情報で「?チェック」とし、人手で確認する。
明日	あした	あした	
	あす	明日	
後	そのあと	[その後 (そのあと) =Y]	「[夏休み後 (なつやすみあと) =Y]」などは Y をつけるが、「直後」「午後」など「後」を含む 1 語はつけない。
	そのご	[その後 (そのご) =Y]	
一日	いちにち	[一日 (いちにち) =Y]	
	ついたち	[一日 (ついたち) =Y]	
入れる	いれる	[入れる (いれる) =Y]	
	はいれる	[入れる (はいれる) =Y]	
お札	おさつ	[お札 (おさつ) =Y]	
	おふだ	[お札 (おふだ) =Y]	
風	ふう	ふう	例) どんなふう
	かぜ	風	例) 風が強い
辛い	からい	[辛い (からい) =Y]	
	つらい	[辛い (つらい) =Y]	
通う	かよって	[通って (かよって) =Y]	
	とおって	[通って (とおって) =Y]	
昨日	きのう	昨日	
	さくじつ	[昨日 (さくじつ) =Y]	
今日	きょう	今日	
	こんにち	[今日 (こんにち) =Y]	
頃	ころ	[八時頃 (はちじころ) =Y]	

	ごろ	[八時頃 (はちじごろ) =Y]	
寂しい	さみしい	[寂しい (さみしい) =Y]	「淋」は使わない
	さびしい	[寂しい (さびしい) =Y]	
七月	しちがつ	[七月 (しちがつ) =Y]	
	なながつ	[七月 (なながつ) =Y]	
十分	じゅうぶん	[十分 (じゅうぶん) =Y]	
	じゅっぶん	[十分 (じゅっぶん) =Y]	
零	ゼロ	[零 (ゼロ) =Y]	
	れい	[零 (れい) =Y]	
何	なに	[何 (なに) =Y]	表現のまとまりでつける。 例) [何語 (なにご) =Y]、[何才 (なんさい) =Y]
	なん	[何 (なん) =Y]	表現のまとまりでつける。 例) [何年生 (なんねんせい) =Y]、[何本 (なんぽん) =Y]
	なんか	なんか	
日本	にほん	日本	
	にっぽん	[日本 (にっぽん) =Y]	[日本人 (にっぽんじん) =Y] [日本語 (にっぽんご) =Y]のように、表現のまとまりでYをつける。
話します	はなします	話します	
	はなしします	[話します (はなしします) =Y]	
降りました	ふりました	[降りました (ふりました) =Y]	
	おりました	[降りました (おりました) =Y]	
方	ほう	[方 (ほう) =Y]	
	かた	[方 (かた) =Y]	「言い方」「使い方」などの時はタグ不要
他	ほか	[他 (ほか) =Y]	
	た	[他 (た) =Y]	
皆	みんな	みんな	
	みな	皆	
良い	いい	いい	
	よい	良い	

私	わたし	私	
	あたし	あたし	
	わたくし	[私（わたくし）=Y]	

表 6. 迷いやすい表記・タグなしでいい口語や縮約形

発音	表記方法	備考
あんた	あんた	
あんまり/あまり	あんまり/あまり	
おっかい	おっかい	
こないだ	こないだ	[こないだ=G=この間]でもよい
sonだけ	sonだけ	
sonとき	[son=R]時	
たぶん	たぶん	
ちっちゃい/ちっさい	ちっちゃい/ちっさい	「ちさい」は[ちさい=G=小さい]
～ちゃ	しなくちゃ/言っちゃだめ など	「ては」の縮約「ちゃ」は G タグ不要
～ちゃう	食べちゃう/見ちゃった など	「てしまう」の縮約「ちゃう」は G タグ不要
～てく	増えてく/広がってった など	「ていく」の縮約「てく」は G タグ不要
～てる	見てる/言ってた など	「ている」の縮約「てる」は G タグ不要
～とく	置いとく/買っという など	「ておく」の縮約「とく」は G タグ不要。
とこ	とこ	「所」の縮約「とこ」は G タグ不要
～なきゃ	行かなきゃ	「なくては」の縮約「なきゃ」は G タグ不要
ほんと	ほんと	
やっぱ/やっぱし	やっぱ/やっぱし	
～れる	食べれる/起きれた など	ら抜き言葉は G タグ不要

4-6. 補足情報タグ（その他データ利用者に有益となりうる情報を記載する）（「?」）

書き起こした文字情報以外に付与しておくべき付属の情報は、補足情報タグを使用して、記載しておく。タグを付与した箇所に補足の情報を付ける場合は、アルファベットのタグの後に「?」を使い、情報を記載する（4-2（X タグ）、4-3-2（N タグ）参照）。

例) 宿題が[もでません=X?もう出ません/持てません/持っていない]

例) 生れたところは[オボネキーン=N?地名]

また、アルファベットのタグは必要ないが、補足情報はつけておいた方が良い箇所もある。その場合は、アルファベットのタグなしで補足情報のみを付与することも可能である。以下のような補足情報を付与した箇所は、自動の形態素解析では誤解析になりやすい箇所であるため、形態素解析後に人手での確認が必要である。その際、付与した補足情報の「?」を手がかりにチェックし、誤解析になっていたら修正する。

(1) 「の」「だ」「な」の過剰使用

「の」「な」「だ」の過剰使用は、おおむね正しく解析されるが、形態素解析で一部、品詞に誤解析を起こす可能性があるため、補足情報タグを付与する。

例) 大きい[な?過剰使用]ビーチ

例) 作りたくない[の?過剰使用]とき

例) 厳しい[の?過剰使用]先生

例) かわいい[だ?過剰使用]から

(2) 方言など

方言は正しく解析される場合とそうでない場合があり、判断が難しいため、「?方言など」という補足情報を付与する。

例) [そうやね?方言など]

例) [あかん=G=だめ?方言など]

(3) 語の説明

固有名詞の表記や発音などの話で、「経済の「経」です」「太陽の「陽」」のように、メタ言語的な説明を行うことがある。その場合も補足情報を付ける。「?語の説明」と表記する。

例) 経済の「[経?語の説明]」です

(4) 品詞情報など

助詞がポーズや発話の切れ目で出現している場合、誤解析になる場合がある。そのため、そのような箇所には補足情報で品詞を記入しておき、人手でのチェック作業で確認し、正しい品詞を付与するようにする。

例) 昨日、[あー=F]、[はー?助詞]、友達一、[と?助詞]、

例) 本を[よ=X]、本を読む時の、[の?助詞]、魅力

① 接続詞・接続助詞の「で」

話の転換などで使われる「で」は「接続詞」の場合が多い。助詞として使用されている場合、前後の文脈があれば、品詞情報は付けなくてもよいが、助詞から始まっている

ような場合は補足情報を付与する。

例) 大変ですねそれは〈はい〉 うん、[で?接続詞]、家の中はどうなってますか？

例) C：小学校と中学校と高校は上海

K：はい、上海です

C：[で?助詞]過ごしましたか

②助詞の「は」

助詞の「は」のうち、「は一」は語末であっても、後に読点があっても、感動詞になりやすいため、補足情報の「?助詞」を付与する。

例) 小説とか[は一?助詞]、そういうことでね

例) その時に[は一?助詞]、家を建てたいです

③助詞・接続詞・接続助詞の「が」

「が」は助詞、接続詞、接続助詞があるので、注意が必要である。音声と文脈から判断し、品詞情報をつける。以下の例は、「が」の前後にポーズがあり、「が」を特に取り立てて、接続詞として使っている。

例) その先生は、厳しいです、[が?接続詞]、自分の存在は特別

(5) 動詞の「ゆう」(言う)

「ゆう(言う)」は、前後の文脈によって誤解析になりやすい。そのため、「そうゆう」「ああゆう」「こうゆう」などは、補足情報の「?チェック」をつけておく。その際、「?チェック」は「ゆう」にのみタグを付与する。「ゆう」が「ゆ」と発音されている場合はGタグを付与する。

例) そう[ゆう?チェック]ことがありました

例) 合わせすぎたって[ゆう?チェック]か、友達に...

例) 応援してくれよって[ゆっ?チェック]て、許可もらって

例) 全然[ゆわ?チェック]ないです

例) 話があるって[ゆっ?チェック]てたけど、

例) そう[ゆ=G=言う]ことがありました

5. 個人情報について

本コーパスは、データを公開することに同意してもらった上で調査を実施しているが、個人を特定するような情報は伏せて公開することになっている。そのため、個人が特定される可能性があると思われる箇所は、書き起こしデータ上では一般名詞を使用して伏字化し、音声データ上ではピー音でマスキングする。個人情報と認定する箇所は以下の表7の通りである。

現住所については、「市」規模の地名は伏せない。しかし、対話タスクの最初の「ここま

でどうやってきましたか」などの話題で、調査協力者の現住所や最寄り駅名が具体的に述べられる場合がある。このような場合は、個人が特定される可能性があるため、「地名」「駅名」などを個人情報として伏字化する。

また、個人情報に関わる語断片は、X タグにはせず、個人情報に含めて【 】に含める。

例) a き、きた、北村先生

→【人名 1】先生

b き、きた[えー=F]北村先生

→【人名 1】[えー=F]【人名 1】先生

表 7. 伏字化する個人情報

	個人情報の種類	(例) 伏せる前 ⇒【伏せる際の一般名詞】
1	調査協力者の名前	マイケル ⇒【SES】: 調査 ID の一部
2	調査実施者の名前	山本 ⇒【調査実施者の名前 1】
3	友人、先生等の具体的な個人名	陳 ⇒【人名 1】
4	現地での所属大学、現所属大学以外の学校名（小中高、日本留学中の学校等）	東京大学 ⇒【大学名 1】 ※「大学」まで含める
5	所属大学の学部・学科名 (大学名が特定されると思われる場合)	日本語学科 ⇒【学科名 1】
6	所属機関名	さくら ⇒【ボランティア教室名 1】
7	誕生日の日付	15 日 ⇒【日付】
8	調査協力者が現在住んでいる場所	緑町 ⇒【地名】

ただし、表 8 にある内容は、インタビュー調査の質問項目になっているため、これらの地名は伏せない。

表 8. 削除しない個人情報

	個人情報の種類
1	故郷、出身地の地名
2	出身地の観光名所などの地名
3	観光スポットなどの固有名詞