自然言語処理 143－6
音声言語情報処理 36－1
（2001. 6. 1）

『日本語話し言葉コーパス』の設計の概要と書き起こし基準について

○ 小磯 花絵†, 前川 喜久雄†

† 国立国語研究所 研究開発部門 第2領域
〒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時間の音声(約700万形態素に相当)，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，本コーパスの設計の概要および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

The Corpus of Spontaneous Japanese: Its Design and Transcription Criteria

○ Hanae Koiso†, Kikuo Maekawa†

† Department of Language Research,
National Institute for Japanese Language
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 Institute for Japanese Language, the Communications Research Laboratory, and Tokyo Institute of Technology. This corpus is designed to contain about 700 hours of speech (about 7 million morphemes), its transcription, and various tagging information such as POS information. In this paper, the transcription criteria designed specifically for CSJ is described after a brief overview of the general architecture of the corpus.

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

1 はじめに

従来，音声認識の分野では朗読音声を中心に研究が進められてきた．しかし我々が日常話す言葉には，不明瞭な発音や言い淀み，言い間違い，非文法的な表現など，朗読音声を対象としてきた今までの音声認識技術では対応の難しい現象が数多く含まれている．その為，今後の音声認識技術の向上の為には，大規模な自発音声コーパスが必要とされる．またこのような自発音声のコーパスは，音声工学の分野に限らず，言語学や音声学，談話分析，日本語学など，さまざまな分野でも求められるようになってきた．

このような状況の下，国立国語研究所，通信総合研究所，東京工業大学では，「話し言葉の言語的・パラ言語的構造の解明に基づく『話し言葉工学』の構築」プロジェクトの一環として，「日本語話し言葉コーパス (Corpus of Spontaneous Japanese:CSJ)」の構築を進めている．CSJは，自発性の高いモノローグ音声を対象としており，日本語の自発音声コーパスとしては最大の規模 (700 時間程度) を目指している．このような規模の音声を書き起こすには，多くの作業者が長期間に渡って作業しなければならない．その為，質の揃ったデーターを作成するには，明確な書き起こしの基準が必要となる．しかし現在のところ，日本語自発音声の厳密な書き起こし基準は存在していない．そこで本プロジェクトでは，書き起こし作業をするにあたり，マニュアルを作成したり，7 万語程度の電子辞書を整備するなど，基準を揃える為の作業を行なってきた．本稿では，本コーパスの設計の概要と書き起こし基準の詳細について紹介する[1]．

2 コーパス設計の概要

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

【言語変種】 CSJ は自発的なモノローグを中心対象とする．ただし比較検討の為，完全に自発的な発話から，ほぼ原稿を読み上げている朗読に近い発話まで，自発性の程度に幅を持たせている．また CSJ は，全国共通語の音声を格納する．ここで言う全国共通語とは，分節音，語彙，文法が東京語に類似した日本語の変種という意味である．韻律が地方色を帯びている場合や，稀に方言語彙が出現する場合などは本コーパスの対象に含める．

【発話内容】 以下に挙げる 3 種類に大別される．

学会講演：多人数の聴衆を前にした改まり度の高い講演．2001 年 3 月までに収録を実施した学会は，日本音響学会，日本音声学会，人工知能学会，国語学会，言語処理学会，社会言語科学会，日本行動計量学会，全国大学国語教育学会の 8 学会である．

模擬講演：本研究の為に派遣された一般の人が行なう 10〜15 分程度の長さのスピーチ．大まかなテーマを与え，その範囲の中で各自自由にスピーチをしてもらう．テーマとしては，自分の住んでいる町の紹介や過去の社会的な出来事の説明・意見といった比較的客観的なものから，過去の自分の経験 (楽しかった，悲しかったことなど) の説明といった個人的な内容に至るまで，幅広く設定している．

その他：一般の講演会や大学等の講義など．

【話者】 模擬講演については，各テーマごとに，年齢 (20-60 歳代) と性別のバランスを取っている．一方学会講演については，学会によって年齢や性別に偏りが見られる．

【規模】 CSJ には 700 万語 (650〜700 時間に相当) の日本語を格納する．学会講演，模擬講演をそれぞれ 300 時間程度，その他を 100〜200 時間程度収録する予定．

【付与情報】 700 万語分の音声を人手で書き起こした後，以下の情報を付与する．

- 形態論情報：短めの単位と長めの単位の境界と品詞を付与[2]．
- 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：拡張された J_ToBI(Venditti, 1995) 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：5 節参照．

分節音情報や韻律情報は，700 万語分のうち 50 万語分にのみ付与する．50 万語の部分をコアと呼ぶ．形態論情報についても，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 転記基本単位の同定作業

音声を文字化した書き起こしテキストが，音声・言語研究に欠かすことのできない重要な資料であることは言うまでもない．しかし書き起こしテキストはあくまでも音声情報の一部を記述したものに過ぎず，文字データーに変換することによって失われる情報は非常に多い．その為，書き起こしテキストから音声情報を容易に参照できるようにコーパスを設計することが望まれる．CSJ では，書き起こし上の基本となる単位 (以下「転記基本単位」) を決定し，その単位ごとに音声との同期を図る．

このような基本単位として，文などの文法的基準が利用されることもあるが，本コーパスが対象とする自発音声は必ずしも文法的に正しく発話される訳ではなく，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，

[1] なお本原稿は，数字やアルファベットなど一部の表記を除き 4.3 節に示す表記法に従って書き記している．表記の実際は本文を参照されたい．

[2] 短めの単位 (短単位) とは，言語の形態的な特徴に着目して設計した単位で，基本語彙の調査や語構成の研究などに適した単位である．例えば「国立国語研究所」という複合名詞は，「国立／国語／研究／所」のように分割される．この例のように，短単位では一般に単語として意識されるものよりも短いものが切りだされることになる．そこでこの点を補うために，CSJ では長めの単位 (長単位) も付与することにした．長単位とは，基本的に文節を自立語と付属語とに分け，それぞれを 1 単位とするもので，「国立国語研究所」は 1 長単位となる．これは音声認識における言語モデルの構築に利用されるほか，特徴語や専門用語などの調査にも適した単位である．

```
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>
0262 00429.446-00429.859 L:
それから                                &    ソレカラ 0263 00430.125-00436.116 L:
(F あー)                                &    (F アー)
三番目の                                &    サンバンメノ
点と                                    &    テント
いたしまして                            &    イタシマシテ<H>
(F あのー)                              &    (F アノー)
こうやって                              &    コーヤッテ
取った                                  &    トッタ
データーが                              &    データガ
全部                                    &    ゼンブ
使えるかというと                        &    ツカエルカトユート
多分                                    &    タブン
使えないだろうと                        &    ツカエナイダロート
```

図 1: 書き起こしテキストの例 (一部抜粋)

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 ミリ秒未満の途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，音声スペクトログラムを表示し，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．文字化作業は，この単位の開始・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, 21, 22 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") がそれぞれ記されている．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．転記基本単位同定作業で確定するのは，この情報部の部分である．

なお本作業では，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 21 行目にある <雑音> がこれに相当する．

[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．

4 文字化作業

4.1 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，音響モデルを構築する為に，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．

4.2 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】 発音形の表記には，表 1 に示す 141 の文字を使用する．周辺的モーラ A, B 系列の

文字については，外来語，擬音語・擬態語，感情表出系感動詞，言い淀みや言い間違いなど，限定された状況でのみ使用する．

【発音の怠けや転訛，言い間違い】　「六義園」を「リッギエン」,「手術」を「シジツ」,「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，5節で述べる (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行目にある (W スコラ; ソコラ) を参照)．

【綴り字における母音の連鎖】　「かあさん (kaasaN)」のように，綴り字において母音が連鎖しており，その母音連鎖部の発音が [ka:saN] のように長音化している場合には，長音記号を利用して「カーサン」のように表記する．長音化し得る母音連鎖のパターンには，表2に示す7つがある．

長音化のパターンには,「カーサン」のように長音化が1つの形態素内で生じるものと,「アブラーゲ」のように2つの形態素 (「油」と「揚げ」) にまたがって生じるものとがある．前者の場合は大抵長音化して発音されるが，1つ1つの音を区切って発音したり，また強調により母音連鎖部の第1母音から第2母音にかけてピッチやパワー等が急激に変化するような場合には，長音化せずに母音がはっきりと発音される (ように聞こえる) ことがある．その場合には，長音記号ではなく，発音2 に示すように母音を記す．一方2つの形態素をまたぐ場合，朗読など丁寧に発音された音声ではあまり長音化しないが，本コーパスが対象とするような自発性の高い音声では長音化することも少なくない．明らかに長音化していると判断された場合については，形態素をまたぐ場合であっても発音1に示すように長音記号を使用して表記する．

【母音・子音の引き延ばし現象】　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) には，長音記号「ー」の変わりに<H> というタグを用いて「コレカラ<H>」や「スゴ<H>イ」のように表記する (5節参照).「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する．

子音の引き延ばし現象についても同様に,「サッスガ」や「ブンッセキ (分析)」のように子音の引き延ばしの存在が意味の対立を引き起こさないような場合には,「ッ」の代わりに タグ<Q> を使用して「サ<Q>スガ」や「ブン<Q>セキ」のように表記する.「シッカリ」や「カット」のように，子音の引き延ばしの素材が意味の対立を引き起こす場合には場合には「ッ」を使用する．

【その他】　助詞の「は」「を」「へ」については，実際の発音である「ワ」「オ」「エ」を使用して表記する．また「ぢ」「づ」を用いて表記する語であっても，発音形では一律「ジ」「ズ」に統一する．

4.3　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や送り仮名の振り方など表記の原則を定め，それを元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を作成した．以下にその詳細を示す．なおここで述べる基準はあくまで現時点での規定である．品詞などの形態論情報が付与された段階で見直される場合があることを予めお断わりしておく．

4.3.1　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 (5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，促音，撥音を使用．片仮名語の表記には，これらに加え表1の周辺的モーラAに示す文字も使用する．ただし片仮名語の固有名詞に限っては，慣習に従い周辺的モーラBも使用する (「ルイ・ヴィトン」など)．

4.3.2　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．これは，形態論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．形態論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．例えば，動詞「切る」を漢字で表記するならば,「割り切り」や「逆切れ」「締め切り」のように，この語を構成要素として持つ語も同様に漢字で表記する．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」と表記を合わせると「飛び切り」と表記されることになる．しかしこの

表 1: 表記に利用する仮名文字のリスト

直音系列	拗音系列			周辺的モーラ A	周辺的モーラ B
アイウエオ	ヤ	ユ	ヨ	イェ	
カキクケコ	キャ	キュ	キョ		クヮ
ガギグゲゴ	ギャ	ギュ	ギョ		グヮ
サシスセソ	シャ	シュ	ショ	シェ	スィ
ザジズゼゾ	ジャ	ジュ	ジョ	ジェ	ズィ
タチツテト	チャ	チュ	チョ	ティ トゥ チェ ツァ ツィ ツェ ツォ	テュ
ダヂヅデド				ディ ドゥ デュ	
ナニヌネノ	ニャ	ニュ	ニョ	ニェ	
ハヒフヘホ	ヒャ	ヒュ	ヒョ	ヒェ ファ フィ フェ フォ フュ	
バビブベボ	ビャ	ビュ	ビョ	ブィ	ヴァ ヴィ ヴ ヴェ ヴォ
パピプペポ	ピャ	ピュ	ピョ		
マミムメモ	ミャ	ミュ	ミョ	ミェ	
ラリルレロ	リャ	リュ	リョ		
ワヲ				ウィ ウェ ウォ	

撥音	促音	長音
ン	ッ	ー

表 2: 長音化に関わる母音連鎖のパターン

母音連鎖パターン	形態素内の場合			複数の形態素をまたぐ場合		
	語彙	発音 1	発音 2	語彙	発音 1	発音 2
aa	母さん	カーサン	カアサン	油揚げ	アブラーゲ	アブラアゲ
ii	小さい	チーサイ	チイサイ	第一	ダイーチ	ダイイチ
uu	空気	クーキ	クウキ	安売り	ヤスーリ	ヤスウリ
ee	姉さん	ネーサン	ネエサン	影絵	カゲー	カゲエ
oo	大きい	オーキイ	オオキイ	お教え	オーシエ	オオシエ
ei	経路	ケーロ	ケイロ/ケエロ	毛色	ケーロ	ケイロ
ou	結構	ケッコー	ケッコウ/ケッコオ	お産まれ	オーマレ	オウマレ

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 1 単語 1 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，1 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，コーパスに出現する全ての語に渡ってなされるものである．日常において個々人が持つ表記の慣習とは食い違う場合もあり得るが，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．

4.3.3 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．そこでこういったパターンごとに表記の方針を整理した．例えば上記の例では，前者の表記 (「ビオラ」「ウインドー」) が採用される．なお,「ドクター (英語由来) ／ドクトル (ドイツ語由来)」のように由来する語が異なるものや,「ストライク／ストライキ」のように同一の語に由来しているが外来語としての使い分けが存

在するものについては，別語彙としてそれぞれ登録し表記を使い分ける．

4.3.4 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．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ない．

【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」等は，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．

4.3.5 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のように表記するかについては，必ずしも一意に決定しない．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．

【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】 用語リストから作成される仮名漢字変換用の辞書．使用可能な表記のみ登録されており，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．

5 タグ付け作業

5.1 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[4]．

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

[4] 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 (5 節参照)．

表 3: 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，その範囲の特徴に言及するタイプ		
◇ (F)	フィラー・感情表出系感動詞	(F あの)，(F うわ)
◇ (W)	言い間違い，転訛，発音の怠け，など	(W ミダリ；ヒダリ)
◇ (D)，(D2)	言い直し	(D こ) これ，これ (D2 は) が
◇ (?)	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 音や事象自体を表現するタイプ		
◇ <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>

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[5]．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．

書き起こしの際には，言い間違いや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 3 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」や「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，および「うわ」や「げげ」，「あーあ」など驚いた時や落胆した時などに発する感情表出系の感動詞に付与するタグ．

フィラーについては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な場合以外はフィラーと判断する．

[5]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．

感情表出系感動詞については語彙を限定しない．また 4.3 節に示したような表記の統一も行なわず，表 1 に示した文字の範囲内で聞いた通り表記する．

【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 (「あたら」) それを別の表現 (「最新の」) 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．以下の例のように，言い掛け部が単語[6]より短い語の断片の場合には (D) を，機能語 (助詞・助動詞の類) の場合には (D2) を付与する．「ここ か から」のように機能語の断片が言い直されている場合には (D) を付与する．

(D あたら) 最新の問題が
(D だい) (D だいが) 大学の学部の会議での
(D じゅう) 精度の上で重要なポイントであるが
そこの (D み) (F あー) 左の方に
評価値 (D2 が) の数値が
組み合わせ (D2 や) (D2 は) については
三 (D2 までの) (F あ) (D か) からの
学習データー (D2 が) (D こん) (F え) (D しゅ) の収集が困難な

「スライド (F えーっと) プロジェクターで」や「それ それについて その問題については」のように，言い掛け部が機能語以外の単語，あるいは複数の単語である場合には付与しない．また「ブン シ セキ (分析)」のように，

[6]ここでいう単語とは，同プロジェクトにおける形態論情報付与作業 (単位分割および品詞付け) で採用している単位「短単位」を指す．

単語内で生じる言い淀みについては，本タグではなく，(W ブンシセキ;ブンセキ) のようにタグ (W) で対応する．

【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．音の聞き取りが曖昧なのか，それとも語彙や漢字の同定が曖昧なのかにより，本タグを基本形と発音形のどちら (あるいは両方) に付与するかが決まる．以下に幾つか例を示す．

音聞き取り曖昧で語彙も不確定： (? 字数) の & (? ジスー) ノ
音聞き取り曖昧，文脈から語彙確定：それで & (? ソレデ)
音は明瞭だが語彙 (漢字) が曖昧： (? 対象) の & タイショーノ

全く候補がない場合には (?) のように，複数の候補がある場合には (? 対照，対象，対称) のように記述する．また (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．

【タグ (M) 】 以下の例に見られるように，音や言葉自体が言及の対象となるような「メタ的な引用」に付与するタグ．

◇ (M あ) という文字は (M め) と非常によく似ている
◇ (M 僕が) の (M が) は格助詞
(M 行って) の (M て) は接続助詞という具合に
◇ (M という) や (M といった) という表現を使って文を作る

このようなメタ的な引用の前後では，通常の単語の連接パターンから逸脱することもあり，後に形態論情報を自動的に付与する際に問題が生じる恐れがある．そこでメタ的な引用のうち，特に後の自動処理で問題となる可能性が高い以下のパターンにのみ本タグを付与する．

(a) 単語未満の要素 (音，文字，語の断片，接辞)，機能語 (助詞，助動詞)，活用系自立語のうち言い切りの形 (終止形，命令形) 以外の語，及び連体詞が単独で引用 される場合．

(b) 「と彼は言った」や「僕は」,「と言うと」のように，引用部の始端 が機能語または語の断片であるか，あるいは 終端 が (a) に挙げた要素 (ただし助動詞の言い切り形を除く) の場合．

【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．例えば「パーソナル」や「ボーカル」のように，外来語として定着している単語であっても，外国語風の発音をしている (通常の日本語の音韻体系から外れている) と考えられるものについてはこのタグを付与する[7]．

【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<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．

6 おわりに

本稿では日本語話し言葉コーパス (CSJ) の設計の概要と書き起こしの基準について紹介した．CSJ の公開はプロジェクトが終了する 2004 年に予定している．また 2001 年度から，毎年 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．本プロジェクトの活動の詳細については，

http://www.crl.go.jp/pub/orc-speech/

を参照されたい．

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.
3. Venditti, J. (1995). *Japanese ToBI Labelling Guidelines*, Manuscript. Ohio State University, Columbus, USA. (http://www.ling.ohio-state.edu/phonetics/J_ToBI/)

[7] 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．