転記テキストの仕様

Version 1.0

小磯花絵[‡]・間淵洋子[‡,†]・西川賢哉[‡]・斎藤美紀[‡,††]・前川喜久雄[‡]
[‡] 国立国語研究所, [†] 東京都立大学大学院, [††] 東京大学大学院

目 次

0 はじめに

本文書では,『日本語話し言葉コーパス』における転記テキスト [拡張子: trn] の仕様について概説する[1]．転記テキストは 3302 講演全てに対して提供される．基本的に全ての転記テキストは同じ仕様に基づいて作成されているが，再朗読と対話については一部特殊な基準が適用されているため，扱いに注意されたい (6 節・7 節参照)．またコア (overview.pdf 参照) は，他に比べて発音形の精度が高い．これは，コアに対してのみ提供される分節音ラベリング (segment.pdf 参照) の結果を，転記の仕様の範囲内で発音形に反映したためである．それ以外の点に関しても，コアは他に比べて精度は高いが，両者の違いは精度のみであり仕様自体に変更はない．

1 転記テキストの概要

転記テキストの例を図 1 に示す．この例に基づきながら転記テキストの基本的な構成について説明する．

【講演 ID・講演の開始終了位置の情報】

- ◇ 講演 ID：ファイル冒頭に「%講演 ID:」の形式で記載.
- ◇ 講演の開始位置：講演開始の直前に「%<SOT>」の形式で記載.
- ◇ 講演の終了位置：講演終了の直後に「%<EOT>」の形式で記載.

【転記基本単位 (転記単位)】 ※ 詳細は 2 節参照

- ◇ 原則，0.2 秒以上のポーズによって挟まれた音声の範囲.
- ◇ 転記単位には以下の 4 種類がある.
  - A：講演者の言語音
  - B：講演者のボーカル音 (笑い声，泣き声，咳，息)
  - C：上記 A・B 以外の音で特に目立つ音 (聴衆の発話や笑い，拍手，発表中のデモの音など)
  - D：その他 (朗読間違いの箇所．再朗読ファイルに限定して付与)
- ◇ A (講演者の言語音) の場合，転記単位情報部と発話部 (下記参照) から構成.
  ⇒ 図 1 中の発話 ID=0265～0269，0271，0273～0277
- ◇ B～D の場合，単位情報部と<咳>のような音声種別を記したタグから構成. ※ タグの詳細は表 6 参照
  ⇒ 図 1 中の発話 ID=0270 の<咳>，0272 の<雑音>

【転記基本単位情報部 (単位情報部)】単位情報部は以下 3 つの要素から構成.

- ◇ 発話 ID：4 桁の通し番号 (転記単位を開始時刻の早い順に並べ，0001 から昇順に ID を付与).
- ◇ 当該転記単位の開始・終了時刻 (秒単位).
  対応する音声ファイルの開始時刻を 0(秒) とした場合の時刻.
  ⇒ 発話 ID=0265 の場合,「00697.054(秒)」が開始時刻,「00701.891」が終了時刻
- ◇ 話者 ID（L/R)：モノローグファイルは L に固定.
  対話 (話者 2 名) の場合は L と R (音声ファイルのチャンネルに対応).

【発話部】

- ◇ 基本形と発音形という 2 つの表記方法を用いて発話内容を記す.
- ◇ & の左側に基本形，右側に発音形が記される.
- ◇ 「文節」に相当する単位で改行. ※ 文節の基準は bunsetsu.pdf 参照

【基本形】漢字仮名交じりで表記．可読性が高く，検索に適する. ※ 基準の詳細は 3 節参照

【発音形】片仮名表記．聞き取れる範囲で発音を忠実に記す. ※ 基準の詳細は 4 節参照

【タグ】談話に生じる様々な現象を表現する為のタグ. ※ タグの詳細は 5 節参照
⇒ 発話 ID=0265 コ (?ノ)：音声不明瞭 タグの一覧は表 6 参照
⇒ 発話 ID=0266 (W ユ;ユウ)：発音のなまけ・言い間違い

【コメント】 コメント行は%で開始．コメントには以下の 3 種類がある.

- ◇ 講演全体に対するコメント：講演 ID と<SOT>の間に記述
- ◇ 転記基本単位に対するコメント：単位情報部の下に記述
- ◇ 局所的な発話に対するコメント：当該発話行の下に記述

[1] 転記テキストに含まれる情報は全て XML 形式の文書 [拡張子:xml] でも表現されている．また転記テキストに含まれる情報のうち，コメントを除く全ての情報が形態論情報データ [拡張子:sdb,ldb] にも表現されている (詳細は xml.pdf, wdb.pdf 参照).

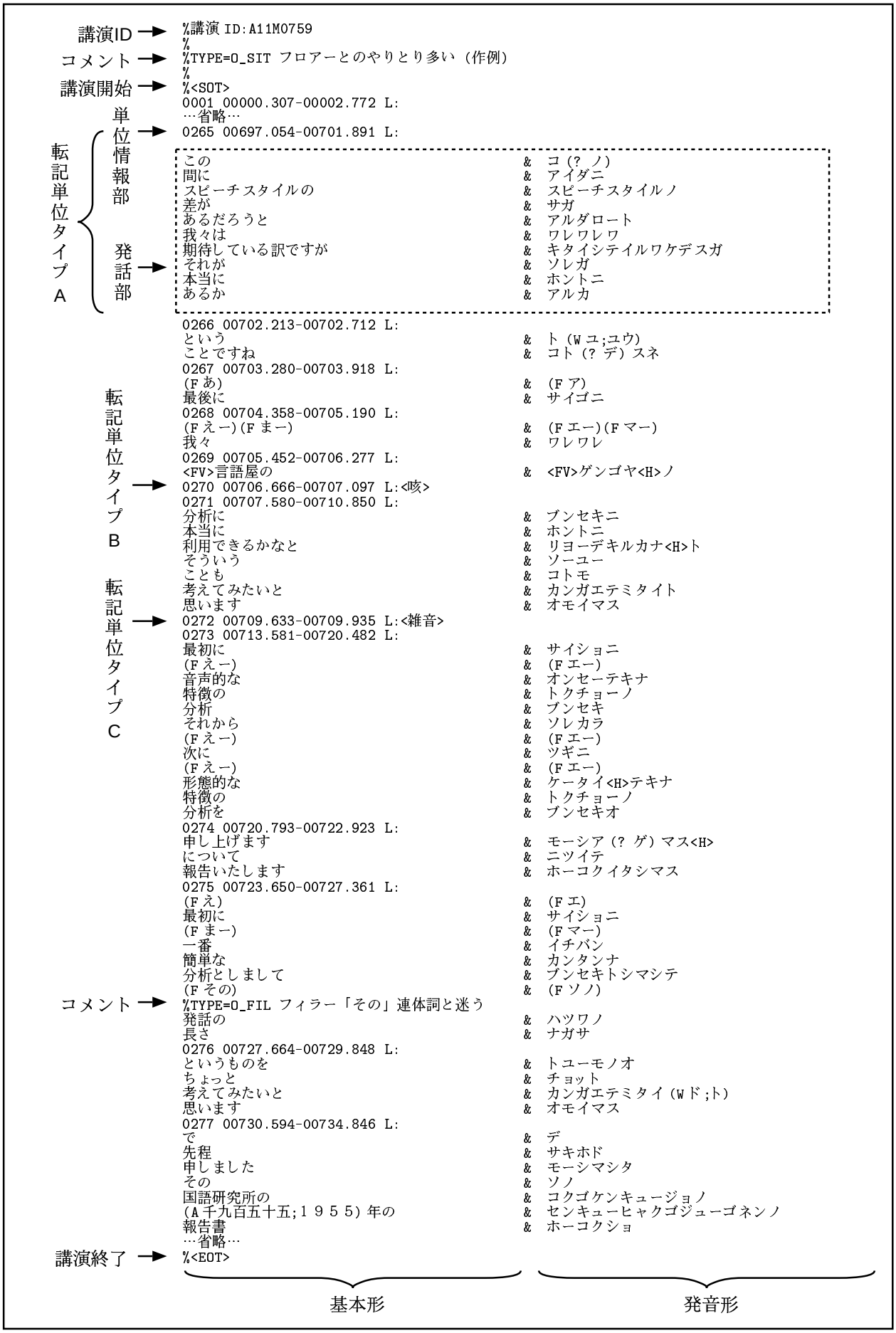

講演ID
%講演 ID:A11M0759
%
コメント
%TYPE=O_SIT フロアーとのやりとり多い（作例）
%
講演開始
%<SOT>
0001 00000.307-00002.772 L:
…省略…
単位情報部
0265 00697.054-00701.891 L:
転記単位タイプA
この & コ (? ノ)
間に & アイダニ
スピーチスタイルの & スピーチスタイルノ
差が & サガ
あるだろうと & アルダロート
我々は & ワレワレワ
期待している訳ですが & キタイシテイルワケデスガ
発話部
それが & ソレガ
本当に & ホントニ
あるか & アルカ
0266 00702.213-00702.712 L:
という & ト (W ユ;ユウ)
ことですね & コト (? デ) スネ
0267 00703.280-00703.918 L:
(F あ) & (F ア)
最後に & サイゴニ
0268 00704.358-00705.190 L:
(F えー)(F まー) & (F エー)(F マー)
我々 & ワレワレ
0269 00705.452-00706.277 L:
<FV>言語屋の & <FV>ゲンゴヤ<H>ノ
転記単位タイプB
0270 00706.666-00707.097 L:<咳>
0271 00707.580-00710.850 L:
分析に & ブンセキニ
本当に & ホントニ
利用できるかなと & リヨーデキルカナ<H>ト
そういう & ソーユー
ことも & コトモ
考えてみたいと & カンガエテミタイト
思います & オモイマス
転記単位タイプC
0272 00709.633-00709.935 L:<雑音>
0273 00713.581-00720.482 L:
最初に & サイショニ
(F えー) & (F エー)
音声的な & オンセーテキナ
特徴の & トクチョーノ
分析 & ブンセキ
それから & ソレカラ
(F えー) & (F エー)
次に & ツギニ
(F えー) & (F エー)
形態的な & ケータイ<H>テキナ
特徴の & トクチョーノ
分析を & ブンセキオ
0274 00720.793-00722.923 L:
申し上げます & モーシア (? ゲ) マス<H>
について & ニツイテ
報告いたします & ホーコクイタシマス
0275 00723.650-00727.361 L:
(F え) & (F エ)
最初に & サイショニ
(F まー) & (F マー)
一番 & イチバン
簡単な & カンタンナ
分析としまして & ブンセキトシマシテ
(F その) & (F ソノ)
コメント
%TYPE=O_FIL フィラー「その」連体詞と迷う
発話の & ハツワノ
長さ & ナガサ
0276 00727.664-00729.848 L:
というものを & トユーモノオ
ちょっと & チョット
考えてみたいと & カンガエテミタイ (W ド;ト)
思います & オモイマス
0277 00730.594-00734.846 L:
で & デ
先程 & サキホド
申しました & モーシマシタ
その & ソノ
国語研究所の & コクゴケンキュージョノ
(A 千九百五十五;１９５５) 年の & センキューヒャクゴジューゴネンノ
報告書 & ホーコクショ
…省略…
講演終了
%<EOT>
基本形
発音形

図 1: 転記テキストの例

2　転記基本単位

本節では，転記テキストの基本単位である「転記基本単位 (以下「転記単位」)」の概要について説明する．

【転記単位の種類】

転記単位Ａ：講演者の言語音．		
転記単位Ｂ：講演者のボーカル音．	<笑>，<泣>，<咳>，<息>	
転記単位Ｃ：Ａ・Ｂ以外の音で特に目立つ音．	<フロア発話>，<フロア笑>，<拍手>，<デモ>，<ベル>	
上記以外で特に目立つ音．	<雑音>	… 音種は特定せず一律「雑音」として扱う
転記単位Ｄ：朗読間違いの箇所．再朗読に限定．	<朗読間違い>	… 扱いはＡに同じ．詳細は 6 節参照．

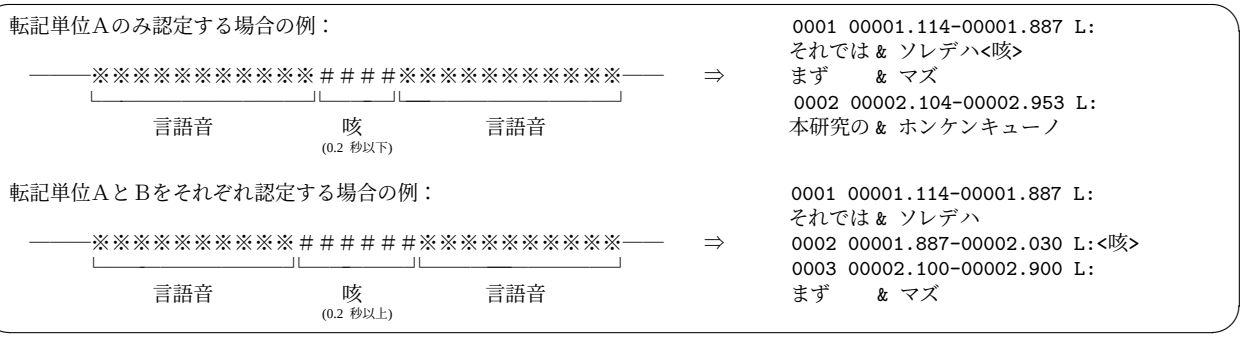
【単位認定：転記単位Ａ（Ｄ）の場合】

- ◇ 原則：言語音が，0.2 秒以上の途切れがなく，連続して生じている区間[2]．
- ◇ 例外：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, 0.05 秒以上 0.2 秒未満の途切れであっても，転記単位を分割する．
- ◇ 時間的関係：転記単位Ａ同士，及び転記単位ＡとＢは，原則として時間的に重複しない．
  対話の場合，二人の話者の言語音は重複し得るが，同一話者内での重複は原則通り存在しない．
  転記単位Ａのうち，タグ<P>で示されるポーズ範囲内に関しては，転記単位Ｂと重複し得る (下記例参照)．

```
0041 00098.407-00100.645 L：
(F えーっと)　&　(F エーッ<P:00098.768-00100.487>ト)
0042 00098.986-00099.228 L:<咳>　　… 上記ポーズ区間と時間的に重複
```

【単位認定：転記単位Ｂの場合】

- ◇ 原則：講演者のボーカル音が 0.2 秒以上の途切れなく連続して生じている区間．同種の音種 (笑い等) 毎に認定．
- ◇ 例外１：転記単位Ａと時間的に重複する場合 (発話の最中に生じる短い笑い等で，言語音が 0.2 秒以上途切れない場合)，独立した転記単位Ｂとして認定せず，転記単位Ａの一部とする (下記例参照)．



- ◇ 例外２：<息>は，転記単位Ａ (言語音) の直後に後続しており，言語音と音声的に切り離せないものに限定する．
- ◇ 時間的関係：転記単位ＡとＢ，転記単位Ｂ同士は，原則として時間的に重複しない．

【単位認定：転記単位Ｃの場合】

- ◇ 原則：転記単位Ｃの対象音が，0.2 秒以上の途切れがなく，連続して生じている区間．同種の音毎に認定．
- ◇ 時間的関係：転記単位Ｃは，Ａ～Ｄのいずれの単位 (自己を含む) とも時間的に重複し得る．
- ◇ 転記単位Ｃはあくまで談話の流れを理解する為の補足的な情報であり，Ａ・Ｂと比べて単位認定の精度は落ちる．例えば，転記単位Ｃの音は, 言語音と重複する等の理由で聞き取りが困難な場合が多い．その為，開始・終了位置が厳密に同定できないこともある．また長いデモンストレーションの音などについては，0.2 秒の途切れで細かく分割せずに，１つのまとまったデモンストレーションとして分割するなど，状況に応じて適宜単位の認定を行なうこともある．

[2] 語中語末の促音，及び語中語末の破裂音・破擦音の閉鎖区間に相当する途切れは除く．

3　基本形の表記

以下に基本形の表記方針の概要を示す[3].

3.1　字種

【基本】基本形の表記には以下の字種を用いる（使用範囲の詳細は表 3 参照）.

漢字： JIS 第 1・2 水準 (JIS X 0208–1990) の漢字を使用.

平仮名： 表 1 に挙げる仮名のうち「ゔぁ」「ゔぃ」「ゔ」「ゔぇ」「ゔぉ」以外の仮名文字.

片仮名： 表 1 に挙げる仮名を，3.5 節で定める範囲内で使用.

【オプション】上記字種に併記する形で，以下の字種も使用．表記の際にはタグ (A) を利用 (5 節参照).

算用数字： 全角の算用数字 (０１２３４５６７８９)

アルファベット： 全角のローマ字・ギリシャ文字 (大文字・小文字)[4]

記号： 表 2 に挙げる記号 (いずれも全角) を,「使用条件」の項に示す範囲内で用いる.

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 ン	促音 ッ	長音 ー	

表 2: 記号の使用範囲

記号	コード 区点・JIS	使用条件	例
.	0105・2125	数字の小数点，節番号等　▷(A) 内	(A 三．一四;３．１４)，(A 二．三;２．３) 節
○	0191・217B	ゼロの意で発話された「マル」　▷(A) 内	(A 一〇八四;１０８４)　×一丸八四
・	0106・2126	(1) 片仮名語：a. 姓名の間，b. 固有名詞で慣習的に必要性高いもの c. 前置詞・冠詞が入る場合はその前後 (2) 複合語で語の切れ目が分かりづらい箇所　▷ 以上基本形で使用	マルコ・ポーロ，チェコ・スロバキア テキスト・トゥー・スピーチ れる・られる型敬語
&	0185・2175	略語 (アルファベット構成) で慣習的に必要性高いもの　▷(A) 内	(A エムアンドエー;M&A)
－	0161・215D	略語 (アルファベット構成) で慣習的に必要性高いもの， 電話・郵便番号，住所の区切り　▷(A) 内	(A シーディーアール;ＣＤ－R)，(A 三の九;３－９) (A 一一三五四二五;１１３－５４２５)
×	0163・215F	講演者名や差別語・誹謗中傷の伏せ字として利用　▷(R) 内	国語研の (R ××) と申します

[3] 本節で述べる方針に従い，揺れが生じないよう統一的に書き記しているが，固有名詞の表記等については若干例外があるので注意されたい.
[4] ローマ字は区点コード第 3 区 (0333–0358，0365–0390)，ギリシャ文字は第 6 区 (0601–0624, 0633–0656) を使用.
ローマ字は原則大文字, ギリシャ文字は原則小文字表記とし，一般的に表記が固定しているものは逆も可とする.

表 3: 基本形における字種の使用範囲[ア]　　記号 A：.(ピリオド)，記号 B：○，記号 C：&ー，記号 D：･(中点)，記号 E：×

文字種 / 出現環境	平仮名 直拗	平仮名 周 A	平仮名 周 B	片仮名 直拗	片仮名 周 A	片仮名 周 B	長音	漢字	算用数字	アルファベット	記号 A	記号 B	記号 C	記号 D	記号 E[ク]
(F) フィラー	○	×	×	×	×	×	○	×	×	×	×	×	×	×	△
(F) 感情表現	○	○	○	×	×	×	○	×	×	×	×	×	×	×	△
(D)	○	○	○	○	○	○	○	○	×	×	×	×	×	×	△
(D2)[イ]	○	×	×	△	△	△	△	○	×	×	×	×	×	×	△
(A) 左	○	×	×	○	○	○	○	○	×	×	○	○	×	×	△
(A) 右	×	×	×	×	×	×	×	×	○	○	○	×	○	○	△
(O) 古語方言	○	○	○	×	×	×	×	○	×	×	×	×	×	○	△
(O) 外国語	×	×	×	○	○	○	○	△[カ]	×	×	×	×	×	○	△
(K) 左	○	×	×	×	×	×	×	×	×	×	×	×	×	×	△
(K) 右	×	×	×	×	×	×	×	○	×	×	○	○	×	×	△
(M) 音や文字の引用[ウ]	○	○	○	○	○	○	○	○	×	×	×	×	×	×	△
その他 和語漢語[エ]	○	△	△	×	×	×	×	○	×	×	×	○	×	△[キ]	△
その他 片仮名語	×	×	×	○	○	△[オ]	○	×	×	×	×	×	×	○	△

※ア. (O 言い(F んー)やる) のように複数のタグが重複する場合，字種の使用範囲は一番内側のタグ (この場合は F) の制約に従う．ただし (?) については独自の制約はない．(D (? シェ)) であれば (D) の制約に従う．タグ (R) の例外については「※ク」参照．
※イ. 片仮名，長音記号の使用は，数字「ゼロ」の言い直し及び外来語系接辞 (「センチ」等) に限定．
※ウ. 音や文字に関するメタ的引用は，原則として平仮名で表記．ただし話者が片仮名語や外国語の音，文字を引用している場合は片仮名で表記．文字の引用も，文字種が文脈から分かる場合は，言及・示唆されている文字種 (漢字や片仮名等) で表記．
※エ.「その他」はタグの付与されていない範囲及び上記以外のタグの範囲 (音・文字以外の引用の (M) 含む．ただし (?),(R) は除く)．また和語，漢語の周辺的モーラA・Bについては，擬音語，擬態語，一部の俗語的表現 (「おとっつぁん」等) に限定．
※オ. 固有名詞に限定 (「ルイ・ヴィトン」など)．
※カ. 中国語など漢字を使用する外国語に限定．
※キ. れる・られる型敬語のように，語の切れ目が分かりづらい複合語の場合に限定．
※ク. タグ (R) 内の伏せ字化でのみ使用．記号 E に関してのみ，外側のタグ (R) の制約が内側のタグ範囲にも影響を与える．(R 国研(D 太)太郎) が発表します→ (R ××(D ×)××) が発表しま す

3.2 漢字と平仮名の使い分け

【原則】表記が漢字と平仮名で揺れるもので，両表記とも一般的に使用されるものについては，原則漢字を採用[5]．

> ○例えば／×たとえば，○全て／×すべて

【実質名詞・形式名詞】「こと (事)」「もの (物)」「ところ (所)」については一律平仮名で表記[6]．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.「やって おく」や「食べて みる」のように，テ形複合動詞[7]として用いられる場合には，平仮名で表記．

【「言う／いう」の使い分け】以下の組み合わせで出現した場合に限定して平仮名で表記[8]．ただしこの条件を満たす場合であっても，明らかに動作性を有するものについては漢字で表記．

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

【当て字】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．

個々の語の表記については，上記の原則に基づきつつ，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[9]．

[5] これは転記テキストに基づく自動形態素解析の精度向上，及び検索の便宜を図って立てた方針である．
[6] 実質名詞の場合は漢字で，形式名詞は平仮名で表記される慣習が高いが，その区別は非常に難しく書き分けが困難な為，平仮名に統一．
[7] 長単位 (pos.pdf 参照) において「助動詞相当句」と認定されたものに限定．
[8]「言う」という動作が形骸化された用法では，平仮名書きされることが多いが，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為，このように形式的な判断基準を採用．
[9] 例えば副詞の「とびきり」を，その語の構成要素である「飛ぶ」と「切る」に表記を合わせ「飛び切り」と表記する，ということまではしない．

3.3 漢字の使い分け

【新字と旧字の揺れ】固有名詞も含め，例外なく新字を採用.

> ○ 万 屋／× 萬 屋，○幕末太陽 伝 ／×幕末太陽 傳

【JIS 第１水準と第 2 水準の揺れ】JIS 第 1 水準の漢字を採用.

> ○ 憧 れ／× 憬 れ，○一 獲 千金／×一 攫 千金

ただし，著名人の氏名に関しては例外的に使用可とする.

> ○小 渕 元首相 (通常は「淵」に統一)

【同音異義語】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，片方の漢字で代用可能である場合に限り表記を統一するが，それ以外の同音異義語については書き分ける.

> 統一の例：　○悲しい／×哀しい，○会う／×逢う，○尊ぶ／×貴ぶ
> 書き分けの例：○表わす／○現わす，○計る／○測る／○図る

3.4 送り仮名

- ◇ 用言で複数の送り仮名の候補がある場合：送り仮名の字数の多い方を採用 (○行なう／×行う)
- ◇ 名詞で送り仮名の有無に揺れがある場合：原則送り仮名を付ける (○後ろ／×後). 慣習等で送り仮名を付与しないものは個別に定義.

> ○合図／×合い図，○立場／×立ち場，○学割／×学割り，○番組／×番組み，○関取／×関取り，など

3.5 片仮名

片仮名表記の対象は以下の通り.

- ◇ 外来語・外国語[10]
- ◇ 専門用語や俗語，固有名詞などで片仮名書きの習慣が強いもの (ダフ屋・タンパク質・ト書き)
- ◇ 動物，植物，魚介，虫の名称 (原則片仮名表記とし，例外は個別に定義)

特に外来語については,「ビオラ／ヴィオラ」のような表記の揺れが多く見られる.「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毎に表記の方針を整理した上で，コーパスに現われる全ての片仮名語を用語リスト (3.7 節参照) に登録.

3.6 口語表現

CSJ では，(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，という 3 つの条件を満たしたものを口語表現として認め，その形式で基本形に書き記してよいものとする[11]. それ以外は後述のタグ (W) を利用し，そのまま基本形に記すことはしない.

> 口語表現：　しゃあないっちゅうか & シャーナイッチューカ
> 非口語表現：しようがないというか & (W チヨー;シヨー) ガナイ (W ティユ;トユウ) カ

3.7 用語リスト

上記の作業方針に従い，実際の語の表記を定めた用語リスト (約 10 万語) を作成し，オンライン検索や仮名漢字変換用の辞書として利用した．本公開データに用語リストは含まれていないが，将来的には公開する予定である.

[10] 中国の人名地名，及び韓国の著名人 (「金日成」等) の名前に関しては漢字表記とする.

[11] 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．表現を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た.

4 発音形の表記

以下に発音形における表記方針の概要を示す.

4.1 字種

実際に発音された音を，表 1 に挙げる片仮名 (ヂヅ以外) を用い，表 4 に示す範囲内でできる限り正確に書き表わす.

表 4: 発音形における字種の使用範囲　　(直音等：直音・拗音・撥音・促音・長音)

	直音等	周辺的モーラ A	周辺的モーラ B
(F) 感情表現，(D)，(O)，(W) 左項，(M) 音や文字の引用，片仮名語	○	○	○
(F) フィラー・上記以外の和語漢語	○	×	×

4.2 綴り字における母音連鎖

「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ついては，以下のように対処する.

タイプ A　長音化が (狭義の) 形態素内で生じるもの[12]：長音表記・母音表記共に認める (表 5 の※ア参照)

○カーサン・カアサン (母さん), ○チーサイ・チイサイ (小さい)・○ケーロ・ケエロ・ケイロ (経路), ○カナシー・カナシイ (悲しい)

タイプ B　長音化が 2 つの形態素に跨がるもの：長音表記を認めず，一律母音表記とする (表 5 の※イ参照)

○ダイイチ・×ダイーチ (第／一), ○ケイロ・×ケーロ (毛／色), ○オウミ・×オーミ (お／産み), ○モノオ・×モノー (もの／を)

表 5: 母音連鎖パターンにおける発音の扱い

	タイプ A：形態素内の場合				タイプ B：形態素を跨ぐ場合			
母音連鎖パターン	語彙例	長音	母音 1※ア	母音 2	語彙	長音	母音 1	母音 2
a a	母さん	カーサン	(カアサン)	—	油揚げ	×アブラーゲ	アブラアゲ	
i i	小さい	チーサイ	(チイサイ)	—	第一	×ダイーチ	ダイイチ	—
u u	空気	クーキ	(クウキ)	—	安売り	×ヤスーリ	ヤスウリ	—
e e	姉さん	ネーサン	(ネエサン)	ネイサン	影絵	×カゲー	カゲエ	×カゲイ※イ
o o	大きい	オーキイ	(オオキイ)	オウキイ	お教え	×オーシエ	オオシエ	×オウシエ※イ
e i	経路	ケーロ	(ケエロ)	ケイロ	毛色	×ケーロ※イ	ケイロ	×ケエロ※イ
o u	講師	コーシ	(コオシ)	コウシ	子牛	×コーシ※イ	コウシ	×コオシ※イ

※ア：括弧内に示す母音表記は，一音一音区切って発音する場合など，分節音のレベルで明らかに母音が発音されている場合に限定
※イ：この発音が実際に生じた場合には，(W) 対処とする：毛色 & (W ケエロ;ケイロ)，お産み & (W オーミ;オウミ)

4.3 非語彙的な母音・子音の引き延ばし

母音や子音の引き延ばし現象のうち,「コレカラー」や「スゴーイ」,「サッスガ」,「カイッセキ (解析)」のように，語彙的には引き延ばしは存在しないが，パラ言語的意味等が付与されることによって，あるいは言い淀みによって，一時的に引き延ばされているものについては，タグ<H>，<Q>を用い,「コレカラ<H>」や「サ<Q>スガ」のように表記する．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」で対処し，基本形にも「やっぱり」のように「っ」を含めて表記する.

4.4 発音の怠けや転訛・言い間違い

「国語研」を「コッコケン」,「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した上で，5 節で詳述する タグ (W) を利用して，丁寧に発音された場合に生じるであろう音を併記する.

4.5 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.「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.

[12] 原則,「最小単位」(pos.pdf 参照) と一致 (最小単位における固有名詞等の例外は除く). また，助動詞の「う」は例外的に本タイプに含め，長音表記を認める.

5 各種情報のタグ付け

転記に用いるタグの一覧を表 6 に示す.

表 6: 転記テキストに用いるタグの一覧

● 転記単位A (講演者の言語音) に出現するタグ：文字範囲を指定し，その範囲の性質に言及するタイプ

タグ	タグの概要	使用例	付与対象 [*1]	範囲 [*2]
(F)	フィラー，感情表出系感動詞	(F あの)，(F うわ)	基・発	ー・ー
(D)	言い直し，言い淀み等による語断片	(D こ) これ，(D ぼい) 子音の	基・発	ー・ー
(D2)	助詞，助動詞，接辞の言い直し	そこ (D2 が) に，(D2 不) 不自然	基・発	ー・ー
(?)	聞き取りや語彙の判断に自信がない場合	(? タオングー)，(? 堆積, 体積)	基・発 [*3]	文　・ー
(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	文・転
(O)	外国語や古語，方言など	(O ザッツファイン)	基・発	文・転
(R)	講演者の名前，差別語，誹謗中傷など	国語研の (R ××) です	基・発	文・転
(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／再朗読	文・ー
(A)	アルファベットや算用数字，記号の表記	(A シーディーアール;ＣＤ－Ｒ)	基・ー	ー・ー
(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・ー	ー・ー
(W)	転訛，発音の怠けなど，一時的な発音エラー	(W ギーツ;ギジュツ)	ー・発	ー・ー
(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)	ー・発	ー・ー
(笑)	笑いながら発話	(笑 ナニガ)	ー・発	文・ー
(泣)	泣きながら発話	(泣 ドンナニ)	ー・発	文・ー
(咳)	咳をしながら発話	シャ(咳 リン) ノ	ー・発	文・ー
(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)	ー・発	文・ー

[*1] 基／発：当該タグが基本形／発音形に出現　[*2] 文／転：当該タグの括弧の範囲が複数の文節／転記単位を跨ぐ ことがある
[*3] 状況に応じて，基本形のみ，発音形のみ，あるいは両形に出現．詳細は (?) の項参照.

● 転記単位A (講演者の言語音) に出現するタグ：音や事象自体を記号で表現するタイプ

<FV>	ボーカルフライ等で母音が同定できない場合	だから<FV> & ダカラ<FV>	基・発	
<VN>	「うん/うーん/ふーん」の音の特定が困難な場合	(F うん) & (F <VN>)	ー・発／対話	
<H>	非語彙的な母音の引き延ばし	ソレデ<H>，ス<H>ゴイ	ー・発	
<Q>	非語彙的な子音の引き延ばし	カイ<Q>セキ，スゴ<Q>イ	ー・発	
<笑>	言語音と独立に講演者の笑いが生じている場合	ガクセー<笑>ノ	ー・発	
<咳>	言語音と独立に講演者の咳が生じている場合	デ<咳>	ー・発	
<息>	言語音と独立に講演者の息が生じている場合	ツマリ<息>	ー・発	
<P>	0.2 秒以上のポーズ	オ<P:00453.373-00454.013>モイ	ー・発	

● 転記単位B (講演者の発するボーカル音) に出現するタグ：音や事象自体を記号で表現するタイプ

<笑>	講演者の笑い声	単位情報部	
<泣>	講演者の泣き声	単位情報部	
<咳>	講演者の咳	単位情報部	
<息>	講演者の息（言語音と連続した息で，波形上切り離せないものに限定）	単位情報部	

● 転記単位C (A・B以外の音で特に目立つ音) に出現するタグ：音や事象自体を記号で表現するタイプ

<フロア発話>	聴衆 (司会者等も含む) の発話	単位情報部	
<フロア笑>	聴衆の笑い	単位情報部	
<拍手>	聴衆の拍手	単位情報部	
<デモ>	講演者が発表中に用いたデモンストレーションの音声	単位情報部	
<ベル>	学会講演時に発表時間を知らせる為に鳴らすベルの音	単位情報部	
<雑音>	上記以外の音で特に目立った音	単位情報部	

● 転記単位D (その他) に出現するタグ：音や事象自体を記号で表現するタイプ

<朗読間違い>	転記単位全体が再度読み直された場合	単位情報部／再朗読	

5.1 各種タグの説明

本節では表 6 に示したタグのうち，転記単位Aに用いるタグについて概説する[13].

【タグ (F)】 本タグは，以下に挙げる「フィラー」と「感情表出系感動詞」に対して付与する[14].

フィラー：

◇ 場繋ぎ的な機能を持つ表現．語彙を以下のように限定し，その範囲内で場繋ぎ的機能を有する場合に付与.
◇ 下記以外の表現は，たとえ場繋ぎ的機能を有するものであっても (F) は付与しない.
◇ 「あの／その」に関して，フィラーと連体詞とで迷う場合には，原則 (F) を付与した上で迷った旨をコメント.

フィラー表現

あ (ー)，い (ー)，う (ー)，え (ー)，お (ー)，ん (ー)，と (ー)*，ま (ー)*，
う (ー) ん，あ (ー)(ん) の (ー)*，そ (ー)(ん) の (ー)*，
う (ー) ん (ー)(っ) と (ー)*，あ (ー)(っ) と (ー)*，え (ー)(っ) と (ー)*，ん (ー)(っ) と (ー)*
※上記全てのフィラーとの組み合わせ：〜ですね (ー)，〜っすね (ー)　[例] あのですね，えーとっすねー
※上記「*」印のフィラーとの組み合わせ：〜ね (ー)，〜さ (ー)　[例] まーねー，うーんとさー
※括弧内は任意　[例] あの，あのー，あーの，あーんのー

感情表出系感動詞：

◇ 驚いた時や落胆した時などに発する感動詞.
◇ 語彙の限定や表記の統一は行なわず，長音や促音も含め聞こえた通り表記.

例：(F あーあ)，(F あっちゃー)，(F ぎょっ)，(F うおー)

【タグ (D)】

◇ 本タグは，以下のケースで生じる「語の断片[15]」に対して付与.

言い直しに伴う語断片：「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」) に付与.

その他の語断片：「その ん こと」など，言い直しに伴う語断片というよりは，発声上の問題で生じたと考えられる断片的な音声に対しても，同様に本タグを付与.

◇ 「スライド (F えーと) 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，言い直された対象が短単位の断片でない場合には，(D) は付与しない.
◇ 「ダイガ ノ カイギ (大学の会議)」の「ダイガ」のように，語の断片であっても，言い直されずにそのまま発話されたものについては,「大学の & (W ダイガ;ダイガク) ノ」のように (W) で対処.

(D 形) 形式の，(D だ)(D だいが) 大学の学部の会議，(D じゅ) 精度の上で重要な
(D んみ) 考えもしないし，(D てん) 元気な子が，洗濯機で (D つ) やった

【タグ (D2)】

◇ 本タグは，言い直しにおける言い掛け部・訂正部が共に助詞・助動詞・接頭辞・接尾辞から構成される場合，及び言い掛け部・訂正部が共に数字の場合 (数字内の数字の言い直し) に付与する.
◇ 「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は (D) を付与する.

評価値 (D2 が) の数値が，組み合わせ (D2 や)(D2 は) については，ここ (D2 です) ですね，広さ (D2 や) たるや
(D2 第) 第一関門を，西洋 (D2 的)(F えー)(D ふ) 風というか，(A 千 (D2 八百) 九百四十;1 9 4 0) 年
学習データー (D2 が)(D こん)(F え)(D しゅ) の収集が困難な

[13] 転記単位Ｂ・Ｃについては 2 節を，また転記単位Ｄについては 2 節と 6 節を参照されたい.
[14] 対話 (講演 ID が D から始まるファイル) では応答表現に対しても付与する．詳細は 7 節参照.
[15] ここでいう「語」とは，原則「短単位」(pos.pdf 参照) を指す．ただし外来語に関してのみ語の定義が異なる.「カード / ゲーム」のような 2 つの単語から構成される複合語に関して，短単位では 1 つの単位と見なすのに対し，転記では 2 つの単位とする.

【タグ（?）】

◇ 本タグは，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．以下 3 つの形式を取る．

1. 値 1 つ (デフォルト)：　（? 字数) の　＆（? ジスー）ノ
2. 値複数 (複数の候補が想定)：(? 次数, 実数)　＆（? ジスー, ジッスー）
3. 値なし (全く不明な場合)：　(?) で　＆（?）で

◇ 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) に付与するか決まる．

(? 字数) の　＆ (? ジスー) ノ　… 音の聞き取り曖昧で語彙も不確定
それで　＆ (? ソレデ)　… 音の聞き取り曖昧だが文脈から語彙確定
(? せんさ) 空間の ＆ センサクーカン　… 音は明瞭だが語彙が曖昧．候補があれば漢字，なければ平仮名
(? 大賞，対象) の ＆ タイショーノ　… 音は明瞭だが語彙が曖昧．複数の候補あり．

【タグ（M）】

◇ 本タグは，音や言葉自体が言及の対象となるようなメタ的引用のうち，特に以下のパターンに限定して付与する[16]．

1. 以下の要素が単独で引用される場合：
   - 単語未満の要素 (音，文字，語の断片，接辞)・非文：(M あ) という文字は (M め) とよく似ている
   - 機能語 (助詞，助動詞)：助詞の (M は) は (M は) と書くが発音は (M わ)
   - 活用する自立語のうち言い切り (終止形，命令形) 以外：(M 走ら) は動詞の未然形
   - 活用する自立語の言い切りが引用の「と/って」以外に接続する場合：(M いらっしゃる) が後続すると
   - 引用部の終端が上記 1 の要素の場合 (「と/って」に接続する終助詞・助動詞言いきり形を除く)：(M 僕が) が来ると
   - 連体詞・副詞・感動詞・接続詞：(M そして) と接続詞を置くことで
2. 引用部の始端が機能語または語の断片の場合：(M という) や (M といった) という表現を使って
3. 引用部の終端が上記 1 の要素 (終助詞・助動詞言い切り形以外) の場合：(M 僕が) が来ることで
4. 記号 (句読点・括弧等) の読み上げ：走ります (M 丸) と句点を付ける，平仮名で (M こっかい (M 中括弧閉じる) では) で
5. 表記や音に関する引用は言及された通り表記した上で (M) を付与：(M カッコ) と片仮名書き，(M えきー) と延ばして発音

【タグ（O）】

◇ 本タグは，外国語や古語，方言など CSJ が対象とする現代共通日本語から外れている (可能性のある) 箇所に付与．

◇ 外来語として定着している語であっても，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本タグを付与．

◇ (O) 内の表記は，3 節に示す表記の原則や用語リストに定める表記から外れることもある．

【タグ（R）】

◇ 本タグは以下の場合に付与する：

1. 講演者が特定できる情報 (講演者・共著者の氏名，学会講演の場合発表タイトル，前後の講演者の氏名[17]等)
2. 一般人が特定できる情報 (著名人研究者を除く一般人の氏名．ファーストネームのみの場合は付与せず)
3. 誹謗中傷・差別語のうち，特に問題になると判断されたもの
4. 講演者が非公開を希望した箇所

◇ 本タグが付与された範囲は伏せ字化される．伏せ字化に際しては，伏せ字化以前の文字の数だけ記号「× (全角)」を繰り返す．(R) 内に含まれる記号はそのままとする．また音声ファイルに対しては，本タグの対象となる転記単位の全体を白色雑音で置換する処理を施している．つまり，転記単位の一部でも本タグの範囲であれば，その転記単位全体が白色雑音で置換される．

伏せ字前：(R 話し言葉の (F えー) イントネーション) というタイトルで国語研の (R 佐藤) が発表します
伏せ字後：(R ×××××(F ××) ××××××××) というタイトルで国語研の (R ××) が発表します

【タグ（X）】

◇ 本タグは，朗読原稿にない余分な発話に対して付与する．詳細は 6 節参照．

[16] メタ的引用の前後では，通常の単語の連接パターンから逸脱することもあり，後の自動形態素解析の処理で問題が生じる恐れがある為，特に問題となる可能性の高いパターンに限定して本タグを付与するという方針をとる．

[17] CSJ に収録されている場合のみ．[例] 先程の (R ××) さんの発表にもありましたように

◇ 再朗読 (講演 ID が R00 から始まるファイル) に限定して付与.

【タグ (W)】

◇ 本タグは,「ガクジツ (学術)」や「コッコケン (国語研)」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.

◇ (W ガクジツ;ガクジュツ) 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.

◇ 本タグの範囲は短単位 (9 頁の注 15 参照) とし，その範囲でタグをくくり直す (下記例「そうすると」参照). ただし，語の融合などが生じて切り離せない場合はその限りでない (「今度は」参照).

共分散	& キョー (W ブサン;ブンサン),	左へ行って	& (W ミダリ;ヒダリ) エイッテ
手術すると	& (W シジツ;シュジュツ) スルト,	けれども	& (W ケード;ケレド) モ
そうすると	& (W ホ;ソー)(W ス;スル) ト	今度は	& (W コンダー;コンド ワ)

◇ タグ (F), <FV>は原則，基本形，発音形の両形に出現するが，(W) の左項 (実際の発音部) に (F), <FV>が生じた場合，下記の通り (W) の右項，及び基本形にこれらのタグは現れない. そのため，(F), <FV>を完全に抽出する為には，基本形 (だけ) ではなく発音形を参照する必要がある.

チェコ・スロバキアの & (W セコスロバ (F アノー) スロバキア;チェコスロバキア) ノ

◇ 「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.

◇ 本タグと (?) との組み合わせで典型的な例を幾つか示す.

(? 手)	& (W エ;(? テ))	… 音は確実. 訂正候補は不確実 (W の右項と基本形の両方に?付与)
(? 手)	& (W (? エ);(? テ))	… 音も訂正候補も不確実 (W の両項と基本形に?付与)
手	& (W (? エ);テ)	… 音は不確実だが訂正候補は確実 (W の左項のみ?付与)
(? 絵, 手)	& (? エ,(W エ;テ))	… 音はAと確実だが語彙がAかBかで迷う (?の複数候補の一部に W)
(? 毛, 手)	& (W エ;(? ケ,テ))	… 訂正候補が不確実で複数の可能性あり (W の右項に?の複数候補)

【タグ (B)】

◇ 本タグは，漢字の読みに関する知識レベルの言い間違いに付与する[18].

◇ (B シブタイ;ジュータイ) のように，セミコロンの左側に実際に発話された読みを書き表わすと同時に，セミコロンの右側には正しい読みを併記する.

◇ 原則として，本タグの範囲は短単位 (9 頁の注 15 参照) とする.

半ば & (B ハンバ;ナカバ)， 借家であるか & (B シャクイエ;シャクヤ) デアルカ

◇ 以下に該当するものには，(B) を付与しない.

- 専門的な慣用読み，及び意味の明確化等の理由で本来の (辞書上の) 読みと異なる読みが慣用化している場合

両耳で & リョージデ	… × (B リョージ;リョーミミ) デ
裏面が & ウラメンガ	… × (B ウラメン;リメン) ガ

- 人名における連濁の有無，音転，読み入れ換え

谷川俊太郎 & タニガワトシタロー	… × (W タニガワ;タニカワ)(B トシタロー;シュンタロー)

- 「日本」における「ニホン」「ニッポン」の入れ替え，行政区画名の「町」における音訓入れ替え

大日本帝国憲法 & ダイニホンテーコクケンポー	… ×ダイ (B ニホン;ニッポン) テーコクケンポーガ
櫛形町 & クシガタチョー	… ×クシガタ (B チョー;マチ)

[18] 「メグスリ (目薬)」を「マグスリ」と発音する (母音 /e/→/a/),「ムジン (無人)」を「ブジン」と発音する (子音/m/→/b/)，といったように，母音１つ，ないし子音１つのみの交替の場合は，語の読みに関する知識レベルの言い間違いではなく音声レベルの言い間違いの可能性もある為，(B) ではなく (W) 対処とする.［例］無人 & (W ブジン;ムジン) × (B ブジン;ムジン)

【タグ (A)】 本タグは，アルファベット・算用数字・記号 (いずれも全角) を表記する為に用いる．これらの字種は，以下に示すのように本タグを利用し 漢字仮名に併記する形で記述.

(A 千九百九十五;１９９５) 年，(A 二十六．三五;２６．３５)，(A シーディーアール;ＣＤ－R)，
(A トリプルエーアイ;ＡＡＡＩ)，(A ハンドレッドベースティー;１００ＢＡＳＥ－T)

【タグ (K)】 本タグは，本来基本形で漢字や一部の記号 (○と．) で表記するはずの語が，(F) や<FV>，値なしの (?) が途中に生じた為に表記できなくなった場合に用いる．

(K ひ (F いー) だり;左)，(A 三 (K ま<FV>る;○) 四;３０４)

【タグ (笑)，(泣)，(咳)】

◇ 本タグは，話者の笑い声・泣き声・咳と発話が，同時もしくは入り混じりながら進行している区間 (笑いながら発話している，など) に付与する．

【タグ (L)】

◇ 本タグは，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．必ずしも独り言であるとは限らない．

【タグ<FV>】

◇ 本タグは，強いきしみ発声 (ボーカルフライ) などによって，母音が明確に同定できない場合に用いる．
◇ 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<VN>】

◇ 本タグは,「うん/うーん/ふーん」の音の認定が困難なものに対して用いる．詳細は 7 節参照.
◇ 対話 (講演 ID が D で始まるファイル) に限定して付与.

【タグ<H>，<Q>】

◇ 本タグは，母音 (<H>)・子音 (<Q>) の引き延ばし現象のうち,「コレカラー」「スゴーイ」や「サッスガ」「カイッセキ (解析)」のように，本来語彙的には引き延ばしは存在しないが，パラ言語的意味等が付与されることによって，あるいは言い淀みによって，一時的に引き延ばされているものに対して付与する (4.3 節参照).

すごい & ス<Q>ゴ<H>イ　　… ×すっごーい & スッゴーイ

◇ タグ (W)，(D)，及び感情表出系感動詞 (F) の内部では，原則本タグは用いず「ー」「っ」で表記 (表 8・10 参照).

解析が & (W カイッシェーキ;カイセキ)　　… × 解析が & (W カイ<Q>シェ<H>キ;カイセキ)
(D ざっつ) 雑音の & (D ザッツ) ザツオンノ　… × (D ざつ) 雑音の & (D ザ<Q>ツ) ザツオンノ
(F うわわっわー) & (F ウワワッワー)　　… × (F うわわわ) & (F ウワワ<Q>ワ<H>)

【タグ<笑>，<咳>，<息>】 本タグは，言語音とは独立して生じる話者の笑い・咳・息に付与する.

【タグ<P>】

◇ 本タグは，短単位 (9 頁の注 15 参照) の内部に生じる 0.2 秒以上のポーズに対して付与する[19].
◇ 以下に示す例のように，ポーズの開始・終了時刻の情報 (ミリ秒単位, 形式は時間情報部と同じ) を値に持つ.

0122 00331.802-00334.403 L:
最近　&　サイキン
→ 半年　&　ハン<P:00333.068-00333.442>トシ<H>
0124 00335.025-00335.170 L:

[19] 本来，0.2 秒以上のポーズで転記単位を分割するが，ポーズが短単位内部に生じた場合には，転記単位を分割せず 1 転記単位とした上で，本タグを利用して 0.2 秒以上のポーズとして時間の情報を残す.

5.2 タグ間の共起関係

本節では，複数のタグが重複して付与される場合の制約について説明する．タグ間の共起関係は，以下に挙げる「同一範囲」と「入れ子」の 2 種類がある．CSJ では，(1)2 つのタグが共起するかしないか，(2) 同一範囲の場合どちらのタグが外側になるか，の 2 点に関して制約を与えている．同一範囲，入れ子毎に表 7，8 にまとめる．

【同一範囲】複数のタグが同じ文字範囲に付与：(M (A エス;S)), (D (? かわ)), (F (W トノー;(? ソノー, アノー)))
【入れ子】あるタグの付与範囲の一部に別のタグが付与：(L (? 市外, 市内) 通話), (K たち (F いー) ばな;橘)

同一範囲のタグの優先順位 (右にある記号が外側)： A, K < ?単 < W < B < F, D, D2 < R < O < M < X < L < 笑

表 7: 【同一範囲】タグ間の共起関係 (○：共起，△：条件付きで共起，×：共起しない，–：タグ制約上そもそも共起しない)

内＼外																				
タグ	A 左	A 右	K 左	K 右	?単	W 左	W 右	B 左	B 右	F フ	F 感	F 応	D	D2	R	O	M	X	L	笑	?複
A	-	-	×	×	○	-	-	-	-	×	×	×	×	×	○	×	○	○	-	-	○
K	-	-	-	-	○	-	-	-	-	×	×	×	×	○	○	○	○	○	-	-	○
?単	-	-	-	-	-	○	○	○	×	○	○	○	○	△	○	○	○	○	○	○	-
W	-	-	-	-	-	-	-	○	×	○	×	○	×	△	○	○	○	○	○	○	○
B	-	-	-	-	-	-	-	-	-	×	×	×	×	○	○	○	○	○	○	○	○
F フ	-	-	-	-	-	-	-	-	-	-	-	-	×	×	×	×	○	○	○	○	○
F 感	-	-	-	-	-	-	-	-	-	-	-	-	×	×	×	×	○	○	○	○	○
F 応	-	-	-	-	-	-	-	-	-	-	-	-	×	×	×	×	○	○	○	○	○
D	-	-	-	-	-	-	-	-	-	-	-	-	-	-	○	×	○	○	○	○	○
D2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	○	×	○	○	○	○	○
R	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	○	○	○	○	○	△
O	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	○	○	○	○	○
M	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	○	○	○	○
X	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	○	○	○
L	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	○	×
笑	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	×
?複	×	×	×	○	-	○	○	×	×	△	△	△	×	×	○	○	○	○	○	○	-
(?)	×	×	×	×	-	○	×	×	×	×	×	×	×	×	×	×	○	×	○	○	-
<FV>	×	×	×	×	×	○	×	×	×	×	×	×	×	×	×	×	○	×	○	○	×
<VN>	-	-	-	-	△	×	×	×	×	×	×	△	×	×	×	×	△	△	△	△	△
<笑>	-	-	-	-	×	×	×	×	×	×	×	×	×	×	×	×	×	×	○	×	×
<H>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<Q>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<P>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

表 8: 【入れ子】タグ間の共起関係 (○：共起，△：条件付きで共起，×：共起しない，–：タグ制約上そもそも共起しない)

内＼外																				
タグ	A 左	A 右	K 左	K 右	?単	W 左	W 右	B 左	B 右	F フ	F 感	F 応	D	D2	R	O	M	X	L	笑	?複
A	-	-	×	×	○	-	-	-	-	×	×	×	×	×	○	×	○	○	-	-	○
K	○	×	-	-	○	-	-	-	-	×	×	×	×	×	○	○	○	○	-	-	○
?単	○	×	×	×	-	○	△	○	×	○	○	○	○	△	○	○	○	○	○	○	-
W	-	-	-	-	○	-	-	×	×	△	×	×	×	△	○	○	○	○	○	○	○
B	-	-	-	-	○	×	×	-	-	×	×	×	×	×	○	○	○	○	○	○	○
F フ	○	×	○	×	○	○	△	○	×	-	-	-	×	×	○	○	○	○	○	○	○
F 感	○	×	○	×	○	○	×	○	×	-	-	-	×	×	○	○	○	○	○	○	○
F 応	○	×	○	×	○	○	△	○	×	-	-	-	×	×	○	○	○	○	○	○	○
D	○	×	×	×	○	×	×	×	×	×	×	×	-	×	○	○	○	○	○	○	○
D2	△	×	×	×	○	×	△	×	×	×	×	×	×	-	○	○	○	○	○	○	○
R	×	×	×	×	○	×	×	×	×	×	×	×	×	×	-	○	○	○	○	○	△
O	×	×	×	×	○	×	△	×	×	×	×	×	×	×	○	-	○	○	○	○	○
M	×	×	×	×	○	×	△	×	×	×	×	×	×	×	○	×	△	○	○	○	○
X	○	×	×	×	○	×	△	×	×	○	×	×	×	○	○	○	○	-	○	○	○
L	-	-	-	-	○	○	×	○	×	○	○	○	○	○	○	○	○	○	-	○	○
笑	-	-	-	-	○	○	×	○	×	○	○	○	○	○	○	○	○	○	○	-	○
?複	×	×	×	×	-	×	△	×	×	△	×	×	×	×	○	○	○	○	○	○	-
(?)	○	×	○	×	-	○	×	○	×	○	○	○	○	○	○	○	○	○	○	○	-
<FV>	○	×	○	×	○	○	×	○	×	○	○	○	○	○	○	○	○	○	○	○	○
<VN>	-	-	-	-	△	△	×	△	×	×	×	×	×	×	△	△	△	△	△	△	△
<笑>	-	-	-	-	○	○	×	○	×	○	○	○	○	○	○	○	○	○	○	○	○
<H>	-	-	-	-	○	△	×	○	×	△	△	○	△	○	○	○	○	○	○	○	○
<Q>	-	-	-	-	○	×	×	○	×	○	×	○	×	○	○	○	○	○	○	○	○
<P>	-	-	-	-	○	○	×	○	×	○	○	○	×	○	○	○	○	○	○	○	×

F フ…タグ F・フィラー；F 感…タグ F・感情表出系感動詞；F 応…タグ F・応答詞 (対話の場合のみ)；<笑>…<息>, <咳>も含む
(?)…値なし (?)；?単…値 1 つ (? ○)；?複…値複数 (? ○○, △△)；左・右…タグの左項・右項 (例：(A 左項;右項))

(?) 複数候補に関する補足説明：

◇ 同一範囲において，(W)(F)(M)(O)(R)(X) と (?) の複数候補との組合せは，(?) が外に来る場合と内に来る場合の両方を許容しているが，以下の条件で使い分けをしている．(O) と (W) を例に説明する．

- (O)・(W) かそれ以外かで迷う場合は，(?) が外，(O)・(W) が内：

(? (O 寺や), 寺屋) & テラヤ, (? 国保, 国庫) & (? (W コクコ;コクホ), コクコ)

- (O)・(W) であることは確かだが，その中で複数の候補が想定される場合，(O)・(W) が外 (?) が内.

(O (? イムニダ, イムニカ)) & (O (? イムニダ, イムニカ))
国保 & (W (? コクコ, コフコ);コクホ), (? 国保, 国庫) & (W コフコ;(? コクホ, コクコ))

◇ 同一範囲，入れ子のいずれの条件においても，(?) の複数候補が外に来るケースを概ね許容 (○) しているが，基本的に (?) の複数候補はあまり多用せず，できる範囲で候補を一意に同定する．

表 9: 【同一範囲】△の場合の使用条件．非許容の場合 (×) の補足説明

	外	内	転記例／許容条件
△	F	?複	音が曖昧な場合，(F (? うー, えー)) のように複数候補も十分あり得るが，多用せずできる限り候補を一つに絞る．
△	F 応	VN	対話の応答詞のうち，基本形が (F うん)，(F うーん)，(F ふーん) の場合に限定．
△	F 応以外	VN	<VN>は必ず (F) と共起して出現する．<VN>とそれ以外のタグとの関係は，(F) 応と他のタグとの関係に依存する．
△	D2	?単	(D2) 中の (?) は発音形の場合のみ．語が曖昧な場合は (D2) でなく (D) 対処：(D (? ご)) 御本人
△	D2	W	助詞の「を」を「ウォ(ー)」と発音した場合のみ：コレ (D2 (W ウォ;オ)) オ
△	?複	R	(? (R ××), 砂糖) のように (R) を複数候補の 1 つにのみ付与すると伏せ字の意味をなさないため原則認めない．
×	AK 左	?複	(A (? エヌ, エム);(? Ｎ, Ｍ)) とはせず，(?) を外に出し (? (A エヌ;Ｎ),(A エム;Ｍ)) とする．
×	A 右	?複	(A スリーエヌ;(? ３Ｎ, ＮＮＮ)) など理論上あり得るが，(A) 内ではどちらかの候補に倒し複数候補とはしない．
×	?単複	FV,	<FV>は母音が曖昧な場合に使用するタグのため，(?) 対処しない．
×	?単複	笑	<笑>等のボーカル音は，そもそも存在が曖昧な場合には表記しない．
×	B 右	?短	右項に自信がない場合には (B) 対処としない．
×	F 感	W	感情表出系感動詞は聞こえた通り記すため (W) 対処しない．
×	D	?複	(D (? のぼ, のご)) などの複数候補は理論上あり得るが，ここまでせず (D (? のぼ)) のように単一候補とする．
×	?複	L, 笑	同一範囲の (L), (笑) は (?) 複数候補の外に出す：(? (L ソレ),(L コレ)) → (L (? ソレ, コレ))
×	D	A	(D (A シー;Ｃ))(A シーディー;ＣＤ) など，理論的にはあり得るが，(D) 内では (A) 対処しない．
×	D	K	(D (K か<FV>い; 回)) 回復など，理論的にはあり得るが，(D) 内では (K) 対処しない．
×	D	W,B	(D) は発話された音をそのまま表記するため，(D 運) 運動 & (D (W オン;ウン)) ウンドー のように (W), (B) 対処せず．

表 10: 【入れ子】△の場合の使用条件．非許容の場合 (×) の補足説明

	外	内	転記例／許容条件
△	A 左	D2	数字内の数字の言い直しに限定：(A 千 (D2 八百) 九百四十;１９４０)
△	W 左,F,D	H	母音の引き延ばしが極めて長い場合に長音記号に追加した形でのみ使用：(W スー<H>;スル),(F アー<H>)
△	W 右	?,F,D2,O,M,X	複数の語が融合して一つの (W) にまとめられた場合に限定：(W コキャラ;ココ (D2 カラ)) マデ
△	F フ	W	(F えっとね) や (F あのですね) のように (F) が複数の短単位 (例：あの/です/ね) から構成される場合に限定．
△	F フ	?複	(F (? あっと, えっと) ですね) のように複数候補も十分あり得るが，多用せずできる限り候補を一つに絞る．
△	D2	?単	(D2) 中の (?) は発音形の場合のみ許容．語が曖昧な場合は (D2) でなく (D)．
△	D2	W	助詞の「を」を「ウォ(ー)」と発音した場合のみ：ナニ (D2 カ (W ウォ;オ)) カガ
△	M	M	括弧や句読点などのメタ記号の読み上げに限定：平仮名で (M こっかい (M 中括弧閉じる) では) で
△	?複	R	R を複数候補の 1 つにのみに付与すると伏せ字化の意味をなさない場合もあるため，原則認めない．
△	–	VN	VN は必ず (F) と共起して出現するため，FV との関係は，F 応と他のタグとの関係に依存．
×	A 左	?複	(A) 内では (A (? エヌ, エム) 三五;(? Ｎ, Ｍ) 3) とはせず (?) を外に出す：(? (A エヌ;Ｎ 3),(A エム;Ｍ 3))
×	A 右	?単	理論上あり得るが，右項に自信がない場合には (A) 表記としない．
×	W 左	D	(W) 左項では (W カイ (D セ) セキュ;カイセキ) のように (D) は使用せず，(W カイセセキュ;カイセキ) とする．
×	W 左	?複	(W ホ (? ト,ド);ホントー) のように左項の一部ではなく，(W (? ホト,ホド);ホントー) のように全体に (?) 付与．
×	B 左	D	(B シブ (D タ) タイ;ジュータイ) のように (D) は使用せず (W) 対処：(B (W シブタタイ;シブタイ);ジュータイ)
×	B 右	?単	右項に自信がない場合には (B) 対処としない．
×	W 左,F 感,D	Q	聞こえた通り表記する為，<Q> は使用せずに「っ」で表記：(D こっ) 話し言葉
×	W 右	R	(W) 右項の一部が (R) の対象であっても (R) は (W) 全体に付与：(W ××;(R ××) ×) → (R (W ××;×××))
×	O	A,M	理論上あり得るが，タグ (O) を付与した時点で (A) や (M) は付与しない．

6 再朗読の転記について

本節では，再朗読 (講演 ID が R00 から始まるファイル) の転記のうち，通常の講演の転記と異なる点について説明する．朗読 (講演 ID が R01,R02,R03 から始まるファイル) については本節で述べる特別な措置は行なっていないので注意されたい．

◇ 転記単位全体が，朗読原稿にない余分な発話の場合[20]：<朗読間違い>タグを付与し，その単位の文字化はしない．

0035 00112.070-00114.322 L:	右のように対処	0035 00112.070-00114.322 L:<朗読間違い>
それでは & ソレデワ	⇒	0036 00115.436-00116.199 L:<朗読間違い>
(F あ) & (F ア)		0037 00117.940-00119.938 L:
0036 00115.436-00116.199 L:		それはですね & ソレワデスネ
やり直します & ヤリナオシマス		
0037 00117.940-00119.938 L:		
それはですね & ソレワデスネ		

◇ 転記単位の一部が，朗読原稿にない余分な発話の場合[21]：文字化した上で (X) を付与

0414 01172.918-01176.613 L:		
(F ま) & (F マ)		
そういったものを & ソーイッタモノオ		
色々 & イロイロ		
(F えー) & (F エー)		
(X 見てて) & (X ミテテ)		← 転記単位の一部を言い直し
見てみたんですが & ミテミタンデスガ		
0090 00382.480-00384.785 L:		
(X (D (? つ))) & (X (D (? ツ)))		← 朗読原稿にない (D)
全ての & スベテノ		
0360 00788.358-00792.626 L:		
バリエーションが & バリエーションガ		
少ない & スクナイ		
為に & タメニ		
(X (F えー) & (X (F エー)		← 転記単位の一部を言い直し
関西に比べて & カンサイニクラベテ		
回答 (D げ)) & カイトー (D ゲ))		
0361 00793.262-00796.283 L:<朗読間違い>		← 転記単位全体を言い直し
0362 00796.763-00797.614 L:<朗読間違い>		← 転記単位全体を言い直し
0363 00799.184-00799.758 L:<朗読間違い>		← 転記単位全体を言い直し
0364 00800.032-00800.486 L:<朗読間違い>		← 転記単位全体を言い直し
0365 00801.043-00806.870 L:		
(F えー) & (F エー)		
関西に比べて & カンサイニクラベテ		
回答 (D げ) 語形の & カイトー (D ゲ) ゴケーノ		
異なりというのは<FV> & コトナリトユーノワ<FV>		← <FV>に関しては (X) を付与せず
少なく & スクナク		
なってるんですけれども & ナッテルンデスケレドモ		

◇ 上記以外の朗読間違いについては，コメントを付与する．

0120 00289.467-00291.674 L:
仕事を & シゴトオ
持っている & モッテイル
方にのみ & カタニノミ
%TYPE=O_ERR 朗読間違い 原稿では「方にのみ」ではなく「方のみに」
0168 00392.749-00395.306 L:
引き続き & ヒキツズキ
百四ページの & ヒャクヨンページノ
%TYPE=O_ERR 朗読間違い 原稿では「百四」ではなく「百十四」.
グラフ一から & グラフイチカラ

[20] 読み間違えたため最初から読み直した場合など．

[21] 基本形のレベルで朗読原稿と異なるものを対象とする．ただし，<FV>, 値なしの (?) のみが異なる場合は本タグの付与対象外とする．

7 対話の転記について

本節では，対話 (講演 ID が D から始まるファイル) の転記のうち，通常の講演の転記と異なる点について説明する．対話の転記テキストの例を図 11 に示す．

◇ **話者 ID：** 話者 2 名の ID は L と R．それぞれ音声ファイルのチャンネルに対応．
インタビューの場合，インタビュアーは L，インタビュイーは R に固定．

◇ **転記単位の認定方法：** 単位の認定は L・R 毎に行なう．その際の基準は通常の講演と同じ．

◇ **転記単位の順番：** L・R に関係なく，開始時間の早い順に並べる．

◇ **タグの付与範囲：** タグは，L・R それぞれの転記単位内で整合的に付与する[22]．

◇ **応答表現の扱い：**

- 応答表現に対し，通常の講演では特にタグは付与していないが，対話では (F) を付与する[23]．
- 対話に実際に出現し (F) の付与対象となった応答表現は以下の通り：

> はい，ええ，うん，うーん，ううん，いえ，いいえ，いや

- 応答表現の表記：基本形における応答表現の表記は，上記リストに示したものに固定する．その為,「ハーイ」のように母音や子音の引き延ばしが生じた場合には，<H>, <Q>を用いて表記する (発話 ID=0026 参照)．
- 「うん」「うーん」「ふーん」「んー」「ん」の区別は概ね以下のように行なう：

> 下降の音調：(F うん)・(F うーん)，　平坦：(F んー)・(F ん)，　上昇：(F ふーん)
> ※ (F ふーん) については，上記の音声的条件と機能的条件 (「感嘆」) の両方を満たした場合に限定

◇ **タグ<VN>：** 応答表現「うん」「うーん」「ふーん」の発音において，鼻音が混ざる・母音カテゴリーが曖昧等の理由で，音 (分節単位) の認定が困難なものについては，発音形を <VN>で表記する (発話 ID=0028,0034,0037)．

◇ **共話の扱い：** 以下の例に示すように，対話では 2 人で単語や文を共同して発話する場合がある (以下「共話」)．タグの付与や文字の表記は，L・R 毎に行なう為，例えば「共話例Ｃ」の (D ざ) に見られるように，語の途中 (この場合は「小遊三」の「ざ」) から発話をしたものについては，例え共話としては語の構成要素であっても，語の断片と見なし (D) が付与される為，扱いに注意されたい．

［共話例Ａ］
0212 00251.164-00252.195 R:
北野　　&　(L キタノ)
0213 00253.762-00254.620 L:
大さん　　&　マサルサン

［共話例Ｂ］
0483 00572.810-00573.463 R:
食堂に　　&　ショクドーニ
0484 00573.358-00574.506 L:
に　　&　ニ
行くんだ　　&　イクン (笑ダ<笑>

［共話例Ｃ］
0332 00429.353-00432.664 R:
小遊三さんに　　&　コユーザサンニ
し (D (? て))　　&　シ (D (? テ))
(F はい)　　&　(F ハイ)
0333 00430.389-00431.390 L:
(D ざ) さん　　&　(D ザ) サン　　… 語の断片と見なす
0334 00431.688-00432.236 L:
しましょう　　&　シマショー
0335 00437.409-00437.949 R:
(F うーん)　　&　(F <VN>)

[22] 例えば，図 11 の発話 ID=0023〜0025 に示すように，(R) が，相手話者の発話 (0024) を狭んで次の転記単位に跨がる場合であっても，発話 ID=0024 は (R) の付与範囲とは見なさない．
[23] 質問に対する肯定，依頼に対する承諾，話し手に対する相槌など，様々な機能があるが，機能に関わらず一律 (F) を付与する．

表 11: 対話の転記テキストの例

0021 00024.323-00026.764 L:			
じゃ	&	ジャ<H>	
浮かぶ	&	ウカブ	
人	&	ヒト	
取り敢えず	&	トリアエズ	
言ってもらっていいですか	&	イッテモラッテイーデスカ	
0022 00026.915-00027.264 R:			
(F ええ)	&	(F エー)	← 応答表現「ええ」にもタグ (F) 付与
0023 00028.979-00030.610 R:			
(F えーとー)	&	(F エートー)	
(R ×××・	&	(R ×××	← タグ (R) の範囲が，話者 L の「はい」を狭んで 0025 まで続く
0024 00030.813-00031.154 L:			
(F はい)	&	(F ハ<H>イ)	
0025 00030.892-00031.778 R:			
×××××) さん	&	×××××) サン	
0026 00031.540-00032.076 L:			
(F はい)	&	(F ハ<H>イ)	← 応答表現「はい」にもタグ (F) 付与
0027 00033.312-00033.837 R:			
でしょう	&	デショー	
0028 00034.008-00034.573 L:			
(F うーん)	&	(F <VN>)	←「うーん」の発音形 <VN>
0029 00034.520-00036.565 R:			
それから	&	(W (? ホ);ソレ)(W カ;カラ)	
水野晴郎さんでしょう	&	ミズノハルオサン (? デ) ショー	
0030 00036.053-00036.521 L:			
(F はい)	&	(F ハ<H>イ)	
0031 00037.377-00038.396 R:			
この	&	コノ	
人	&	ヒ (? ト)	
誰だっけ	&	ダレダッケ	
0032 00039.707-00042.703 R:			
(F あー)	&	(F アー)	
あたし	&	アタシ	
結構	&	ケッコー	
名前	&	ナマエ	
知らない	&	シラナイ	
水沢アキさんか	&	ミズサワアキサンカ	
0033 00042.938-00043.613 L:			
(F はい)(F はい)(F はい)(F はい)	&	(F ハイ)(F ハイ)(F ハイ)(F ハイ)	← 1 つの応答表現毎に括り直す
0034 00043.636-00044.010 R:			
(F うーん)	&	(F <VN>)	
0035 00044.577-00045.449 R:			
それと	&	ソレト<H>	
0036 00046.146-00047.869 R:			
この	&	コノ	
駄洒落	&	ダジャレ	
言う	&	ユー	
人	&	ヒト	
和田勉さんだ	&	ワダベンサンダ	
0037 00048.011-00048.753 L:			
和田勉	&	ワダベン	
(F うん)	&	(F <VN>)	
0038 00048.550-00048.910 R:			
(F はい)	&	(F ハイ)	
0039 00049.402-00049.652 L:<笑>			