第 3 章

形態論情報

小椋秀樹

本章では，CSJ の転記テキストに付与した形態論情報について，その設計方針や単位認定・情報付与の基準等を述べる。

CSJ の形態論情報は，単位境界・代表形・代表表記・品詞・活用の種類・活用形・その他の情報の七つから成る。以下，3.1 節では単位の設計方針と認定基準，3.2 節では代表形・代表表記の必要性とその付与基準，3.3 節では品詞等の情報の設計方針と品詞・活用の種類・活用形・その他の情報の概要と品詞の判定基準について述べる。

ここでまず，本論に入る前に，本章の内容に関し，二つの点について断っておく。

1 章に述べたとおり，CSJ の転記テキストに形態論情報を付与する作業（形態論情報解析）は，次の二つの方法によって行った。

○ 人手による形態論情報解析（以下，人手解析とする。）
　国語研究所が担当した。解析したデータの規模は約 100 万語。
○ 計算機による自動形態論情報解析（以下，自動解析とする。）
　情報通信研究機構が担当した。解析したデータの規模は約 650 万語。

本章で述べる形態論情報の設計方針・単位認定基準等は，こうした解析の手法を問わず，基本的に CSJ 全体に共通するものである。ただし，3.1.5 節に述べた単位認定の際のタグの扱いのうち，具体的な手順に関する記述は人手解析を前提としたものである。しかし，タグの扱いに関する考え方（タグを単位として切り出さない等）は，やはり CSJ 全体に共通するものである。

次に，本章と 4 章との関係について説明しておく。本書で CSJ の形態論情報について述べた章としては，この 3 章のほかに 4 章がある。このように形態論情報に関して二つの章を設けたのは，次のような理由による。

国語研究所では形態論情報解析のうち，人手解析を担当した。しかし，人手解析とは言っても，単位認定から品詞等の情報付与までのすべてを人手だけで行ったわけではない。代表形・代表表記・品詞等の情報付与，データの修正等に当たっては，計算機を活用し，作業の効率化を図った。

コーパス構築の方法ということを考えた場合，形態論情報の設計・付与基準等とともに，計算機を活用してどのように作業を実施していったかということも非常に重要な問題である。そこで，CSJ の形態論情報について，主として設計に関することを説明する章と，その仕様に基づいて具体的にどのような手法を用いて形態論情報解析の作業を実施したのかについて説明する章とを設けることとし，3 章では設計について述べ，4 章では形態論情報解析の作業について，特に計算機の活用等を中心に述べることとしたのである。

3.1　長単位・短単位

3.1.1　単位の設計

3.1.1.1　語彙調査の調査単位

国語研究所は，これまでに，マスメディアにおける書き言葉や話し言葉を中心に，数度にわたり大規模な語彙調査を実施してきた。この語彙調査に当たっては，当然 語というものを規定することが必要となる。しかし，語の定義については研究者によって様々な立場があるため，語彙調査において語をどのように規定するかということは常に大きな問題となる。

国語研究所がこれまでに行ってきた語彙調査では，調査単位 (語) の設計に当たって，語とは何かという本質的な議論の上に立って調査単位を設計するという立場は取っていない。それぞれの語彙調査の目的に応じて最もふさわしい単位を設計するという方針の下に，一貫して操作主義的な立場を取ってきた[*1]。そのため，表 3.1 に示すように，複数の調査単位が使われてきた[*2]。

表 3.1　主な調査単位

	単位の名称	語彙調査名
長い単位の系列	α単位	婦人雑誌の用語
	W単位	高校教科書の語彙調査，中学校教科書の語彙調査
	長い単位	雑誌用語の変遷，テレビ放送の語彙調査
短い単位の系列	β単位	総合雑誌の用語，現代雑誌九十種の用語用字，雑誌２００万字言語調査
	M単位	高校教科書の語彙調査，中学校教科書の語彙調査

【調査単位の概略】

(1) 長い単位の系列：主として構文的な機能に着目して考えた単位で，基本的に文節に相当するようなものである。

○ α単位　文節を基にした単位。「｜小学校｜卒業｜」「｜男児用｜外出着｜」のように長い語を分割する規定を設けている。

○ 長い単位　文節に相当する単位。なおテレビ放送の語彙調査の長い単位は，複合辞を助詞・助動詞として扱っていること，人名・地名以外にも固有名詞を広く取っていることから，雑誌用語の変遷で採用した長い単位よりも長くなっている。

○ W単位　非活用語及び活用語のうち終止・連体形，命令形，中止用法・修飾用法の連用形を１W単位とする。また，それらに接続する付属語も１W単位とする。

[*1] ここで言う「操作主義的な立場」とは，「これこれこういうものを「～単位」とする，という規定をするだけで，その「～単位」が言語学的にどのようなものなのか，単語なのか，単語でないとすれば，どこが単語とちがうのか，といった問題には，まったくふれない」(国立国語研究所 1987:11) という単位設計上の立場を指す。

[*2] 単位の概略・切り方の例については，林 (1982:582-583)，中野 (1998:171-172) を基にした。

(2) 短い単位の系列：主として言語の形態的な側面に着目して考えた単位。

○ β単位　原則として，現代語において意味を持つ最小の単位 (最小単位と呼ぶ。) 二つが，文節の範囲内で 1 回結合したものを 1 単位とする。

○ M単位　β単位と同様に最小単位を基にした単位。漢語は，β単位と同様に二つの最小単位が文節の範囲内で 1 回結合したものを 1 単位とするが，和語・外来語は 1 最小単位を 1 単位とする。

【調査単位の例】

(1) 長い単位の系列

α単位：	型　紙｜どおり　に｜裁断　し　て｜外出　着　を｜作り　まし　た｜
W単位：	型　紙　どおり｜に｜裁断　し　て｜外出　着｜を｜作り　まし　た｜
長い単位 (雑誌用語の変遷)：	型　紙　どおり｜に｜裁断　し｜て｜外出　着｜を｜作り｜まし　た｜
長い単位 (テレビ放送の語彙調査)：	型　紙　どおり　に｜裁断　し　て｜外出　着　を｜作り　まし　た｜

(2) 短い単位の系列

β単位：	型　紙｜どおり｜に｜裁断｜し｜て｜外出｜着｜を｜作り｜まし｜た｜
M単位：	型｜紙｜どおり｜に｜裁断｜し｜て｜外出｜着｜を｜作り｜まし｜た｜

調査単位の設計に当たって，操作主義的な立場を取ってきたのは，「必要以上に学術的な議論に深入りして，実際上の作業がすすまないことをおそれたため」(国立国語研究所 1987:12) であり，「学者の数ほどもある「単語」の定義について，まず，意見を一致させてから，というのでは，見とおしがたたない。」(同:12) からである。

このような立場に対しては，当然のことながら「語というのは何なのか，調査のため便宜的に設けられた単位にすぎないのかという問題が残る。」(前田 1985:740) という批判がある。しかし，語とは何かという本質的な議論を積み重ねていくことは確かに重要ではあるが，国立国語研究所 (1987:12) にも，「原則的にただしい定義に達したとしても，それが現実の単位きり作業に役立たないならば，無意味である。語い調査というのは，現象の処理なのだから。」と述べられているように，語彙調査においては対象とする言語資料に現れた個々の事象を，的確に処理するということも極めて重要なのである。そして，結局のところ，これまでの語彙調査においては，この言語現象の処理ということの方をより重視してきたということなのである。

このような立場の下，各種の語彙調査を進めてきたことにより，「同じ資料の語彙調査を短単位と長単位との両方で行ってみてどのような違いが出てくるかを検討したことなどは，単位の区切り方を曖昧にしたまま「語彙調査」を行なうことに対する反省を促す」(前田 1985:740) など，国語の計量的な研究を進める上で先駆的な役割を果たしてきたと言うことができる。国語研究所の語彙調査における調査単位の設計方針には，批判もあるが，それにより現実の言語事象を的確に処理してきたことは，十分に意味があったと言える。

3.1.1.2 CSJ の単位

CSJ の単位の設計に当たっては，語彙調査と同様に，まず目的を設定した上で，その目的に適した単位を設計することとした。このような立場を取ったのは，語とは何かという本質的な議論の重要性はもちろん認めるところではあるが，時間的な制約等を考えた場合，CSJ に現れた言語事象を的確に処理できる単位を設計する

ことの方が，まずは重要であると考えたからである。また，そのようにして大規模な話し言葉データを処理した結果をまとめておくことは，今後，言語単位論を進める上での基礎的な資料になるとも考えたのである。

単位の設計に当たり語彙調査と同様に目的を設定するとしたが，この場合の目的というのは，我々が CSJ を使ってどのような国語研究を行うのかということである。CSJ を利用した国語研究として，我々は次の 2 点を掲げた。

(1) CSJ から用例を採集し，話し言葉の語彙・語法の研究を行う。
(2) 品詞の分布などの計量研究によって CSJ の言語的な特徴を明らかにする。

もちろん，CSJ を使った研究はこの二つに限られるものではない。しかし様々な研究を想定し，それらすべてに適した単位を設計することは不可能に近い。そこで，我々はひとまず上記の二つの目的に絞って，それに適した単位を設計することとした。以下，それぞれの目的にふさわしい単位について考えていくこととする。

まず，目的 (1) のためには，合成語を語構成要素に分割したような短い単位が求められる。表 3.1 に示した単位で言えば，「短い単位の系列」に属する単位が望ましいということになる。しかし，語構成要素に分割すると言っても，語構成要素をすべて切り出してしまうような単位では，取り出した単位の意味が文脈から離れすぎてしまうこともあり，結果的に不要な用例まで検索してしまうという問題がある。

例えば，「気持ち」という語を例に考えてみよう。「気持ち」を語構成要素に分割すると，「気」と「持ち」の二つの要素に分割できる。しかし，「気持ち」はこれ全体で「心の在り方」などという意味を表しているものであり，「気」と「持ち」とに分割して，「持ち」を取り出しても，その「持ち」には動詞「持つ」が本来持っている「手の中に入れて保つ」などという意味は認め難い。そのため，「気持ち」の「持ち」を一つの単位として切り出して，「荷物を持つ」という例のような，実質的な意味を持つ動詞「持つ」と同様に扱い，見出し等の情報を付与しても，付与した情報と実際に文脈の中で使われている意味との間に大きなずれが生じることになる。また，動詞「持つ」を検索した結果に，「気持ち」の語構成要素として用いられた「持ち」が含まれることは望ましいこととは言えない。つまり，(1) の目的のために短い単位が求められるとは言っても，語構成要素にすべて分割してしまうような単位では問題があるということになる。

次に，目的 (2) のためには，CSJ の資料的な性格を反映するような単位であることが求められる。一般に単位を短くすればするほど，取り出した単位はいわゆる基本的な語となる。その反対に，より長い単位とすれば，当該資料の性格を反映するような特徴語などを取り出せるようになる。したがって，表 3.1 で言えば，「長い単位の系列」に属する単位が適当ということになる。

このことについて，「言語」という語を例に少し説明しておく。「言語」は，CSJ に収録された幾つかの学会講演に用例が見られるが，その用いられ方 — 特にどのような語と結合するか — については，学会によって差異が見られる。例えば，音声関係の工学系学会 (A 学会) と国語関係の人文系学会 (B 学会) での「言語」の例を比較してみよう[*3]。

A 学会・B 学会ともに，「言語」が単独で用いられた例のほか，以下のように合成語の語構成要素として用いられた例がある。

[*3] ここで「言語」の用例を採集するために用いたデータは，人手解析データである。

【A学会】 音声言語 音声言語概念 各言語 各言語モデル 各種言語モデル 確率的言語モデル 言語重み 言語音 言語音カテゴリー判断 言語音モード 言語音声 言語解析 言語学的 言語カテゴリー 言語間 言語形成期 言語圏 言語刺激 言語習得時 言語情報 言語情報処理 言語条件 言語スコア 言語制約 言語生活 言語的 言語的規則 言語的情報 言語伝達 言語特有 言語非依存 言語非言語刺激 言語モデル 第二言語学習者 聴覚運動性言語野 聴覚性言語野 聴覚的言語判断 統計的言語情報 特異性言語障害者 パラ言語情報 パラ言語的 パラ言語的意味 非言語 非言語音 非言語音モード 非言語刺激 非言語情報 文字言語 融合言語モデル

【B学会】 一言語体系 音声言語 音声言語重視 各言語 簡易言語 基本的言語単位 言語外 言語学 言語研究者 言語現象 言語作品 言語社会 言語習得 言語政策的 言語体系 言語的研究 言語内 言語表現 西洋言語学 第二言語習得 第二言語習得者 他言語 比較言語学

ここで注意したいのは，A学会で下線を付した語 (「言語音声概念」「言語刺激」「言語モデル」など) はB学会には用いられておらず，B学会で下線を付した語 (「一言語体系」「言語作品」「言語表現」など) はA学会には用いられていないということである。つまり「言語音声概念」「言語刺激」「言語モデル」などはA学会を特徴付ける語であり，「一言語体系」「言語作品」「言語表現」はB学会を特徴付ける語であると言うことができる。このような各分野に特徴的な語を把握するためには，「言語モデル」を「言語」と「モデル」とに，「言語作品」を「言語」と「作品」とに分割するのではなく，全体で一つとして扱うような単位が必要となる。

なお，(1)(2) いずれの目的のためにも，不統一のない単位とすることが必要である。同じ種類の単語が異なる分割のされ方をしていては，効率的な検索ができない。また計量的な研究では，計量される対象である単位が等質であることが求められるので，不統一のない単位にすることが重要である。

以上のことを踏まえて，CSJ の単位について検討した結果，次のような結論を得た。

まず，二つの目的を掲げたが，目的 (1) については「短い単位の系列」に属する単位，目的 (2) については「長い単位の系列」に属する単位というように，それぞれの目的にとって望ましい単位が異なっている。そこで，CSJ では単位を一つに限ることはせず，長短 2 種類の単位を採用することとした。また，今回は新たに単位を設計するのではなく，表 3.1 に挙げた語彙調査の調査単位の中から，それぞれの目的に適した単位を採用し，必要に応じて拡張等を行うこととした。

その結果，長い単位 (以下，長単位と呼ぶ。) については，テレビ放送の語彙調査で採用された長い単位を基にして設計を行うこととした[*4]。一方，短い単位 (以下，短単位と呼ぶ。) については，現代雑誌九十種の用語用字で用いられた β 単位を採用し，必要に応じて話し言葉の処理用に拡張することとした[*5]。

[*4] 長い単位については，国立国語研究所 (1995:49-63) を参照。

[*5] β 単位については，国立国語研究所 (1962:6-14) を参照。

3.1.2　長単位の認定基準

以下，本節で長単位の認定基準，次の 3.1.3 節で短単位の認定基準について説明することとする。その際単位等の境界を示すために，次の記号を用いる。

文節の境界	…… ‖	例：‖国立国語研究所の‖
長単位の境界	…… ┃	例：┃国立国語研究所┃の┃
短単位の境界	…… ｜	例：｜国立｜国語｜研究｜所｜の｜
最小単位の境界	…… ／	例：／国／立／国／語／研／究／所／の／

※注目している単位が分かりにくい場合は，その単位に下線を施すことがある。また，切らないことを示す場合には「=」(例：西が=丘) を用いる。

長単位・短単位の認定は，まず転記テキストの基本形を対象として行い，その後，基本形の単位認定結果を基に自動で発音形の単位認定を行った。そのため，以下に述べる単位認定基準は，転記テキストの基本形を例にしたものとなっている。転記テキストの発音形を対象とした単位認定については，4 章を参照されたい。

以下，長単位の認定基準を説明するが，長単位は文節を基にした単位であり，認定に当たっては，文節の認定を行った上で，各文節の内部を規則に従って自立語部分と付属語部分とに分割していくという手順を踏む。そのため，長単位の認定基準は，文節認定基準と長単位認定基準の二つの基準から成る。

3.1.2.1　文節認定基準

長単位の認定に当たっては，まず文節の認定を行う。この文節は，テレビ放送の語彙調査で用いられた長い単位を基にしたものである。

テレビ放送の語彙調査と同様に，付属語には複合辞も含めた。複合辞は，現代語の研究や日本語教育ではよく取り上げられるものである。国立国語研究所 (2001) では，助詞的複合辞 83 語，助動詞的複合辞 42 語を取り上げ，用例を示すとともに解説を加えている。このように現代語の研究等では，多くの複合辞が認定されているところではあるが，CSJ では，それらすべてを複合辞として認定することはしなかった。これは複合辞の認定には意味の問題が絡んでくるため，その認定自体が極めて難しいということによる。CSJ では，付録 3.1・付録 3.2 に挙げた助詞相当句・助動詞相当句のみを複合辞として認定した。なお，今回複合辞として認定したものの範囲は，テレビ放送の語彙調査と全く同じというわけではない。また，複合辞については敬語形式のもの (「について」に対する「につきまして」など) をどのように扱うかが問題となる。これについては，CSJ では敬語形式になっているものも複合辞として認定することとした。

以下，文節を認定する上で，問題となる点について簡単に触れておく。

文節は，一般に付属語又は付属語連続の後で切れるが，以下のように，固有名・動植物名・「—の～」「—が～」の体言句・分数の読み上げの内部にある助詞・助動詞の後では切らないこととする。

固有名： 西が=丘　国立少年自然の=家　蛤御門の=変
動植物名： タツノ=オトシゴ　サキシマスオウノ=キ
「—の〜」の体言句のうち，以下に挙げるもの： 麻の=葉　味の=素　ありの=まま　絵の=具　男の=子
思いの=丈　思いの=外　女の=子　髪の=毛　上の=句　気の=毒　木の=芽　木の=下
下の=句　茶の=間　念の=為　日の=出　目の=当たり　身の=上　身の=程　身の=回り
目の=敵　山の=手　世の=中
「—が〜」の体言句のうち，以下に挙げるもの： 万が=一
分数の読み上げ： 三分の=二　後続単語種類数分の=先行単語頻度

また，次のように，2 文節以上から成る形式全体を受ける，若しくはそれに係る接辞及び体言的な形式は，その前後で切ることとする。この規則によって，以下の例のように「等」「型」「各」のような接辞が単独で一つの文節を構成する場合もあり得るということになる。

‖円形劇場とか‖水路‖ 等 ‖　‖への‖字‖ 型 ‖　‖ 各 ‖日本語の‖文章‖

なお，ここで述べた文節は，長単位の認定を行うために規定するものであり，長単位の認定のために必要な概念として持っておくという性質のものである。したがって，この文節の境界は CSJ のデータには示されていない。また，転記テキストにおける改行基準としての文節とは細部において一致しないところがある。転記テキストにおける文節の詳細については，2.8 節を参照されたい。

3.1.2.2 長単位認定基準

長単位は，以下に示した規則によって文節（3.1.2.1 節の文節認定基準によって認定されたもの。）を分割し，それによって得られたものを 1 単位とするような単位である。以下，長単位の認定基準を示す。

［1］ 付属語 (付録 3.1・付録 3.2 に示した複合辞を含む。) は 1 長単位とする。

|今|は|ファックス|とか|そう|いう|の|が|ある|んです|けれども|

(1) 形容動詞及び形容動詞活用型の助動詞 (そうだ・みたいだ・ようだ) の活用語尾は助動詞として扱い，1 長単位とする。

|統一的| な |視点|で|切り|ましょ|う|　|涙|が|出|そう| に |なる|
|エンジニア|な|んだ|そう| です |　|駅員さん|が|いる|みたい| だ |
|使える|よう| に |し|たい|

(2) 文節の認定の際に一続きとして扱うこととした固有名・動植物名・「—の〜」「—が〜」の体言句・分数の読み上げの内部にある助詞・助動詞は切り出さない。

|西=が=丘|　|サキシマスオウ=ノ=キ|　|絵=の=具|　|万=が=一|
|三分=の=二|　|後続単語種類数分=の=先行単語頻度|

［2］ 並列及び同格の関係にある語は互いに切り離す。

| 安心 | 確実 | な | 方法 |　　| 塩 | こしょう | を | かける |　　| 機関誌 | 計量国語学 |

並列及び同格の関係にある体言連続のうち，並列された体言全体に係る体言・接辞がある場合は切らない。また並列された体言全体を受ける体言・接辞・形式的な意味の「する」「できる」「なさる」「いたす」がある場合は切らない。

| 平成=九年=十年 |　　| 関東=東北=地方 |　　| 機関誌=計量国語学=発行 |
| 観察=整理=する |

［3］ 体言（合成語）の一部分が連体修飾語を受けている場合，その部分の後で切る。

| 項構造 | の | 曖昧性 | 解消 |

「以降」「間 (かん)」「ごと」「自体」「達」が付いた場合は切らない。

| 文章 | の | 途中=以降 |　　| 住ん | でる | 人=達 |

［4］ 体言及び副詞に形式的な意味の「いたす」「する」「できる」「なさる」が直接続く場合，体言及び副詞と「いたす」「する」「できる」「なさる」との間は切らない。

| 許容=する |　　| 演出=できる |　　| 体験=なさる |　　| きらきら=する |　　| きちんと=する |

ただし，前の体言が連体修飾を受けている場合は用言部分を切り離す。

| 面白い | 説明 | する | 人 |

［5］ 「お (ご) ＋動詞連用形 (名詞) ＋する・くださる・いただく・なさる・いたす・ねがう・もうしあげる・あそばす」は全体で一続きとする。

| お=会い=する |　　| お=与え=ください |　　| お=電話=なさる |　　| 御=登場=願う |

［6］ 数量を表す要素を含む自立語は，以下のように処理する。

(1) 前の要素に関する順序・番号を直後の要素が表している場合，両者を切り離さない。

| 昭和十三年=八月=八日 |　　| 朝=八時 |　　| 予稿集=八十七ページ |　　| 入所=二十年目 |

(2) 上記の規則に該当しない場合，数量を表す要素とその直前の要素とを切り離す。

| 果汁 | 百パーセント |　　| バニラエッセンス | 少々 |　　| 山の手線 | 京浜東北線 | 二本 |
| 一箱 | 三万 |　　| 週 | 二通 |　　| 一学年 | 上 |　　| 十年以上 | 前 |　　| 延べ | 百二十九文 |

ただし，数量を表す要素が前で列挙された要素の個数を表しているものについては，数量を表す要素と前の要素とを受ける体言がある場合，切り離さない。

| 果汁=百パーセント=オレンジジュース |

3.1.3 短単位の認定基準

短単位は，言語の形態的側面に着目して規定した単位である。認定に当たっては，現代語において意味を持つ最小の単位（最小単位と呼ぶ。）を規定した上で，その最小単位を，一つの長単位の範囲内で，短単位認定基準に定めた条件を満たす形で結合させていく（又は結合させない（これを 0 回結合と考える。））という手順を踏む。そのため，短単位の認定基準は，最小単位認定基準と短単位認定基準の二つの基準から成る。

3.1.3.1 最小単位認定基準

短単位の認定に当たっては，まず最小単位というものを認定する。最小単位は，現代語において意味を持つ最小の単位であり，和語・漢語・外来語・記号・人名・地名の種類ごとに次のように認定される。

和　語：　／話し／言葉／　　／豊か／な／暮らし／
　　／お／話し／し／ます／　　／雨／が／降る／みたい／だ／
　　／大／雨／が／降っ／た／の／で／
漢　語：　／国／語／　　／研／究／所／
外来語：　／データー／ベース／　　／ネット／ワーク／
記　号：　／図／Ａ／　　／ＮＨＫ／
人　名：　／星野／仙一／　　／ジェフ・／ウィリアムス／
　　※ 姓と名がそれぞれ 1 最小単位。
地　名：　／大阪／府／豊中／市／待兼山町／　　／六甲／山／　　／神崎／川／
　　※ 地形名の名を表す部分は 1 最小単位。

「豊かだ」などのいわゆる形容動詞，「みたいだ」「そうだ」「ようだ」の形容動詞活用型の助動詞については，その活用語尾を「／豊か／だ／」「／みたい／だ／」「／そう／だ／」「／よう／だ／」のように 1 最小単位として分割する。

「だが」「では」などの助詞・助動詞から転化した接続詞も「／だ／が／」「／で／は／」のように分割する。また接続助詞「ので」や副助詞「とか」のような複数の助詞・助動詞が結合してできた助詞についても，「／の／で／」「／と／か／」のように最小単位を認定する。

「ていく」「について」などの複合辞は，長単位においては一つの付属語として扱ったが (3.1.2.2 節参照)，最小単位においては「／て／いく／」「／に／つい／て／」のように分割する。

以上のように認定した最小単位について，短単位を認定する必要上，表 3.2 のように分類する。

以下，「付属要素」「数」「助詞・助動詞」について説明しておく。

「付属要素」とは，接頭辞・接尾辞のことである。ただしすべての接頭辞・接尾辞が付属要素として扱われるわけではない。CSJ に出現したものの中から，造語力が高いなど特に注目されるものを「付属要素一覧」という一覧表に挙げ，その一覧表に挙げられたもののみを付属要素として扱うこととした。この「付属要素一覧」を，付録 3.3・付録 3.4 として示した。

「数」には，一・十・百・千などの数詞のほか，「数十」「何百」「幾千」の「数」「何」「幾」なども含めた。ま

表 3.2　最小単位の分類

分　類	例
一　　般	和　語　：　山　川　白い　話す　言葉　……
	漢　語　：　社　会　用　研　究　所　……
	外来語　：　オレンジ　ボックス　アルゴリズム　……
付 属 要 素	接頭的要素　：　相　お　各　御（ご）　……
	接尾的要素　：　合う　致す　っぽい　性　的　……
記　　号	Ａ　Ｂ　ω　イ　ロ　ア　ＮＨＫ　ＪＲ　……
数	一　二　十　百　千　……
	幾　数　何　……
人名・地名	星野　仙一　大阪　六甲　……
助詞・助動詞	う　た　です　ます　か　から　て　も　……

た数詞のうち，数え進むことができないと考えられるもの (例えば「一応」の「一」や「百科」の「百」など) については，「一般」に分類した。

「助詞・助動詞」には，いわゆる形容動詞及び形容動詞活用型の助動詞 (そうだ・みたいだ・ようだ) の活用語尾も含めた。また，「だが」「では」などの助詞・助動詞から転化した接続詞は，先に示したように「／だ／が／」「／で／は／」と最小単位が認定されることから，その「だ」「が」「で」「は」はそれぞれ「助詞・助動詞」に分類した。

なお，ここで述べた最小単位は，短単位の認定を行うために規定するものであり，短単位の認定のために必要な概念として持っておくという性質のものである。したがって，この最小単位の境界は CSJ のデータには示されていない。

3.1.3.2　短単位認定基準

短単位の認定基準は，表 3.2 の分類ごとに適用すべき規則が定められている。この規則に基づいて，最小単位を結合させていく（又は結合させない）ことによって，短単位が認定されるのである。なお，最小単位の結合に当たっては，長単位境界を超えて結合させることはしないという制約を設けている。これによって，1 文節の中には 1 個以上の長単位があり，1 長単位の中には 1 個以上の短単位があるというように，CSJ の言語単位が階層的な構造を持つことになる。

以下，短単位の認定基準を示していくが，その規則のうち，短単位認定の基本原則に当たるのが，「一般」に分類した最小単位に適用される以下の規則［1］である。

［1］　「一般」に分類した最小単位 2 個の 1 次結合は 1 短単位とする。

｜母親｜　　｜食べ歩く｜　　｜音声｜　　｜レーザープリンター｜　　｜無口｜　　｜オレンジ色｜

「一般」に分類した最小単位であっても，それ単独で 1 短単位になるものや 3 最小単位以上の結合であっても全体で 1 短単位とするものがある。それを以下に示す。

［2］ 1 最小単位を 1 短単位とするもの。

(1) 最小単位が三つ以上並列した場合の各最小単位。

｜衣｜食｜住｜　　｜松｜竹｜梅｜　　｜都｜道｜府｜県｜

(2) 重複形の擬音語・擬態語で，重複が奇数回の場合の，その重複されている要素。

｜さく｜さく｜さく｜と｜混ぜ｜　　｜ちょこ｜ちょこ｜ちょこ｜動く｜

なお，偶数回の繰り返しの場合は規則［1］を適用する。

｜さくさく｜と｜混ぜる｜　　｜さくさく｜さくさく｜と｜混ぜる｜

(3) 類概念を表す部分と名を表す部分とが結合してできた固有名詞のうち，類概念を表す部分と名を表す部分とが共に 1 最小単位の場合の，それぞれの部分。

｜さくら｜屋｜　　｜リクルート｜社｜　　｜ハーバード｜大｜　　｜のぞみ｜号｜　　｜キリスト｜教｜
｜タイムズ｜紙｜　　｜キャノン｜カメラ｜

ただし，名を表す部分が 1 字の漢語で，類概念を表す部分が 1 最小単位である場合は，その 1 次結合体を 1 短単位とする。

｜仏教｜　　｜儒教｜　　｜阪大｜

(4) 外来語の最小単位のうち英語の接続詞・前置詞・冠詞に当たるもの。

｜アウト・｜オブ・｜ドメイン｜　　｜ショアーズ・｜アット・｜ワイコロア｜
｜基本｜レフト・｜トゥー・｜ライト｜構造｜　　｜コール・｜フォー・｜ペーパー｜

(5) 外来語の最小単位 2 個の 1 次結合体が 11 拍以上になる場合の各最小単位。

｜インサーション｜ペナルティー｜　　｜スペクトル｜パラメーター｜

(6) 外国語（転記テキストにおいてタグ (O) を付与されたもの。）。

｜アイ｜ノウ｜ザット｜シープ｜キャン｜スイム｜

(7) 規則［1］，［2］の (1)〜(6)，［3］，［4］，［5］によって得られた短単位に，前又は後ろから結合した最小単位。

｜内閣｜府｜　　｜副｜大統領｜　　｜光ファイバー｜網｜

(8) 単独で文節を構成する最小単位。

｜やっぱり｜これ｜も｜一｜つ｜の｜　　｜オレンジ｜を｜食べる｜

［3］ 3 最小単位以上の結合であっても全体で 1 短単位とするもの。

(1) 三つ以上の最小単位からなる組織名等の略称。

｜日経連｜　　｜通総研｜

(2) 切る位置が明確でないもの，あるいは切った場合と一まとめにした場合とで意味にずれがあるもの。

｜大統領｜　　｜不可解｜　　｜明後日｜　　｜殺風景｜　　｜輸出入｜　　｜国内外｜　　｜町村長｜
｜原水爆｜　　｜市町村長｜　　｜大袈裟｜　　｜大雑把｜　　｜大丈夫｜　　｜一辺倒｜　　｜十文字｜
｜二枚目｜　　｜十八番｜

(3) 文節の認定の際に一続きとして扱うこととした「—の～」「—が～」の体言句（3.1.2.1 節参照）。

「—の～」の体言句　：　｜麻の葉｜　　｜味の素｜　　｜絵の具｜　など
「—が～」の体言句　：　｜万が一｜

以下，「一般」以外の最小単位に適用する規則を示す。

［4］ 「記号」「人名・地名」「付属要素」「助詞・助動詞」は，1 最小単位を 1 短単位とする。

記　　号　　：　｜図｜A｜　　｜NHK｜
人　　名　　：　｜星野｜仙一｜
地　　名　　：　｜大阪｜府｜豊中｜市｜待兼山町｜　　｜六甲｜山｜　　｜神崎｜川｜
付属要素　　：　｜お｜母｜さん｜　　｜見｜にくい｜
助詞・助動詞　：　｜単位｜に｜切り｜ましょ｜う｜　　｜それ｜に｜つい｜て｜
　　｜とても｜きれい｜だ｜

［5］ 「数」は，「数」以外の最小単位と結合させない。「数」どうしの結合については，結合の次数にかかわらず，一・十・百・千のとなえを取るけたごとに 1 短単位とする。「万」「億」「兆」などの最小単位は，それだけで 1 短単位とする。小数部分は 1 最小単位を 1 短単位とする。

｜十｜二｜月｜二十｜三｜日｜　　｜七百｜万｜語｜　　｜五｜分｜の｜二｜　　｜何十｜倍｜
｜一二｜年｜前｜　　｜二三十｜回｜　　｜零｜.｜四｜五｜

3.1.4　話し言葉特有の現象の単位認定

話し言葉には，書き言葉にはない様々な現象が見られる。このうち，単位認定の際に問題となる現象として，次に挙げるような現象がある。

(1) 融合：二つ以上の単位が音の転訛等によって，一つになったもの。

*括弧内に示したのは融合する前の形。
そりゃ（それ＋は）　面白きゃ（面白けれ＋ば）　食べりゃ（食べれ＋ば）　じゃ（で＋は）
てる（て＋いる）

(2) 省略：前接又は後接の単位の影響によって，単位の一部が略されたもの。

> *括弧内に示したのは元の形。下線部が省略された部分。
> やんだっけ（や<u>る</u>んだっけ）　　そうっす（そう<u>で</u>す）

(3) フィラー：転記テキストでタグ (F) を付けられたもの。

> (F えー)　　(F あのね)　　(F んーと)

(4) 語断片：言い直し等に伴って語が断片化したもの。転記テキストでは，タグ (D) が付けられている。

> (D す) すると　　(D テニ) 昨日のテニスは　　(D 情) 情報が

(5) 助詞・助動詞等にかかわる言い直し：転記テキストでタグ (D2) が付けられたもの。

> 評価値 (D2 が) の数値が　　桜 (D2 です)(D か) ですから　　(D2 第) 第一関門を　　懐疑 (D2 的) 的な
> (A 一．(D2 五) 五六;1． 5 6) キロ

(6) 言い直し：上記 (4)(5) 以外のもので，転記テキストでタグが付けられていないもの。

> *下線部が言い直し。
> 国立 <u>日本語</u> 国語研究所

以下，(1)〜(6) に示した話し言葉特有の現象の単位認定について説明する。

【融合】

融合を処理する方法としては，まず元の語形に戻した上で，単位認定するという方法がある。例えば，「面白きゃ」を「面白ければ」，「じゃ」を「では」に戻した上で単位認定するというものである。この方法は，過去の国語研究所の語彙調査で取られたものでもある[*6]。このような処理は，基礎語の選定等を目的とした語彙調査においては，妥当なものと言えよう。しかし，話し言葉コーパスにおける処理方法としては，話し言葉の特徴である融合という現象を分からなくするという点で問題がある。また CSJ では融合現象が多く見られることが予想されるため，すべて元の形に戻していたのでは，作業が煩雑になるという問題もある。そこで，CSJ では，融合を元の形に戻さずに単位認定を行う。例えば，「面白きゃ」「じゃ」「てる」は，長単位・短単位ともにそれぞれ 1 単位となる。

【省略】

省略についても，元の形に戻すことなく，可能な範囲で単位を認定する。例えば，「やんだっけ」は，「や」を「やる」の活用語尾が省略された形，「ん」を準体助詞「の」の撥音化したものと考え，長単位では「┃や┃んだ┃っけ┃」，短単位では「｜や｜ん｜だ｜っけ｜」と分割する。

[*6] 国立国語研究所 (1962:7) を参照。

【フィラー】

フィラーについては，長単位・短単位ともに，以下に示した出現形を 1 単位とする。

あ (ー)，い (ー)，う (ー)，え (ー)，お (ー)，ん (ー)，と (ー)，ま (ー)，
う (ー) ん，あ (ー) (ん) (ー) の (ー)，そ (ー) (ん) (ー) の (ー)，
う (ー) ん (ー) (っ) と (ー)，あ (ー) (っ) と (ー)，え (ー) (っ) と (ー)，ん (ー) (っ) と (ー)

実際に単位認定を行った例を示すと，以下のようになる。

| (F えー) |　　| (F んーと) |　　※短単位も長単位と同様の単位認定となる。

上記の要素に助詞・助動詞が結合した場合，長単位では助詞・助動詞も含めて 1 単位とするが，短単位では，助詞・助動詞を切り出す。

長単位　：　| (F あのですね) |　　| (F あのね) |
短単位　：　| (F あの | です | ね) |　　| (F あの | ね) |

またフィラーが，単位の中に現れる場合がある。例えば，以下のような例である。

味わうことが (F えー) できま (F えー) せん　　ここでもメタ (F あ) 言語行動表現てものを手掛かりに

「ま (F えー) せ」は，長単位・短単位いずれにおいても 1 単位となる助動詞「ます」の未然形の中にフィラーが現れたものであり，「メタ (F あ) 言語行動表現」は 1 長単位となる「メタ言語表現行動」の中にフィラーが現れたものである。このような例について，テレビ放送の語彙調査の長い単位では，「単位の中に割り込んでいる要素は，その単位には含めない。適宜，位置を変える」(国立国語研究所 1995：61) という規則を設け，

あの | 百 ‖ ま ‖ 六十度以上にね　　→　あの ‖ ま ‖ 百六十度以上にね
強攻策に | 出た | 現 ‖ えー ‖ 執行部に対する　→　えー ‖ 現執行部に対する

というような形で単位認定している。つまり，単位認定がしやすいように，音声を書き起こしたテキストにおいてフィラーなどの位置を適宜変えるのである。このような方法は，切り出した単位を最終的に実際の音声とは切り離し，語彙表という形にまとめる語彙調査だからこそできるものと言えよう。実際の音声との対応関係が保たれている CSJ のような音声言語コーパスでは，出現した単位語のテキスト上の位置を変えるということは適当な処理とは言い難い。

そこで，CSJ では，長単位・短単位いずれにおいても，フィラーを無視して単位認定を行うこととした。先に挙げた二つの例は，次のように単位が認定されることになる。

| 味わう | こと | が | (F えー) | でき | ま (F えー) せ | ん |
※短単位も長単位と同様の単位認定となる。

| ここ | で | も | メタ (F あ) 言語行動表現 | て | もの | を | 手掛かり | に |
※短単位については，「メタ言語表現行動」が「 | メタ | 言語 | 表現 | 行動 | 」と分割されるので，ここで問題としている 1 単位の中に現れるフィラーには当たらない。短単位では，「 | メタ | (F あ) | 言語 | 行動 | 表現 | 」と分割される。

つまり，「ま (F えー) せ」については，助動詞「ます」の中に現れたフィラーを 1 長単位・1 短単位と認めることをせず，「┃ま (F えー) せ┃」全体で 1 長単位・1 短単位として認定するということである。別の言い方をすると，「ま (F えー) せ」全体を助動詞「ます」の未然形の異形態と考えて処理するということでもある。

「メタ (F あ) 言語行動表現」についても同様で，「メタ言語行動表現」の中に現れたフィラーを 1 長単位と認めることをせず，「メタ (F あ) 言語行動表現」を「メタ言語行動表現」の異形態として扱うこととし，「┃メタ (F あ) 言語行動表現┃」全体で 1 長単位として認定する。

なお，上記の (F えー) や (F あ) のように，一つの単位として認定されなかったフィラーには，品詞等の情報は付与しない（3.3.5 節を参照）。そのため，品詞情報を基にフィラーを数えても，CSJ に出現したフィラーの総数とは一致しないことになる。この点，CSJ の形態論情報を利用する際に注意する必要がある。

【語断片】

語断片は，単独で 1 長単位又は 1 短単位とする。例えば，次のように単位認定される。

┃ (D す) ┃する┃と┃　　┃ (D テニ) ┃昨日┃の┃テニス┃は┃　　┃ (D 情) ┃情報┃が┃
※短単位も長単位と同様の単位認定となる。

1 長単位又は 1 短単位の中に現れる語断片については，フィラーと同様に無視して単位認定を行う。

┃それ┃と┃ポライト┃ノン (D プロ) ポライト ┃と┃いう┃
※短単位も長単位と同様の単位認定となる。
┃それ┃を┃ 利用 (D す) する ┃の┃も┃
※短単位については，「利用する」が「｜利用｜する｜」と分割されるため，ここで問題としている 1 単位の中に現れる語断片には当たらない。短単位では，「｜利用｜ (D す) ｜する｜」と分割される。

つまり，「ノン (D プロ) ポライト」については，「ノンポライト」の中に現れた語断片を 1 長単位・1 短単位と認めることをせず，「┃ノン (D プロ) ポライト┃」全体で 1 長単位・1 短単位として認定するということである。別の言い方をすると，「ノン (D プロ) ポライト」全体を「ノンポライト」の異形態として処理するということでもある。「利用 (D す) する」についても同様の考え方によって単位を認定する。

なお，先に述べたフィラーと同様に，上記の (D プロ) や (D す) のように，一つの単位として認定されなかった語断片には，品詞等の情報は付与しない（3.3.5 節を参照）。そのため，長単位・短単位で語断片として認定されたものだけを数えても，CSJ に出現した語断片の総数とは一致しないことになる。この点についても，CSJ の形態論情報を利用する際に注意する必要がある。

【助詞・助動詞等にかかわる言い直し】

数詞・助詞・助動詞・接頭辞・接尾辞にかかわる言い直し（タグ (D2) が付けられたもの）については，通常の数詞・助詞・助動詞・接頭辞・接尾辞と同様に単位認定を行う。

長単位　：　┃実験三┃ (D2 の) ┃として┃は┃　　┃国内┃ (D2 で ┃ も) ┃の┃選手┃も┃
　　┃ (D2 未) ┃未観測┃だっ┃た┃
短単位　：　｜六十｜ (D2 二) ｜二｜パーセント｜の｜　　｜実験｜三｜ (D2 の) ｜と｜し｜て｜は｜
　　｜国内｜ (D2 で ｜ も) ｜の｜選手｜も｜　　｜ (D2 未) ｜未｜観測｜だっ｜た｜

また，長単位については，1 単位内にタグ (D2) を付けられた要素が出現する場合がある。

六十 (D2 二) 二パーセントの 懐疑 (D2 的) 的な

このような 1 長単位内に出現したものについては，1 単位内に現れたフィラーや語断片と同様に，タグ (D2) を付けられたものを無視して単位認定を行う。

|六十=(D2 二)=二パーセント|の| |懐疑 (D2 的) 的|な|

つまり，「六十 (D2 二) 二パーセント」については，「六十二パーセント」の中に現れた (D2 二) を 1 長単位と認めず，「|六十 (D2 二) 二パーセント|」全体で 1 長単位として認定するということである。「懐疑 (D2 的) 的な」についても同様である。

また上記の (D2 二) や (D2 的) のように，一つの単位として認定されなかったものに対する情報付与及びそれに関連する注意事項は，1 長単位・1 短単位の中に現れた語断片におけるそれと同様である。

なお，短単位については，助詞・助動詞・接頭辞・接尾辞が，ほかの最小単位と結合することなく，常に単独で 1 短単位となることから，短単位の内部に，助詞・助動詞等に関する言い直しが現れることはない。数詞については，タグ (D2) が短単位又はその連鎖に対して付与されるため（2.5.3 節参照），短単位内に数詞の言い直しが現れることはない*7。

【言い直し】

ここで取り上げる言い直しは，既に述べた語断片や助詞・助動詞にかかわる言い直しとは異なり，以下の例（下線部）のように転記テキストにおいてタグを付けられていないものである。

益岡・田窪氏の 基本日本語基礎日本語文法(D2 の) での

このような言い直しについては，短単位では，

|益岡・|田窪|氏|の|基本|日本|語|基礎|日本|語|文法| (D2 の) |で|の|

と単位が認定されるため，単位認定の際に問題となることはない。しかし長単位については，言い掛け部である「基本日本語」を一つの長単位（合成語）として認めるかどうかなど，言い掛け部・訂正部に関連する箇所の扱いについて種々の問題がある。

CSJ では，上記のような言い直しを，まず次に示すように四つに分類した上で，長単位の認定基準を作成した。

*7 「千九百九十七」と言おうとして「千 九，九百九十七」と言った場合，下線部を言い掛け部として扱う。具体的には，「九」を「九百」という短単位の語断片として扱い，転記では「(A 千 (D 九) 九百九十七;1 9 9 7)」のように，タグ (D2) ではなくタグ (D) を付ける。短単位では「(D 九)」が単独で 1 単位となる。

このようなタグの付与基準・短単位の認定基準から，数詞の言い直し（タグ (D2) を付けられたもの）が短単位内に現れることはない。

［1］ 分類

(1) 語の一部を述べたところで，語全体を言い直している場合。

益岡・田窪氏の 基本日本語基礎日本語文法

(2) 前に述べた語の一部のみを直後で言い直している場合。

阪倉篤義さん篤義先生 の　　国語 についてつきまして

(3) 前に述べた語全体を言い直している場合。

向こうで 教育機関教育事業 始めたいということで

(4) 1 長単位の内部に言い直しがある場合。

国立日本語国語研究所 で

［2］ 長単位認定基準

(1)〜(3) については，まず言い掛け部と訂正部との間に長単位の境界を設け，その境界の前後の要素に長単位認定基準（3.1.2.2 節参照）を適用して，長単位を認定する。

(1) 語の一部を述べたところで，語全体を言い直している場合。

┃益岡・田窪氏┃の┃ 基本日本語 ┃ 基礎日本語文法 ┃
※長単位認定基準には，言い掛け部「基本日本語」，訂正部「基礎日本語文法」を更に短く分割するような規定はない。そのため，言い掛け部・訂正部がそれぞれ 1 長単位となる。

(2) 前に述べた語の一部のみを直後で言い直している場合。

┃ 阪倉篤義さん ┃ 篤義先生 ┃の┃　　┃国語┃ について ┃ つき ┃ まし ┃ て ┃
※「についてつきまして」のうち，言い掛け部「について」は，長単位で複合辞として扱われるため全体で 1 長単位となる。一方「つきまして」は，この形では複合辞として認められないため，付属語「まし」「て」を分割して「つき｜まし｜て｜」のように 3 長単位とする。
　なお，「についてにつきまして」という言い直しであれば，「┃について┃につきまして┃」のように長単位を認定する（以下の (3) に該当する。）。

(3) 前に述べた語全体を言い直している場合。

┃向こう┃で┃ 教育機関 ┃ 教育事業 ┃始め┃たい┃という┃こと┃で┃
※長単位認定基準には，言い掛け部「教育機関」，訂正部「教育事業」を更に短く分割するような規定はない。そのため，言い掛け部・訂正部がそれぞれ 1 長単位となる。

(4) については，言い掛け部と訂正部との間に長単位境界を設けることはしない。1 単位中に出現した語断片と同様，言い掛け部を 1 長単位として切り出さずに，無視して単位認定を行う。

(4) 1 長単位の内部に言い直しがある場合。

> ┃国立=日本語=国語研究所┃で┃
> ※下線を施した訂正部「日本語」を 1 長単位として認定することはしない。

なお，情報付与に関連して，(4) の扱いについて補足しておく。長単位では，言い掛け部「日本語」を切り出さず，「国立=日本語=国語研究所」を 1 単位として認定するため，言い掛け部「日本語」には品詞等の情報は付与されない（3.3.5 節参照）。一方，短単位では，

｜国立｜日本｜語｜国語｜研究｜所｜で｜

のように言い掛け部も単位として認定されるため，言い掛け部にも品詞等の情報が付与される。

3.1.5　タグの扱い

CSJ の転記テキストにおいては，フィラー・言いよどみ等，種々の談話現象にタグが付けられている（2.5 節参照）。単位認定に当たっては，これらのタグをどのように扱うかが問題となる。これらタグの扱いについては，既に 3.1.2〜3.1.4 節で挙げた用例の中に示されているものもあるが，本節において改めて解説することとする。

なお本節で述べるのは，人手で長単位・短単位を認定する場合に，タグが付けられたテキストをどのような考え方に基づいて単位に分割していくかということである。そのため，取り上げるタグは転記テキストの基本形に記載されるタグのみである。

人手で単位を認定した後，タグが付いた単位に対してどのように品詞等の情報を付与していくか，また発音形を単位に分割していく際に，発音形に記載されたタグをどのように扱うかについては，4 章を参照されたい。

単位認定の際のタグの扱い方としては，例えば，「｜(F ｜えー｜) ｜」のようにタグを単独で切り出すという方法が考えられる。しかし，転記テキストと実際の音声との対応関係が保たれている CSJ のような音声言語コーパスにおいて，音声と対応しないタグを単独で切り出すことには問題がある。

そこで，長単位・短単位の認定に当たっては，タグを単独で切り出すことはせず，直前若しくは直後の単位に含めることとした。具体的には，次のような手順で単位認定を行うことになる。

なお，以下に示す手順の中で「タグを取り外す」「タグを戻す」といった文言が出てくるが，これは，飽くまで作業者がそのような状態を想定して単位認定を行うということを述べたものであり，人手による単位認定に当たって，実際にそのような手順を踏んでいるわけではない。また，手順を示す際に挙げる例は，煩雑になるのを避けるため，基本的に短単位の例のみとした。ただし必要に応じて長単位の例も示すこととする。

【タグ (F)・タグ (D)・タグ (D2)・タグ (M)・タグ (O)・タグ (X)】 タグ (F)・(D)・(D2)・(M)・(O)・(X) が付けられた範囲については，以下のような手順で単位認定を行う。

(1) 転記テキスト基本形からタグを取り外して，単位認定を行う。

〈**転記テキスト**〉※下線部が取り外す対象。
その後 (F えー) そのエコーによって
(F あのねー) 何かこの眼鏡を
これは (D ん) キー押しの場合には
融合バイグラム (D2 が) の提案手法のが
後半の (M 彼にあげた) の部分では
英語の日常会話に現われる (O アイシンク) が
色々 (F えー) (X 見てて) 見てみたんですが

〈**単位認定**〉※タグを外した状態を想定して，単位認定を行う。下線部は，元々タグが付けられていた単位。
｜その後｜ えー ｜その｜エコー｜に｜よっ｜て｜
｜ あの｜ねー ｜何｜か｜この｜眼鏡｜を｜
｜これ｜は｜ ん ｜キー押し｜の｜場合｜に｜は｜
｜融合｜バイグラム｜ が ｜の｜提案｜手法｜の｜が｜
｜後半｜の｜ 彼｜に｜あげ｜た ｜の｜部分｜で｜は｜
｜英語｜の｜日常｜会話｜に｜現われる｜ アイ｜シンク ｜が｜
｜色々｜ えー ｜ 見｜て｜て ｜見｜て｜み｜た｜ん｜です｜が｜

(2) タグを元の位置に戻す。

例えば，「｜彼｜に｜あげ｜た｜」の場合，元にあったタグ (M) のうち，「(M 」(記号の後の半角スペースも含む。) は直後の単位「彼」とその前の単位境界との間に戻し，閉じ括弧は直前の単位「た」とその後の単位境界との間に戻す。

「えー」の場合，「(F 」の直後の単位と閉じ括弧の直前の単位とが一致しているが，タグを戻す際の考え方は上記のタグ (M) の扱いと同じである。

｜その後｜ (F えー) ｜その｜エコー｜に｜よっ｜て｜
｜ (F あの｜ねー) ｜何｜か｜この｜眼鏡｜を｜
｜これ｜は｜ (D ん) ｜キー押し｜の｜場合｜に｜は｜
｜融合｜バイグラム｜ (D2 が) ｜の｜提案｜手法｜の｜が｜
｜後半｜の｜ (M 彼｜に｜あげ｜た) ｜の｜部分｜で｜は｜
｜英語｜の｜日常｜会話｜に｜現われる｜ (O アイ｜シンク) ｜が｜
｜色々｜ (F えー) ｜ (X 見｜て｜て) ｜見｜て｜み｜た｜ん｜です｜が｜

【タグ (A)】 タグ (A) については，「(A 」・閉じ括弧に加えて，可読性を高めるために記載されたタグ内の右項も取り外した上で，単位認定を行う。

(1) 転記テキスト基本形からタグを取り外して，単位認定を行う。

〈**転記テキスト**〉※下線部が取り外す対象。
音圧は大体 (A 零．四五 ; ０．４５) から (A 二．二四 ; ２．２４) パスカルで
(A エヌエイチケー ; ＮＨＫ) 出版から出ております

〈**単位認定**〉※タグを外した状態を想定して，単位認定を行う。下線部は，元々タグが付けられていた単位。
｜音圧｜は｜大体｜ 零｜．｜四｜五 ｜から｜ 二｜．｜二｜四 ｜パスカル｜で｜
｜ エヌエイチケー ｜出版｜から｜出｜て｜おり｜ます｜

(2) タグを元の位置に戻す。

戻し方は，タグ (F) 等と同様である。なお長単位では，下記の「エヌエイチケー出版」のように，タグを戻す位置が単位境界と隣接しない場合があるが，いずれにしても元の転記テキストの位置に戻すことに変わりはない。

｜音圧｜は｜大体｜(A 零｜．｜四｜五;０．４５)｜から｜(A 二｜．｜二｜四;２．２４)｜パスカル｜で｜
｜(A エヌエイチケー;ＮＨＫ)｜出版｜から｜出｜て｜おり｜ます｜
※長単位：｜(A エヌエイチケー;ＮＨＫ) 出版｜から｜出｜ており｜ます｜

【タグ (K)】 タグ (K) の扱いも基本的にタグ (A) と同様である。ただし，以下に示すように，タグ (K) の中にあるタグ (F) 等を外すという手順が加わる。

(1) 転記テキスト基本形からタグ (K) を取り外した後，さらにタグ (F) とフィラーを取り外して，単位認定を行う。

〈転記テキスト〉 ※下線部が取り外す対象。タグ (K)，その中のタグ (F) 及びフィラーを外す。
そんなお (K かん (F あー) が ; 考) えを申し上げました

〈単位認定〉 ※タグを外した状態を想定して，単位認定を行う。
｜そんな｜お｜ かんがえ ｜を｜申し上げ｜まし｜た｜

(2) タグ (K) 及びタグ (F)・フィラーを元の位置に戻す。

｜そんな｜お｜ (K かん (F あー) が ; 考) え｜を｜申し上げ｜まし｜た｜

【タグ (?)】 以下のように，タグ内の値の数によって扱いが異なる。

《値が一つの場合》　タグ (F) 等と同様の方法で単位認定を行う。

(1) 転記テキスト基本形からタグを取り外して，単位認定を行う。

〈転記テキスト〉 ※下線部が取り外す対象。
有効である (? って) (F ま) これ前回と同じ結果です

〈単位認定〉 ※タグを外した状態を想定して，単位認定を行う。
｜有効｜で｜ある｜ って ｜ ま ｜これ｜前回｜と｜同じ｜結果｜です｜

(2) タグを元の位置に戻す。

｜有効｜で｜ある｜ (? って) ｜ (F ま) ｜これ｜前回｜と｜同じ｜結果｜です｜

《値が複数の場合》　列挙された候補のうち，先頭に挙げられたものを単位解析の対象とする。2 番目以下の候補は，タグ (A) の右項と同様の扱いをする。

(1) 転記テキスト基本形からタグを取り外して，単位認定を行う。

> **〈転記テキスト〉**※下線部が取り外す対象。
> 真剣に (? やれ , なれ) なかったことの一つが
>
> **〈単位認定〉**※タグを外した状態を想定して，単位認定を行う。
> ｜真剣｜に｜ やれ ｜なかっ｜た｜こと｜の｜一｜つ｜が｜

(2) タグを元の位置に戻す。

> ｜真剣｜に｜ (? やれ , なれ) ｜なかっ｜た｜こと｜の｜一｜つ｜が｜

《値なしの場合》　タグ (?) を 1 単位として切り出す。ただし語の同定ができないため，品詞等の情報は付与されない。

> ｜飛行｜経路｜を｜ (?) ｜限定｜さ｜せ｜

《タグ内にある場合》　ほかのタグの中にタグ (?) が出現する場合は，まず外側のタグを外した上で，上記のタグ (?) の扱いに従って，タグ (?) の付けられた範囲を単位認定する。

(1) 転記テキスト基本形からタグを取り外して，単位認定を行う。

> **〈転記テキスト〉**
> ※タグ (F)・タグ (D) を取り外す。
> (F (? あ)) すいません談話に含まれる段落
> (D よっ(? つ)) (F あのー) 今説明した
>
> ※タグ (?) を取り外す。
> (? あ) すいません談話に含まれる段落
> 今のよっ (? つ) あのー今説明した
>
> **〈単位認定〉**※タグを外した状態を想定して，単位認定を行う。
> ｜ あ ｜すい｜ませ｜ん｜談話｜に｜含ま｜れる｜段落｜
> ｜今｜の｜ よっつ ｜ あのー ｜今｜説明｜し｜た｜

(2) タグを元の位置に戻す。

> ｜ (F (? あ)) ｜すい｜ませ｜ん｜談話｜に｜含ま｜れる｜段落｜
> ｜今｜の｜ (D よっ(? つ)) ｜ (F あのー) ｜今｜説明｜し｜た｜

【タグ (R)】公開されたデータでは，タグ (R) が付けられた範囲は，「(R ××)」のように伏せ字化されているが，作業段階では「(R 小椋)」のように発話どおりに記載されている。単位認定をはじめとする形態論情報解析は，伏せ字化される前の段階で行うため，タグ (R) は，タグ (F) 等と同様の扱いをして，単位認定を行う。

(1) 転記テキスト基本形からタグを取り外して，単位認定を行う。

〈**転記テキスト**〉※下線部が取り外す対象。
国語研究所の (R 小椋秀樹) です

〈**単位認定**〉※タグを外した状態を想定して，単位認定を行う。
｜国語｜研究｜所｜の｜ 小椋｜秀樹 ｜です｜

(2) タグを元の位置に戻す。

｜国語｜研究｜所｜の｜ (R 小椋｜秀樹) ｜です｜

以上，3.1.2〜3.1.5 節では長短 2 種類の単位の認定基準等について述べた。その規定により長単位・短単位を認定した例を次に示す。

〔長単位〕 ┃(F えー)┃パラ言語情報┃という┃こと┃な┃んです┃が┃(F あ)┃簡単┃に┃最初┃に┃(F えー)┃復習┃を┃し┃ておき┃たい┃と┃思い┃ます┃(F ま)┃(F あのー)┃こう┃やっ┃て┃(D あっ)┃話し┃ており┃ます┃と┃それ┃は┃勿論┃(F あの)┃言語的情報┃を┃伝える┃という┃こと┃が┃

〔短単位〕 ｜(F えー)｜パラ｜言語｜情報｜と｜いう｜こと｜な｜ん｜です｜が｜(F あ)｜簡単｜に｜最初｜に｜(F えー)｜復習｜を｜し｜て｜おき｜たい｜と｜思い｜ます｜(F ま)｜(F あのー)｜こう｜やっ｜て｜(D あっ)｜話し｜て｜おり｜ます｜と｜それ｜は｜勿論｜(F あの)｜言語｜的｜情報｜を｜伝える｜と｜いう｜こと｜が｜

3.2　代表形・代表表記

単位認定基準に基づいて長単位・短単位を認定した後，各単位に対して付加情報を与えていく。CSJ では，各単位に与える付加情報として，代表形・代表表記・品詞・活用の種類・活用形・その他の情報という六つの情報を設けた。

このうち代表形は，単位認定基準に基づいて認定した各単位に与える見出し語の情報で，CSJ では片仮名で表記する。代表表記は，片仮名表記の代表形を漢字等で表したものである。この代表形と代表表記とは，国語辞典の見出しと見出しに与えられた漢字表記として位置付けられるものである。

本節では，付加情報のうち代表形と代表表記とを取り上げ，各単位に代表形を与えるための作業（同語異語判別）が必要な理由と，代表形・代表表記の付与基準とについて述べることとする。

3.2.1　同語異語判別の必要性

単位認定基準によって認定した一つ一つの単位は，活用変化・音の転訛・ゆれ・省略・融合等によって生じた異形態や異表記形そのままの形のものである。このような切り出されたままの単位について，どの単位とどの単位とが同じ語であるかを判断し，同じ語と判断した語群に対して，一つの見出し（代表形）を与えるという作業を行っていく。この作業を同語異語判別と呼ぶ。

言語単位の認定に当たっては，語の長さをどうするかということとともに，どこまでを一つの語と認めるかということも重要な問題である。しかしながら，語の長さをどうするかということが，目に見える具体的な形

で扱うことのできる問題であるのに対し，どこまでを同じ語とみなすかということは，意味という実態のつかみにくいものを扱う問題であるため，語の長さの問題と比べて議論しにくい面がある。そのため，同語異語判別の方法等については，単位の長さの問題ほどにはその重要性が余り強く意識されていないように思われる。

そこで本節では，まず同語異語判別を行うことの必要性について述べておくこととする。

CSJ の単位は，3.1.1.2 節で述べたように，

(1) CSJ から用例を採集し，話し言葉の語彙・語法の研究を行う。
(2) 品詞の分布などの計量研究によって CSJ の言語的な特徴を明らかにする。

という二つの目的に適したものとなるように設計したものである。したがって，各単位に与える付加情報についても，この二つの目的を念頭に置いて考えていく必要がある。

そこで，まず (1) の用例採集という観点から同語異語判別の必要性を考えてみたい。

国語研究においてコーパスを利用する際，最も基本的な利用法として挙げられるのは，文字列検索によって用例を採集するということであろう。この時に問題となることとして，活用変化や異形態・異表記形がある。動詞・形容詞・助動詞は，テキストの中では，未然形・連用形等の活用変化した形で出現するため，ある活用語について，その全用例を検索しようとすると，正規表現を用いて検索文字列を指定する必要がある。このこと自体，コーパスの活用に慣れた人にとっては，問題にならないことかもしれないが，そうでない人にとっては煩雑なことである。

また，話し言葉の場合は，融合・省略といった異形態が数多く出現する。これらは活用変化とは異なり，必ずしも規則的ではないため，出現する形態をあらかじめ想定し，網羅的に検索するということは難しい。

表記の面から見ると，日本語には「常用漢字表」「現代仮名遣い」等の表記の目安・よりどころがあるものの，正書法は完全には確立していない。そのため，例えば，形容詞〈アオイ〉には，「青い」「蒼い」「碧い」「あおい」「アオイ」など複数の表記があり得る。また，動詞〈イク〉について，本動詞の場合は「行く」と漢字表記し，補助動詞の場合は「いく」と平仮名表記するなどといった，意味・用法による表記の使い分けが行われることもある。したがって，用例を検索する場合には，こうした異表記形についても，想定し得る限り列挙して検索しないと，目的とする語のすべての用例を集めることができない[*8]。

このような問題を解決し，CSJ を使った用例採集を効率的に行うことができるようにするためには，同じ語と認められる異形態・異表記形をまとめた上で，代表形を与えることが必要である。そうしておけば，例えば〈チイサイ〉という代表形を検索することで，「ちいさい」のほか，「ちっちゃい」「ちっさい」といった異形態の用例も得ることができるようになり，効率良く検索ができるようになる。

次に，(2) の計量研究という観点から検討する。

従来の国語研究所の語彙調査において，必ずと言ってよいほど同語異語判別の作業が行われてきていたことから，計量研究における同語異語判別の必要性は極めて高いと言えよう。もし，同語異語判別を行わなければ，「行か」「行こ」「行く」といった活用形や，「おんなじ」「同じ」，「ちっちゃい」「小さい」といった異形態が，それぞれ見出しとして立てられることになる。そうなると，幾つかのテキストについて異なり語数を比較しよう

[*8] 異表記の問題については，CSJ の場合，転記テキストを作成する段階で，表記について一定の方針を立てているため，問題が起こらないのではないかという見方もあろう。しかし，本動詞と補助動詞とで漢字表記・仮名表記を使い分けるということなどが行われているため，同一語の異表記形が全く存在しないわけではない。したがって，異表記形の検索という面からも代表形が必要となるのである。

とした場合，活用語が多く使われているために異なり語数が多くなっているのか，使われている語の種類自体が多いのか把握できなくなる。

また，見出し語の品詞比率について検討する場合，活用語の異なり語数が名詞等の無活用語よりも多くなるため，活用語の品詞比率が高くなり，その結果，量的な構造を正確に把握することができなくなってしまう。

このような問題を回避するためには，「行か」「行こ」「行く」といった活用形や，「おんなじ」「同じ」，「ちっちゃい」「小さい」といった異形態を一つの見出し語の下にまとめておくことが求められる。つまり，同語異語判別を行うことで初めて，延べ語数の面と異なり語数の面とから，調査対象とした資料における語彙の量的な構造を把握できるのである。

以上述べたように，CSJ の形態論情報を設計する際に掲げた二つの目的 — 用例採集と計量研究 — をより良い形で実現するためには，同語異語判別の作業が必要であると言うことができる。

なお，CSJ では代表形に加えて，代表形を漢字等で表記した代表表記を各単位に与えることとした。代表形は片仮名で表記されるため，同語異語判別を行っても，代表形だけでは同音異義語の区別が付かなくなってしまうが，代表表記を与えることで同音語の区別が可能となる。これによって，同語異語判別の結果をより厳密にデータとして示すことができるようになる。このことは，計量的な研究にとって有用であるとともに，代表表記を活用した検索が可能になり，用例検索の効率が向上するという点で，用例採集にとっても有用である。

3.2.2　代表形・代表表記の付与基準

ここでは，長単位・短単位に対して，どのような基準によって代表形・代表表記を与えるのかについて述べることとする。

3.2.2.1　代表形の付与基準

CSJ の代表形は，次に述べる基準に基づいて付与する。

［1］ 代表形は，転記テキストの発音形に基づいて定める。

(1) 発音形の表記を国語表記の基準に基づいて改めるなどする。

［基本形］	［発音形］	［代表形］
衣装	イショー	イショウ
或いは	アルイワ	アルイハ
続く	ツズク	ツヅク

(2) 発音形にタグが付けられている場合は，以下のようにタグを取り除く。

［基本形］	［発音形］			［代表形］
時刻	ジコ<H>ク	→	ジコク	ジコク
(F えー)	(F エー)	→	エー	エー※
厳しい	(L キビシー)	→	キビシー	キビシイ

※フィラーの代表形は，上記 (1) の規定にかかわらず，長音を表すのに長音符号（ー）を用いる。

タグ (W)・タグ (B) が付けられている場合は，タグ及び左項に示した語を取り除き，右項に示した語を代表形として採用する。タグ (?) で値が複数あるものは，タグ及び２番目以降に挙げられた語を取り除き，１番目に挙げられた語を代表形として採用する。

［発音形］				［代表形］
(W アリーワ;アルイワ)	→	アルイワ	→	アルイハ
(B カボン;ゲボン)	→	ゲボン	→	ゲボン
(? サイシュー,サイシュ)	→	サイシュー	→	サイシュウ

複数のタグが付けられている場合は，次のようにタグを内側から順次取り除き，代表形を決定する。

［基本形］	［発音形］			［代表形］
(? 対峙,対置)	(? (W タイチ;タイジ),タイチ)	→	(? タイジ) タイチ	→	タイジ

［２］ 活用語は，終止形を代表形とする。

［基本形］	［発音形］	［代表形］
続き (ます)	ツズキ	ツヅク
白かっ (た)	シロカッ	シロイ
(述べ) ましょう	マショー	マス

ただし，連用形転成名詞は終止形に戻さず，連用形を代表形とする。

［基本形］	［発音形］	［代表形］
動き (が)	ウゴキ	ウゴキ

［３］ 連濁した語は，清音の形を代表形とする。

［基本形］	［発音形］	［代表形］
(衣装) 箱	バコ	ハコ
(条件) 付け	ズケ	ツケ
(沈ん) だ	ダ	タ
(八) 本	ポン	ホン
(知らん) ぷり	プリ	フリ

名詞用法と接尾辞用法とを持つ語については，名詞として使われている場合は清音の形を代表形とし，接尾辞として使われている場合は濁音の形を代表形とする。

［基本形］	［発音形］	［代表形］	
(活用) 型	ガタ	カタ	※名詞
(連続) 型	ガタ	ガタ	※接尾辞

連濁によって濁音化したのか，元々濁音なのか判断が付きにくいもののうち，清音の単独用法がないもの，又は単独用法があっても濁音が自然と考えられるものは，濁音の形を代表形とする。

［基本形］	［発音形］	［代表形］
(合成) 側	ガワ	ガワ

［4］ 語頭音が清音化した語は，元の形（濁音形）を代表形とする。

［基本形］	［発音形］	［代表形］
(そんなこっ) たろ (う)	タロ	ダ
右っ側 (ミギッカワ)	カワ	ガワ

［5］ 上記［3］［4］以外の転訛形・省略形で，意味変化が生じていない場合は，元の形を代表形とする。

［基本形］	［発音形］	［代表形］
(言わん) こっ (ちゃない)	コッ	コト
こっ (から)	コッ	ココ
そん (で)	ソン	ソレ
そい (から)	ソイ	ソレ
もん	モン	モノ
(山) ん (中)	ン	ノ
(見) や (しない)	ヤ	ハ
おっ (さん)	オッ	オジ
おんなじ	オンナジ	オナジ
おっきい	オッキー	オオキイ
何 (だ)	ナン	ナニ※
何 (本)	ナン	ナン※
(どう) す (んだ)	ス	スル
やっぱ	ヤッパ	ヤハリ
あんた	アンタ	アナタ

※代名詞の場合は「ナニ」を代表形とし，数詞の場合は「ナン」を代表形とする。

ただし機関の略称などは省略形を代表形とする。

［基本形］	［発音形］	［代表形］
国研	コッケン	コッケン
通総研	ツーソーケン	ツウソウケン

［6］ 数の代表形は，以下の規定により決める。

(1) 短単位の数の代表形は，音の系列と訓の系列とを区別した上で，それぞれ表 3.3 に挙げるものを代表形とする。

表 3.3　数の代表形（短単位）

		一	二	三	四	五	六	七	八	九	十
代表形	音の系列	イチ	ニ	サン	シ	ゴ	ロク	シチ	ハチ	キュウ，ク	ジュウ
	訓の系列	ヒト	フタ	ミ	ヨン	イツ	ム	ナナ	ヤ	ココノ	トオ

［基本形］	［発音形］	［代表形］
(十) 一 (日)	イチ	イチ
一 (回)	イッ	イチ
一 (つ)	ヒト	ヒト
四 (人)	ヨ	ヨン
六 (日)	ムイ	ム
八 (本)	ハッ	ハチ
九 (回)	キュー	キュウ
(四十) 九 (日)	ク	ク
三十 (歳)	サンジュッ	サンジュウ

(2) 長単位の数の代表形は，原則として発音形に基づいて決める。

［基本形］	［発音形］	［代表形］
三本	サンボン	サンボン
四十名	ヨンジューメイ	ヨンジュウメイ

「十」に関して，「十回」「十本」等が「ジュッカイ」「ジュッポン」と発音されていても，「ジッカイ」「ジッポン」を代表形とする。

［基本形］	［発音形］	［代表形］
十種類	ジュッシュルイ	ジッシュルイ
二十キロヘルツ	ニジュッキロヘルツ	ニジッキロヘルツ

「六回」の発音形が「ロッカイ」と「ロクカイ」，「八本」の発音形が「ハッポン」と「ハチホン」というように，発音形にゆれが見られる場合，原則として促音の方を代表形として採用する。ただし，ゆれが見られない場合は，原則どおり発音形に基づいて代表形を決める。

［基本形］	［発音形］	［代表形］
六回	ロッカイ	ロッカイ
六回	ロクカイ	ロッカイ

［7］ アルファベット・記号は，タグ (A) の左項に記された読みを代表形とする。

［基本形］	［代表形］
(A エイチエムエム;ＨＭＭ)	エイチエムエム
(A エスオメガエル;Ｓ ω Ｌ)	エスオメガエル
(A ケー;Ｋ)	ケー
(A ニスト;ＮＩＳＴ)	ニスト
(A エスアイドット;Ｓｉ．)	エスアイドット

3.2.2.2　代表表記の付与基準

CSJ の代表表記は，次に述べる基準に基づいて付与する。

［1］ 代表表記には，書き起こしテキストの基本形の表記を採用する。ただし，［2］以下の規定に該当するものについては，その規定によって代表表記を定める。

［基本形］	［代表形］	［代表表記］
コウモリ	コウモリ	コウモリ
パラメーター	パラメーター	パラメーター
行なう	オコナウ	行なう
表わす	アラワス	表わす

［2］ 以下に挙げるものは，転記テキストの基本形の表記によらず，各規定に基づいて代表表記とする漢字表記を定める。

(1) 数字の代表表記は，すべて漢字とする。

［基本形］	［代表形］	［代表表記］
（A 九十	キュウジュウ	九十
四;９４）	ヨン	四

(2) 助詞・助動詞の代表表記は，すべて平仮名とする。

［基本形］	［代表形］	［代表表記］
程	ホド	ほど

(3) 以下に挙げる語については，転記テキストにおける表記の使い分けにかかわらず，以下のとおり代表表記を統一する。

［基本形］	［代表形］
現われる，表われる	現われる
替える，代える	替える
言葉，詞	言葉
箱，匣	箱
咄本，噺本	咄本
張る，貼る	張る
汎化，般化	汎化
平行，並行	平行
混ぜる，交ぜる	混ぜる
巡り会える，巡り合える	巡り会える
食料，食糧	食料
ゼロ，〇，零	ゼロ
戦う，闘う	戦う
取れる，捕れる	捕れる
引き延ばす，引き伸ばす	引き延ばす
柔らかい，軟らかい	柔らかい※

※「地盤が軟らかい」のように，「柔らかい」で表記するのが極めて不自然である場合は，「軟らかい」を代表表記とする。

(4) 同一語で漢字表記と仮名表記との使い分けが行われている場合は，すべて漢字表記を代表表記とする。

［基本形］	［代表形］	［代表表記］
行く，いく	イク	行く
置く，おく	オク	置く

(5) 和語・漢語のうち，転記テキストの基本形において一部又は全部が平仮名で表記されているものについては，『岩波国語辞典』第 5 版（岩波書店）及び『国語大辞典』（小学館）を参照して，可能な限り漢字表記を当てる。

［基本形］	［代表形］	［代表表記］
あなた	アナタ	貴方
これ	コレ	此れ
無理やり	ムリヤリ	無理矢理
あるいは	アルイハ	或いは
おる	オル	居る
する	スル	為る
うまい	ウマイ	甘い

『岩波国語辞典』第 5 版及び『国語大辞典』を参照しても漢字表記を当てることができない，又は漢字表記を当てることが適当でない場合は，転記テキストの基本形の表記（平仮名表記）を代表表記とする。

［基本形］	［代表形］	［代表表記］
とんかち	トンカチ	とんかち
とんでも	トンデモ	とんでも
もう	モウ	もう
やや	ヤヤ	やや
にこやか	ニコヤカ	にこやか

［3］ タグ (A) が付された記号等については，タグ (A) の右項に記載された表記を代表表記とする。

［基本形］	［代表形］	［代表表記］
(A エイチエムエム;HMM)	エイチエムエム	HMM
(A エスオメガエル;S ω L)	エスオメガエル	S ω L
(A ケー;K)	ケー	K
(A ニスト;N I S T)	ニスト	N I S T
(A エスアイドット;S i .)	エスアイドット	S i .
ラージ (A ラムダ;λ)	ラージラムダ	ラージλ
ラージ (A シー;C)	ラージシー	ラージC

アルファベット 1 文字が 1 短単位となる場合，タグ (A) の右項に小文字で表記されていても，代表表記は大文字にする。

［基本形］	［代表形］	［代表表記］
(A エス アイ;S i)	エス アイ	S I

3.2.3 話し言葉特有の現象に対する代表形・代表表記の付与

話し言葉特有の現象のうち，融合・省略・フィラー・語断片・言い直しに対する代表形・代表表記の付与基準について説明する。

【融合】

融合は，本来複数の単位に分割されるものが音の転化等によって一まとまりになったものである。長単位・短単位認定基準では，融合について，元の形に戻すことなく，融合した単語連続全体で 1 単位とする。そのため，どのように代表形・代表表記を付与するかということが問題となる。

CSJ では，活用語の融合と非活用語の融合とに分けた上で，それぞれ以下のように代表形・代表表記を付与することとした。

［1］ 活用語の融合は，活用語の終止形を代表形として付与する。代表表記についても同様とする。

［基本形］	［発音形］	［代表形］	［代表表記］
面白きゃ	オモシロキャ	オモシロイ	面白い
(行か) なきゃ	ナキャ	ナイ	ない
(加味さ) れりゃ	レリャ	レル	れる
じゃ	ジャ	ダ	だ

［2］ 非活用語の融合は，元の単語連続の形を代表形として付与する。代表表記についても同様とする。

［基本形］	［発音形］	［代表形］	［代表表記］
こた	コタ	コトハ	事は
こら	コラ	コレハ	此れは
そりゃ	ソリャ	ソレハ	其れは
ちゃ	チャ	テハ	ては
じゃ	ジャ	デハ	では

【省略】

省略は，3.2.2.1 節に述べた代表形の付与基準［5］にあるとおり，元の形を代表形として与える。また，代表表記は，その代表形を基にして付与する。以下に例を示す。

［基本形］	［代表形］	［代表表記］
(どう) す (んだ)	スル	為る
や (んだっけ)	ヤル	遣る

【フィラー】

フィラーのうち，助詞・助動詞を含まないもので，単独で 1 長単位又は 1 短単位となるものについては，CSJ に出現する様々な形態を表 3.4 のように分類した上で，各分類ごとに，代表形・代表表記を定めた。

表 3.4 フィラーの代表形・代表表記

基　本　形	代表形	代表表記
あ（ー）	アー	あー
あ（ー）（ん）（ー）の（ー）	アノ	あの
あ（ー）（っ）と（ー）	アート	あーと
い（ー）	イー	いー
う（ー）	ウー	うー
う（ー）ん	ウン	うん
う（ー）ん（ー）（っ）と（ー）	ウント	うんと
え（ー）	エー	えー
え（ー）（っ）と（ー）	エート	えーと
お（ー）	オー	おー
そ（ー）（ん）（ー）の（ー）	ソノ	その
と（ー）	ト	と
ま（ー）	マー	まー
ん（ー）	ン	ん
ん（ー）（っ）と（ー）	ント	んと

長単位で 1 単位となる助詞・助動詞を含むフィラー (例：(F あのですね)，(F あのね)) は，まず助詞・助動詞以外の部分について，表 3.4 として示した一覧表に基づいて代表形・代表表記を定めた上で，それに助詞・助動詞の代表形・代表表記を結合させるという手順で，代表形・代表表記を定める。例を以下に示す。

［基本形］	［発音形］	［代表形］	［代表表記］
(F あのですね)	(F アノデスネ)	アノデスネ	あのですね
(F あのね)	(F アノネ)	アノネ	あのね

1 長単位又は 1 短単位の中に出現するフィラーについては，単位認定の際に，

味わう	こと	が	(F えー)	でき	ま (F えー) せ	ん
ここ	で	も	メタ (F あ) 言語行動表現	て	もの	を

のように，フィラーが 1 単位として認定されないため，代表形・代表表記も付与されない。つまり「ま(F えー)せ」「メタ (F あ) 言語行動表現」は，フィラーのない「ませ」「メタ言語行動表現」と同様に，次のように代表形・代表表記が与えられることになる。

［基本形］	［発音形］	［代表形］	［代表表記］
ま (F えー) せ	マ (F エー) セ	マス	ます
メタ (F あ) 言語行動表現	メタ (F ア) ゲンゴコードーヒョーゲン	メタゲンゴコウドウヒョウゲン	メタ言語行動表現

【語断片】

語断片は，文字どおり単語の断片であり，意味を持たないため，そもそも同語異語判別という作業の対象に

はならない。長単位認定基準・短単位認定基準では，

長単位 ： ┃処理速度┃や┃ (D お) ┃その┃大きさ┃など┃
短単位 ： ｜処理｜速度｜や｜ (D お) ｜その｜大き｜さ｜など｜

のように語断片を 1 単位として認定するが，同語異語判別の対象外ということから，代表形・代表表記は与えない。

1 長単位・1 短単位の中に現れる語断片は，3.1.4 節で述べたとおり 1 単位として認定されないため，代表形・代表表記は付与されない。つまり，

長単位 ： ┃それ┃を┃ 利用 (D す) する ┃の┃も┃
短単位 ： ｜それ｜と｜ポライト｜ ノン (D プロ) ポライト ｜と｜いう｜

という例の「利用 (D す) する」や「ノン (D プロ) ポライト」は，語断片を含まない「利用する」「ノンポライト」と同様に，次のように代表形と代表表記を付与する。

［基本形］	［発音形］	［代表形］	［代表表記］
利用 (D す) する	リヨー (D ス) スル	リヨウスル	利用する
ノン (D プロ) ポライト	ノン (D プロ) ポライト	ノンポライト	ノンポライト

【助詞・助動詞等にかかわる言い直し】

数詞・助詞・助動詞・接頭辞・接尾辞の言い直し（タグ (D2) が付けられたもの）は，通常の数詞・助詞・助動詞・接頭辞・接尾辞と同様に単位認定を行うので，代表形・代表表記の付与についても，やはり通常の数詞等と同様に行う。以下に，例を示す。

［基本形］	［発音形］	［代表形］	［代表表記］
(D2 の)	(D2 ノ)	ノ	の
(D2 未)	(D2 ミ)	ミ	未
(D2 二)	(D2 ニ)	ニ	二

【言い直し】

ここで取り上げる言い直しは，3.1.4 節と同様，転記テキストにおいてタグを付けられていないものである。3.1.4 節では，この種の言い直しを，

(1) 語の一部を述べたところで，語全体を言い直している場合。

益岡・田窪氏の 基本日本語基礎日本語文法

(2) 前に述べた語の一部のみを直後で言い直している場合。

阪倉篤義さん篤義先生 の　　国語 についてつきまして

(3) 前に述べた語全体を言い直している場合。

向こうで 教育機関教育事業 始めたいということで

(4) 1 長単位の内部に言い直しがある場合。

国立日本語国語研究所 で

という 4 種類に分類した上で，(1)〜(3) と (4) との二つに分けて，単位の認定方法を示した。代表形・代表表記についても，単位認定と同様に (1)〜(3) と (4) とに分けて付与基準を定めた。

上記の分類 (1)〜(3) については，言い掛け部と訂正部との間に長単位境界を設定した上で，言い掛け部・訂正部それぞれに対して長単位認定基準を適用して単位認定を行っている。

このように言い掛け部・訂正部ともに，単位認定に当たって通常の単位と同様の扱いを受けていることから，代表形・代表表記を与えるに当たっても，特別な規定等を設けることなく，通常の単位と同様に代表形・代表表記を与えることとする。

以下，各分類ごとに単位認定の例と代表形・代表表記を付与した例とを示す。

(1) 語の一部を述べたところで，語全体を言い直している場合。

| 益岡田窪氏 | の | 基本日本語 | 基礎日本語文法 |

［基本形］	［代表形］	［代表表記］
基本日本語	キホンニホンゴ	基本日本語
基礎日本語文法	キソニホンゴブンポウ	基礎日本語文法

(2) 前に述べた語の一部のみを直後で言い直している場合。

| 阪倉篤義さん | 篤義先生 | の | | 国語 | について | つき | まし | て |

［基本形］	［代表形］	［代表表記］
阪倉篤義さん	サカクラアツヨシサン	阪倉篤義さん
篤義先生	アツヨシセンセイ	篤義先生
について	ニツイテ	について
つき	ツク	就く
まし	マス	ます
て	テ	て

(3) 前に述べた語全体を言い直している場合。

| 向こう | で | 教育機関 | 教育事業 | 始め | たい | という | こと | で |

［基本形］	［代表形］	［代表表記］
教育機関	キョウイクキカン	教育機関
教育事業	キョウイクジギョウ	教育事業

1 長単位の内部に言い直しがあるものについては，言い掛け部を 1 長単位として切り出さずに，言い掛け部を内部に含む形で 1 長単位として認定する。

そこで，代表形・代表表記の付与においても，言い掛け部に対して代表形・代表表記を付与することはしな

い。つまり，以下の例で言えば，「国立日本語国語研究所」を，言い直しを含まない「国立国語研究所」と同様に見なして，代表形・代表表記を付与する。

以下，単位認定の例と代表形・代表表記を付与した例とを示す。

(4) 1 長単位の内部に言い直しがある場合。

| 国立=日本語=国語研究所 | で |

［基本形］	［代表形］	［代表表記］
国立日本語国語研究所	コクリツコクゴケンキュウジョ	国立国語研究所

3.3　品詞等の情報

CSJ で，長単位・短単位に付与する品詞等の情報は，(1) 品詞・(2) 活用の種類・(3) 活用形・(4) その他の情報 1・(5) その他の情報 2・(6) その他の情報 3 の 6 種類である（以下，本節では，これらの情報を一括して品詞情報と呼ぶ。）。

CSJ で採用した品詞情報は，いわゆる学校文法に基づくものである。また，多くの自動形態素解析システムで採用しているものとは異なり，品詞や活用の種類・活用形について詳細な分類を行っていない。品詞の判定方法についても，自然言語処理の分野における方法とは異なっている。

以下，本節では，CSJ における品詞情報の設計方針と品詞情報の概略について説明することとする。

3.3.1　品詞情報の設計

ここでは，CSJ における品詞情報を設計する際の方針について，学校文法に基づいたこと，品詞等について詳細な分類を行わなかったこと，そして品詞の判定方法のうち，特に自然言語処理の分野における判定方法との違いについて述べておく。

まず，品詞情報を学校文法に基づくものとしたことについて述べる。

学校文法については，これまでに多くの批判がなされており，学校文法とは異なる文法論も，既に幾つも提起されているところである。特に，日本語教育においては，学校文法とは異なる文法論に基づいて教育が行われており，自然言語処理の分野においても，この日本語教育における品詞・活用等の考え方に近いものを基盤として，品詞情報を設計している。

このように学校文法については，いろいろな批判が出されているが，それでもなお現在の我が国における事実上の標準的な文法論であることも，否定できないであろう。そして，このような我が国における学校文法の位置付けは，CSJ の形態論情報をより多くの研究者の間で共有していくということを考えた場合，大きな利点となる。

また，CSJ では，自動形態素解析システムの学習用データとして使うために，750 万語のうち約 100 万語について，人手で形態論情報の解析作業を行った。このような人手による作業を円滑に進める上でも，広く知られている学校文法を基にすることは大きな利点となる。というのは，先に述べたように，学校文法は日本における事実上の標準的な文法論となっていて，既に作業者が一定の理解に達していることから，作業者に対する教育の手間が軽減されるとともに，作業の一貫性も保ちやすいからである。

なお，学校文法は基本的に標準的な書き言葉を基にしたものであるため，CSJ のような実際の話し言葉を解

表 3.5 CSJ と UniDic との品詞（名詞）の比較

CSJ

品詞	その他 1
名詞	
	固有名詞
	数詞

UniDic

大分類	中分類	小分類	細分類
名詞	普通名詞	一般	
		サ変可能	
		形状詞可能	
		副詞可能	
	固有名詞	一般	
		人名	一般
			姓
			名
		地名	一般
			国
		組織名	
	数詞	一般	
		桁	

析していくに当たって対応できない部分もある。そこで CSJ では，学校文法に完全に依拠するのではなく，必要に応じて修正・拡張等を行って，現代における話し言葉の単位解析に適したものとするようにした。この修正・拡張等については，3.3.2 節以下を参照されたい。

次に，自然言語処理における品詞情報との違いのうち，品詞等について詳細な分類を行わなかったことについて述べる。

CSJ の品詞情報を自動形態素解析システムで採用している品詞情報と比較した場合，CSJ の方が詳細な分類を行わない，粗いものとなっている。ここで，CSJ の品詞情報と自動形態素解析システムで採用している品詞情報とを具体的に比較するために，茶筌用解析辞書の UniDic を例に取って CSJ の品詞情報との違いを見てみることとする。

CSJ と UniDic との品詞情報の違いについては，伝・山田・宇津呂 (2004:42) に，次のように述べられている（以下の引用文中の「体系」とは「品詞情報」のことである。）。

- 短単位体系の品詞は概ね UniDic の中分類まで，活用形は概ね UniDic の行分類までである。
- 短単位体系には，人名・地名以外の固有名詞はない。
- 短単位体系では，UniDic の「名詞-形状詞可能」「名詞-副詞可能」を，使用された文脈に応じて「名詞」と「形状詞」「副詞」に分類している。
- 短単位体系では，UniDic の「基本形」を，使用された文脈に応じて「終止形」と「連体形」に分類している。

品詞の分類について，名詞を例にして両者の比較をすると，表 3.5 のとおりである。

また，活用の種類 (UniDic では活用型) について，カ行五段活用を例にして比較すると，表 3.6 のとおりである。

名詞の比較，活用の種類の比較から分かるように，CSJ よりも UniDic の方がかなり詳細な分類を行っている。自動形態素解析システムで，高精度な解析を実現するためには，これぐらい詳細な品詞情報が必要になる

表 3.6　CSJ と UniDic との活用の種類（五段動詞）の比較

CSJ	UniDic			
活用の種類	大分類	行分類	段分類	小分類
カ行五段	五段	カ行		一般
				イク
				ユク
ガ行五段		ガ行		
サ行五段		サ行		
タ行五段		タ行		
ナ行五段		ナ行		
バ行五段		バ行		
マ行五段		マ行		一般
				済ム
ラ行五段		ラ行		一般
				アル
				サル
ワア行五段		ワア行	ア段	一般
				アウ
				カウ
				ガウ
				タウ
				ダウ
				ナウ
				ハウ
				バウ
				マウ
				ヤウ
				ラウ
				ワウ
			イ段	一般
				イウ
			ウ段	一般
				ツウ
			エ段	
			オ段	

ということであろう。

しかしながら，CSJ の長単位・短単位は，3.1.1.2 節でも述べたように，用例採集と計量研究という二つの国語に関する研究を目的としたものであり，代表形・代表表記についても，この目的を踏まえて必要性等について検討を行ったところである。したがって，品詞情報について，どこまで詳細な分類を行うかについても，国語研究，特に用例採集と計量研究においてどのように活用していくかという観点から検討を加える必要がある。

用例採集という面から言えば，品詞情報は，用例を検索する際の検索条件として利用するものである。また，集めた用例を分類・整理していく際にも，利用するものである。このような利用の仕方を考えた場合，UniDic で採用しているような詳細な分類がどこまで必要かということが問題になる。

ただ，この問題は，結局，個々の研究者の研究方法等に依存する面が強いということ，またそもそも現時点では，コーパスを活用した国語研究が，それほど進んでおらず，そういう意味でも，どのような品詞情報が有用なのかということについて共通の考え方ができ上がっていないということから，判断するための材料が極めて乏しい状況にある。

このような状況を踏まえて，CSJ においては，ひとまず自動形態素解析システムで採用しているような詳細な分類は行わないこととした。まずは，最低限必要と思われる品詞情報を付与しておき，実際に研究に活用していく中で，どのような品詞情報が望ましいか検討していくという方針を取ったということでもある。

最後に，自然言語処理における品詞情報との違いのうち，品詞の判定方法の違いについて述べる。

品詞の判定方法で主に問題となるのは，名詞のうち，形状詞や副詞としても使われる語である。例えば，次のような語である。

形状詞としても使われる語　：　安全，インターナショナル，可能，自然
副詞としても使われる語　　：　今日，昨日，今年

国語学では，このような語について，主語や目的語として使われていれば，名詞として扱い，連体修飾語として使われていれば形状詞，連用修飾成分として使われていれば副詞として扱うというのが，一般的であろう。つまり，その語が実際に使われている文脈ごとに，名詞か形状詞か副詞かという判定を行い，品詞を定めていくのである。その結果，同じ「自然」という語であっても，テキストの中では，名詞と判定されたもの，形状詞と判定されたものの 2 種類が存在することになる。

一方，自然言語処理の分野では，このような品詞の判定方法は取っていない。上記の「安全」や「今日」のような語について，どのような文脈で使われているかにかかわらず，常に一つの品詞を与えている。

表 3.5 に示した UniDic の名詞の分類を見ると，名詞の下位分類として，「形状詞可能」「副詞可能」という分類を設けている。これが，名詞のうち，形状詞や副詞としても使われ得る語に付与するために設けられた分類である。具体的には，「インターナショナル」「自然」などには「名詞-普通名詞-形状詞可能」，「今日」「今年」などには「名詞-普通名詞-副詞可能」という品詞を，実際の使用例にかかわらず与えていくのである。

CSJ では，先に引用した伝・山田・宇津呂 (2004:42) にもあるとおり，UniDic のような品詞の判定法をとらず，その語が使われている文脈等に基づいて，名詞・形状詞・副詞の判定を行っていくことにした。このような判定方法をとったのは，3.1.1.2 節に示した目的の一つ，計量研究によってデータの言語的な特徴を明らかにすることに適した品詞情報にするためである。この点について，以下，先行研究の調査結果を引用しながら説明しておく。

次に示す表 3.7 は，国立国語研究所 (1955) における品詞比率の調査結果を基に，日常談話とニュース・

ニュース解説の品詞比率を表にまとめたものである。

表 3.7　日常談話・ニュース解説・ニュースの品詞比率

	日常談話	ニュース解説	ニュース
名　詞	17.9%	27.6%	35.9%
代名詞	2.6%	0.9%	0.5%
形容動詞	1.2%	1.5%	0.9%
連体詞	0.8%	1.2%	1.6%
副　詞	6.1%	2.5%	1.3%
接続詞	1.9%	2.6%	1.0%
感動詞	4.7%	0.3%	0.0%
動　詞	12.2%	16.0%	14.9%
形容詞	2.7%	0.9%	0.4%
助　詞	34.7%	34.3%	33.0%
助動詞	12.9%	12.3%	10.6%
その他	2.3%	0.0%	0.0%

この表を見ると，日常談話とニュース・ニュース解説とを比べた場合，名詞の比率はニュース・ニュース解説の方が高く，副詞・形容詞の比率は日常談話の方が高いという差異が見られる。つまり，日常談話とニュース・ニュース解説の差異が，名詞・副詞・形容詞の比率という点に現れているのである。

こうした調査結果を踏まえると，名詞のうち，副詞・形状詞としても使われる語について，どのような形で品詞の判定を行うかということが，計量研究の調査結果に影響を与えるものと考えられる。

計算機による自動解析と同様の品詞判定の方法をとると，実際の文脈では形状詞・副詞として使われているものが，一律に名詞として扱われることになる。その結果，形状詞・副詞の比率が下がり，名詞の比率が高くなる。そうなると，データの言語的な特徴が見えにくくなってしまう可能性があり，品詞比率を基にデータの言語的な特徴を把握しようとする立場からは，望ましいとは言えないであろう。

このような，計量研究への活用という目的に適した品詞の判定方法という観点から，名詞・形状詞・副詞の品詞判定を，できる限り文脈に基づいて行うこととしたのである。

以上述べたような方針に基づいて品詞情報の設計を行った結果，付録 3.5 に一覧したような品詞情報を採用することとした。これらの品詞情報について，次の 3.3.2 節以下で説明することとする。

3.3.2　品詞

品詞は，長単位・短単位ともに共通で，付録 3.5 に示す 15 種類である。以下，学校文法の品詞と異なる点について説明を加える。

(1) 形状詞

形状詞は，いわゆる形容動詞の語幹のことである。長単位・短単位ともに，形容動詞は「｜きれい｜だ｜」「｜新鮮｜だ｜」のように活用語尾が分割されるので，その語幹に当たる部分に付与する品詞として形状詞を用意した。なお活用語尾は，断定の助動詞「だ」として扱うので，助動詞という品詞を付与する。

(2) 接頭辞・接尾辞・言いよどみ

接頭辞・接尾辞・言いよどみに伴う語断片（タグ (D) を付けられたもの）は，本来，品詞とされるものではない。そういう意味では，品詞の分類に名詞・動詞などと並べて，接頭辞・接尾辞・言いよどみという分類を立てることは，適当でないということになる。

しかし，例えば短単位では，「国立国語研究所」が「｜国立｜国語｜研究｜所｜」と 4 単位に分割され，接尾辞に当たる「所」が「国立」「国語」「研究」という単語と同様に 1 短単位として切り出されることになる。また長単位においても，文節間の係り受けを厳密に考えていこうとする立場から，

┃ローマ時代┃の┃円形劇場┃とか┃水路┃等┃

のように，接尾辞「等」を 1 長単位として認定する場合がある。

また，実際の話し言葉を収録した CSJ では，言いよどみに伴う語断片も数多く出現している。この言いよどみは，そもそも意味を持たず，文の成り立ちや単語の構成にもかかわることがないという点で，明らかに単語とは呼べないものである。しかし長単位・短単位の認定基準では，3.1.4 節に述べたとおり，原則として一つの単位として認定する。

このように CSJ の単位認定基準では，単語よりも短い要素である接辞や，そもそも意味を持たず，単語とは呼べない語断片が，単語と同様に 1 単位として認定される。したがって，これらに対して付与するための品詞情報を用意しておく必要がある。

そこで CSJ では，品詞の分類項目として，接頭辞・接尾辞・言いよどみという三つを設けることとした。

(3) 記号

記号は，文法的な機能に基づく分類ではなく，むしろ文字種という観点からの分類と言うべきものである。そういう意味では，品詞の分類に記号を設けることは，適切ではないという意見もあろう。また，書き言葉なら，記号として扱うべき単位があるかもしれないが，話し言葉において記号という扱いをしなければならないような単位があるのかという疑問もあろう。しかしながら，CSJ のデータを解析していくと，名詞・副詞といった一般的な品詞分類では対応しきれない，記号と扱うより仕方がない単位も出てくるのである。

CSJ に収録した音声には学会講演が含まれているため，例えば，次のような選択肢や箇条書きの項目の読み上げや，語・文字・発音についてのメタ的な言及が見られる（下線部）。

項目の読み上げ： (A ビー;B) 相手の状態を見て話し掛けている例
メタ的な言及 ： それから (M 布) を字母とした字

このような「B」「布」に対して，どのような品詞を付与するのかということが問題になる。「B」「布」について，活用しないということ，上記の例で「布」が格助詞「を」で受けられており，「B」も格助詞で受けることができるということから，名詞とする立場もあろう。しかし，そのような扱いをしていくと，結果的に名詞が「品詞のゴミ箱」となってしまうおそれがある。

結局，このような問題を解決する手段の一つとして，記号という品詞を設けることにしたのである。また，名詞をいわゆる「品詞のゴミ箱」のような扱いにしないという観点から，選択肢等の項目の読み上げ，メタ的な言及のほか，数式の読み上げ（タグ (O) を付けられたもの）に出てくる記号類に対しても記号という品詞を与えることとした。

3.3.3　活用の種類・活用形

活用する語のうち，動詞・形容詞・動詞性接尾辞・形容詞性接尾辞・文語助動詞には，活用の種類と活用形とを，口語の助動詞には活用形を付与した。これらも基本的に学校文法に基づくものであるが，一部変更を加えるなどしている。以下，そのような点について説明する。

(1) 形容詞

口語形容詞には活用の種類は存在せず，すべて一つの活用の種類に属するが，文語の形容詞にはク活用・シク活用という活用の種類が存在する。

CSJ では，まず文語の形容詞に，活用の種類としてク活用・シク活用の種別を与えることとし，ク活用形容詞には文語形容詞型 1，シク活用には文語形容詞型 2 という情報を与えることとした。なお，文語形容詞のうち「多し」は，終止形に「多かり」という活用形を取ることがある。そこで，「多し」は，ク活用と区別して文語形容詞型 3 とした。

口語形容詞については，先にも述べたように本来活用の種類を付与する必要はないが，文語形容詞との釣り合いを考えて，形容詞型という活用の種類を付与することとした。

(2) 接尾辞

接尾辞のうち，動詞性接尾辞・形容詞性接尾辞には，動詞・形容詞に準じて活用の種類を付与した。活用の種類は，活用語尾の形態に基づいて，動詞・形容詞と同じものを付与した。以下に例を示す。

［基本形］	［代表形］	［代表表記］	［品詞］	［活用の種類］
(嬉し) がる	ガル	がる	接尾辞	ラ行五段
(認め) 難い	ガタイ	難い	接尾辞	形容詞型
(長男) らしから (ぬ)	ラシ	らし	接尾辞	文語形容詞型 2

(3) ザ行変格活用

活用の種類で，学校文法と大きく異なるのは，ザ行変格活用を設けた点である。これは，ザ行上一段活用とサ行変格活用のうちザ行に活用するものとを併せたものである。活用表には，次のようにザ行上一段活用の活用語尾とサ行変格活用の活用語尾との両方が登録されている。見出し語形は，サ行変格活用の語形とした。

表 3.8　ザ行変格活用の活用表

見出し	未然形	連用形	終止形	連体形	仮定形	命令形
ずる	じ ぜ	じ	じる ずる	じる ずる	じれ ずれ	じよ ぜよ

このような活用の種類を設けたのは，次のような理由による。

サ行変格活用に分類される「信ずる」「禁ずる」などは，一段化が起こり，「信じる」「禁じる」というザ行上一段活用の形で使われることがある。このような語については，一般的に，その活用語尾の形態を基にして「サ行変格」「ザ行上一段」という活用の種類を付与していくことになる。しかし，そのようにしていくと，サ

行変格活用とザ行上一段活用とで活用語尾の語形が一致する未然形と連用形とについて，どちらの活用の種類に属するものとするかという問題が生じる。

このような場合，語形が同じである以上，一定の方針を立てて一律にどちらかの活用の種類を付与していくより方法がない。しかし，例えば飽くまでサ行変格活用が上一段化したものという立場を取って，未然形・連用形は一律にサ行変格活用として扱うという方針を立てると，ザ行上一段活用動詞の未然形・連用形の例が存在しなくなるという問題がある。一方，その反対に，一律にザ行上一段活用として扱うという方針を立てれば，サ行変格活用動詞の未然形のうち「〜じ」という活用語尾の例と連用形の例とが存在しなくなる。

CSJ では，このような問題を解決する方策として，サ行変格活用とザ行上一段活用とを統合したザ行変格という活用の種類を設けることとした。これにより，サ行変格活用の「信ずる」「禁ずる」などとザ行上一段活用の「信じる」「禁じる」などには，いずれも「ザ行変格」という活用の種類を与えることになる。

3.3.4 その他の情報

その他の情報は，品詞・活用の種類・活用形以外の品詞の下位分類，語形等に関する種々の情報である。各種類ごとに 1〜3 の三つに分けて整理した。以下，各種類ごとに説明する。

(1) その他の情報 1

その他の情報 1 は，名詞・助詞に対して付与する情報である。名詞のうち人名・地名には固有名詞，数詞には数詞という情報を付与する。助詞には，格助詞・準体助詞・接続助詞・係助詞・副助詞・終助詞という助詞の下位区分を付与する。

一般に固有名詞とするものの範囲としては，人名・地名以外にも，会社名・学校名・乗り物の名のほか，元号などが考えられる。しかし，固有名詞の範囲を広げていくと，固有名詞とするか普通名詞とするかについて，判断に迷う例が非常に多くなり，一貫性を持って情報付与を行うことが難しくなるという弊害が生じる。

そこで，CSJ では，固有名詞の範囲を人名・地名に限ることとした。これによって，最小単位認定基準，短単位認定基準においても人名・地名に限って，一般の単位とは異なる扱いをしていることとも対応が取れ，CSJ 内で固有名詞の扱いに一貫性が確保されることとなった。

(2) その他の情報 2

その他の情報 2 は，音便等の語形変化に関する情報である。イ音便・ウ音便・促音便・撥音便のほかに，融合又は省略という現象の見られる単位に付与する融合・省略がある。この融合・省略については，3.3.5 節を参照されたい。

音便には，動詞「行く」の連用形「行っ」のように，活用形の一つとして認められている音便のほかに，「知らない」が「知んない」となるような活用形の一つとして認められていない音便がある。「知んない」のような活用形の一つとして認められていない音便については，撥音便 A のように末尾に A という記号を付けて，活用形の一つとして認められている音便と区別している。つまり，「知んない」の「知ん」には撥音便 A という情報を付与する。

(3) その他の情報 3

その他の情報 3 には，転記でタグ (M) を付けられたものに付けるメタ，転記でタグ (D2) を付けられたものに付ける言いよどみ，長単位で認定される複合辞に付ける連語がある。言うなれば，その他の情報 1 及びその他の情報 2 に分類できない情報が，その他の情報 3 として扱われているのである。

3.3.5　話し言葉特有の現象に対する品詞情報の付与

ここでは，話し言葉特有の現象に対する品詞情報について述べることとする。

【融合】

活用語の融合については，活用語の品詞を付与する。また，活用形については，元の形に戻した場合の活用形を付与する。またその他の情報 2 には融合という情報を付与する。

［基本形］	［代表形］	［品詞］	［活用形］	［その他 2］
面白きゃ	オモシロイ	形容詞	仮定形	融合
(行か) なきゃ	ナイ	助動詞	仮定形	融合
(加味さ) れりゃ	レル	助動詞	仮定形	融合
じゃ	ダ	助動詞	連用形	融合

活用語の融合のうち，以下に挙げるものは助動詞として扱った。

「てる」(「ている」の融合)　　「てらっしゃる」(「ていらっしゃる」の融合)
「てく」(「ていく」の融合)　　「とく」(「ておく」の融合)
「とる」(「ておる」の融合)　　「ちまう・ちゃう」(「てしまう」の融合)
「たげる」(「てあげる」の融合)　「たる」(「てやる」の融合)
「つう (っつう・(っ) ちゅう)・ってえ)」(「という」の融合)

これらは，例えば表 3.9 のように規則的に活用するものとしてとらえることが可能だからである。
情報付与の例を次に示す。

［基本形］	［代表形］	［代表表記］	［品詞］	［活用形］
(走っ) てる	テル	てる	助動詞	終止形
っつう (のは)	ツウ	つう	助動詞	連体形

※その他の情報 2 として融合という情報を付与することはしない。

非活用語の融合については，先行語の品詞を付与する。また，その他の情報 2 として融合という情報を付与する。

［基本形］	［代表形］	［品詞］	［その他 1］	［その他 2］
こた	コトハ	名詞		融合
こら	コレハ	代名詞		融合
そりゃ	ソレハ	代名詞		融合
ちゃ	テハ	助詞	接続助詞	融合
じゃ	デハ	助詞	格助詞	融合

表 3.9 「てる」「てく」の活用表

見出し	未然形	連用形	終止形	連体形	仮定形	命令形
てる	て	て	てる	てる	てれ	てろ
てく	てか てこ	てっ	てく	てく	てけ	てけ

【省略】

省略については，その他の情報 2 として省略という情報を与える。それ以外は，通常の単位に対する情報付与と同様である。

［基本形］	［代表形］	［品詞］	［その他 2］
や (んだっけ)	ヤル	動詞	省略

【フィラー】

表 3.4 に一覧したフィラーの品詞は，「感動詞」とする。

「(F あのですね)」のような，助動詞・助詞と結合しているものについては，長単位と短単位とでフィラーとして認定している範囲に違いがあるため，感動詞として扱う範囲にも違いが生じる。長単位では，「あのですね」を 1 長単位とするため，「あのですね」全体を感動詞とする。一方，短単位では，「｜(F あの｜です｜ね)｜」と分割するため，「あの」のみを感動詞とし，「です」は助動詞，「ね」は助詞とする。

1 長単位又は 1 短単位の中に現れるフィラーについては，単位認定の際に 1 単位として切り出さないことから，代表形・代表表記もフィラーには与えられない。例えば，

｜味わう｜こと｜が｜(F えー)｜でき｜ま (F えー) せ｜ん｜
｜ここ｜で｜も｜メタ (F あ) 言語行動表現｜て｜もの｜を｜

という例の「ま (F えー) せ」「メタ (F あ) 言語行動表現」には，次のように代表形・代表表記が与えられている。

［基本形］	［代表形］	［代表表記］
ま (F えー) せ	マス	ます
メタ (F あ) 言語行動表現	メタゲンゴコウドウヒョウゲン	メタ言語行動表現

そこで品詞情報については，その代表形に基づいて行うこととし，「ま (F えー) せ」はその代表形「マス」を基に助動詞，「メタ (F あ) 言語行動表現」はその代表形「メタゲンゴコウドウヒョウゲン」を基に名詞とした。

【語断片】

語断片のうち，

｜処理｜速度｜や｜(D お)｜その｜大き｜さ｜など｜

とある「(D お)」のように，単独で 1 単位となるものには，言いよどみという品詞を与える。

1 長単位又は 1 短単位の中に現れる語断片については，単位認定に当たって，

| それ | を | 利用 (D す) する | の | も |
| それ | と | ポライト | ノン (D プロ) ポライト | と | いう |

のように 1 単位として切り出されないことから，代表形・代表表記は与えられない。例えば，上記の「利用 (D す) する」「ノン (D プロ) ポライト」には，次のように代表形・代表表記が与えられている。

[基本形]	[代表形]	[代表表記]
利用 (D す) する	リヨウスル	利用する
ノン (D プロ) ポライト	ノンポライト	ノンポライト

そこで品詞情報については，その代表形・代表表記に基づいて行うこととし，「利用 (D す) する」はその代表形「リヨウスル」を基に動詞，「ノン (D プロ) ポライト」はその代表形「ノンポライト」を基に名詞とする。

【助詞・助動詞等にかかわる言い直し】

数詞・助詞・助動詞・接頭辞・接尾辞にかかわる言い直し（タグ (D2) を付けられたもの）は，通常のものと同様に単位認定，代表形・代表表記の付与が行われているので，品詞情報についても通常の単位と同様に行うこととした。ただし，その他の情報 3 として，言いよどみという情報を与える点が，通常の単位と異なる。以下に，例を示す。

[基本形]	[代表形]	[代表表記]	[品詞]	[その他 3]
(D2 の)	(D2 ノ)	の	助詞	言いよどみ
(D2 未)	(D2 ミ)	未	接頭辞	言いよどみ
(D2 二)	(D2 ニ)	二	名詞	言いよどみ

【言い直し】

ここで取り上げる言い直しは，3.1.4 節と同様，転記テキストにおいてタグを付けられていないものである。単位認定及び代表形・代表表記の付与に当たっては，この言い直しを，

(1) 語の一部を述べたところで，語全体を言い直している場合。

益岡・田窪氏の 基本日本語基礎日本語文法

(2) 前に述べた語の一部のみを直後で言い直している場合。

阪倉篤義さん篤義先生 の　　国語 についてつきまして

(3) 前に述べた語全体を言い直している場合。

向こうで 教育機関教育事業 始めたいということで

(4) 1 長単位の内部に言い直しがある場合。

国立日本語国語研究所 で

という 4 種類に分類した上で，(1)〜(3) と (4) との二つに分けて，単位の認定方法，代表形・代表表記の付与基準を示した。そこで，品詞情報の付与に当たっても，(1)〜(3) と (4) とに分けて基準を示すこととする。

分類 (1)〜(3) については，言い掛け部と訂正部との間に長単位境界を設定した上で，言い掛け部・訂正部それぞれに対して長単位認定基準を適用して単位認定を行う。つまり，言い掛け部・訂正部とも，通常の単位認定の方法で単位を認定するのである。そして，代表形・代表表記の付与に当たっても特別な規定等を設けることなく，通常の単位と同様に代表形・代表表記を与える。

そこで，品詞情報も，代表形・代表表記を基にして，通常の単位と同様に与えることとする。以下，各分類ごとに単位認定の例と品詞を付与した例とを示す。

(1) 語の一部を述べたところで，語全体を言い直している場合。

| 益岡・田窪氏 | の | 基本日本語 | 基礎日本語文法 |

［基本形］	［代表形］	［代表表記］	［品詞］
基本日本語	キホンニホンゴ	基本日本語	名詞
基礎日本語文法	キソニホンゴブンポウ	基礎日本語文法	名詞

(2) 前に述べた語の一部のみを直後で言い直している場合。

| 阪倉篤義さん | 篤義先生 | の | | 国語 | について | つき | まし | て |

［基本形］	［代表形］	［代表表記］	［品詞］
阪倉篤義さん	サカクラアツヨシサン	阪倉篤義さん	名詞
篤義先生	アツヨシセンセイ	篤義先生	名詞
について	ニツイテ	について	助詞
つき	ツク	就く	動詞
まし	マス	ます	助動詞
て	テ	て	助詞

(3) 前に述べた語全体を言い直している場合。

| 向こう | で | 教育機関 | 教育事業 | 始め | たい | という | こと | で |

［基本形］	［代表形］	［代表表記］	［品詞］
教育機関	キョウイクキカン	教育機関	名詞
教育事業	キョウイクジギョウ	教育事業	名詞

1 長単位の内部に言い直しがあるものについては，言い掛け部を 1 長単位として切り出さずに，言い掛け部を内部に含む形で 1 長単位として認定する。これを受けて，代表形・代表表記の付与においても，言い掛け部に対して代表形・代表表記を付与することはせず，上記の例で言えば，「国立日本語国語研究所」を，言い直しを含まない「国立国語研究所」と同様に見なして，代表形・代表表記を付与する。

品詞情報は，言い直しを含まない「国立国語研究所」と同様に見なして与えた代表形を基に付与する。以下，品詞を付与した例を示す。

(4) 1 長単位の内部に言い直しがある場合。

| 国立=日本語=国語研究所 | で |

［基本形］	［代表形］	［代表表記］	［品詞］
国立日本語国語研究所	コクリツコクゴケンキュウジョ	国立国語研究所	名詞

3.4 今後の検討課題

以上述べてきたように，CSJ では，用例採集・資料研究という二つの研究目的を設定した上で，用例採集のための短単位，資料の言語的特徴を明らかにするための長単位というように，その目的に応じて 2 種類の単位を設計した。また，長短 2 種類の単位に付与する情報も，用例採集・資料研究という研究目的を念頭に置いて設計した。

その結果，単位の設計については，これまでの国語研究所の語彙調査と同様に，単語とは何かという議論をひとまず棚上げしている。この点については，今回の成果を基に改めて考えていく必要がある。

また，今回設計した長短 2 種類の単位，代表形・代表表記・品詞情報について，本当に上記の二つの研究目的に適した単位となっているのかどうか検証する必要もある。これについては，CSJ の形態論情報を活用した研究を進め，その結果を基に考えていくことが求められよう。

なお以下，現時点において単位認定基準等の中で再検討を要すると思われる事項について述べておく。

3.4.1 単位に関する検討課題

3.4.1.1 長単位に関する検討課題

長単位については，規則［3］として「体言連続の一部分が連体修飾語を受けている場合，その部分の後で切る。」という規定を設けたことが挙げられる。この規定は，語と語との係り受けを厳密に考えたところから作られた規定であり，その意味では問題はない。しかし実際にテキストを単位に分割していく際には，体言連続の一部分が連体修飾語を受けているかどうかの判定が難しいものもあった。また，その結果，特に判定の難しい「以降」「間」「ごと」「自体」「達」が付いた体言連続について，「「以降」「間」「ごと」「自体」「達」が付いた場合は切らない。」という例外規定を設けることにもなった。このようなある意味 煩雑な規則を設けることは，複数の作業者で大量のデータを不統一のないように処理するということを考えた場合，作業上の大きな負担になる。今後はこのような規則を設けないということも考えてよかろう。

また，ここで問題にした規則に見られるように，一般に長単位の認定規則は短単位の認定規則に比べて煩雑になる傾向がある。したがって，長単位自体をもっと単純なものにするということも検討する必要があろう。

3.4.1.2 短単位に関する検討課題

短単位については，まず第一に付属要素の認定の問題が挙げられよう。付属要素の認定の難しさについては，短単位の基となった β 単位においても既に指摘されているところである。CSJ では，一般に接頭辞・接尾辞とされるもののうち，付録 3.3・付録 3.4 に掲げたものを付属要素とすることとしたが，その判断についてはやはり迷う点もあった。今後，さらにほかの資料について短単位での解析を進める際にも問題になることが予想されることであり，付属要素の認定について何らかの指針を設ける必要もあろう。

次に挙げられるのは，外来語の処理についてである。理系の学会講演に出現する専門用語の中には，「インサーションペナルティー」「スペクトルパラメーター」などのような長い語が見られた。そこで，外来語の最小単位 2 個の 1 次結合体が 11 拍以上になる場合には，二つの最小単位を結合させずに単独で 1 短単位とするという例外規則を設けた。このように拍数によって最小単位の結合に制約を与えるという規則は，β 単位の認定基準でも設けられているものである[*9]。

しかしながら，CSJ について言えば，この規則は和語・漢語の短単位の長さとの釣り合いを考えて設けたという性質のものであり，11 拍で線を引くことに言語学的な意味があるわけではない。したがって，今後はこのような例外規則を設けずに一律に最小単位 2 個の 1 次結合を 1 短単位とするか，外来語の最小単位の扱いについて別の規則を考えることが必要であろう。

3.4.2 代表形・代表表記に関する検討課題

CSJ においても，国語研究所がこれまでに行ってきた語彙調査を踏まえ，同語異語判別を行った上で，代表形・代表表記を付与した。これにより，自動形態素解析システムによる解析とは異なり，「ちっちゃい」「ちっさい」といった異形態と「小さい」という規範的な語形とを〈チイサイ〉という代表形の下に統合することとなり，用例採集や計量研究に，より適したデータとなったと言うことができる。

しかしながら，この同語異語判別については，大きな課題を残しているのも事実である。それは，同語異語判別が完全にはできておらず，本来，別語と判定されるべき語を，結果的に同一語として扱うことになったということである。

CSJ では，各長単位・短単位に対して，代表形・代表表記・品詞情報を付与したため，同音異義語を代表表記や品詞によって区別できたり，同音・同表記で意味が異なる語について品詞で区別できたりしている。しかし，中には同音・同表記・同品詞という語がある。例えば，次に挙げる「大 (ダイ)」が，それである。

ケーキ屋はやはり女子 大 とか多いせいかですね (F えー) お洒落なお店が
酢飯の方を (F えーっとー) 手で取りまして一口 大 の大きさに丸めます

「女子大」の「大」,「一口大」の「大」は，共に代表形「ダイ」，代表表記「大」，品詞「接尾辞」であり，同音・同表記・同品詞となっている。しかし，「女子大」の「大」は大学という意，「一口大」の「大」は大きさという意というふうに意味が異なるため，本来は別語として扱うべきものである。しかし，CSJ の形態論情報には，意味の違いを示すための情報を用意していなかったため，上記の「大」は，結果的に同じ語として扱われている。このような語は，余り多くはないが，より高精度な研究を行っていくためには，すべての語について同語異語判別ができていることが求められよう。

ただ，もし人手解析作業において，同語異語判別を完全に行ったとしても，現在の自動形態素解析システムは，意味の判別をできないため，自動解析で作成する約 650 万語のデータについては同語異語判別ができないということになる。今回，同語異語判別を完全な形で行うということをあきらめた背景には，このような自動形態素解析システム側の事情もある。

代表形の付与については，もう一つ課題がある。それは，長単位・短単位の認定基準に比べて，同語異語判

[*9] β 単位の規則では，外来語の最小単位どうしの結合では 7 拍，その他の結合では 6 拍を超える場合，最小単位を結合させずに単独で 1 短単位とするように定めている。なお，活用語の場合，動詞は連用形，形容詞は語幹で拍数を数えることとしている (国立国語研究所 1962：12-13)。

別の基準が十分に整備できなかったという点である。同語異語判別の基準については，国語研究所の語彙調査の報告書を見ても，単位認定基準よりも比較的簡単な記述にとどまっている。主として言語の形態的な面から規定していく単位の規定に比べると，同語異語判別は，意味の面に踏み込む作業であるため，規定が立てにくい面がある。

結局，CSJ においては，3.2.2.1 節に示したような基準を定めた上で，実際のデータを見ながら，同一語とするか，異なる語とするかの判定を行っていった。このような形である程度一貫性を持って作業を行うことができたのは，人手解析分のデータ量が延べ語数で約 100 万語，異なり語数で約 2 万語という規模であったからであろう。今後，より大規模なコーパスを構築していくためには，同語異語判別についても，より明確な基準を作成できるよう検討を行っていく必要がある。

代表表記については，CSJ では基本的に転記テキストの基本形の表記を採用することとした。ただし転記テキストが仮名書きされている場合等は，『岩波国語辞典』第 5 版（岩波書店），『国語大辞典』（小学館）の見出し語の漢字表記を基にして，できる限り漢字を与えていった。その結果，「トテモ」に対する「迚も」，「ウワゴト」に対する「譫言」などのように，一般には余りなじみのない漢字表記を代表表記として採用する結果となった。また，「ツライ」「カライ」ともに代表表記は「辛い」となるなど，代表形は異なるが代表表記が一致するという語も見られる。

代表表記をどのように定めるかということについて基準を立てることは非常に難しいが，少なくとも一般になじみのない漢字表記を採用したり，代表形が異なる語どうしで代表表記が一致するというような例がないようにしていく必要がある。

3.4.3　品詞情報に関する検討課題

CSJ の品詞情報は，学校文法に基づいて設計した。このこと自体には，基本的に問題はないが，今後は UniDic などで採用しているような詳細な分類を取り入れていくことを検討する必要があろう。

UniDic などで採用している詳細な情報を CSJ で採用しなかったのは，3.3.1 節でも述べたように，国語研究にとって必要な情報かどうかの判断が難しかったためである。しかし，CSJ の完成後，これを使って用例の採集・整理・分析等を行ってみると，UniDic の品詞情報で採用されている細分類が，国語研究，特に用例の分類・整理などに有用ではないかと感じることがしばしばあった。

また，UniDic の品詞情報にある細かな分類は，いずれも特定の語形等と対応したものであるため，付与する際に，意味・用法等について判断する必要はない。そういう意味では，作業上の負担もそれほどかからないと考えられる。

今後構築していくコーパスでは，UniDic をはじめとして，自然言語処理で採用している品詞情報等を参考にして，より詳細な品詞情報を設計していく必要がある。ただ，自動形態素解析システムで採用している品詞等の情報すべてが，国語研究に有用とは言えないであろう。その中から，国語研究にとって有用な情報を取捨選択していくことが求められる。また，名称について，自動形態素解析システムで採用しているものには，分かりにくいものがある。より分かりやすい名称を考えていくことも必要である。

上記以外にも，CSJ の形態論情報には，見直しを要する点があろう。先にも述べたように，今後 CSJ を利用した研究を進めつつ，国語研究にとって有用な形態論情報についても検討を行い，より良い形態論情報を提案していきたいと考えている。

［付録 3.1］　助詞相当句

普通形	連用形	丁寧形	連体修飾型	
			普通形	丁寧形
でもって				
にあたって	にあたり	にあたりまして		
にあって		にありまして		
に至る				
において		におきまして	における	におけます
に応じて		に応じまして	に応じた	
に関して	に関し	に関しまして	に関する	
に比べて	に比べ	に比べまして		
に際して				
に従って	に従い		に従った	
に対して	に対し	に対しまして	に対する	に対します
について	につき	につきまして		
につれて	につれ	につれまして		
にとって		にとりまして		
にとっては				
に伴って	に伴い		に伴う	
に基づいて	に基づき	に基づきまして	に基づく に基づいた	
によると		によりますと		
によって	により	によりまして	による	によります
によっては				
にわたって	にわたり	にわたりまして	にわたる	
として		としまして といたしまして		
を通じて		を通じまして		
を通して				
をもって				
をもとにして	をもとに	をもとにしまして をもとにいたしまして		
をめぐる				
という			という ていう っていう	
といった			といった ていった っていった	

［付録 3.2］　助動詞相当句

種　類	普　通　形	丁　寧　形
肯定・否定（肯定）	である	
	でございます	
	のだ	のです
		のである
		のでございます
肯定・否定（否定）	でない	
	ではない	
		ではありません
		ではございません
	のではない	
		のではありません
許可・依頼・勧誘	てもいい	
		てもよろしい
	てほしい	
禁止・当然・義務	てはいけない	
		てはいけません
	てはならない	
	ないといけない	
		ないといけません
	なければいけない	
		なければいけません
	なければならない	
		なければなりません
	なくてはいけない	
		なくてはいけません
	なくてはならない	
	ねばいけない	
	ねばならない	ねばなりません
	ざるを得ない	
		ざるを得ません
推　量	かもしれない	
		かもしれません
	かもわからない	
		かもわかりません
試　行	てみる	
やりもらい	てもらう	
	てもらえる	
	ていただく	
	ていただける	
	てやる	
	てあげる	
	てくれる	
	てくださる	
アスペクト	てある	てございます
	ている	ていらっしゃる
	ておる	
	てしまう	
	ておく	
	ていく	てまいる
	ていける	
	てくる	てまいる

［付録 3.3］　付属要素（接頭的要素）

語	備　　考
相	※「相乗り」は除く。
御（お）	※次の場合は後の部分と併せて 1 最小単位とする。 おかげ　おかず　おかま　おさげ　おしゃれ　おたふく　おでき　おとぎ　おなか　おにぎり おふくろ　おまえ　おまけ　おまわり（さん）　おむつ　おもらし　おやつ
御（おん）	
各	※ 1 字漢語と結合したものは除く。
今	※ 1 字漢語と結合したものは除く。
御（ご）	※次の場合は後の部分と併せて 1 最小単位とする。 御殿　御飯　御免　御覧
諸	※ 1 字漢語と結合したものは除く。
全	※ 1 字漢語と結合したものは除く。
対	※ 1 字漢語と結合したものは除く。
本	※ 1 字漢語と結合したものは除く。
御（み）	※次の場合は後の部分と併せて 1 最小単位とする。 神籤　巫女　神輿　大御

［付録 3.4］　付属要素（接尾的要素）

語	備　　考
合う	※「共に〜する」「互いに〜する」という意のもの。
上がり	
致す	
上 (うえ)	
得 (え) る	※「…することができる」という意のもの。
終える	
遅れる	
終わる	※「すっかり〜する」という意のもの。
化	※ 1 字漢語と結合したものは除く。
掛かる	※動作・作用があるものに向けられるという意のもの。
がかる	
掛ける	※「途中でやめる」「〜し始める」という意及び動作や作用をあるものに向けるという意のもの。
方 (かた)	※「しかた（仕方）」は除く。
型 (がた)	※ 1 字漢語及び和語の 1 最小単位と結合したものは除く。
方 (がた)	※複数を表すもの。おおよそそのぐらいであることを表すもの。
難 (がた) い	
勝 (が) ち	
がてら	
兼ねる	
がましい	
がる	※助動詞「たがる」は除く。
交わす	※「互いに〜する」という意のもの。
間 (かん)	※　1 字漢語と結合したものは除く。
切る	※「すっかり〜し終える」という意のもの。
臭い	※望ましくない意を強める用法のもの。「かび臭い」「焦げ臭い」は除く。
下さる	
君 (くん)	
気 (げ)	
系	※ 1 字漢語と結合したものは除く。
後 (ご)	※ 1 字漢語と結合したものは除く。
ごと	※「…も一緒に」の意。
毎 (ごと)	※そのもの一つ一つ，その時その時という意のもの。

［付録 3.4］　付属要素（接尾的要素）（続き 1 ）

語	備　　考
熟 (こな) す	※「うまく～する」という意のもの。
さ	※「なさ」「よさ」は除く。ケシ型形容詞語幹に接続する「さ」は除く。
様 (さま)	
さん	
時 (じ)	※ 1 字漢語と結合したものは除く。
式	※形式・方法などの意のもの。 1 字漢語と結合したものは除く。
染 (じ) みる	
中 (じゅう)	※ 1 字漢語と結合したものは除く。
上 (じょう)	※ 1 字漢語と結合したものは除く。
状	※「～の形」という意のもの。 1 字漢語と結合したものは除く。
過ぎる	
尽くめ	
為る	※ 1 字漢語と結合したものは除く。
性	※ 1 字漢語と結合したものは除く。
そう	※一般に，様態の助動詞「そうだ」及び伝聞の助動詞「そうだ」の語幹とされるもの。
損なう	
そびれる	
対	※ 1 字漢語と結合したものは除く。
出す	※動作を始める意のもの。
達	
給う	
だらけ	
たらしい	
ちゃん	
中 (ちゅう)	※ 1 字漢語と結合したものは除く。
尽くす	※「十分に～する」という意のもの。
付き	
っこ	※「…比べ」及び「互いに…すること」という意のもの。
っこい	
続く	
続ける	
辛 (づら) い	
的	※ 1 字漢語と結合したものは除く。
出来る	
等 (とう)	
同士	
通す	※「ずっとし続ける」という意のもの。
所 (ところ)	
殿 (どの)	
共 (とも)	※全部の意のもの。
共 (ども)	※へりくだる意味を表すものも含む。
内 (ない)	※ 1 字漢語と結合したものは除く。
乍ら	
為さる	
並 (なみ)	※その類と同じ，あるいは同じ程度であることを表すもの。
形 (なり)	※そのもの相応である様の意のもの。「～するまま」「～するに従う様」という意のもの。
慣れる	
難 (にく) い	※醜悪の意の「みにくい」は除く。
抜く	※「終わりまでする」という意のもの。

［付録 3.4］　付属要素（接尾的要素）（続き２）

語	備　　考
始める	※その動作をやり出すという意のもの。
果たす	※「すっかり～し終える」という意のもの。
果てる	※「すっかり…する」「…し終わる」「完全に…してしまう」という意のもの。
放し	
版	※１字漢語と結合したものは除く。
風 (ふう)	※様子の意のもの。１字漢語と結合したものは除く。
振 (ぶ) り	※時日の過ぎ去った程度の意のもの。形・姿・様子の意のもの。
分 (ぶん)	
ぽい	※形容詞に接続するものは除く。
ぽっち	
前 (まえ)	
捲る	
間違う	
間違える	
周り	
みたい	
向き	
向け	
目	※順序を示すもの。中心となる点や場所の意及び物の程度の意のもの。
めく	※擬態語的なものは「めく」を切り出さない。
易い	
良い	
様 (よう)	※一般に助動詞「ようだ」の語幹とされるもの。方法の意。
用	※１字漢語と結合したものは除く。
等（ら）	※複数を表す。事物をおおよそに指す。
らしい	※助動詞「らしい」は除く。
流	※１字漢語と結合したものは除く。
類	※１字漢語と結合したものは除く。
忘れる	
渡る	※「辺り一面～にする」という意のもの。

［付録 3.5］　品詞情報一覧

品　詞	活用の種類	活　　用　　形	その他の情報		
			1	2	3
名　詞			固有名詞 数詞	融合	メタ
代名詞				融合	メタ
形状詞				融合	メタ
連体詞				融合	メタ
副　詞				融合	メタ
接続詞				融合	メタ
感動詞					メタ
動　詞	○行五段 ○行上一段 ○行下一段 カ行変格 サ行変格 ザ行変格	未然形・連用形・終止形・連体形・ 仮定形・命令形・語幹		○音便 ○音便A 融合 省略	メタ
	文語○行四段 文語○行上二段 文語○行下二段 文語カ行変格 文語サ行変格 文語ナ行変格 文語ラ行変格	未然形・連用形・終止形・連体形・ 已然形・命令形・語幹			
形容詞	形容詞型	未然形・連用形・終止形・連体形・ 仮定形・命令形・語幹		○音便 ○音便A 融合	メタ
	文語形容詞型 1 文語形容詞型 2 文語形容詞型 3	未然形・連用形・終止形・連体形・ 已然形・命令形・語幹		○音便	メタ
助動詞		未然形・連用形・終止形・連体形・ 仮定形・命令形・語幹		○音便 ○音便A 融合 省略	言いよどみ メタ 連語
	文語	未然形・連用形・終止形・連体形・ 已然形・命令形・語幹		○音便 融合 省略	言いよどみ メタ
助　詞			格助詞 準体助詞 接続助詞 係助詞 副助詞 終助詞	融合	言いよどみ メタ 連語
接頭辞					
接尾辞	〔無活用の接尾辞〕				言いよどみ メタ
	〔動詞性接尾辞〕 ○行五段 ○行上一段 ○行下一段	未然形・連用形・終止形・連体形・ 仮定形・命令形・語幹		○音便 ○音便A 融合 省略	言いよどみ メタ
	〔形容詞性接尾辞〕 形容詞型	未然形・連用形・終止形・連体形・ 仮定形・命令形・語幹		○音便 ○音便A 融合	言いよどみ メタ
	文語形容詞型 1 文語形容詞型 2	未然形・連用形・終止形・連体形・ 已然形・命令形・語幹		○音便 融合 省略	言いよどみ メタ
記　号					メタ
言いよどみ					