話し言葉コーパスにおける単位切りと品詞付与の方法

柏野和佳子　小椋秀樹　田中牧郎　加藤安彦
国立国語研究所

1．はじめに

我々は，科学技術振興調整費開放的融合研究推進制度課題「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」（平成 11〜15 年度）［前川・他 2000］の一環で，言語学的に整備された話し言葉コーパスの構築を進めている。ここで言語学的に整備されたコーパスとは，言語学的に裏づけされ，工学的応用上も扱いやすい単位で統一的に区切られ，体系的に品詞が付与されているものをいう。すなわち，十分な単位切りと品詞付与の検討が重要になる。そこで本稿では，単位切りと品詞付与について検討する。

2．単位切りの方針

2.1 さまざまな単位

単位切りは，コーパスの利用目的によって望ましいものが異なる。たとえば，国語研究所のこれまでの語彙調査では，調査対象や調査目的にあわせて単位が検討され異なる単位が用いられてきた。表1に代表的なものを挙げる。

単位は，短めの単位と長めの単位とに大別される。長めの単位では全体で1つの単位になる「国語研究所」は，短めの単位では「国語」「研究」「所」と，3つの単位から構成されると捉えられる。なお，表1のとおり，新聞3紙と中学・高校教科書の調査では，その長短両方の単位を同時に使用している。

2.2 短めの単位の特徴

M単位，β単位は短めの単位である。単位切りの規則が明解であるためゆれが少ないといわれてきた。その規則は，まず最小単位を規定し，そして，その最小単位が単独で，もしくは，2つの最小単位が一次結合したもので1単位になると定めるものである。最小単位とは，現代語で意味を担う最小の言語単位であり，ほぼ形態素に相当するものである。なお，漢語については，漢字一字で表されるものが1最小単位と規定されている。また，結合とは，最小単位どうしの意味的なつながりで，その順番により，一次結合，二次結合…となっていくものである。

たとえば，「接辞」の場合は，1最小単位単独で1単位になると定められている。「思いかねる」という派生動詞は，「思い」で1単位，接辞「かねる」で1単位と分割される。一方，漢語の場合は，一次結合で1単位になり，二次結合以上は分割される。「自転車」という漢語は，「自」と「転」という最小単位が一次結合し，さらに「車」という最小単位が二次結合した語であるとみる。よって，「自転」で1単位，「車」で1単位である。

短めの単位は基本語彙を選定することを目的として規定されたものである。たとえば，「自転車」が「自転」と「車」とで各1単位になるのと同様に，「自動車」の場合は「自動」と「車」とで各1単位になる。単位ごとに使用率を調査し，「自転」「自動」に比べ「車」の使用率が高く出れば，そこから「車」という語は基本度が高いといったことが導き出せる。

表1　国立国語研究所の語彙調査の単位

	単位	単位切りの例[1]	語彙調査
短め	M単位	型｜紙｜どおり｜に｜裁断｜し｜て｜外出｜着｜を｜作り｜まし｜た｜.	中学・高校教科書［国語研 1983, 1986］
	β単位[2]	型紙｜どおり｜に｜裁断｜し｜て｜外出｜着｜を｜作り｜まし｜た｜.	総合雑誌［国語研 1957］, 雑誌 90 種［国語研 1962］, 新聞3紙［国語研 1970］
長め	長単位	型紙どおり｜に｜裁断し｜て｜外出着｜を｜作り｜ました｜.	新聞3紙［国語研 1970］
	W単位	型紙どおり｜に｜裁断して｜外出着｜を｜作りました｜.	中学・高校教科書［国語研 1983, 1986］
	α単位	型紙｜どおりに｜裁断して｜外出着を 作りました｜.	婦人雑誌［国語研 1953］
	長い単位	型紙どおりに｜裁断して｜外出着を 作りました.	雑誌用語［国語研 1987］, テレビ放送［国語研 1995］

[1] ［中野 1998］より引用。
[2] 新聞の調査では「短単位」と呼ばれていたが，内容はほぼβ単位に同じなのでまとめた。

以上みたように，短めの単位というのは定義が明確であり，基本語や語構成について調べる場合にふさわしい単位であることがわかる。しかしながら，我々が単語という概念で直感的に捉えるものは，「自転車」や「自動車」という一かたまりの単位であって，「自転」，「自動」，「車」などと分けた単位ではない。短めの単位は，しばしば単語という概念よりも短かすぎる印象をもたせるものが多い。

M単位とβ単位とでは，和語や外来語の扱いに違いがある。β単位では漢語と同じく最小単位の一次結合で1単位になるが，M単位では1最小単位で1単位になる。よって，β単位では「母親」「行き過ぎる」「カラーコピー」で各1単位になるのに対し，M単位では「母」「親」「行き」「過ぎる」「カラー」「コピー」で各1単位になる。M単位は細か過ぎるため，文意から離れてしまう場合もある。

2.3 長めの単位の特徴

長単位，W単位，α単位，長い単位はいずれも長めの単位である。これらの基本的な規則は，まず文節に切るというものである。文節を客観的に認定することは難しいため定義が必要になる。たとえばW単位の規定の中では，文節とは「修飾・並列・接続・中止・独立などの構文上の機能を持った最小の要素」であると定義されている[国語研 1983, 1986]。

文節に続けてどう切るかの具体的な規則は単位によって異なる。長単位とW単位では，自立語と付属語とで必ず切られる。他方，α単位と長い単位では，文節がそのまま単位になり，自立語と付属語とで必ずしも切られない。さらに，助詞，助動詞の扱いが単位ごとに異なっている。しかし，いずれの長めの単位も先に述べたとおり，「国語研究所」などの語のまとまりは切らずに1単位になるよう定められている。

長めの単位は，語の使用の実態を正確に捉えることを目的として規定されたものである。たとえば，「携帯電話」という言葉は近年生まれた新規の言葉であるが，「携帯」と「電話」とに切ると，それぞれの単語には新規性がなくなる。新語を把握するには長めの単位が適している。また，「国語研究所」などの固有名や，「音声認識処理」「二酸化炭素」などの専門用語の場合も，その語が使用されている本来の性格を捉えるためには長いままに1単位にすることが望まれる。

しかしながら，基本語や語構成を調べるには，この長めの単位は適さない。

2.4 本コーパスで採用する単位

以上みてきたとおり，短めの単位，長めの単位には，それぞれに利点と欠点とがある。そこで，新聞3紙や，中学・高校教科書調査のように（表1参照），本コーパスでも両者を併用することで，利点を生かし，欠点を補いあうという方法をとることにする。具体的には，過去の単位を参考にし，「短単位」と「長単位」を新たに規定して使用する。

「短単位」の規則は現在ほぼ完成している。文意から離れない程度に短く切るβ単位とほぼ同じに規定している。「長単位」の規則は現在作成中である。自立語と付属語とを別々に扱いたいため，α単位や長い単位ではなく，長単位とW単位とを参考にしている。

なお，音声認識などでは，その処理精度を上げるために，できるだけ長い単位の設定が望ましいといわれている。たとえば「せざるを得ない」で1単位としたいといった場合，文節を超える場合の定義を新たに定める必要が出てくる。さらに長い単位の検討は今後の課題である。

3．品詞付与の方針

品詞は，短単位，長単位のそれぞれに付与する。そうすることによって，単位によって品詞が異なるような場合も問題にならない。たとえば，「国語研究所」が，短単位で「国語」「研究」「所」と3つに分割されているところには，各々の品詞にすべて普通名詞を付与する。そして，長単位で1つになっているところには，その品詞に固有名詞を付与する。

現在，短単位についての品詞体系案から先に作成している。その大枠は，表2のとおりである。

表2　短単位の品詞体系の大枠

名詞［固有名，人名，サ変，形状，副，メタ（引用）］				
代名詞	副詞	形容詞	動詞	連体詞
接続詞	感動詞	助動詞	助詞	接頭辞
接尾辞	数詞	助数詞	記号	
無品詞［フィラー，言いよどみなど］				

全体では，上記に加え，文語，活用型，活用形，縮約，音便などまであわせた，3階層からなる体系を組んでいる。

品詞体系を設計する際に特に問題になるのは，文法体系によって扱い方が異なりやすい，次のところであろう。

・ 名詞と副詞の線引き
・ 形容動詞という品詞の認定
・ 接頭辞，接尾辞という品詞の認定とその範囲
・ 助詞・助動詞の範囲

これに加えて，「用法」の明示という問題がある。そもそも，品詞というものは語を文法的性質によって分類したときの「語分類」を示すものである。しかし，語によっては品詞をまたがる用法をもつものがあるため，その明示をどうするかが問題になる。

先に，名詞と副詞の線引きとをあわせ，用法の明示の問題を取り上げ，その対応方針を述べる。その後，残りの問題に関する方針を述べる。

3.1 名詞と副詞の線引きと用法の明示

たとえば，『岩波国語辞典（第5版）』によれば，「実際」と「しばしば」という語は次のように扱われている。

#1 【名詞】［実際］の利益
#2 【名詞】《副詞的に》［実際］そうなる
#3 【副詞】［しばしば］しかられた

このような場合に，「語分類」を示すという態度だけをとるなら，語ごとに同じ品詞を付与することになる。しかし，「実際」のように名詞と副詞的用法の両方をもつ語の品詞を，「名詞」とするか「副詞」とするか，副詞用法をもつものとして名詞を細分類する「名詞-副」とするかで，次の3通りの付与があり得る。

A: #1 名詞 #2 名詞 #3 副詞
B: #1 副詞 #2 副詞 #3 副詞
C: #1 名詞-副詞 #2 名詞-副詞 #3 副詞

形態素解析ツール「茶筌(version 2.0 for Windows)」[3]や「実践 JUMAN」[4]の解析では，岩波国語辞典とは副詞とする語の範囲が違うため，上記Bのパターンで出力された。

一方，「用法」を明示するという姿勢を完全にとろうとするなら，次のようになるだろう。

D: #1 名詞 #2 副詞 #3 副詞

[3]http://cl.aist-nara.ac.jp/lab/nlt/chasen/distribution.html
[4]http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/juman-form.html

我々は，実際の用法を明示した方が研究利用には望ましいと考えるため，その点では#2を#1とは区別したい。しかし，品詞体系はあくまでも「語分類」を示すものであるところに意義があると考える。そこで，副詞用法をもつ語については，他に連体修飾をする用法や，連体修飾を受ける用法のない語のみ「副詞」とし，他に名詞の用法がある語は「名詞」の下位分類の「名詞-副」とする品詞体系を組んでいる。「しばしば」や「やはり」「すっかり」といった語は前者に，「実際」や「かなり」「たくさん」といった語は後者に該当する。よって，先の#1，#2，#3，はすべて区別して，次のように付与する。

E: #1 名詞 #2 名詞-副 #3 副詞

なお，品詞体系の設計では，実際の品詞付与作業の実現性や安定性について考慮することも重要である。機械付与と違って，人手付与の場合は，実際の用例が「名詞」の用法か「副詞」の用法かを判断することはそう難しいことではない[荻野・他 1997]。しかし，実際の用例が副詞的である場合に，常に，辞書引きや内省によって「名詞-副」か「副詞」かの判定を強いることは，作業者へ負担を与えてしまうことである。このような品詞判定については機械的に支援する方法を検討中である。

なお，[水谷・星野 1994]は，名詞から副詞にわたる1271見出しの用法を分析し，94にものぼる詳細な語類を提示している。そのような詳細な語類に基づいて品詞体系を設計することは一つの理想であるが，当面の作業効率を考えた場合に実現は難しいため，我々は，副詞用法のある語については「名詞-副」か「副詞」かの二分類を設けるのみにとどめる。詳細な語類の検討はコーパス構築ののちの研究課題としたい。

3.2 形容動詞という品詞の認定

いわゆる形容動詞については，名詞と助動詞に分割して扱い，形容動詞という単独の品詞は認めない。当該用例が，「だ，な，に，の，さ」に後続する場合に，名詞の下位品詞として設ける「名詞-形状」を付与する。それ以外に，格助詞などに後続する名詞用法の場合は「名詞」を付与する。以下に例を示す。

#4［困難］が生じた　　：名詞
#5［困難］な話しだ　　：名詞-形状

3.3 接頭辞，接尾辞という品詞の認定とその範囲

接頭辞（接頭語），接尾辞（接尾語）は品詞ではないという立場もあるが，短単位の場合は，単独で1単位になるものとし，その語に付与する一つの品詞として扱う。長単位では，すべての接頭辞・接尾辞は，連なる語に接続させて1単位になるのでこの品詞はなくなる。名詞，動詞，形容詞，副詞につく，生産性の高いものを接頭辞，接尾辞としようとしているが，その認定範囲は，以下に示すとおり，現時点ではかなり狭めており，検討途中である。

《接頭辞》〈未完〉
御［お，おん，み，ご］，あい（相なる）
《接尾辞》〈未完〉
敬称：さま，さん，はん，ちゃん，どの，どん，くん，うえ（母上），ぎみ（父君）
複数：たち，かた，がた，ども，ら
関係：どち，どし，どうし，じゅう（うち～）
状態：そう（あり～だ、ある～だ），だらけ，がち，ぶり（男～），らしい，めかしい，がましい，くさい，ぽい，がたい，づらい，にくい，かねる
変化：めく，じみる，しな
行為：めかす，ぶる，がる，やがる
他：あて，ごと，ぐるみ，ずく，がてら，だてら

3.4 助詞・助動詞の範囲

助詞・助動詞とする語の範囲は，国語研の過去の語彙調査でもゆれの大きなところである。今回は，短単位と長単位という2種類の単位を併用するため，短単位での範囲は以下に示すとおり，狭めに考えている。助詞相当の連語類等は，長単位の方で多く認める予定でいる。

《助詞》〈未完〉
格助詞：が，から，って，で，と，に，の，へ，より，を
並立助詞：か，と，なり，に，の，も，や，やら
副助詞：か，すら，ぞ，だけ，だに，たり，ながら，なぞ，など，のみ，ばかり，ほど，まで，も
係助詞：こそ，たら，って，は，ば，も，や
接続助詞：が，から，けれども(けれど・けど)，し，って，て，で，と，ど，に，ば，も
終助詞：い，え，か，かしら，かな，け，ぜ，ぞ，って，な，の，やら，わ
間投助詞：さ，たら，な，ね，や，よ
《助動詞》〈未完〉
せる，させる，しめる，れる，られる，たい，らしい，だ（形容動詞活用語尾を含む），です，ます，ない，ぬ，た，う，よう，まい，べし，ごとし

4．おわりに

本稿で述べた，長短の2種類の単位切りを施し，品詞もそれぞれに付与するという方法によって，言語学的に整備された，研究利用価値の高い話し言葉コーパスの構築を目指す。規模としては短単位で数えた時，約700万形態素分のものを想定している。

今後の課題は，長単位の規則と品詞の設定と，長短ともに，話し言葉特有の問題への対処を検討することである。

謝辞

日ごろからご指導を頂く，本プロジェクトの総括責任者である古井貞熙教授（東京工業大学），宮島達夫教授（京都橘女子大学），ほか，共同研究者のみなさまに感謝申し上げます。

参考文献

[荻野・他1997]荻野紫穂・他(1997)「RWCテキストデータベースにおける口語・古語等の扱い」『言語処理学会第3回年次大会発表論文集』pp. 341-344.

[国語研1953]国立国語研究所(1953)『婦人雑誌の用語―現代語の語彙調査』秀英出版.

[国語研1957]国立国語研究所(1957)『総合雑誌の用語 前編―現代語の語彙調査』秀英出版.

[国語研1962]国立国語研究所(1962)『現代雑誌九十種の用語用字 第1分冊―総記および語彙表』秀英出版.

[国語研1970]国立国語研究所(1970)『電子計算機による新聞の語彙調査』秀英出版.

[国語研1983]国立国語研究所(1983)『高校教科書の語彙調査 I』秀英出版.

[国語研1986]国立国語研究所(1986)『中学校教科書の語彙調査 I』秀英出版.

[国語研1987]国立国語研究所(1987)『雑誌用語の変遷』秀英出版.

[国語研1995]国立国語研究所(1995)『テレビ放送の語彙調査1―方法・標本一覧・分析』秀英出版.

[中野1998]中野洋(1998)「4言語の統計」『岩波講座言語の科学9 言語情報処理』pp149-199, 岩波書店.

[前川・他2000]前川喜久雄・他(2000)「共通日本語話し言葉コーパスの設計」『音響学会2000年春季講演論文集』.

[水谷・星野1994]水谷静夫・星野和子(1994)「名詞から副詞まで―語類の新しい枠づけ」『計量国語学』Vol. 19, No. 7, pp. 331-340.