国立国語研究所学術情報リポジトリ

『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

メタデータ	言語: Japanese 出版者: 国書刊行会 公開日: 2019-03-25 キーワード: 『日本語話し言葉コーパス』, 文節, 長単位, 短単位, 最小単位 作成者: 小椋, 秀樹, 山口, 昌也, 西川, 賢哉, 石塚, 京子, 木村, 睦子 メールアドレス: 所属:
URL	https://doi.org/10.15084/00002132

『日本語科学』16(2004年10月) 93-113

『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

小椋 秀樹 (国立国語研究所)
山口 昌也 (国立国語研究所)
西川 賢哉 (国立国語研究所)
石塚 京子 (埼玉大学大学院生)
木村 睦子 (国立国語研究所)

キーワード
『日本語話し言葉コーパス』, 文節, 長単位, 短単位, 最小単位

要 旨

『日本語話し言葉コーパス』では, 形態論的な単位として, 品詞の分布などの計量研究によって資料の特徴を明らかにするための長単位と, 用例を採集し, 話し言葉の語彙・語法の研究を行うための短単位の2種類の単位を採用した。本稿では, この2種類の単位の設計方針及び認定基準の概略について述べることとする。

1. はじめに

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, 平成11年～15年の5年間にわたって,「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, 本稿の標題にもある『日本語話し言葉コーパス』(*The Corpus of Spontaneous Japanese* 以下 CSJ とする) である[1]。

本稿では, CSJ で採用した長短2種類の単位について, その設計方針や認定基準の概略を紹介することとする。

2. 単位の設計

2.1. 語彙調査の調査単位

国語研究所は, 現在進行中のものも含めると, これまでに合計10回の語彙調査を実施してきた。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り得るため, 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

表1　国語研究所の語彙調査の調査単位

語　彙　調　査	長い単位の系列	短い単位の系列
語彙調査―現代新聞用語の一例―	名称なし	
現代の語彙調査　婦人雑誌の用語	α単位	
現代の語彙調査　総合雑誌の用語		β単位
現代雑誌九十種の用語用字		β単位
電子計算機による新聞の語彙調査	長単位	短単位
高校教科書の語彙調査	W単位	M単位
中学校教科書の語彙調査	W単位	M単位
雑誌用語の変遷	長い単位	
テレビ放送の語彙調査	長い単位	
現代雑誌200万字言語調査		短単位

語彙調査の目的に対し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[2]。そのため，表1に示すように，語彙調査ごとに異なる単位が使われてきた[3]。

調査単位の設計に当たって，このよう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現象を，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現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。

2.2. 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に現れた言語現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」と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がたい。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」の例を比較してみることとする[4]。

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。

【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，国語研究所がこれまでに行った語彙調査の調査単位(表1に挙げた調査単位)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[5]。一方，短い単位(以下，短単位と呼ぶ。)については，現代雑誌九十種の語彙調査で用いられた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[6]。

3. 長単位・短単位の認定基準

ここでは，CSJ において採用した長単位・短単位の単位認定基準について，その概略を説明することとする。なお以下，単位等の境界を示すために，次の記号を用いることとする。

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。

3.1. 長単位の認定基準

3.1.1. 文節の認定

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１・付録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同じではない。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。

文節では，一般に付属語又は付属語連続の後で文節が切れることになるが，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した。

【例】 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　有りの=儘　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文節以上からな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した。この規則によって，以下の例の「等」「型」「各」のように接辞のみで一つの文節が構成される場合もあり得るということになる。

【例】　‖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，西川ほか(2004)を参照されたい。

3.1.2. 長単位の認定

長単位は，以下に掲げた規則によってテキストを分割し，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。ただし，3.1.1で規定した文節を超えないものとする。以下，長単位の認定の際に問題となる点について説明しておく。

[1]　付属語（付録１・付録２に示した複合辞を含む。）は１長単位とする。

【例】　|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う。

【例】　|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・分数の読み上げ・動植物名及び「—の～」「—が～」の体言句の内部にある助詞・助動詞は切り出さない。

【例】　|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。

【例】　|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，その体言全体に係る体言・接辞がある場合は切らない。また並立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。

【例】 ｜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。

【例】 ｜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。

【例】 ｜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。

【例】 ｜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。

【例】 ｜御=会い=する｜ ｜御=与え=ください｜ ｜御=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。

【例】 ｜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。

【例】 ｜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※印を付けた例と同様の形式については，数量を表す要素と前の要素とを受ける体言がある場合は，切り離さない。

【例】 ｜果汁=百パーセント=オレンジジュース｜

3.2. 短単位の認定基準

3.2.1. 最小単位の認定

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。

【例】 和　語　：　／話し／言葉／　　／お／話し／し／ます／
／大／雨／が／降っ／た／の／で／
漢　語　：　／国／語／　　／研／究／
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／　※　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。

「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。「ていく」「について」などの複合辞も「／て／いく／」「／に／つい／て／」のように最小単位を認定する。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。

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。

3.2.2. 短単位の認定

まず，3.2.1で規定された最小単位を表２(次ページ)のように分類する。

以下，付属要素，数，助詞・助動詞について少し説明しておく。

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」(付録３・付録４)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。

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」も含めた。また数詞のうち，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，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。

表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　御　各　御　……
	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　……

短単位の認定基準は，表2の各分類ごとに適用すべき規則が定められている。その規則のうち，短単位認定の基本原則に当たるのが，一般の最小単位に適用される以下の規則である。

[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。

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

この結合に当たっては長単位を超えないという制約を設けている。これによって，長単位の下位に短単位が位置付けられるという階層構造を持つことになる。

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。

【例】　｜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。

【例】　｜ぐる｜ぐる｜ぐる｜っと｜回る｜　｜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。

【例】　｜ぐるぐる｜っと｜回る｜　｜ぐるぐる｜ぐるぐる｜っと｜回る｜

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。

【例】　｜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。

【例】 ｜仏教｜　　｜儒教｜　　｜阪大｜

④ 外来語の最小単位うち英語の接続詞・前置詞・冠詞に当たるもの。

【例】 ｜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。

【例】 ｜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。

【例】 ｜ホワット｜アー｜ユー｜ドゥーイング｜ヒア｜　　｜イッツァ｜ペン｜

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。

【例】 ｜内閣｜府｜　　｜副｜大統領｜　　｜光ファイバー｜網｜　　｜自衛｜隊｜

⑧ 単独で文節を構成する最小単位。

【例】 ｜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。

【例】 ｜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。

【例】 ｜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。

【例】「―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対する短単位認定規則を示す。

[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。

【例】 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。

【例】 ｜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｜

3.3. 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，次のような融合・省略・フィラー・言いよどみ・言い直しという現象がある。

融　合　：　そりゃ　面白きゃ　食べりゃ　じゃ　てる
省　略　：　やんだっけ　そうっす
フィラー　：　えー　んーと
言いよどみ　：　こ ここから
言い直し　：　国立日本語 国語研究所

以下，これら話し言葉特有の現象の単位認定について説明する。

3.3.1. 融合・省略の処理

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採られたものでもある[7]。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することとした。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。

省略についても，元の形に戻すことなく，可能な範囲で単位分割した。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。

3.3.2. フィラー・言いよどみの処理

フィラーについては，「(F あの)」「(F えーと)」のようにＦタグが付されているので[8]，長単位・短単位ともに，そのＦタグが付された範囲を１単位とした。ただし，以下のように助詞・助動詞を含む場合，長単位ではＦタグが付されている範囲全体で１単位としたが，短単位では，助詞・助動詞を切り出した。

【例】 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。

【例】 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，「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）という規則を立て，

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というものである。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，出現した単位のテキスト上の位置を変えるということは不可能である。

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。つまり，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。

【例】 ｜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。

言いよどみについても，フィラーと同様にタグ(D タグ)の付された範囲全体を１長単位又は１短単位とした。また，１長単位又は１短単位の中に現れる言いよどみについても，フィラーと同様に無視して単位認定を行った。

【例】 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｜風｜に｜

なお，言いよどみのうち，数詞・助詞・助動詞・接頭辞・接尾辞の言いよどみ(D2タグが付されたもの)については，通常の助詞・助動詞・接頭辞・接尾辞と同様に単位認定を行った。

【例】 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｜　｜六十(D２二)二パーセント｜の｜

短単位　：　｜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｜
｜六十｜(D2二)｜二｜パーセント｜の｜

3.3.3. 言い直しの処理

言い直しについては，言いよどみ，特にD2タグが付されたものと重なる部分がある。しかし，ここで取り上げる言い直しは，例えば以下の下線部，

益岡・田窪氏の基本日本語基礎日本語文法(D2の)での

のように，D2タグが付されていないものを指す。

言い直しについては，以下の四つに分類し，それぞれについて単位認定の方法を定めた。

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。

【例】 ｜益岡田窪氏｜の｜基本日本語｜基礎日本語文法｜
｜太平洋開戦｜太平洋戦争開戦｜の｜年｜に｜
｜高原農家｜高原野菜農家｜で｜働い｜ている｜

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。

【例】 ｜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。

【例】 ｜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。

【例】 ｜国立=日本語=国語研究所｜で｜
｜(A エイチピーエスジー；HPSG)｜に=基づい=(D だ)=基づいた｜

短単位では，上記の言い直しの例は，例えば，

｜益岡｜田窪｜氏｜の｜基本｜日本｜語｜基礎｜日本｜語｜文法｜
｜阪倉｜篤義｜さん｜篤義｜先生｜の｜
｜向こう｜で｜教育｜機関｜教育｜事業｜始め｜たい｜と｜いう｜こと｜で｜
｜国立｜日本｜語｜国語｜研究｜所｜で｜

のように分割されるため，言い直しをどのように処理するかという問題は起こらない。

以上，本節では長短2種類の単位認定基準の概略について述べた。その規定により長単位・短単位を認定した例を次に示す。

【例】

〔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|

4. まとめ

以上，CSJ における長単位・短単位の設計方針と認定基準の概略について紹介した。

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，その目的に応じて二種類の単位を設計した。しかし，その設計に当たっては，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。

なお最後に，現時点において単位認定基準の中で再検討を要すると思われる事項について，簡単に述べておく。

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。

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・付録 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全く別の規則を考えることが必要であろう。

上記以外にも，長単位・短単位の認定基準について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，単位認定の問題についても検討を行い，より良いコーパスの単位を提案していきたいと考えている。

注

1 CSJの概要については，前川(2004)を参照。

2 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。

3 表1に挙げた各調査単位の概要については，林(1982:582-583)，中野(1998:171-172)を参照。なお，表1を見ると，一部の語彙調査で同じ名称の調査単位が用いられていることがあるが，高校教科書の語彙調査・中学校教科書の語彙調査で同じW単位・M単位を採用している以外は，同じ名称であっても全く同じ単位というわけではない。例えば，総合雑誌の用語と現代雑誌九十種の用語用字とは，共にβ単位を採用しているが，総合雑誌の用語では助詞・助動詞を調査対象外としているのに対し，現代雑誌九十種の用語用字では助詞・助動詞も調査対象としているという違いがある。また，雑誌用語の変遷とテレビ放送の語彙調査とは，共に長い単位を採用しているが，テレビ放送の語彙調査では助詞・助動詞に複合辞を含めていること，人名・地名以外にも固有名詞を広く採っていることから雑誌用語の変遷の調査単位よりも長くなっている。

4 ここで「言語」の用例を採集するために用いたデータは，CSJのうち手作業で単位解析を実施した約100万語である。

5 長い単位については，国立国語研究所(1995:49-63)を参照。

6 β単位については，国立国語研究所(1962:6-14)を参照。

7 国立国語研究所(1962:7)を参照。

8 CSJの書き起こしテキストの仕様については，小磯ほか(2001)を参照。

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）

参考文献

小磯花絵・土屋菜穂子・間淵洋子・斉藤美紀・籠宮隆之・菊池英明・前川喜久雄(2001)「「日本語話し言葉コーパス」における書き起こしの方法とその基準について」『日本語科学』9，43-58，国書刊行会.

国立国語研究所(1962)『国立国語研究所報告21　現代雑誌九十種の用語用字(1)』秀英出版

国立国語研究所(1987)『国立国語研究所報告89　雑誌用語の変遷』秀英出版.

国立国語研究所(1995)『国立国語研究所報告112　テレビ放送の語彙調査Ⅰ』秀英出版.

国立国語研究所(2001)『現代語複合辞用例集』国立国語研究所.

中野洋(1998)「言語の統計」『岩波講座言語の科学9　言語情報処理』149-199，岩波書店.

西川賢哉・小椋秀樹・相馬さつき・小磯花絵・間淵洋子・土屋菜穂子・斉藤美紀「文節の仕様について　Version 1.0」(『日本語話し言葉コーパス』公開版添付文書)，国立国語研究所.

林大監修(1982)『角川小辞典9　図説日本語』角川書店.

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』15，111-133，国書刊行会.

前田富祺(1985)『国語語彙史研究』明治書院.

小椋　秀樹（おぐら　ひでき）
国立国語研究所研究開発部門
115-8620　東京都北区西が丘3-9-14
ogura@kokken.go.jp

山口　昌也（やまぐち　まさや）
国立国語研究所研究開発部門

西川　賢哉（にしかわ　けんや）
国立国語研究所研究開発部門

石塚　京子（いしづか　きょうこ）
埼玉大学大学院文化科学研究科博士後期課程

木村　睦子（きむら　むつこ）
国立国語研究所日本語教育部門

［付録１］ CSJ で認定した複合辞（助詞相当句）

基本形	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	をもとにした	
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２］ CSJ で認定した複合辞（助動詞相当句）

種　類	基　本　形	丁　寧　形	その他の異形態
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	んだ んです んである んでございます
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	じゃない じゃありません じゃございません
	のではない	のではありません	のでない のじゃない んではない んじゃない
許可・依頼・勧誘	てもいい	てもよろしい	ていい たっていい
	てほしい		
禁止・当然・義務	てはいけない	てはいけません	ちゃいけない ちゃいけません
	てはならない		てはならぬ ちゃならない ちゃならぬ
	ないといけない	ないといけません	ないといけぬ
	なければいけない	なければいけません	なきゃいけない なけりゃいけない なきゃいけません
	なければならない	なければなりません	なきゃならない なきゃならぬ なけりゃならない なきゃなりません
	なくてはいけない	なくてはいけません	なくちゃいけない
	なくてはならない		なくちゃならない
	ねばいけない		ねばいけぬ
	ねばならない	ねばなりません	ねばならぬ
	ざるを得ない	ざるを得ません	ざる得ない
推　量	かもしれない	かもしれません	かもしんない
	かもわからない	かもわかりません	かもわかんない かもわからぬ
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		てける
	てくる	てまいる	

［付録３］ CSJ で認定した付属要素（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４］ CSJ で認定した付属要素（接尾的要素）

語	備　　考
合う	※ 「ともに～する」「たがいに～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」「～しはじめる」という意及び動作や作用をあるものにむ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	※ 「たがいに～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。「かびくさい」「こげくさい」は除く。
下さる	
君（くん）	

語	備　　考
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	※「…もいっしょに」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。
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	※「…くらべ」及び「たがいに…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	

語	備　　考
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」「～するに従うさま」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。
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	※　「あたり一面～にする」という意のもの。