「現代日本語書き言葉均衡コーパス」の長単位認定基準について

冨士池優美† 小椋秀樹† 小木曽智信† 小磯花絵† 内元清貴‡ 相馬さつき† 中村壮範†

† 独立行政法人国立国語研究所
‡ 独立行政法人情報通信研究機構（NICT）

1. はじめに

国立国語研究所では，1976 年から 2005 年までの 30 年間に出版された日本語の書き言葉を対象とする「現代日本語書き言葉均衡コーパス」(以下 BCCWJ）を構築している。BCCWJ の形態論情報については，言語単位として，コーパスからの用例収集に適した「短単位」と，格納したサンプルの言語的特徴の解明に適した「長単位」の 2 種類を採用した。この 2 種類の言語単位に基づいて，更に代表形・品詞等の情報を与える。

長短 2 種類の言語単位のうち，短単位の概要については既に小椋ほか（2007），小椋（2007）で報告した。そこで本稿では，長単位の認定基準の概要等について述べる。また併せて，BCCWJ に格納する中央省庁刊行の白書の長単位解析の現状について報告する。

2. 長単位の概要

BCCWJ の短単位・長単位は，いずれも「日本語話し言葉コーパス」（以下 CSJ）で採用した単位を書き言葉用に修正・拡張したものである。

長単位は文節を基にした単位である。長単位の認定は，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に分割していくという手順で行う。そのため，長単位の認定基準は，文節と長単位，二つの認定基準から成る。

本節では，文節の認定基準，長単位の認定基準，CSJ からの変更点及びコーパスの言語単位としての長単位の長所について述べる。以下，例文中の文節の境界を「┃」，長単位の境界を「｜」とし，注目している境界を「//」，切らないことを示す場合には「-」を，中でも注目している部分には「=」を用いる。また，注目している単位には下線を付す場合がある。

2.1 文節の認定基準

長単位の認定に当たっては，まず文節の認定を行う。

文節は，一般に付属語又は付属語連続の後で切れる。BCCWJ では，CSJ と同様に複合辞も付属語として認めた。文節を認定する上で問題となることの一つに，固有名，動植物名，「―の～」「―が～」で１短単位と認める体言句がある。これらについては，内部にある助詞・助動詞の後では切らないこととする。

┃源=頼朝┃ ┃虎の=門交差点┃
┃タツノ=オトシゴ┃ ┃ユキノ=シタ┃
┃案の=定┃ ┃油絵の=具┃ ┃万が=一┃

2.2 長単位の認定基準

長単位は，文節を規定に基づいて分割する，あるいはしないことによって得られた要素を１単位とする形式であり，文節を超えることはない。

文節と長単位の関係を**表１**に示す。

表１ 文節と長単位との関係

文 節：	また 、┃市街地 と┃耕地 が┃共存し ている┃地域 で は 、┃
長単位：	また｜、｜市街地｜と｜耕地｜が｜共存し｜ている｜地域｜で｜は｜、｜
	いわゆる┃地産地消 や┃肥飼料化さ れ た┃生ごみ の┃活用 が┃行わ れ 、┃
	いわゆる｜地産地消｜や｜肥飼料化さ｜れ｜た｜生ごみ｜の｜活用｜が｜行わ｜れ｜、｜
	地域内 で の┃食 と┃農 の┃連携 が┃進み ます 。┃
	地域内｜で｜の｜食｜と｜農｜の｜連携｜が｜進み｜ます｜。｜

以下，長単位認定基準の概要を示す。

〔1〕 区切り符号は1長単位とする。
｜湾岸戦争後｜、｜英｜、｜仏｜など｜と｜
ただし，区切り符号のうち，①中点等，②数字連続の中に現れるもの，③全体が1短単位となるものの中に現れるものは，1長単位としない。
｜官=・=財｜　｜１７=．=３％｜
｜小=、=中学生｜

〔2〕 語と同じ働きをする記号・記号連続及びそれらを含む結合体は，全体で1長単位とする。
｜２，０００=㎡｜　｜WHO｜　｜ＰＨＳ｜

〔3〕 付属語(複合辞を含む。)は1長単位とする。
｜公害紛争処理法｜における｜公害紛争処理｜の｜手続｜は｜，｜原則｜として｜紛争当事者｜から｜の｜申請｜によって｜開始さ｜れる｜。｜

〔4〕 体言及び副詞に形式的な意味の「する」「できる」「なさる」「いたす」が直接続く場合，体言及び副詞と「する」「できる」「なさる」「いたす」とを切り離さない。
｜往復運動=し｜ている｜
｜きちんと=できる｜

〔5〕 並列の関係にある語は切り離す。
｜公正｜妥当｜な｜実務慣行｜

(1) 並列された語のうち，①中点でつなげている場合，②漢語の最小単位の並列，③和語の最小単位二つが並列した語のうち，『岩波国語辞典』第6版（岩波書店），『国語大辞典』（小学館）のいずれか一方で見出し語になっている語は切り離さない。
｜官=・=財｜　｜前=後｜　｜市=町=村｜
｜あち=こち｜

(2) 並列の関係にある体言連続のうち，並列された体言全体を受ける，若しくはそれら全体に係る体言的な形式や接辞がある場合及び形式的な意味の「する」「できる」「なさる」「いたす」がある場合は切らない。
｜英語=日本語-間｜　｜芸術家=，=文化人等｜
｜新-学年=・=学期｜　｜在学=・=在校する｜

〔6〕 同格の関係にある体言連続は切り離さない。
｜機関誌=計量国語学｜が｜発刊さ｜れ｜

〔7〕 数を表す要素を含む自立語は，以下のように長単位を認定する。

(1) 数を表す要素は，単位の変わり目の後ろで切る。
｜平成｜１５年／／９月／／１５日｜午後｜７時／／３３分｜

(2) 数を表す要素の前で切る。
｜延べ／／２３時間｜３０分｜
ただし，数を表す要素と前の要素とを受ける体言がある場合，数を表す要素と前の要素との間に中点がある場合には，数を表す要素と前の要素とを切り離さない。
｜果汁=百パーセント-オレンジジュース｜
｜７業種=・=４２品目｜

(3) 数を表す要素とそれに続く体言・接辞とは切り離さない。
｜週｜４０時間=勤務｜
｜９６年｜3月｜３１日=以前｜

2.3 CSJからの変更点

（1）記号に関する規定の追加

CSJ の書き起こしテキストには用いられていなかった句読点等，区切り符号を含む記号を1長単位にする規定を追加し，書き言葉に対応した。

（2）数量を表す要素に関する変更

CSJ では数量を表す要素は分割せず一続きとしていたが，長すぎるという指摘があった。

CSJ：　｜１m=８０ｃｍ｜

BCCWJ では前述のとおり，単位の変わり目の後ろで分割することとした。

BCCWJ：｜１ｍ／／８０ｃｍ｜

（3）係り受けが関係する規定の簡素化

CSJ では「体言連続の一部分が連体修飾語を受けている場合，その後ろで切る」「2 文節を受ける，若しくは 2 文節以上に係る接辞はその前後で切る」という規定があった。

CSJ：　｜項構造｜の｜曖昧性／／解消｜
｜円形劇場｜とか｜水路／／等｜

これらは，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ものである。しかし実際に単位分割をする際には，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があり，特に判定が難しい体言連続については例外規定を設ける等，煩雑な規則であった。これが単位認定のゆれにつながっていたため，BCCWJ では規定を簡素化することとした。

BCCWJ：｜項構造｜の｜曖昧性=解消｜
｜円形劇場｜とか｜水路=等｜

2.4 複合辞・連語

BCCWJ では，CSJ と同様に複合辞・連語を1長単位と認めた。複合辞・連語は，現代語の研究や日本語教育でよく取り上げられるものである。国立国語研究所（2001）では複合辞として助詞相当句 83 語，助動詞相当句 42 語を挙げている。またグループ・ジャマシイ（1998）では大見出しとして 1,087

語を挙げており，そのうち，空見出し・活用語尾（例：かろう）・活用形（例：よかろう）・呼応の副詞（例：ぜんぜん…ない）・定型的な表現（例：をして…させる）・短単位に合致するもの（例：ばあい）等を除くと，複合辞・連語が約 600 語ある。この中に類似形態・異形態が多く含まれる（例：なきゃ・なくては・なくちゃ・なくてはいけない）としても，複合辞・連語が多く認定されていると言える。

BCCWJ では複合辞・連語の選定に当たって，ゆれがなく認定できること，長単位は短単位を基に自動解析するため，この自動解析で高い精度が維持できることという方針を立て，複合辞・連語とするものを先行研究よりも限定した。

具体的な手順としては，まずグループ・ジャマシイ（1998）の大見出しについて，短単位に合致する見出し語や文節を超える見出し語を削除し，類似形態・異形態を整理した上で，国語辞典等での採録状況を確認し，採録されていない語を削除した。このように絞り込んだ見出し語について，生産実態サブコーパスの入力済み書籍データ[1]（約 500 万語）を対象に用法の調査を行い，形式の面から複合辞・連語としてゆれなく判定できるものを選んだ。これに CSJ で認定されていた複合辞・連語を加えた上で，書籍データで頻度 200 以上の語を抽出し，BCCWJ における複合辞とした。

ここで，頻度 200 としたのは，複合辞を高精度で自動解析するためには，学習用データとなる人手修正済みデータ 100 万語の中に最低 50 例（使用率 0.005%）出現することが必要だからである。書籍データ 500 万語で使用率 0.005%に当たる 250 例よりも若干低く基準を設定し，200 例とした。

その結果，CSJ では助詞相当句 79 語，助動詞相当句 57 語，その他連語 90 語を 1 長単位とする複合辞・連語として認めていたが，BCCWJ で選定した複合辞・連語は，現時点で助詞相当句 24 語，助動詞相当句 39 語，その他連語 12 語である。

2.5 長単位の長所

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。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特徴語を取り出せるようになる。短単位は基準がわかりやすくゆれが少ないため，用例収集を行う上では便利な単位であるが，合成語を構成要素に分割してしまうという問題点がある。

中央省庁刊行白書の人手修正済み短単位データ（約 20 万語）を基に，白書を安全・科学技術・外交・環境・教育・経済・国土交通・農林水産・福祉に分類した場合，どのような語と結合するかという点から，ジャンル別の差異を見る。以下，「生活」という語を例に説明する。20 万語中，「生活」は 211 例見られる。そのうち「生活」単独で使われた例が 42 例，合成語の構成要素として使われている例が 169 例と，「生活」という短単位は，合成語の構成要素として使われることが多いことがわかる。

経済と福祉，それぞれのジャンルでの「生活」を見てみよう。経済では「生活」は 7 例使われており，そのうち，「生活」単独で使われた例は 1 例である。一方，福祉では「生活」が 126 例用いられており，そのうち「生活」単独で使われた例が 27 例である。以下に，「生活」が合成語の構成要素として使われている例を示す。

【経済】（全例）

国民生活選好度調査　消費生活　人間生活　生活不安度指数　労働者生活

【福祉】（一部）

家庭生活　共同生活　国際生活機能分類　国民生活センター　国民生活選好度調査　自立生活　社会生活　消費生活　消費生活センター　障害者生活訓練　食生活環境　生活コスト　生活する　生活できる　生活確保体制　生活環境　生活教養テレビ番組　生活支援　生活支援体制　生活施設　生活実態　生活上　生活水準　生活全般　生活相談　生活満足度　精神障害者地域生活支援センター　地域生活　地域生活支援　日常生活　日常生活支援体制　日常生活上　避難生活　別居生活

上に挙げた中で，下線を付した語はそれぞれ経済のみ，福祉のみに出現しているものである。つまり，「生活不安度指数」「労働者生活」などは経済の白書を特徴付ける語であり，「障害者生活訓練」「生活コスト」「地域生活」などは福祉の白書を特徴付ける語であると言うことができる。このように「労働者生活」を「労働」と「者」と「生活」とに，「生活コスト」を「生活」と「コスト」とに分割するのではなく，全体で一つとして扱う長い単位を使うことで，各分野の特徴的な語を把握することができる。長単位は各ジャンルの言語的特徴を解明するという目的にかなう，各媒体・各分野の資料的な性格を反映する単位と言える。

3. BCCWJ の長単位解析の現状

先に述べたように，長単位の自動解析に当たっては，短単位から長単位を自動構成する解析器

[1] 現時点では，学習に用いる予定の人手修正データが未整備であるため，入力済みデータ量が多く，幅広い分野をカバーすると考えられる生産実態サブコーパスの調査を基に，複合辞・連語を選定した。生産実態サブコーパスについては山崎（2007）を参照。

（Uchimoto, K., & Isahara, H. , 2007）を用いる。中央省庁刊行白書の人手修正済み短単位データ約 20 万語を基に，文節等の情報を使わずに長単位の自動解析を行った。解析器の学習には京都大学テキストコーパス Version3.0 の 1 月 1 日分の記事 1,129 文を用いた。以下，解析結果から長単位境界の認定に関する典型的な誤りについて，具体的にどのような事例があるのかを見ていく。

（1）連用修飾成分

長単位では，助詞・助動詞を伴わない自立語に関して，連用修飾成分の後ろで文節を切るという規定がある。長単位は文節を超えないので，長単位でも連用修飾成分の後ろで切り離されることになる。しかし以下のように連用修飾成分の後ろで切れずに後続の語と結合して 1 長単位となる場合がある。

｜それぞれ和解｜の｜仲介｜の｜制度｜が｜規定さ
｜れ｜てい｜た｜　→｜それぞれ／／和解｜

この要因は連用修飾成分の品詞にある。「それぞれ」の品詞は「名詞-普通名詞-副詞可能」だが，副詞可能という情報だけでは連用修飾成分かどうかの判断はできない。なぜなら，「単身赴任」の「単身」のように，副詞可能であっても名詞として使われることがあるからである。この場合は「単身」と「赴任」とが結合して 1 長単位になる。

正：｜単身赴任｜を｜始める｜サラリーマン｜

BCCWJ では今後，副詞可能を文脈での用いられ方に応じて，名詞と副詞に分類する予定である。上の例では「それぞれ」は副詞，「単身」は名詞となる。この分類作業をした後に長単位解析することで，連用修飾成分の判断に関する問題はある程度解決するだろう。

ただし，次に挙げる「一定期間」の場合，「期間」の品詞は「名詞-普通名詞-一般」であり，「一定期間」という合成語になることで連用修飾成分になる。

｜海外｜に｜一定期間滞在する｜
→｜一定期間／／滞在する｜

短単位の品詞情報だけでは連用修飾成分に関する問題は解決されない。今後の課題としたい。

（2）複合辞

複合辞の認定は形式からの判断だけではなく，意味の問題が絡んでくることで難しくなる例がある。例えば「について」の場合，複合辞とする場合と，「に（格助詞）＋就く（動詞）＋て（接続助詞）」に分割する場合とがある。「について,」のように「について」の後に読点がある形式は，複合辞として誤りなく解析される。しかし，それ以外の形式では次のように誤解析となる場合もある。

｜定期借地権｜の｜利用｜に｜つい｜てみる｜と｜
｜既に｜職｜に｜つい｜ている｜技術者｜が

これは本来「について＋見る（動詞）」と解析すべきものである。このような誤解析が生じるのは，のように，「について」の前後の品詞が同じであっても，「つく」が実質的な動詞として機能している例があるためだと推測される。このように同じ形式であっても，複合辞と認定する場合とそうでない場合とがあるため，自動解析による複合辞の認定が困難になっている。この対応は今後の技術的な課題である。

以上のように，長単位境界に関する誤解析には，様々な要因がある。今後，精度向上のために学習用データを追加・整備していくほか，解析結果の分析を継続的に行っていく。

4. 終わりに

本稿では，BCCWJ で採用した長短 2 種類の言語単位のうち長単位の認定基準の概要について説明した。また，長単位の自動解析の現状についても報告した。

長単位の認定基準については，今後 BCCWJ の構築を進めていく中で，適宜修正・追加を行っていく必要がある。自動解析については，学習コーパスの整備，それに基づく学習等を進めていく予定である。

参考文献

Uchimoto, K., & Isahara, H. (2007). Morphological annotation of a large spontaneous speech corpus in Japanese. , In Proceedings of IJCAI, 1731-1737.

小椋秀樹ほか（2007）「「現代日本語書き言葉均衡コーパス」の短単位解析について」『言語処理学会第 13 回年次大会発表論文集』, 720-723.

小椋秀樹（2007）国立国語研究所内部報告書『『現代日本語書き言葉均衡コーパス』短単位規程集 version1.2』

グループ・ジャマシイ（1998）『日本語文型辞典』

国立国語研究所（2001）『現代語複合辞用例集』

国立国語研究所（2006）国立国語研究所報告 124『日本語話し言葉コーパスの構築法』

山崎誠（2007）「『現代日本語書き言葉均衡コーパス』の基本設計について」『特定領域「日本語コーパス」平成 18 年度公開ワークショップ（研究成果発表会）予稿集』, 127-136.

付記　本研究は，文部科学省科研費特定領域研究「日本語コーパス」による補助を得た。