「日本語話し言葉コーパス」における書き起こしの方法とその基準について

著者	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

雑誌名	日本語科学
巻	9
ページ	43-58
発行年	2001-04
URL	http://doi.org/10.15084/00002055

「日本語話し言葉コーパス」における書き起こしの方法とその基準について

小磯　花絵（国立国語研究所）
土屋　菜穂子（青山学院大学／国立国語研究所）
間淵　洋子（東京都立大学／国立国語研究所）
斉藤　美紀（東京大学／国立国語研究所）
籠宮　隆之（国立国語研究所）
菊池　英明（国立国語研究所）
前川　喜久雄（国立国語研究所）

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

要　旨

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『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』プロジェクトにおいて，自発性の高い話し言葉の情報処理基盤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700万語に相当）の音声，書き起こしテキスト，および品詞や分節音，韻律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の方法とその基準について紹介する。

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，『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』プロジェクトにおいて，自発性の高い話し言葉の情報処理基盤技術の確立を目標に活動を進めている。本プロジェクトの柱の１つである音声認識の分野では，従来，予め与えられたテキストを読み上げるタイプの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，また非文法的な表現などが多く含まれている。このような現象には，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することができない。その為，大規模な自発音声コーパスへの需要が高まりつつある。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ている。しかし，一定の規模や品質を備えた日本語の自発音声コーパスはようやく構築され始めた段階にあり，現在のところ自由に利用できるものは殆ど存在しないと言ってよい。

このような状況の下，国立国語研究所では，本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（700～800時間）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，各種マニュアルを作成し，更に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】 CSJは自発的なモノローグを対象とする。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には，全国共通語の音声を格納する。ここで言う全国共通語は厳密な意味ではなく，分節音，語彙，文法が東京語に類似しているという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，20歳代から60歳代までの５つの年齢層ごとに男女各５名ずつ収録することで，年齢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（700～800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である。

【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(小椋，2001)[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：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，全体700万語のうち50万語分にのみ付与する。この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．書き起こし作業の流れ

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，以下に示す方法で音声と同期した精度の高い書き起こしテキストを効率的に作成している。

3.1．転記基本単位の同定作業

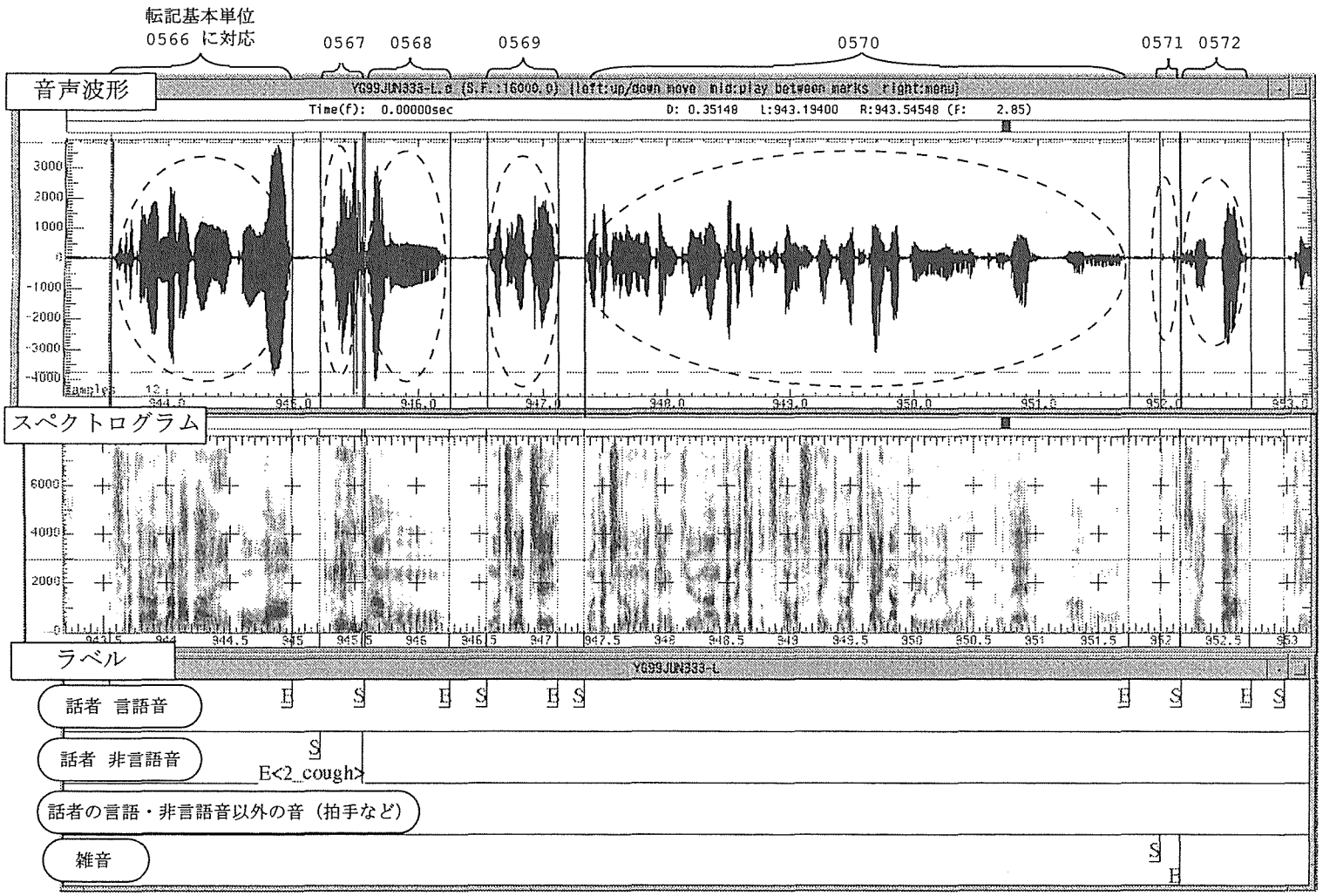
CSJでは，書き起こし上の基本となる単位（以下「転記基本単位」）を決定し，その単位ごとに音声との同期を図る。このような基本単位として文などの文法的基準が利用されることもあるが，本コーパスが対象とする自発音声は必ずしも文法的に正しく発話される訳ではなく，文の認定は容易ではない。そこでCSJでは，転記基本単位の認定基準として物理的な指標を採用した。原則として，200ミリ秒以上のポーズ（言語音の途切れに相当。息・咳・笑い声などの非言語音もポーズに含む）に挟まれた範囲を転記基本単位とする。ポーズの長さが200ミリ秒未満の場合には，そこで単位は分割せず，ポーズも１単位内に含める。ただし言語的な文末形式（述語の終止形や終助詞など）が存在している場合には，ポーズが50ミリ秒以上200ミリ秒未満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，および転記基本単位の始端・終端位置をマークする為のラベル用ウインドーを表示し，音を聴取しながら行なう。図１[A]に作業画面の例を示す。ラベル用のウインドーは４つのレイヤーからなる。このうち１段目のレイヤーで，講演者の発話の転記基本単位の開始・終了位置をマークする。この図では，Sが各単位の開始位置，Eが終了位置を意味する。CSJでは，この（1）講演者の発話（言語音）に加えて，（2）講演者の非言語音（笑い声や泣き声，咳など），（3）講演者の言語音・非言語音以外の音で特に談話の流れを把握する上で重要なもの（拍手の音や司会者の音声，デモ音など），および（4）雑音（以上３種類以外の音で特に目立つ音）についても同様にその範囲を同定する。ウインドーの下３つのレイヤーがそれぞれ（2）～（4）の音に対応している。

3.2．文字化の作業

前節に示した作業により，各転記基本単位の開始・終了時刻が確定する。この情報を元に，図

[A] 転記基本単位の始端・終端位置のラベリング作業画面



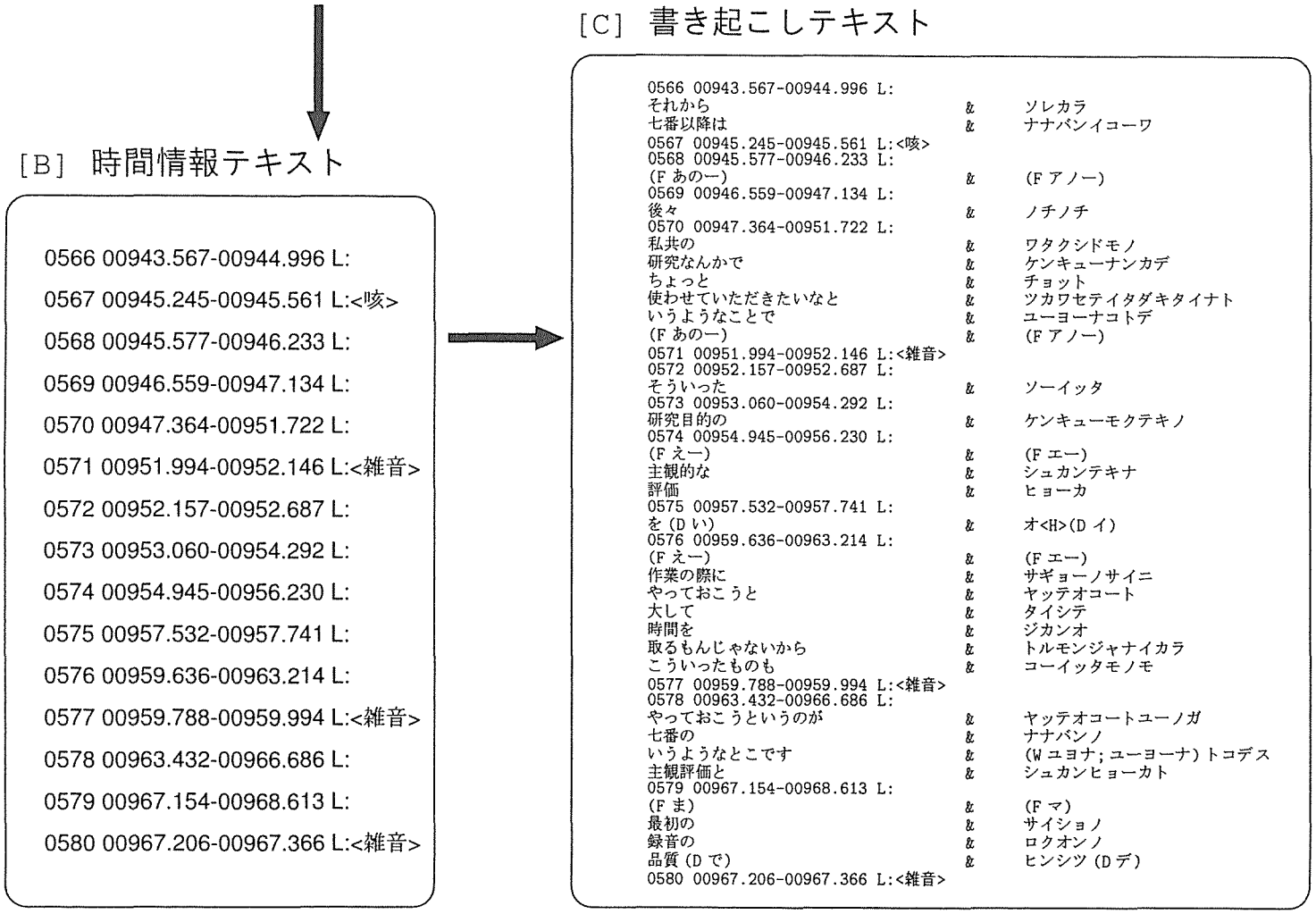


図１：書き起こし作業の流れ

1[B]に示す時間情報テキストを自動的に作成する。1行が1つの転記基本単位に対応する。1列目から順に，転記基本単位の通し番号（4桁の数字），開始・終了時刻（それぞれ秒単位），音声チャンネルのIDである。講演者の言語音以外の音は，転記基本単位同定作業でその音種が特定される（この例では「咳」と「雑音」が同定されている）為，話者IDの後に，その音を示すタグ（〈咳〉，〈雑音〉など）が付与される。

講演者の発話の文字化作業，および各種情報のタグ付け作業は，この時間情報テキストをエディター上に表示し，主話者の言語音（話者IDの右が空欄のもの）を対象にその範囲の音を聴取しながら行なう[4]。この作業の結果，図1[C]に示すような文字化テキストが作成される。文字化，およびタグ付け作業の詳細については，4節，5節でそれぞれ述べる。

4．文字化作業の基準

4.1．2種類の表記法：「発音形」と「基本形」

本プロジェクトの柱の1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2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［C］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。

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					

【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[C]の下から８行目にある（W ユヨナ;ユーヨーナ）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「ー」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，以下の７つがある。

	形態素内の場合			複数の形態素を跨ぐ場合		
母音連鎖パターン	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	お生まれ	オーマレ	オウマレ

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」）に跨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に示すように発音された母音を記す。一方２つの形態素を跨ぐ場合，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，形態素を跨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，本来語彙的には短母音であるが，パラ言語的意味等が付与されることにより一時的に引き延ばされているものについては，長音記号の代わりに〈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）。基本形には「これから」「凄い」のように〈H〉を除いたものを表記する。一方「オネーサン（お姉さん）」や「コート」のように，長音の存在が語彙的に決まっている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「スッゴイ」のように子音の引き延ばしの存在がパラ言語的意味等に関連した一時的なものである場合には，「ッ」の代わりにタグ〈Q〉を使用して「サ〈Q〉スガ」や「ス〈Q〉ゴイ」のように表記する。「ガッコウ（学校）」や「カット」のように，子音の引き延ばしの存在が語彙的に決まっている場合（促音）には「ッ」を使用する。ただし，「ヤハリ」が「ヤッパリ」，「ヨホド」が「ヨッポド」となるように，周囲の音が規則的に転訛しているものについては，〈Q〉ではなく「ッ」で対処する。基本形にも「やっぱり」や「よっぽど」のように「っ」を含めて表記する。

【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」を用いて表記する。また，「縮む（ちぢむ）」や「続く（つづく）」など，現代仮名遣いの上では「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（辞書）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“—”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水準を採用する。ただしJIS第２水準の漢字であっても，

「完璧」の「璧」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，撥音を使用する。また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。２節で述べたように，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，「割り切る」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推し進め，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用語リストに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」「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，一般に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。

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

「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平仮名か漢字のいずれかに統一した。(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う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】 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。

4.4. 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

表２：動詞ラ行音の撥音化にかかわる口語表現（マニュアルから一部抜粋）

◇ ～んない	(← ～らない)	[例] 知んない・やんないね・取んないよ
◇ ～んない	(← ～れない)	[例] かもしんない
◇ ～んな	(← ～るな)	[例] やんなよ・見んなよ・取んな・すんな・あんな
◇ ～んの	(← ～るの)	[例] やんの・見んのか・取んのね・すんの・あんの
◇ ～んじゃん	(← ～るじゃん)	[例] やんじゃん・見んじゃんか・取んじゃんね・すんじゃん・あんじゃん
◇ ～んだろ	(← ～るだろ)	[例] やんだろ・見んだろ・取んだろ・すんだろ・あんだろ
◇ ～んでしょ	(← ～るでしょ)	[例] やんでしょ・見んでしょ・取んでしょ・すんでしょ・あんでしょ
◇ ～んじゃ	(← ～るんじゃ)	[例] やんじゃ・見んじゃ・取んじゃ・すんじゃ・あんじゃ
◇ ～んだ	(← ～るんだ)	[例] やんだ・見んだよ・取んだよね・すんだ・あんだ
◇ ～んです	(← ～るんです)	[例] やんです・見んですよ・取んですね・すんです・あんです

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)，口語表現とは考えない。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)。

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３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[5]。その際，口語化のパターンを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についてはできるだけ統一的な扱いをするように心掛けた。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で，動詞の種類にかかわらず出現し得る。また，このような動詞活用語尾のラ行音が撥音化する現象は，「ない」が後続する場合だけでなく，「見んな」のような禁止の「な」や「すんの」のような疑問の「の」などが後続する場合にも同様に見られる。このような類似した現象については，それぞれ個別に口語表現として登録するか否かを判断するのではなく，統一的に判断するようにしている（以上の例はいずれも口語表現として登録)。また作業者が理解し易いよう，作業マニュアルでは表２に示すように類似した現象をまとめて記す，例を付与するなどの工夫をしている。

５．各種情報のタグ付け

5.1. タグの概要

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。タグは大きく２種類に分けられる。

Ⅰ　文字化された発話の一部を指定しその範囲の特徴に言及するタイプ：図１[C]の６行目の (F あのー) のように，半角の丸括弧で範囲を指定し，開き括弧の後の記号で特徴の種類を示す。(F あのー) の場合は，「あのー」の範囲がフィラーであることを意味する。

Ⅱ　タグ自体が音や事象を表現するタイプ：図１[C]の４行目の〈咳〉のように，〈 〉括弧を用い，その中の記号で事象の種類を示す。〈咳〉の場合，咳の存在を意味する。

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	
	◇(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	(?)
	◇(F)	フィラー・感情表出系感動詞	(F あの)，(F うわ)
	◇(M)	音や言葉に関する引用	(M わ)は(M は)と表記する
	◇(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〉

これらの記号は原則として基本形と発音形の両方に付与されるが，声が小さい，笑いながら発話されている，母音や子音が引き延ばされているといった，音声にかかわるタグについては発音形にのみ，漢字等の表記にかかわるタグについては基本形にのみ付与される。

同じ文字列に対して２つ以上のタグが付くことがある。例えば小さい声で発話しており音の聞き取りに自信がない場合には「(L (? ジャナクテ))」のように，また小さい声で笑いながら発話しており語尾が引き延ばされている場合には「(笑 (L ウソ〈H〉))」のように表記される。

5.2. 各種タグの説明

本節では表３に示したタグのうち幾つかのタグについて説明する。

【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）や接辞の場合には (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	評価値(D2 が)の数値が
(D だい)(D だいが)大学の学部の会議での	組み合わせ(D2 や)(D2 は)については
(D じゅう)精度の上で重要なポイントであるが	西洋(D2 的)(F えー)(D ふ)風というか
そこの(D み)(F あー)左の方に	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 あーっと)プロジェクターで」や「それ それについて その問題について考えてみると」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(D)は付与しない。また「ブン シ セキ（分析)」のように，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)】「ガクジツ（学術)」や「リッギエン（六義園)」のように，発音の怠けや音の転訛，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。「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については修正の対象としない。

【タグ(?)】音の聞き取りや語彙の同定，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。

◇音の聞き取り曖昧で語彙も不確定：	(? 字数)の	& (? ジスー)ノ
◇音の聞き取り曖昧だが文脈から語彙確定：	それで	& (? ソレデ)
◇音は明瞭だが語彙（漢字）が曖昧：	(? 長)単位の	& チョータンイノ
◇複数の候補がある場合：	(? 次数,実数)	& (? ジスー,ジッスー)
◇全く分からない場合：	(?)ので	& (?)ノデ

音の聴取が曖昧な場合には，本タグと先述のタグ(W)とを組み合わせて適用することもある。典型的な例を幾つか以下に示す。

◇Aのように聞こえるが文脈からBの言い間違いと推測：	(? 手)	& (W (? エ);(? テ))
◇Aのように聞こえるが文脈からBの言い間違いと確信：	手	& (W (? エ);テ)
◇確かにAと聞こえるが文脈からBの言い間違いと推測：	(? 手)	& (W エ;(? テ))

【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。一方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(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という課題では

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。

- 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，および連体詞が単独で引用される場合。
- 「と言う」や「僕が」，「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，あるいは終端が上記要素（ただし終助詞除く）および活用語の言い切り形以外である場合。

【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。

【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。これらの字種は以下に示すように本タグを利用し漢字仮名に併記する形で記述する。

(A 千九百九十五;１９９５)年,	(A 二十六．三五;２６．３５)
(A 四ダブリューディー;４WD),	(A シーディーロム;ＣＤ−ＲOM)
(A テンベースティー;１０ＢＡＳＥ−Ｔ),	(A トリプルエーアイ;ＡＡＡＩ)

6．おわりに

本稿では日本語話し言葉コーパス（CSJ）の書き起こし方法およびその基準について紹介した。現在進行中の研究の中間報告という性格上，最終的に採用する基準が本稿の内容と若干異なってくる可能性があることをお断りしておく。なお，CSJ の公開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2001年度から，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://www2.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さまざま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 Japanese," *Proc. LREC-2000*, pp.947-952, Athens.
3．小椋 秀樹 (2001)「話し言葉コーパスの単位認定基準について」『話し言葉の科学と工学ワークショップ講演予稿集』，pp.21-28.
4．Venditti, J. (1995) *Japanese ToBI Labelling Guidelines*, Manuscript. Ohio State University, Columbus, USA. (http://www.ling.ohio-state.edu/phonetics/J_ToBI/)

注

1　本稿で紹介する基準は，必ずしも最終的なものではなく，今後数年に渡る作業の過程で多少の変更が加えられる可能性があることをお断りしておく。６節でも述べるように，我々はコーパスについても構築中のものを少しずつ公開していく予定である。このように，基準や成果物を随時公開することにより，多くの方から御意見を頂戴し，最終段階に向けてより良いものを構築していきたいと考えている。

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出されることになる。そこでこの点を補う為に，CSJでは長めの単位（長単位）も付与する。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。長単位は音声認識における言語モデルの構築に利用される他，特徴語や専門用語などの調査にも適した単位である。

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　音声の再生は，転記基本単位の開始・終了時刻を参照して，各単位ごとの音声を，あるいは複数の単位に渡る範囲の音声を聴取することができる。

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。

6　ここでいう単語とは，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。詳細は注の２を参照。

（投稿受理日：2000年12月26日）

小磯　花絵（こいそ　はなえ）
国立国語研究所　東京都北区西が丘3-9-14　koiso@kokken.go.jp
土屋　菜穂子（つちや　なおこ）
青山学院大学大学院／国立国語研究所　tutiya@kokken.go.jp
間淵　洋子（まぶち　ようこ）
東京都立大学大学院／国立国語研究所　mabuchi@kokken.go.jp
斉藤　美紀（さいとう　みき）
東京大学大学院／国立国語研究所　itaike@kokken.go.jp
籠宮　隆之（かごみや　たかゆき）
国立国語研究所　kagomiya@kokken.go.jp
菊池　英明（きくち　ひであき）
国立国語研究所　kikuchi@kokken.go.jp
前川　喜久雄（まえかわ　きくお）
国立国語研究所　kikuo@kokken.go.jp