音声言語情報処理 34－30
(2000. 12. 22)

「日本語話し言葉コーパス」の書き起こし基準について

○ 小磯 花絵[†], 土屋 菜穂子[†‡], 間淵 洋子[†‡‡], 斉藤 美紀[†*],
籠宮 隆之[†], 菊池 英明[†], 前川 喜久雄[†]

[†] 国立国語研究所　　[‡] 青山学院大学大学院
[‡‡] 東京都立大学大学院　　[*] 東京大学大学院

〒115–8620 東京都北区西が丘 3–9–14

あらまし　国立国語研究所，通信総合研究所，東京工業大学では，科学技術振興調整費開放的融合研究制度の下，自発性の高い話し言葉の情報処理技術の確立を目標に活動を進めている．現在国立国語研究所では，このプロジェクトの一環として，モノローグを対象とした大規模な日本語話し言葉コーパスを作成している．このコーパスには，約 700 時間 (約 700 万形態素に相当) の音声，書き起こしテキスト，および形態素などの情報が含まれる予定である．本稿では，本コーパスの書き起こし基準の詳細について紹介する．

キーワード　話し言葉コーパス，自発的発話，モノローグ，書き起こし基準

Transcription Criteria for *the Corpus of Spontaneous Japanese*

○ Hanae Koiso[†], Naoko Tsuchiya[‡†], Yoko Mabuchi[‡‡†], Miki Saito[*†],
Takayuki Kagomiya[†], Hideaki Kikuchi[†], Kikuo Maekawa[†]

[†] The National Language Research Institute　　[‡] Aoyama Gakuin University
[‡‡] Tokyo Metropolitan University　　[*] Tokyo University

3–9–14 Nishigaoka, Kita–ku, Tokyo, 115–8620 Japan

Abstract　A large–scale corpus of spontaneous Japanese speech is being compiled as a joint work of the National Language Research Institute, the Communications Research Laboratory, and Tokyo Institute of Technology. This corpus is designed to contain about 700 hours of speech (about 7 million morphemes), a transcription, and various tagging information such as POS information. This article describes the details of the transcription criteria for this corpus.

Key words　spoken corpus, spontaneous speech, monologue, transcription criteria

1 はじめに

国立国語研究所，通信総合研究所，東京工業大学では，科学技術振興調整費開放的融合研究制度の下，自発性の高い話し言葉の情報処理技術の確立を目標に活動を進めている．本稿では，国立国語研究所が本プロジェクトの一環として作成している「日本語話し言葉コーパス (Corpus of Spontaneous Japanese: CSJ)」の書き起こし基準について紹介する．

CSJは，日本語の自発音声コーパスとしては例を見ない規模 (700〜800 時間) を目指している．この規模の音声を書き起こすには，多くの作業者が長い年月に渡って作業しなければならず，質の揃ったデーターを作成することは非常に難しい．本プロジェクトでは，各種マニュアル (6 種類，計 200 頁弱) を作成したり，7 万語程度の辞書を整備するなど，基準を揃える為の準備を行なってきた．本稿では特に，表記の揺れを統制する為の基準や口語表現の扱い，また言い淀みなど談話に生じるさまざまな現象を体系的に表現する為の枠組みについて述べる[1]．

2 コーパスの設計の概要

CSJ の設計の要旨を以下に示す (詳細は前川他 (2000), Maekawa, et al.(2000) を参照)．

言語変種： 自発性の高いモノローグ．全国共通語[2]中心．

規模： 700 万形態素 (700〜800 時間に相当)．

発話内容：

- 学会講演 (全国大会，研究会などの発表や講演)．
- 模擬講演 (本研究の為に派遣された一般の人が行なう 10 分程度のスピーチ)．
- その他 (一般の講演会や大学の講義など)．

付与情報：

- 形態素情報 (長めの単位と短めの単位を付与)．
- 分節音情報 (分節単位ラベルとその時間情報を付与)．
- 韻律情報 (J_ToBI に準拠したラベリングを予定)．
- フィラーや言い淀みなどの情報．

分節音情報や韻律情報は，700 万形態素分のうち 50 万形態素分にのみ付与する．この 50 万形態素の部分をコアと呼ぶ．形態素情報についても，全体に対しては自動で処理できる範囲にとどめ，コアのみ人手修正を行ない高い精度を確保する．

3 書き起こしの基準

3.1 基本単位

書き起こしテキストの基本となる単位 (以下「転記基本単位」) として，文などの文法的基準が利用される場合もある．しかし自発音声では，必ずしも文の認定は容易ではない．そこで CSJ では，転記基本単位の認定基準として物理的な指標を採用した．原則として，言語音が 200 ミリ秒以上の途切れなく連続して生じている区間を転記基本単位とする．ただし，言語的な文末形式（述語の終止形や終助詞など）が存在している場合には，50 ミリ秒以上 200 ミリ秒未満の区切れであっても，その文末形式の後で転記基本単位を分割する[3]．

転記基本単位の同定作業は，計算機上に音声波形を表示し音を聴取しながら行なう．この作業により各単位の開始・終了時刻が確定する．次に，単位の開始・終了時刻を参考に，エディター上で音を聞きながら文字化作業を行なう．

書き起こしテキストの例を図 1 に示す．テキスト中の 1，4，9，13，15 行目には，各転記基本単位の情報，

(1) 転記基本単位の通し番号 (4 桁の数字)

(2) 開始，終了時刻 (それぞれ秒単位)

(3) 話者 ID(この例では"L")

が記されている．これら情報部の後に，その単位の発話内容が記されている．

言語音以外の音であっても，笑い声や拍手など，談話の流れを把握する上で重要と考えられるものについては，言語音と同様に単位の認定を行ない，書き起こしテキストに書き表わす．図 1 の最後の行がこれに相当する．

3.2 二種類の表記法：『発音形』と「基本形」

本プロジェクトの柱の一つである音声認識研究では，書き起こしテキストを用いて音響モデルと言語モデルを構築する．音響モデルを構築する為には，音声データーと実際の発音情報が必要である．今回対象とする自発音声は，朗読音声とは異なり発音の怠けや言い間違いなどが頻繁に生じる為，忠実な発音の記録が重要となる．

また，言語モデルの構築には，通常漢字仮名交じりテキストが利用されるが，その際重要なことは，同一の語や句の表記が統一されていること，つまり表記の揺れが存在しないことである．

CSJ では，上記二つの目的に沿った書き起こしテキストを，共に人手で作成する．前者を発音形，後者を基本形と呼ぶ．図 1 の "&" の左側に記されているのが基本形，右側が発音形である．両表記の対応が容易に取れるよう，概ね文節に相当する単位で改行されている．

3.3 発音形の表記法

発音形では，実際に発音された音を，片仮名を利用してできる限り正確に書き表わす．表記の概要を以下に記す．

使用可能な文字の範囲 外来語や擬音語，擬態語，言い間違いなどでは，「ヴァ」や「ツィ」など，直音，拗音系列以外の音 (以降「周辺的モーラ」) も頻繁に出現する．その為現在では，直音，拗音，促音，撥音，長音 (記号) に加え，35 の周辺的モーラも発音形の表記に利用している．

[1] 本原稿は，3.4 節に示す表記法に従って書き記している．

[2] 分節音，語彙，文法が東京語に類似していれば全国共通語と認定する．その為，韻律が地方色を帯びている場合なども本コーパスの対象となる．

[3] 文末形式が出現したか否かの判断を行なうだけであり，実際にそこが文末であるかどうかの判断は行なわない．

0256 00417.057-00419.269 L:		
情報処理絡みのものが	&	ジョーホーショリガラミノモノガ
多くなりますので	&	オークナリマスノデ<H>
0257 00419.774-00421.920 L:		
そこら辺で	&	(W スコラ; ソコラ) ヘンデ
内容が	&	ナイヨーガ
(D かたか)	&	(D カタカ)
偏るだろうと	&	カタヨルダロオト
0258 00422.527-00423.569 L:		
で	&	デ
ここに	&	ココニ
関しても	&	カンシテモ
0259 00424.783-00425.401 L:		
(F えー)	&	(F エー)
0260 00425.626-00428.147 L:		
何か	&	ナニカ
ちょっと	&	チョット
考えなきゃ	&	カンガエナキャ
いけないかもしれないと	&	イケナイカモシレナイト
思っております	&	オモッテオリマス
0261 00429.016-00429.198 L:<雑音>		

図 1: 書き起こしテキストの例 (一部抜粋)

助詞の「は」「を」「へ」の表記　実際の発音である「ワ」「オ」「へ」を使用.

「ヅ」「ヂ」の表記　一律「ズ」「ジ」に統一.

母音の連鎖　「かあさん」のように母音が連鎖しており，実際の発音が長音化している場合には，長音記号 (ー) を利用して「カーサン」のように表記.

発音の怠けや転訛，言い間違い　「六義園」を「リッギエン」,「あります」を「アリマ」,「共分散」を「キョーブサン」と発音するなど，発音の怠けや転訛，言い間違いなどが生じた場合には，実際に発音された音を可能な限り正確に書き表わす．ただし，3.6 節で詳述する (W) タグを利用して，丁寧に発音された場合に生じるであろう音も併記する (図 1 の 5 行目 (W スコラ; ソコラ) を参照).

3.4　基本形の表記法

先に述べたように，基本形では表記の揺れを極力抑える必要がある．そこで CSJ では，漢字と平仮名の使い分け (例：「全て／すべて」) や，送り仮名の振り方 (例：「行なう／行う」) など，表記の原則を定め，それを基に表記マニュアルを作成した．またその原則に従い，実際の語の表記を定めた用語リスト (辞書) を作成した．以下にその詳細を示す.

3.4.1　字種およびその使用範囲

使用する字種　基本形の表記には，原則として漢字，平仮名，片仮名を使用する．ただし，これらの字種に併記する形で，アルファベット (ローマ字・ギリシャ文字) や算用数字，幾つかの記号 (". "や"ー"など) も使用する (3.6 節のタグ (A) の項を参照).

漢字の使用範囲　原則として JIS 第 1 水準を採用．ただし JIS 第 2 水準の漢字であっても，「曖昧」の「曖」や「天丼」の「丼」のように，一般に漢字でよく表記されているものについては，その使用を認める.

仮名の使用範囲　和語や漢語の表記には，直音・拗音系列の平仮名，および促音 (っ)，撥音 (ん) を使用する．また片仮名語の表記には，これらに加え 25 の周辺的モーラも使用する[4].

3.4.2　漢字・平仮名語の表記原則

漢字表記か平仮名表記か　「例えば／たとえば」や「全て／すべて」のように，表記が漢字と平仮名で揺れるものが数多く存在する．漢字・平仮名表記のどちらも頻繁に使用されるものについては，原則として漢字を優先して採用するという方針を取る (上記の例では，「例えば」「全て」が採用される).

これは，形態素情報を自動的に付与する際に，平仮名で表記するよりも，漢字で表記した方が，高い処理精度が期待できるからである．2 節で述べたように，形態素情報の付与については，全体の 9 割強が自動処理の範囲で行なわれる．その為この方針は，コーパス全体の精度にかかわるものであると言える.

個々の語の表記については，上記の原則に基づき，関連する語との整合性を検討しながら決定する．ただし，関連語との表記の一致を強く押し進め，無理に表記を統一することはしない．例えば，「とびきり上等な」の「とびきり」は，その語の構成要素である「飛ぶ」「切る」と表記を合わ

[4]発音形で使用する 35 の周辺的モーラのうち，基本形では「ヴァ」や「テュ」など，10 のモーラを使用しない.

せると，「飛び切り」と表記されることになる．しかし，この語が「飛び切り」と漢字表記されることはめったにない．このようなものまで無理に漢字に統一することはしない．

なお，当て字に関しては，常用漢字表の付表に記された熟字訓（「玄人」や「相撲」など）のみ使用可能とし，それ以外（「蕎麦」や「矢張り」など）は用いない．

複数種類の漢字表記が可能な場合　原則，一単語一表記とする．例えば，「憧れ／憬れ」や「一獲千金／一攫千金」のように，JIS第1水準の漢字とJIS第2水準の漢字の両方で表記されるような場合には，JIS第1水準の漢字を採用する．

「悲しい／哀しい」や「会う／逢う」，「尊ぶ／貴ぶ」のように，厳密には同義語ではなく，微妙なニュアンスの差があるものもある．このうち，書き分けが困難で表記の揺れが生じ易いものについては，それが片方の漢字で代用可能である場合に限り，一種類の表記（この例では前者）に統一する．片方の漢字で代用できないものについては，無理に統一しない．その際，揺れをでできるだけ抑える為に，多く出現する語については，書き分けに関する基準を整備した．

また「表わす／現わす」や「計る／図る」のように，明らかな同音異義語に関しては，表記を書き分ける．

送り仮名の統一　「行なう／行う」のように，用言で複数の送り仮名の候補がある場合には，一律，送り仮名の字数の多い方を採用する．また「書き留め／書留」のように，名詞で送り仮名の有無に揺れがあるものについては，原則として送り仮名を付ける方を採用する．ただし，「関取」や「取締役」など，慣習的に送り仮名を付けないものについてはその限りでない．

以上に挙げてきたような表記の統一は，出現する全ての語に渡ってなされるものである．したがって，日常において個々人が持つ表記の直観とは若干食い違う場合もありうる．あくまで表記の統一を目的としたものであること，および各種国語辞書を参照してできる限り無理のない範囲で統一したことを申し添えておく．

3.4.3　片仮名語の表記原則

片仮名で表記するものは，外来語，外国語，専門用語や俗語などで慣習的に片仮名表記をするもの（「ト書き」や「ダフ屋」など），および一部の動植物名（「リス」や「カバ」など）に限定している．それ以外のものを片仮名表記することはない．

上記片仮名語の中でも，特に外来語については，「ビオラ／ヴィオラ」や「ウインドー／ウィンドー」などのように，表記の揺れが非常に多く見られる．その為，漢字・平仮名表記の場合と同様に表記を統一する必要がある．片仮名語の場合，上記の例のように，「ビ」と「ヴィ」，「ウイ」と「ウィ」など，表記の揺れが起き易いパターンが数多く存在する．そこで，パターン毎に表記の方針を整理し，マニュアル化した．

3.4.4　統一/書き分けの一例

以下に，基本形の表記の統一・書き分けの規定のうち，特に重要なものの例を幾つか示す．

実質名詞・形式名詞　「こと」や「もの」「ところ」は通常，実質名詞の場合には漢字で，形式名詞の場合には平仮名で表記される慣習が高い．しかし，その区別は非常に難しく，書き分けが困難である．そこでこれらの語については，実質名詞・形式名詞にかかわらず，一律平仮名表記に統一するという方針を取る[5]．

本動詞とテ形複合動詞　「行く」「来る」「置く」「見る」「上げる」「貰う」「参る」は，単独で本動詞として出現する場合，漢字で表記する．一方，「やって おく」や「食べて みる」のように，テ形複合動詞の後項に現われる場合には，平仮名で表記する．

「言う」と「いう」　動詞の「言う」は，通常漢字で表記されるが，「山田と いう 人」や「そう いった 問題」など，「言う」という動作が形骸化されたような用法では，平仮名書きされることが多い．しかし，形骸化しているか否かの判断は非常に難しく，その書き分けは揺れを招き易い．

そこで，特に形骸化が多く見られる以下の組み合わせパターンで出現した場合に限り，平仮名表記とする（上記2つの例文も参照）．ただし，この条件を満たした場合であっても，明らかに動作性を有することが判断できる場合には，漢字で表記する．

$$\left\{\begin{array}{l}\text{指示副詞：ああ／こう／そう／どう}\\ \text{引用の助詞：と／って}\end{array}\right\}+\left\{\begin{array}{l}\text{いう}\\ \text{いった}\end{array}\right\}+\{\text{体言}\}$$

「言う／いう」の例に見られるように，使い分けが微妙な場合には，前後の語との共起関係を見るなど，できるだけ客観的な基準を構築するようにした．また，客観的な基準の確立が難しく，使い分けの揺れが頻繁に生じるような場合には，以下のように対処した．(1) 実質名詞・形式名詞の項に挙げた「こと」や「もの」「ところ」のように，どちらかの表記に統一してもそれ程違和感のないものについては，無理に使い分けることはせずに統一した．(2) 表記を統一すると違和感が生じる為，使い分けがどうしても必要な場合には，使い分けの基準を明記し用例を示すようにした．

3.4.5　用語リスト・辞書の整備

上記のように表記の基本原則を確立し，それをマニュアルに示しても，具体的にある語をどう表記するかについては，必ずしも一意には決定しない．そこで，実際の作業における表記の決定・統一を支援する為に，以下の作業環境を整備した．

[5]ただし，「事柄」や「物語」のように，単語の構成要素である場合にはその限りではない．

用語リストの作成　表記の基本原則に従い，実際の語の表記を定めた用語リスト (現時点で 7 万語程度) を作成した．このリストから，オンラインで用語を検索する為の辞書と，仮名漢字変換用の辞書が生成される．これらの辞書については後述する．

用語リストは，語句の読み，表記，品詞情報，および備考から構成される．備考には，間違い易い表記についての注意事項や関連語に関する情報，また略語や口語・縮約形における元の形などの情報が記載されている．この辞書には，使用可能な表記に加え，使用不可能な表記についても，使用の可否が区別できる形で登録されている．

書き起こし作業の過程でリストに存在しない語句 (未知語) が出現した場合には，表記に関する責任者が，表記原則や慣用等に照らし合わせ，表記を決定した上で登録する．未知語の登録を含め，作業者が本リストに変更を加えることは許されていない．

オンライン辞書　前掲の用語リストから，語句の読み，表記，使用の可否，品詞情報，備考を，可読性の高い形式で表現した辞書．書き起こし作業を行なっているエディター上で，本辞書を対象に語句の言い切り形から用語を検索することができる．

仮名漢字変換用辞書　前掲の用語リストから仮名漢字変換用の辞書が作成される．使用可能な表記のみが登録されており，使用できない語は変換候補として現われないようになっている．また，例えば一般名詞の場合には使用できない表記が，固有名詞の場合に限り使用が認められている，といったように，状況に応じて使用の可否が変わるものがある．そこで，固有名詞などに特例的に認められている表記については，作業者の注意を促す為に，それを示す記号と共に変換候補に現われるようになっている．

3.5　口語表現

自発性の高い話し言葉には，「こりゃすげえ (これは凄い)」や「見たげる (見てあげる)」といった，くだけた表現が数多く出現する．CSJ では，このような口語表現を積極的に基本形に書き表わすという方針の下で作業を進めている．

本コーパスで扱う口語表現は，(1) 音の転訛を伴い，(2) くだけた場面で (意図的に) 使用される表現で，(3) 一個人に限らず幅広く観察されるものに限定する．例えば「リッギエン (六義園)」や「コレア (これは)」などは，あくまで発音上の問題であり，場面に応じた使い分けがなされている訳ではないと考えられる為 (条件 2 への抵触)，ここでは口語表現とは考えない[6]．

CSJ では，80 時間のデーターを書き起こした段階で，そこに出現した口語調の表現を抽出し，上記三つの条件と照らし合わせながら，口語表現として登録する語の選別を行なった[7]．

その際，表現をそれぞれ個別に登録するのではなく，ある程度体系的に整理した上で，同じ，あるいは類似した現象は，できるだけ同様の扱いをするように心掛けている．例えば「知んない」「やんない」「取んない」などは，動詞活用語尾「ら」に否定の助動詞「ない」が後続する場合に撥音化する，というパターンの口語表現である．このような場合，それぞれの表現をマニュアルに個別に登録するのではなく，上記のパターンを (具体例と共に) 示すようにしている．

3.6　各種情報のタグ付け

書き起こしの際には，言い直しや言い間違い，フィラーといった談話現象や，笑いながら話したり母音を通常よりも引き延ばすといった音声的現象など，談話に生じるさまざまな現象を体系的に表現する必要がある．CSJ では，表 1 に示すようなタグを書き起こしテキストに付与している．本節では幾つかのタグについて簡単に説明する．

タグ (F)　「あのー」や「その」,「えっと」といった，言い淀み時に生じる場繋ぎ的な機能を持つフィラーに付与するタグ．語彙を限定し，その範囲内で付与する．「あの」や「その」は，フィラーか連体詞かで迷うことが多い．前後の文脈から指示する対象が明らかである場合以外はフィラーと判断する．

タグ (D), (D2)　「あたら 最新の研究で」の例に見られるように，何かを言い掛けそれを別の表現で言い替えた場合の，言い掛けの部分 (「あたら」) を対象に付与するタグ．「あたら 最新の」のように，単語より短い語の断片の場合には (D) を，「ここ の に集めて」のように，機能語 (助詞・助動詞の類) が言い直されている場合には (D2) を付与する．「学校 大学の」のように，機能語以外の単語，あるいはそれ以上の長さの単位が言い直されているものは対象としない．

タグ (W)　「リッギエン」や「キョーブサン」のように，発音の怠けや転訛，言い間違いなどが生じた場合に付与するタグ．(W リッギエン; リクギエン) のように，セミコロンの左側に，実際に発音された音を可能な範囲で正確に書き表わすと同時に，セミコロンの右側には，丁寧に発音された場合に生じる (と予想される) 音を併記する．また,「アメリカの大統領 エリツィン は」や「これ が やります」のように，世界知識や文法のレベルで間違っている，あるいは適格でないものは，修正の対象としない．

タグ (?)　音の聞き取りや漢字表記などに自信がない場合に付与するタグ．(? 対照, 対象, 対称) のように複数の候補を指定したり，(W コッコ;(? コクゴ)) のように他の記号と組み合わせて使用することもできる．

[6] これらは，タグ (W) で対処し基本形には「六義園」「これは」と表記する (3.6 節参照)．

[7] 条件の (2) や (3) の判断は，厳密には現段階で確定できるものではない．コーパス全体の書き起こしが終了した時点で，再度検討する必要があるだろう．

表 1: 書き起こしテキストに使用されるタグ一覧

I 文字範囲を指定し，その範囲の性質に言及するタイプ		
◇ (F)	フィラー・感情表出系感動詞	(F あの)，(F うわ)
◇ (D)，(D2)	言い直し	(D こ) これ，これ (D2 は) が
◇ (W)	言い間違い，転訛，発音の怠け，など	(W ミダリ; ヒダリ)
◇ (?)	聞き取り，語彙同定，漢字表記に自信なし	(? タオングー)
	・複数の候補がある場合	(? あのー，あんのー)
	・全く分からない場合	(?)
◇ (M)	音や言葉に関する引用	(M わ) は (M は) と表記
◇ (O)	外国語や古語，方言など	(O ザッツファイン)
◇ (R)	個人名，差別語，誹謗中傷，など	国語研の (R 小林) さんが
◇ (A)	基本形で漢字仮名以外の文字を使用する場合	(A イーユー; EU)
◇ (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 んー) ばな; 橘)
◇ (S)	未登録の口語表現が出現した場合	(S こりゃ)
◇ (笑)，(泣)，(咳)，(あくび)	非言語音との共起	(笑 ナニソレ)
◇ (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)
II 音や事象自体を記号で表記するタイプ		
◇ <H>	母音の引き延ばし	ソレデ<H> …[sorede:]
◇ <Q>	子音の引き延ばし	カイ<Q>セキ …[kais:eki]
◇ <FV>	母音不確定音	ソレデ<FV>
◇ <息>，<笑>，<泣>，<咳>	非言語音	アルワケデ<息>

タグ (O) 外国語や古語，方言など，現代共通日本語から逸脱している (可能性のある) 箇所に付与するタグ．このタグの付いた範囲は，現代共通日本語を対象とした研究をする際に特に注意する必要がある．

タグ (R) 個人名や差別語，誹謗中傷が生じている部分に付与するタグ．コーパス公開の際には，この部分は伏せ字にする等の処理を施す予定．

タグ (A) アルファベットや算用数字を表記する為に使用するタグ．これらの字種は，(A シーディー; CD) や (A 千九百九十五; 1995) 年のように，本タグを利用し漢字仮名に併記する形で記述．

タグ (笑)，(泣)，(咳)，(あくび) これらの非言語行動と発話が，同時もしくは入り混じりながら進行している区間に付与するタグ．

タグ (L) 前後の音声と比べて，かなり小さな声で発話されている区間に付与するタグ．必ずしも独り言であるとは限らない．

タグ<H> 母音の引き延ばし現象のうち，「コレカラー」や「スゴーイ」のように，母音伸張の存在が意味の対立を引き起こさないような場合 (「コレカラ」や「スゴイ」と同じ) に，長音記号に代わって使用するタグ．「オバーサン」や「コート」のように，長音の有無が意味の対立を引き起こす場合には長音記号を使用．

タグ<Q> 「サッスガ」や「ブンッセキ (分析)」のように，意味の対立を引き起こさない子音の引き延ばし現象に対して使用されるタグ．

タグ<FV> 強いボーカルフライ (きしみ発声) などによって，母音が明確に同定できない場合に用いるタグ．きしみ音であっても，母音が同定できる場合には，本タグは使用せずその母音を記す．

タグ<息>，<笑>，<咳>，<泣> (話し手の身体によって生成される) 非言語音が発話中に発話とは独立して出現した場合に用いるタグ．

4 おわりに

CSJ の公開については，プロジェクトが終了する 2004 年 3 月に予定している．また本年度から，年度末に 100 時間程度のデーター (音声と書き起こしテキストのみ) をモニター公開することを予定している．

謝辞 本コーパスに音声を提供していただいた話者の皆様に感謝いたします．また古井貞熙代表を始め，本プロジェクトの関係者の皆様には，書き起こし基準を作成する上で様々な御意見をいただきました．ここに感謝いたします．

参考文献

1. 前川 喜久雄，籠宮 隆之，小磯 花絵，小椋 秀樹，菊池 英明 (2000)．「日本語話し言葉コーパスの設計」『音声研究』，4(2), pp.51–61.
2. Maekawa, K., Koiso, H., Furui, S. and Isahara, H. (2000). "Spontaneous speech corpus of Janapese," *Proc. LREC-2000*, pp.947–952, Athens.