『日本語話し言葉コーパス』書き起こしの為の用字用語辞書の作成

間淵洋子 †,‡，西川賢哉 †，土屋菜穂子 ††，相馬さつき †，籠宮隆之 †，小磯花絵 †，前川喜久雄 †
† 国立国語研究所，‡ 東京都立大学大学院，†† 青山学院大学大学院

1 初めに

独立行政法人国立国語研究所，独立行政法人通信総合研究所 (現・独立行政法人情報通信研究機構)，東京工業大学の 3 機関は，1999-2003 年の 5 年間に渡って,「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」という研究課題を共同で実施した．この研究課題において，国語研究所が中心となって構築したのが,『日本語話し言葉コーパス』(The Corpus of Spontaneous Japanese；以下 CSJ) である (概要については，前川 2004 を参照).

CSJ は，自発性の高いモノローグ音声を主な対象としており，日本語の自発音声コーパスとしては最大規模 (661 時間) である．CSJ には，種々の研究用付加情報が含まれるが，その中で最も基本的な情報が，音声を書き起こした転記テキストである．これは，661 時間全体に渡って提供されるもので，全て人手で書き起こされている．

本稿では，この書き起こし作業を効率的に行なう為に作成した用字用語辞書，およびそこから派生させた仮名漢字変換用辞書，表記確認用辞書について述べる (具体的な派生方法および実装については，本大会予稿集所収の籠宮他を参照).

2 転記テキスト：二つの表記法

CSJ の転記テキストは，漢字仮名交じりで表記される「基本形」と片仮名で表記される「発音形」の二つの表記法を用いて記述されている．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．転記テキストで利用される主なタグを表 1 に示す.

```
0300 00786.419-00787.869 L:
我々の                          & ワレワレノ
データーを                      & データーオ
見ますと                        & ミマスト
0301 00788.076-00788.783 L:
(F えー)                        & (F エー)
(D がっ)                        & (D ガッ)
0302 00789.488-00793.099 L:
模擬講演                        & モギコーエン
ある                            & アル
種の                            & シュノ
ある                            & アル
模擬講演が                      & (W モリ;モギ) コーエンガ
(A 三．一二;３．１２) 文節 & サンテンイチニ<H>ブンセツ
0303 00793.471-00796.785 L:
から                            & カラ<H>
音響学会や                      & オンキョーガッカイヤ
国語学会では                    & コクゴガッカイデワ
このように                      & コノヨーニ
やや                            & ヤヤ
ちょっと                        & チョット
長いと                          & ナガイト
0304 00797.225-00798.356 L:
いう                            & ユー
程度に                          & テードニ
        基本形                          発音形
```

図 1: 転記テキストの例 (一部抜粋)

CSJ が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．発音形には，人が聞き取ることのできる範囲で正確に発音が記される.

一方基本形は，漢字仮名交じりで表記されるものであり，人が読んだり語句を検索する際に利用することができる．また CSJ の主な応用領域の一つである音声認識研究では，CSJ を用いて音響モデルや言語モデルを構築するが，このうち言語モデルの構築においては，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が求められる．その為基本形の設計および実際の書き起こし作業において，表記の統一は最も重要な課題となる.

3 用字用語辞書作成の目的

2 節で述べたように，書き起こし作業において重要なことは表記の揺れが存在しないことである．しかしながら CSJ のように大規模な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ず，作業者間ないし同一作業者内での揺れをいかにして抑えるかが作業上大きな問題となる.

この問題に対処する為，書き起こし作業を始めるにあたり，表記方針を決定しマニュアルを作成した．同時に，実際の作業における，表記の効率的な決定・統一を支援する為に，次の二つのタイプの辞書を作成し，作業時に利用することとした：

1. CSJ の表記原則に従った仮名漢字変換用辞書
2. 作業者が表記を決定する際に参照する為の可読性の高い表記確認用辞書

これらの辞書に記載される情報には，一方にのみ必要な情報だけでなく，両者共通に必要な情報も存在する．そこで，この二種類の辞書を効率的に作成する為，必要となる全ての情報を含む用字用語辞書 (約 11 万語) を作成し，そこから二つの辞書を生成することにした.

4 節において転記テキストの基本形を記す為に採用された表記方針について簡単に触れた後，5 節で各種辞書について述べる.

4 基本形の表記方針

本節では基本形の表記方針の概要を示す (詳細は，小磯他 2004 を参照) [1].

4.1 字種

基本形の表記には，漢字 (JIS 第 1・2 水準)，平仮名，片仮名を用いる．ただしこれらの字種に併記する形で，算用数字，アルファベット (ローマ字・ギリシャ文字)，一部の記号 (以上いずれも全角) も使用することができる[2].

[1] 基本形は，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されているが，固有名詞の表記等については一部例外もある.

[2] 具体的には，(A シーディーロム;ＣＤ－ＲＯＭ) のように規定のタグを用いて記す.

表 1: 転記テキストに使用されるタグの一覧 (一部省略)

【文字範囲を指定し，その範囲の性質に言及するタイプ】		例
◇ (F)	フィラー・感情表出系感動詞	(F あの)，(F うわ)
◇ (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
◇ (D2)	助詞・助動詞・接辞の言い直し	そこ (D2 の) まで，(D2 不) 不自然
◇ (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),(W ミダリ;ヒダリ)
◇ (B)	語の読みに関する話者の知識レベルの言い間違い	(B シブタイ;ジュータイ)
◇ (?)	聞き取りや語彙の同定などに自信がない場合	(? タオングー)，(? 堆積，体積)
◇ (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
◇ (O)	外国語や古語，方言など	(O ザッツファイン)
◇ (R)	講演者名，差別語，誹謗中傷など	国語研の (R ××) です
◇ (A)	基本形でアルファベットや算用数字・記号を使用する場合	(A シーディーアール;ＣＤ－Ｒ)
【音や事象自体を記号で表記するタイプ】		
◇ <H>,<Q>	母音 (<H>)・子音 (<Q>) の引き延ばし	ソレデ<H>，カイ<Q>セキ
◇ <FV>	強いボーカルフライ等で母音が同定できない場合	ダカラ<FV>

4.2 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，両表記共，一般的に使用されるものについては，原則漢字を採用する (○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する (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習慣が強いが，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). 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は漢字で,「やって おく」や「食べて みる」のようにテ形複合動詞として用いられる場合には平仮名で表記する.

【「言う／いう」の使い分け】 以下の組み合わせで出現した場合のみ平仮名で表記する (「言う」という動作が形骸化された用法は平仮名で表記されることが多いが，その判断は非常に難しく揺れを招き易い為，このように形式的な判断基準を採用). 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.

{指示副詞：ああ／こう／そう／どう, 引用の助詞：と／って} + {いう, いった} + {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[3].

4.3 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する (○ 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 水準の漢字を採用する (○ 憧 れ／× 憬 れ，○一 獲 千金／×一 擭 千金). ただし著名人の氏名は例外的に第 2 水準も使用可とする (○小 渕 元首相…通常は「淵」に統一).

【同音類義語】 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 (○悲しい／×哀しい，○会う／×逢う)，それ以外の類義語については書き分ける (○表わす／○現わす，○計る／○測る／○図る).

4.4 送り仮名

- 用言で複数の送り仮名の候補がある場合は送り仮名の字数の多い方を採用する (○行なう／×行う).
- 名詞で送り仮名の有無に揺れがある場合は，原則送り仮名を付ける (○後ろ／×後). ただし送り仮名を付与しない習慣の強い語は個別に定義する (○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).

4.5 片仮名

片仮名表記の対象は以下の三つに限定する：(1) 外来語・外国語 (ただし中国の人名地名，および韓国・北朝鮮の著名人の名前は漢字表記)，(2) 専門用語や俗語，固有名詞などで片仮名表記の習慣が強いもの (例：フッ素，ト書き，ダフ屋)，(3) 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 (1) の外来語については,「ビオラ／ヴィオラ」のような表記の揺れが多く見られる. そこで,「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ごとに表記の方針を定めた上で，CSJ に現われる全ての片仮名語を 5 節で述べる辞書に登録し，統一を図った.

4.6 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，という三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た[4].

5 辞書

前節に示した表記原則に基づき，実際の語の表記を定めた用字用語辞書を作成した. この辞書から，日本語入力システム「かんな」用の日本語仮名漢字変換用辞書，および表記確認用辞書を自動的に生成した. 本節ではそれぞれの辞書について説明する.

5.1 用字用語辞書

5.1.1 基本構成

用字用語辞書には，仮名漢字変換用辞書，および表記確認用辞書を作成する為に必要な情報が全て含まれている. 具体

[3] 例えば副詞の「とびきり」を，構成要素である「飛ぶ」と「切る」に合わせて「飛び切り」と表記する，ということまではしない.

[4] 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した. 実際には，ある程度の量の書き起こしを行ない口語的な表現を抽出した上で，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.

的には，(1) 使用の可否，(2) 語句の読み，(3) 表記，(4) 品詞ラベル，(5) 品詞記号，(6) 注記から構成される．

○ おこな:行な:あわ行 (#W5r)
× おこな:行:あわ行(#W5r)〔→「行な (う)」〕
(1) (2) (3) (4) (5) (6)

本辞書では，表記の揺れが想定される語について，個々の表記の可否を明示する目的で，使用可能な表記だけでなく，使用不可能な表記も，CSJ での正表記などを示す注記を付けた上で，積極的に登録している．その区別を (1) において「○ (可)」「× (不可)」で示す．仮名漢字変換用辞書では○のみ，表記確認用辞書では○×双方のデーターが利用される．

(2)，(3) は，語句の読みとそれに対応する表記を示したもので，日本語入力システムにおける入力文字列および変換文字列に対応するものである．活用語は活用語幹のみ，非活用語や特定の連語表現 (例：「国語研究所」「に基づいて」) は，語や表現の全体を登録してある[5]．また，基本形において必ずタグ付きで表記される語 (＝ (A) タグ付きで表記されるアルファベットや算用数字を含む語) についても，タグを含めた形全体で登録している．このように，見出し項目の長さや単位が統一されていない，あるいは特殊な表記形態が存在するのは，本辞書があくまで CSJ の書き起こし作業を支援する目的で整備されたものであり，適切な表記を正確かつ効率的に変換・検索するのに必要なものは積極的に登録するという方針を取っている為である．

(4)，(5) は，品詞情報を示したものである．日本語入力システム「かんな」の品詞体系に準拠した仮名漢字変換用の品詞記号 (5) に，対応する品詞ラベル (4) を併記した形で表わす．(4) は人が読む為に付加したもので，表記確認用辞書に変形して利用され，(5) は仮名漢字変換用辞書にそのまま利用される．なお，口語表現を中心に従来の「かんな」の品詞体系に収まらないものについては，新たな品詞を立てた (5.2 節 b を参照)．

(6) の注記は，表記確認用辞書に記載するもので，表記の統一や選択の補助となる情報が記されている (5.3 節 d を参照)．

5.1.2 構築の手続き

用字用語辞書の作成にあたっては，フリーの仮名漢字変換用辞書である Pubdic+ などをベースにした．これらは一般の仮名漢字変換用辞書である為，表記の統一はなされていない．そこで，見出し項目それぞれについて，CSJ の表記原則に従って使用可能な表記と不可能な表記に分類した上で，使用の可否の情報を記した．また品詞についても，表記原則や連接を考慮し，適宜変更を加えた．

更に，書き起こし作業の過程で，辞書に存在しない語句 (未知語) が出現した場合には，その都度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，表記や品詞を決定した上で，新たな項目として本辞書に登録した．

5.2 仮名漢字変換用辞書

「かんな」用の仮名漢字変換用辞書は，用字用語辞書から自動的に作成される．書き起こし作業の効率と精度を高める為に，以下のような工夫を施した[6]．

a. 使用不可の表記は変換候補に現われない

仮名漢字変換用辞書には，用字用語辞書内で使用可能とされたもののみを登録する．これにより，適当な単語・文節ごとに変換を行なえば，表記可能なものだけが変換候補として現われることになり，誤表記を防ぐことができる．

＜例＞「おこなう」を変換した場合
1 行なう 2 おこなう 3 オコナウ
※ 使用不可の表記「行う」は変換候補に現われない．
※ 変換候補のうち，最後の二つは，「かんな」の仕様により現われるものであり，基本形表記においては使用できない．

b. 口語表現が適切に変換できる

前述の通り，CSJ では一定の基準を設けた上で口語表現を積極的に採用するという方針を取ったが，口語表現の中には，通常の「かんな」システムでは適切に変換できないものも多い．そこで CSJ で認めた口語表現を適切に変換する為に，以下のような処理を行なった．

【品詞の変更・追加】

＜例＞ 「歩く」の連用形促音化「歩っ(て)」
変更前： 連用形がイ音便となる品詞記号のみ．
変更後： カ行五段活用で連用形が促音便になる語 (「行く」など) に当てる品詞記号を追加．

【文法定義ファイルの変更】

(1) 接続情報の変更

＜例＞ ラ行五段動詞終止連体形の撥音化
(「やるの」→「やんの」など)
変更前： 撥音形の接続情報として，助動詞「ない」「ねえ」のみを規定．
変更後： 撥音形の接続情報に，助動詞「だ」「だろう」「です」「じゃん」「じゃ」，終助詞「な」「の」を追加．

※同様の方法で対応した口語表現の例
- 一段・サ変・カ変動詞終止連体形の撥音化
(例：「来るの」→「来んの」)

(2) 新品詞の規定

＜例＞ 形容詞終止連体形の音融合型口語表現
(「痛い」→「いてえ」など)
形容詞終止連体形の音融合型口語形を新たに一品詞として規定し，形容詞終止連体形と同様の接続情報を付与．

※同様の方法で対応した口語表現の例
- 「落っこちる」の連用形促音化「落っこって」
- 形容詞連用形語尾「く」＋助詞「は」の融合形「か」
(例：「良くは (ない)」→「良か (ない)」)
- 「てあげる/であげる」の融合形「たげる/だげる」
- 「ちゃった」の口語形「ちった」

c. タグ付きの基本形表記に変換できる

基本形において，必ずタグ付きで表記される語 (＝ (A) タグ付きで表記されるアルファベットや算用数字) は，タグも含めた形で変換できるようにしている．これにより，表記統一が徹底できると共に，作業効率も上げることができる．

＜例＞「しーでぃーろむ」の入力で，
「(A~シーディーロム; ＣＤ－ＲＯＭ)」と変換
※「~」は処理の都合上，1 バイトスペースの代わりとして使用．

d. 固有名詞など特定の場合のみ用いることができる表記を，特定の記号で示す

CSJ の表記原則においては，同音同義の普通名詞と固有名詞で表記の使い分けがある場合が少なくない．その為，辞書に

[5] したがって，本辞書で登録されている単位は，CSJ における形態論的単位 (短単位・長単位) とは必ずしも一致しない．形態論的単位については，小椋他 (2004) を参照．

[6] 以下に挙げるものには，用字用語辞書の段階で対応しているものと，仮名漢字変換用辞書への派生時に対応したものの両者が含まれるが，いずれも仮名漢字変換用辞書の為に施した工夫である為，本節でまとめて述べることとする．

表 2: 注記の例

【表記候補を指示する注記】
◇基本形表記の指示
× 逢う あう [動詞] 〔→「会（う）」〕
× まるっきり まるっきり [副詞] 〔>「まるきり」; 口語・促音<Q>表記「マル<Q>キリ」〕
◇発音形表記の指示
× 学校 がっこ [名詞] 〔口語・「がっこう」の読みのみ可．発音形 (W) 表記〕
◇同一表記複数読み情報・読みのデフォルト情報
○ 愛想 あいそ [名詞] 〔「あいそう」の読みでも登録；デフォルト「あいそ」〕
○ 愛想 あいそう [名詞] 〔「あいそ」の読みでも登録；デフォルト「あいそ」〕
【適切な表記の選択を補助する注記】
○ 伊勢佐木 いせざき [地名] 〔横浜市中区の繁華街〕
○ 伊勢崎 いせざき [地名] 〔高知県高知市伊勢崎町〕
【文節の切り方を指示する注記】
○ 天の川 あまのがわ [固有名詞] 〔文節・AのB／天の川／〕
○ 数多く かずおおく [名詞] 〔文節・その他／数／多く／〕

は，普通名詞の表記には使用できないが，固有名詞の表記には使用できるというような表記が数多く存在する．また，単独では使用できないが，何らかの語を構成する文字として使用する可能性のある単漢字も，未知語の表記用に必要となる為，多く登録している．このような使用に注意が必要となる表記や文字は，以下の通り括弧付きで変換されるようにし，注意を促す．

《 》: 固有名詞
［ ］: JIS 第 1 水準の単漢字 (使用に注意が必要な文字)
【 】: JIS 第 2 水準の単漢字 (使用に特に注意が必要な文字)

＜例＞「あしずり」を変換した場合
1《足摺》 2 足ずり 3 あしずり 4 アシズリ
※ 「足摺」は地名でのみ使用可．普通名詞は「足ずり」を使用．

e. 不適当な仮名遣いの入力でも変換できる
誤り易い仮名遣いについては，誤った形で入力されても，基本形に表記すべき形に変換されるようにしており，表記の統一を図っている．

＜例＞「うなづく」の入力で「頷く」と変換
※ 正しい仮名遣いは「うなずく」．

5.3 表記確認用辞書

5.1 節で述べた用字用語辞書から，使用の可否，語句の読み・表記，品詞情報，注記を可読性の高い形で表現した表記確認用辞書が生成される．本辞書は，書き起こし作業時にエディター上で用語を検索することを目的に作成したものである．

以下で，辞書を構成する項目について，解説する．

a. 使用の可否 (第 1 項)
CSJ で使用できる表記を「○」，できない表記を「×」で示している．×の項目を辞書に掲載することにより，使用できない表記からの検索，およびその表記が使用不可能であることの明示が可能となる．

＜例＞ ○ 全て すべて [副詞]
× すべて すべて [副詞] 〔→「全て」〕
× 総て すべて [副詞] 〔→「全て」〕
× 凡て すべて [副詞] 〔→「全て」〕

b. 語句の表記 (第 2 項) と読み (第 3 項)
語句の表記とそれに対する読みを示している．活用語は言い切りの形で登録してある．

＜例＞ ○ 行ない おこない [名詞]
○ 行なう おこなう [動詞]

c. 品詞情報 (第 4 項)
品詞情報を 1 バイト角括弧“ [ ] ”によって示している．5.1 節で示した用字用語辞書の品詞情報を，語の同定に必要となる程度の粗い品詞体系に変換し，品詞を付与する．

【品詞一覧 (一部省略)】動詞，形容詞，形容動詞，名詞，名詞 (する)，人名，地名，固有名詞，形式名詞，数詞，連体詞，副詞，接続詞・感動詞他，接頭語，接尾語・助数詞・単位など

d. 注記 (第 5 項)
基本形または発音形の表記に関して注意の必要な語について，基本形表記候補の指示や使い分けの情報，発音形表記の指示，文節の切り方，その他の情報を 2 バイト亀甲括弧“〔〕”によって示している．例については，表 2 参照．

6 終わりに

CSJ の書き起こし作業の為に作成した用字用語辞書と，そこから生成される仮名漢字変換用辞書，表記確認用辞書について述べた．

揺れのない，質の揃った転記テキストを作成する為には明確な書き起こしの基準が必要となるが，今まで，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かった．その意味で，本稿で紹介した CSJ 書き起こし基準およびそれに基づく各種辞書は，今後，話し言葉のコーパスを作成する上で重要な役割を果たすと思われる．

本稿で述べた辞書のうち，仮名漢字変換用辞書と表記確認用辞書については，後日公開の予定である．

また，この発表全体を通じて紙幅の制限の為に十分な例を挙げることができなかったが，作業マニュアルには豊富な例が挙がっている．マニュアルについても，要請があれば公開する．

付記

CSJ の公開情報等については，以下のページを参照：
`http://www.kokken.go.jp/katsudo/kenkyu_jyo/corpus/`

参考文献

1. 小椋秀樹，山口昌也，西川賢哉，石塚京子，木村睦子 (2004)「『日本語話し言葉コーパス』における単位認定基準について」『日本語科学』16, pp.93-113.
2. 籠宮隆之，間淵洋子，土屋菜穂子，西川賢哉, 小磯花絵 (2005)「書き起こし作業用用字用語辞書の仮名漢字変換システムへの実装と計算」言語処理学会第 11 回年次大会予稿集.
3. 小磯花絵，間淵洋子，西川賢哉，斉藤美紀，前川喜久雄 (2004)「転記テキストの仕様 Version 1.0」『日本語話し言葉コーパス』付属電子文書.
4. 前川喜久雄 (2004)「『日本語話し言葉コーパス』の概要」『日本語科学』15, pp.111-133.