話し言葉コーパスの転記タグ：『多言語母語の日本語学習者横断コーパス』と『日本語話し言葉コーパス』の比較

話し言葉コーパスの転記タグ：『多言語母語の日本語学習者横断コーパス』と『日本語話し言葉コーパス』の比較

西川 賢哉（国立国語研究所コーパス開発センター）†

Tags in Speech Corpora: A Comparison between `International Corpus of Japanese As a Second language' and `the Corpus of Spontaneous Japanese'

Ken'ya NISHIKAWA (National Institute for Japanese Language and Linguistics)

要旨

近年の話し言葉コーパスにおいては，発話を書き起こした転記テキストに，タグ（転記タグ）が付与されることが多い。本発表では，『多言語母語の日本語学習者横断コーパス』(I-JAS) および『日本語話し言葉コーパス』(CSJ) を対象に，そこで用いられているタグの種類・形式・目的・実際の用例を整理したうえで，両者の比較を行なう。比較の結果，(i)一方にしか存在しないタグもあるが，両コーパスでほぼ同様の機能を果たすタグも少なからず存在する（例えば，フィラー，語断片，発音誤りを表すタグ），(ii)ただし，同様の機能を果たすタグとはいえ，タグの適用範囲（転記テキストのどこからどこまでにタグを付与するか）や，タグの適用対象（タグをそもそも付与するか否か）など，細かい点では違いもある，ということが判明した。

１．はじめに

複数のコーパスを統一的な枠組みで扱うことができれば，利用者の利便性は高まる。その可能性を検討するための第一歩として，本稿では，話し言葉コーパスの転記テキストに付与されているタグ（転記タグ）の比較・検討を行なう。対象とするコーパスは次の二つである：

- I-JAS :『多言語母語の日本語学習者横断コーパス(International Corpus of Japanese As a Second language)』（cf. 迫田他 2016）
- CSJ :『日本語話し言葉コーパス(the Corpus of Spontaneous Japanese)』(cf. 前川 2006)

I-JAS の検索には，コーパス検索アプリケーション『中納言』を用いる（引用にあたり，迫田他(2016)等で用いられている形式に従ってタグを表記する)。CSJ については，現在のところ CSJ 中納言に転記タグの情報が格納されていないため，DVD 版の転記テキストを使用する（引用時，基本形のみを改行なしで表示することがある)。引用箇所中特に着目している個所を示すために，適宜下線を施す。

† nishikawa[@]ninjal.ac.jp

なお，I-JAS は現在も開発が進められており，今後タグの仕様が変更される可能性があることをあらかじめお断りしておく。

２．タグ概観

まず，I-JAS と CSJ のタグ対応表を表 1（次ページ）に示す。同種の機能を果たす，あるいは機能の一部が重複していると思われるタグは並べて表示してある。参照の便のため，両端に行番号を付す（本稿で「n 行目」とある場合，表 1 の行番号を指すことにする）。I-JAS において「タグ」とされているのは，厳密には表 1 の 1 行目から 10 行目までに挙げたものに限られ，11 行以降に「記号」として挙げたものは「タグ」とは呼ばれない。しかし，I-JAS でいう「記号」も実質的にはタグとみなしうること，また，CSJ との比較という点では「記号」と「タグ」を区別する必要はないこと，といった理由により，ここでは「記号」もタグとみなすことにする（必要な場合，「記号」も含めた I-JAS のタグ全体を「広義のタグ」，I-JAS で「タグ」とされているものを「狭義のタグ」と呼んで区別する）。

３．一方のコーパスにのみ存在するタグ

表 1 を見ると，一方のコーパスにのみ存在するタグも，両コーパスで類似の機能を果たすタグも存在することが分かる。本節ではまず，一方のコーパスにのみ存在するタグについて，両コーパスの目的・方針・使用に触れながら，簡単に見ておく。

３．１　タグ笑，泣，咳，L

24-27 行目のタグ笑，泣，咳，L は，CSJ にのみ存在し，I-JAS にはそれに対応するタグが存在しない。これは，両コーパスの目的・方針の違いによるものと思われる。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を目的として書き起こしされる」（迫田編 2016: 170）ものであり，書き起こしの基本方針の一つに「（略）発話はできるだけ発音に忠実に書き起こす。しかし，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。（略）」とある（迫田編 ibid；下線は引用者による）。笑，泣，咳，L に相当するタグが I-JAS に用意されていないのは，こうした基本方針を反映したものと見られる。一方，CSJ 構築の背景には，自発音声の自動認識システムの開発があった（前川 2006: 2）。この目的からすると，特殊な発声（笑いながら／泣きながら／咳をしながら／小声で…）については，タグ付けする価値のある個所だということであろう。

３．２　タグ Y

次に，9 行目のタグ Y を取り上げる。このタグは，I-JAS にのみ存在するタグで，次のように使用される：

ドアが[開いて（ひらいて）=Y]ました　【出典】I-JAS サンプル ID：JJC35-D

この例の場合，何の対応も施さなければ「開いて」の読みに曖昧性が発生するが（「あいて／ひらいて」），タグ Y を用いて読みを添えることで，その曖昧性が解消される。CSJ にこ

表 1. I-JAS, CSJ タグ対応表：迫田編(2016: 173, 176)，小磯他(2006: 80）をもとに作成

	I-JASのタグ				CSJのタグ				
	概要	内容	タグ表記	タグの由来	タグ	内容	使用例	付与対象	
1	(1)解析用の品詞を指定	フィラーを感動詞に指定	[α=F]	フィラー	(F)	フィラー，感情表出系感動詞，（応答表現）	(F あの)，(F うわ)，(F うーん)	基・発	1
2		外国語を名詞に指定	[α=N]	Noun	(O)	外国語、古語、方言、複雑な数式の読み上げ	(O ザッツファイン)	基・発	2
3		連体詞に指定	[α=R]	連体詞					3
4	(2) 解析用の語を指定	PC入力時の変換ミス	[α=K=β]	漢字・仮名					4
5		語中の長音、ポーズ	[α=T=β]	訂正	(K)	何らかの原因で漢字表記できなくなった場合	(K たち(F えー)ばな;橘)	基・－	5
6					(B)	語の読みに関する知識レベルの言い間違い	(B シブタイ;ジュータイ)	－・発	6
7		語や活用や発音の誤り	[α=G=β]	誤用	(W)	転訛や発音の怠けなど，一時的な発音エラー	(W ギーツ;ギジュツ)	－・発	7
8	(3) 解析から除外	意味不明語、語の断片	[α=X]	バツ	(D)	言い直し・言い淀み等による語断片	(D こ)これ，(D チ)チーズ	基・発	8
9	(4) 曖昧性への対応	複数の読みがある漢字語α	[α（読み）=Y]	読み					9
10		発音不明瞭（α1かα2）	[α1/α2=H]	発音	(?)	聞き取りや語の判断に自信がない場合	(? タオングー)，(? 堆積, 体積)	基・発	10
11	記号	聞き取り不能	*				(?)	基・発	11
12		間・ポーズ	、		<P>	短単位の内部に生じる0.2 秒以上のポーズ	オ<P:00453.373-00454.013>モイ	－・発	12
13		長音	ー		<H>	非語彙的な母音の引き延ばし	ソレデ<H>，スゴ<H>イ	－・発	13
14					<Q>	非語彙的な子音の引き延ばし	カイ<Q>セキ，ス<Q>ゴイ	－・発	14
15		個人情報	【 】		(R)	話者の名前・差別語・誹謗中傷など	国語研の(R ××)です	基・発	15
16		非言語情報	{ }		<笑>	言語音と独立に生じる話者の笑い	ガクセー<笑>ノ	－・発	16
17					<咳>	言語音と独立に生じる話者の咳	ソレデ<咳>	－・発	17
18					<息>	言語音と独立に生じる話者の息	ツマリ<息>	－・発	18
19		あいづち	〈 〉						19
20		上昇イントネーション	?						20
21		直接引用	「 」						21
22		書名、映画名、ドラマ名	『 』						22
23		複数の読みがある場合のフリガナ	（ ）						23
24					(笑)	笑いながら発話している箇所	(笑ナニガ)	－・発	24
25					(泣)	泣きながら発話している箇所	(泣ドンナニ)	－・発	25
26					(咳)	咳をしながら発話している箇所	シャ(咳リン)ノ	－・発	26
27					(L)	ささやき声や独り言などの小さな声	(L アレコレナンダッケ)	－・発	27
28					(A)	アルファベット・算用数字・記号の併記	(A シーディーアール;C D － R)	基・－	28
29					(D2)	助詞・助動詞・接辞・数字の言い直し	そこ(D2 が)に，(D2 不)不自然	基・発	29
30					(M)	音や言葉に関するメタ的な引用	助詞の(M は)は(M わ)と発音	基・発	30
31					(X)	非朗読対象発話（朗読における言い間違い等）	(X 実際は)実際には，	基・発	31

の種のタグが存在しないのは，転記テキストの仕様による。CSJ では，図 1 に示すように，漢字仮名を中心に書き表される「基本形」と，実際の発音を仮名の範囲で忠実に書き表した「発音形」という 2 種の表記法が採用されている。そのため，基本形の表記において読みの曖昧性が発生するケースであっても，対応する発音形を参照することで，その曖昧性は解消される。CSJ においては，わざわざタグによって読みを与える必要はないわけである。

0285 00642.999-00645.100 L:		
十一時ぐらいに	&	ジューイチ(W イ;ジ)グライニ
うちに	&	ウチ(W ン;ニ)
帰ってきたんですけど	&	カエッテキタンデスケド
0286 00645.894-00649.273 L:		
(F その)	&	(F ソノ)
門は	&	モンワ
閉まってるんですが	&	シマッテルンデスガ
玄関が	&	ゲンカンガ
薄く	&	ウスク
開いていまして	&	アイテイマシテ

図 1. CSJ 転記テキスト例(S02M0198)：
&の左側が基本形，右側が発音形

なお，基本形と発音形という二種類の表記法が存在する CSJ においては，タグを(i)基本形に付与するのか，(ii)発音形に付与するのか，(iii)基本形と発音形の両方に付与するのか，を規定しておく必要がある。表 1 では右端の「付与対象」欄にその情報が記されている。

３．３　タグＡ

I-JAS におけるタグ（ここでは「記号」を含まない，狭義のタグ）は，もっぱら形態素解析の精度向上のために用いられる（迫田編 2016: 173ff.）。一方，CSJ においては，同様の目的で使用されるタグも多くあるが，それに限定されるわけではない（小磯他 2006: 27-28）。表 1 の 28 行目のタグ(A)は，転記テキストの可読性を高めるために導入されたタグである。このタグを用いることにより，算用数字やアルファベットを表記に添えることができる。

(A 三．ゼロ六;３．０６)　　& サンテンゼロロク　【出典】CSJ ID：A06F0075
(A エイチエムエム;HMM)は　& エイチエムエムワ　【出典】CSJ ID：A01M0099

I-JAS にはタグ A に対応するタグは用意されていない。I-JAS の表記ルールでは，数字は漢数字で，アルファベット通りの発音をしている場合はアルファベット全角で記すことにな

っている（迫田編 2016: 170-171）。

４．類似の機能を果たすタグ

本節では，両コーパスで類似の機能を果たすと思われるタグを比較する。

４．１　フィラー

フィラーおよび感情表出系感動詞には，どちらのコーパスにおいても，タグ F が付与される（表 1 の 1 行目）。

はい、[あー=F]島田店長，ちょっとよろしいですか？　【出典】I-JAS サンプル ID：JJC46-RP1
発表内容ですが(F あー)まず研究の背景として　【出典】CSJ ID：A01M0065

相違点として，フィラーが連続した場合，I-JAS では全体に一つのタグを付与する場合があるのに対し，CSJ では個々のフィラーにそれぞれタグ F を付与する。

でも実際はそんなにー、[あの、んー=F]私，　【出典】I-JAS サンプル ID：JJC32-RP1
この例に見られますように(F あの)(F んー)一方が　【出典】CSJ ID：A06F0073

I-JAS におけるこの対応は，「学習者の多様な場つなぎ的表現に関して，何を一語とするかという判断が難しいため」である（迫田編 2016: 178）。このような対応の違いにより，両コーパスをタグ F で検索した場合，フィラーの頻度や種類に差が出てくることが予想される（例えば，タグ F が付与された「あの」を I-JAS で検索しても，[あの、んー=F]はヒットしないおそれがある）[1]。

４．２　語の断片等

語の断片に対しては，I-JAS ではタグ X，CSJ ではタグ D が付与される（表 1 の 8 行目）。

今のーシフトはー[あの=F]、[し=X]週三回，　【出典】I-JAS サンプル ID：JJC32-RP1
(F えー)(D し)下の項が少し変わってきます　【出典】CSJ ID：A01M0097

I-JAS のタグ X は，語の断片の他に，不明語，学習者によるオリジナルの語に対しても付与される点で，CSJ のタグ D とは異なる。

[みやまー=X]　（迫田 2017: 183）

[1] このことは，迫田編(2016: 178)にある通り，I-JAS 開発者によって既に認識されている問題である。なお，「タグ F の内部に”、”がある場合，そこでタグ F を括り直す」という処理（例：[あの、んー=F]→[あの=F]、[んー=F]）を検討中であるとのご連絡を最近 I-JAS 開発者からいただいた。

お父さんが[ピャーピャー=X?擬音語・擬態語]怒って　（迫田 2017: 183）

また，前述のタグ F と同様の相違点だが，I-JAS ではタグ X を付与する箇所が連続して現れる場合は，まとめてタグ X を付与しているのに対し（迫田 2017: 182-183），CSJ では複数の語の断片の連続と解釈される場合には，それぞれにタグ D を付与している。

でもほんとに、[かな、悲し=X]、悲しくなった、　【出典】I-JAS サンプル ID：JJC28-I
住所とか聞きますから(D き)(D き)(D き)聞きたいんですけどって
【出典】CSJ ID：S01F0183

４．３　発音誤り

発音の誤りについては，I-JAS ではタグ G，CSJ ではタグ W が用いられる（表 1 の 7 行目）。

雰囲気が、ちょっと、[ちょと=G=ちょっと]違うと友達、
【出典】I-JAS サンプル ID：JJC28-I
ちょっと ＆(W チョト;チョット)
早めに　＆ ハヤメ(W ン;ニ)
出る　＆ デル
時に　＆ トキニ　【出典】CSJ ID：S02M0198

両コーパスの重要な相違点として，CSJ ではタグ W の範囲（スコープ）は原則として短単位なのに対し[2]，I-JAS ではそのような制約はないということが挙げられる。

記事を[かきて=G=書いて]います　【出典】I-JAS サンプル ID：GAT39-I
書いてきたんですね ＆(W カエ;カイ)テキタンデスネ　【出典】CSJ ID：S01F1522

「書いて」は短単位としては，「書い」と「て」に分割され，CSJ では「書い」の部分にだけタグ W が付与されているが，I-JAS では「書いて」全体にタグ G が付与されている。日本語学習者の日本語という観点から独自にタグ G の範囲を規定することには大きな意義があると思われるが，少なくとも，中納言のような短単位ベースの検索システムでコーパス検索をする時には，個々の短単位にこの種のタグが付与されていた方が扱いやすい，ということは言えるであろう。別の例を挙げると，

[使えな、なかた=G=使えなかった]　【出典】I-JAS サンプル ID：JJC28-I

[2] 例外として，短単位境界で音の融合などが生じそこで切り離しがたい場合，複数の短単位にまとめてタグ(W)を付与することを CSJ では認めている（小磯他 2006:104）。例：僕は ＆(W ボクワ;ボカー)

は，CSJ 方式ではおそらくタグ D とタグ W を使って，次のように書き起こされる：

使えなかった & ツカエ(D ナ)(W ナカ;ナカッ)タ

また，適用範囲とは別の問題として，そもそもこの種のタグを適用すべきか否かはっきりしないケースがある。

[助けてくれ、くれません=G=助けてくれません] 【出典】I-JAS サンプル ID：TTH27-I
入ろうと[思いですけど=G=思いますけど] 【出典】I-JAS サンプル ID：KKD20-ST2

これらの例は確かに誤用ではあるが，形態素解析に関して致命的な失敗を招くとは考えられず，その点ではタグ G は不要であるとも思える。CSJ の基準では，「助けてくれ」の部分にはいかなるタグも付与されない。また，「思いです」の類については，おそらく CSJ には明示的な規定はない―母語話者による発話のコーパスなので，そもそも想定していない―が，この部分にもタグは付与されないと思われる。

４．４　外国語

外国語については, I-JAS ではタグ N が, CSJ ではタグ O が付与される（表 1 の 2 行目）。

一番高いところは[シャンライ=N]のほう 【出典】I-JAS サンプル ID：JJC09-I
(O カナディアンレイジング)でもって 【出典】CSJ ID：A05F0039

一見同種の機能を果たすように思われるタグではあるが，実際のところ，両者は適用対象もその導入目的も大きく異なる。I-JAS のタグ N は，外国語で表現された語に加え，アニメなどの架空のカタカナ語の固有名詞にも付与される。このタグの目的は，タグが付与された要素の品詞を「名詞」とすることである（ただし，解析用辞書に登録されている語であれば，その語の情報が与えられる）。一方で，CSJ のタグ O は，外国語に加え，古語，方言，複雑な数式の読み上げに付与される。その目的は，必要に応じて CSJ の利用者が分析から除外できるようにすることである。前述のとおり，CSJ 構築の背景には，自発音声の自動認識システムの開発があった。そのためには，現代共通日本語の体系から外れた個所は除外するほうがよい。この際，タグ O の情報が利用できる。

このような目的の違いにより，CSJ ではタグ O が期待される個所で，I-JAS ではいかなるタグも付与されない，というケースが生じる。I-JAS 書き起こしマニュアル（内部資料）によると，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ことになっている。

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

しかし, CSJ では下線部にタグ O が必要である。このような事例があるとすると,「外国語」という観点から，I-JAS のタグ N と CSJ のタグ O を単純にまとめて扱うことは難しい。

５． おわりに

I-JAS と CSJ で用いられているタグの比較を行なった。そもそも転記タグは，コーパスの目的や方針に基づいて設定されるものであり，一見機能的に類似しているタグであっても，本稿で見たように，細部においては違いもありうる。しかし，冒頭で述べた通り，コーパスの違いに関わらず，統一的な枠組みで扱えるのであれば，そちらのほうが利用者にとっての利便性という点では望ましい。今後は，本稿で明らかになった問題点を解消しつつ，複数のコーパスを統一的に扱う仕組みを検討する。また，他の話し言葉コーパス―例えば，現在国立国語研究所で構築中の『日本語日常会話コーパス』(川端他 2017) ―で用いられているタグについても，今後の比較の対象としたい。

謝　辞

本研究は国立国語研究所コーパス開発センターの共同研究「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」の成果である。I-JAS 開発者の佐々木藍子氏および小西円氏からは，I-JAS の仕様について数々のご教示をいただいた。記して感謝する。

文　献

川端良子・臼田泰如・西川賢哉・徳永弘子・小磯花絵 (2017)「『日本語日常会話コーパス』の転記基準と作業工程」言語資源活用ワークショップ 2016 発表論文集.

小磯花絵・西川賢哉・間淵洋子 (2006)「第 2 章 転記テキスト」『日本語話し言葉コーパスの構築法』国立国語研究所, pp.23-132.（http://pj.ninjal.ac.jp/corpus_center/csj/k-report-f/02.pdf よりダウンロード可能）

迫田久美子・小西円・佐々木藍子・須賀和香子・細井陽子 (2016)「多言語母語の日本語学習者横断コーパス」『国語研プロジェクトレビュー』6:3, pp.93-110.（https://ninjal-sakoda.sakura.ne.jp/lsaj/?page_id=364 よりダウンロード可能）

迫田久美子（編）(2016)『海外連携による日本語学習者コーパスの構築―研究と構築の有機的な繋がりに基づいて―I-JAS 構築に関する最終報告書』平成 24-27 年度科学研究費助成事業（基盤研究 A）研究成果報告書（https://ninjal-sakoda.sakura.ne.jp/lsaj/?page_id=364 よりダウンロード可能）

前川喜久雄 (2006)「第 1 章 概説」『日本語話し言葉コーパスの構築法』国立国語研究所, pp.1-21.（http://pj.ninjal.ac.jp/corpus_center/csj/k-report-f/01.pdf よりダウンロード可能）

関連 URL

コーパス検索アプリケーション『中納言』　https://chunagon.ninjal.ac.jp/