「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

小椋秀樹　　小木曽智信　　小磯花絵　　冨士池優美　　相馬さつき

独立行政法人国立国語研究所

1. はじめに

国立国語研究所は，明治時代から現代に至るまでの日本語の全体像を解明するため，大規模言語コーパスKOTONOHAの構築を開始した。この構築計画では，まず2006年度から2010年度までの5か年計画で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）を構築する[1]。

BCCWJには，国語学・情報工学をはじめとする幅広い分野での活用を目指して，様々な研究用の付加情報を与える。このうち形態論情報については，まず言語単位として，コーパスからの用例収集に適した「短単位」とBCCWJに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。

本稿では，BCCWJにおける言語単位の設計方針，長短2種類の言語単位のうち短単位の認定規定等について述べる。また併せて短単位解析のための自動解析用辞書の作成状況，BCCWJに格納する中央省庁刊行の白書のデータを解析した結果も報告する。

2. BCCWJにおける言語単位の設計

語の定義については研究者によって様々な立場がある。そのため，コーパスの言語単位をどのように規定するかについても様々な立場があり，容易に決めることはできない。

BCCWJの言語単位の設計に当たっては，まずBCCWJを日本語研究に利用するために，どのような言語単位が必要か整理した。その上で設計方針を立て，その方針に基づいて単位を設計した。

2.1 言語単位の設計方針

我々は，BCCWJの言語単位の設計方針として，次の三つを掲げた。

方針1：コーパスに基づく用例収集，各ジャンルの言語的特徴の解明に適した単位を設計する。

コーパスの日本語研究への活用としてまず考えられるのは，コーパスから用例を集めることである。そのため，BCCWJを日本語研究で幅広く利用できるようにするためには，用例収集に適した単位を設計する必要がある。またBCCWJは，新聞・雑誌・書籍といった複数の媒体を対象としたコーパスであり，内容も政治・経済・自然科学・文芸等と多岐にわたっている。このようなBCCWJの構成から，媒体別・分野別の言語的な特徴を明らかにしていくことが重要な研究テーマになると考えられる。したがって，そのような分析に適した単位を設計することが必要になる。

方針2：「日本語話し言葉コーパス」と互換性のある形態論情報を設計する。

国立国語研究所が既に構築したコーパスとして，現代の話し言葉を対象とした「日本語話し言葉コーパス」（以下CSJ）がある。KOTONOHAの計画では，BCCWJ・CSJは，KOTONOHAを構成するコーパスの一つとして位置付けられている。そのため，BCCWJとCSJとを統一的に扱うことのできるような，互換性を持った単位を設計する必要がある。

方針3：国立国語研究所の語彙調査における知見を活用する。

国立国語研究所は，1949年の『語彙調査 ―現代新聞用語の一例―』以来，合計10回の語彙調査を実施した。その中で，調査単位（言語単位）の設計や言語事象の処理に関して，様々な知見を蓄積している。そこで，BCCWJの言語単位の設計や単位認定の際に，これら語彙調査の知見を活用していく。語彙調査の結果は，日本語研究でも様々に活用されており，言語単位の設計等に語彙調査の知見を活用していくことは，BCCWJを使った日本語研究を進めていくためにも有用であると考えられる。

2.2 BCCWJの言語単位

以上の方針の下，BCCWJの言語単位について検討した結果，次のような結論を得た。

BCCWJの言語単位には，方針1で挙げた，用例収集・各ジャンルの言語的特徴の解明という二つの利用目的に応じて，次に示す2種類を採用する。

(1) 用例収集を目的とした短単位

(2) 言語的特徴の解明を目的とした長単位

この短単位・長単位は，いずれもCSJで採用した単位である。また短単位は国立国語研究所が行った現代雑誌九十種調査のβ単位を，長単位はテレビ放送の語彙調査の長い単位を基に設計したものである。

このようにして，CSJとの互換性の保持と，国立国語研究所の持つ語彙調査の知見の活用とを図る。

3. 短単位の概要

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。短単位の認定は，現代語において意味を持つ最小の単位（以下，最小単位）を規定し，その最小単位を文節の範囲内で短単位の認定規定に基づいて結合させる（又は結合させない）という手順で行う。

以下，最小単位の認定規定，短単位の認定規定，短単位の付加情報の概略及びコーパスの言語単位としての短単位の長所について述べる。

3.1 最小単位の認定規定

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，次のように認定する。

和　語：/豊か/な/暮らし/に/つい/て/
　　/大/雨/が/降っ/た/の/で/
漢　語：/国/語/　/研/究/所/
外来語：/コール/センター/　/オレンジ/色/
記　号：/図/Ａ/　/ＪＲ/
人　名：/星野/仙一/　/アンディー/・/シーツ/
地　名：/大阪/府/豊中/市/待兼山町/
　　/六甲/山/

上記のように認定した最小単位を短単位認定の必要上，**表1**のように分類する。

表1 最小単位の分類

分　類		例
一　般		和　語：豊か　大　雨… 漢　語：国　語　研　究　所… 外来語：コール　センター　オレンジ…
数		一　二　十　百　千…
その他	付属要素	接頭的要素　：　相　御　各… 接尾的要素　：　兼ねる　がたい　的…
	助詞・助動詞	う　だ　ます　か　から　て　の…
	人名・地名	星野　仙一　大阪　六甲…
	記　号	Ａ　Ｂ　ω　イ　ロ　ア　ＪＲ…

上記の分類のうち「付属要素」とは，接頭辞・接尾辞・補助用言のことである。ただし，すべての接頭辞・接尾辞・補助用言を付属要素に分類するわけではない。コーパスに出現したものの中から造語力が高いなど注目されるものを付属要素に分類する。

なお，最小単位は短単位認定のために必要な概念として規定するものである。そのため，BCCWJのサンプルを最小単位に分割することはしない。

3.2 短単位の認定規定

短単位の認定規定は，表1の分類ごとに適用すべき規定が定められている。その規定に基づいて最小単位を結合させる（又は結合させない）ことにより，短単位を認定していく。なお，最小単位を結合させる際には，文節境界を超えないという制約を設け，文節と短単位とが階層構造を持つようにしている。

以下，「一般」・「数」・その他に分けて，短単位認定規定の概略を示す。

(1)一般

《和語･漢語》2最小単位の1次結合を1短単位とする。
　|母=親|　|食べ=歩く|　|音=声|　|本=箱|
《外来語》原則として1最小単位を1短単位とする。
　|コール|センター|　|オレンジ|色|
例外①：省略された外来語の最小単位と外来語の最小単位との1次結合は1短単位とする。
　|エア=コン|　|マス=コミ|
例外②：省略された外来語の最小単位は，和語・漢語の最小単位と同様に扱う。
　|パソ=コン|　|塩=ビ|　|ピン=ぼけ|

(2)数

「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，一・十・百・千のとなえを取る桁ごとに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|五十|二|万|語|
　|五|分|の|二|　|二三十|回|　|〇|．|四|五|

(3)その他

1最小単位を1短単位とする。
《付属要素》|扱い|兼ねる|
《助詞･助動詞》|豊か|な|暮らし|に|つい|て|
《人名》|星野|仙一|　|アンディー|・|シーツ|
《地名》|大阪|府|豊中|市|待兼山町|　|六甲|山|
《記号》|図|Ａ|　|ＪＲ|

なお，「一般」に分類した外来語に関する短単位の認定規定について補足しておく。

CSJの短単位や現代雑誌九十種調査のβ単位では，「一般」の外来語の最小単位も，和語・漢語と同様，2個の1次結合を1短単位としていた。つまり，「コールセンター」「オレンジ色」を1単位としていた。ただし，①欧米語の冠詞・前置詞に当たるものは1最小単位を1短単位とする，②β単位では最小単位2個の1次結合が7拍を超える場合，短単位では同じく10拍を超える場合，結合させずに1最小単位を1短単位とするという例外規定を設けていた。

しかし，外来語の最小単位2個の1次結合を1短単位とすることについては，CSJの構築当初から和語・漢語に比べて長すぎるのではないかという指摘があった。このような指摘を踏まえ，上記②の拍数による例外規定を設けたが，10拍を超える場合としたことに言語学的な意味がある訳ではなく，そういう意味でこの例外規定にも問題があった。

以上のことから，BCCWJでは，「一般」の外来語の最小単位は，原則として1最小単位を1短単位とするというように，和語・漢語とは異なる扱いにした。

3.3 付加情報

3.2節に示した規定によって認定した短単位には，次に挙げる情報を与える。

代表形　代表表記　品詞　活用型　活用形

代表形は国語辞典の見出しに，代表表記はその見出しに与えた漢字表記に相当するものである。

品詞・活用型・活用形は，CSJを基に，BCCWJの短単位解析に用いる解析用辞書unidicの品詞・活用型・活用形を参考にして細分化を行った[2]。

例えば品詞は，CSJの品詞を基に，次の16種類に分類した。これは学校文法に準ずるものである。

名詞　代名詞　形状詞　連体詞　副詞　接続詞
感動詞　動詞　形容詞　助動詞　助詞　接頭辞
接尾辞　記号　補助記号　空白

さらに，これらの品詞をunidicを参考にして用法等の観点から細分化した。例えば名詞については，次の11種類に細分化した。

普通名詞　サ変可能　数詞　助動詞語幹
固有名詞-一般　固有名詞-人名　固有名詞-姓
固有名詞-名　固有名詞-国　固有名詞-地名
固有名詞-組織名

3.4 短単位の長所

ここでは，短単位がコーパスの言語単位として，どのような長所を持つのかについて述べておく。

短単位の長所としては，次の2点が挙げられる。

長所1:基準が分かりやすく，ゆれが少ない。

これは，短単位の基礎となる最小単位の認定に当たり，個人によってとらえ方に幅のある要素を基準に持ち込んでいないことによる。

ここで，基準が分かりやすく，ゆれが少ないという長所を裏付ける事例として，CSJの構築過程で行った人手による短単位認定作業を紹介する。

CSJでは，約752万語のうち約100万語について人手による解析を行った。この作業では，人手でテキストに短単位境界を入力した後，その短単位に基づき自動解析システム「茶筌」を使って品詞情報を付与するという方法を取った。この人手による短単位の認定作業では，1人日当たり約8,000短単位と，効率良く，大量のデータを処理することができた。さらにその精度は約99%と，非常に高い精度であった。また，作業者を新規に採用した場合に掛かる教育期間も比較的短く，5〜8日程度の実習で99%以上の精度で短単位を認定できるようになった。このように，作業者一人が1日に処理できる短単位の数が多く，精度も高いこと，また作業者の教育も比較的短期間で行えることは，基準が分かりやすく，ゆれが少ない単位であることを示すものである。

ところで，基準が分かりやすく，ゆれが少ないという短単位の長所は，作業効率の向上につながるだけでなく，コーパスの使いやすさにもつながる。基準が分かりやすければ，利用者が語を検索する際，どのように検索条件を指定すればよいか迷うことが少なくなる。また，ゆれの少なさ，つまりデータの精度の高さは，分析結果の確かさにもつながる。

長所2:取り出した単位が文脈から離れすぎない。

上で短単位はゆれが少ない単位であると述べたが，実は最もゆれが少ない単位は，短単位ではなく，その基礎となっている最小単位である。それにもかかわらず，最小単位を言語単位として採用しなかったのは，最小単位は文脈から離れすぎるため，日本語の研究に使いにくいからである。

例えば，短単位「気持ち」は「気」と「持ち」の二つの最小単位に分割することができる。もしこのような最小単位でコーパスが解析されていると，動詞「持つ」を検索した際に，「荷物を持つ」などの「持つ」とともに，「気持ち」の「持ち」も検索結果として得られることになる。

しかし，動詞「持つ」の分析を行う際に，「気持ち」の「持ち」まで検索結果に含まれるのは望ましいとは言い難い。それは，実際の文脈の中では，動詞「持つ」として機能していないからである。したがって，コーパスから用例を収集し，分析することを考えた場合，正確に単位認定ができるとしても，最小単位のような単位では問題が多いということになる。

このように考えた場合，短単位は，基準の分かりやすさ・ゆれの少なさという条件を満たしつつ，用例を収集して分析を行うという利用目的にもかなう単位と言うことができる。

4. BCCWJの短単位解析の現状

短単位解析のための自動解析用辞書の作成状況，BCCWJに格納する中央省庁刊行の白書のデータ(500万語)を解析した結果について報告する。

4.1 解析用辞書の作成状況

BCCWJの短単位解析は，自動解析エンジンに「茶筌」，解析用辞書にunidicを使う。unidicで採用している単位は，短単位とほとんど一致しており，品詞体系もBCCWJの品詞体系と互換性がある。

国立国語研究所では，2006年度の研究計画において，unidicの見出し語追加作業を実施した。この作業では，千葉大学の伝康晴氏が中心になって構築したunidic（見出し語:約46,000語）を基に，国語辞典や国立国語研究所の語彙調査等を基に作成されたデータから，unidicにない見出し語を，短単位の認定規定に基づいて分割し，順次登録した。その結果，2007年1月末時点で，約10万4,000語まで拡充できた。

4.2 白書データの解析精度

BCCWJに格納する中央省庁刊行の白書のデータ(500万語)を，2006年10月時点のunidic（見出し語・約10万語）を使って解析した。

その解析結果から，数詞を除く自立語1万語を無作

為抽出し，短単位境界の認定，代表形・品詞・活用型・活用形の情報付与が正しく行われているかを調べた。結果は誤解析が536例で，精度は94.64%であった。この536例の誤解析を短単位境界に関する誤りなどの5種類に分けて，その数を**表2**に示した。この表から，誤解析の約半数が短単位境界に関する誤りであることが分かる。

表2　白書データの誤解析の分類

短単位境界	代表形	品詞	活用型	活用形	合計	精度
253	163	112	0	8	536	94.64%

以下，短単位境界の誤りについて，要因は何か，具体的にどのような事例があるのかを見ていく。

短単位境界の誤りのうち，unidicに正解の短単位が登録されていなかったことに起因するものは115例ある。このうち，未登録の漢語が構成要素に分割される誤り，特に短単位の規定から漢語の大半を占める2字漢語が1字ずつに分割される誤りが54例あり，ほぼ半数を占めている。例えば，「各年」「同法」を「|各|年|」「|同|法|」と誤解析した例がある。また，外来語の一部を既登録の語として認定してしまうことによって，正解より短く分割するという誤りも36例見られた。例えば，経済用語「スタグフレーション」を「|ス|タグ|フレーション|」，外国人名「アーメド」を「|アー|メド|」と誤解析した例がある。前者はその一部を既登録の「タグ」(tag)としたことによる誤り，後者はその一部を既登録の「メド」（目処の片仮名表記）としたことによる誤りである。

unidicでは，表記が異なっても同じ語であれば，一つの見出し語にまとめるという方針を取り，語を階層化した形で登録している。単位境界の誤りの中には，正解の短単位は登録されているが，白書に出現した表記形が登録されていなかったことに起因するものが38例あった。

例えば，《モノヅクリ》という語は「物作り」「モノづくり」「モノ作り」という表記が，《マチヅクリ》という語は「町作り」「町づくり」「街づくり」という表記がunidicに登録されていた。しかし，平仮名表記の「ものづくり」「まちづくり」が登録されていなかったため，白書に出現したこの平仮名表記形が「|もの|づくり|」「|まち|づくり|」と誤解析された。

このほか，表記に関するものとしては，交ぜ書きされた漢語，送り仮名が省略された語の誤解析が多く見られた。漢語の交ぜ書きは，常用漢字表にない漢字は使わないという公用文の表記の基準，送り仮名の省略は，活用のない語のうち読み誤るおそれのないとされる特定の語について送り仮名を省略するという公用文の表記の基準によるものである。

誤解析となった交ぜ書き漢語としては，「研さん」「洗じょう」「ヨウ素」などがあった。これらはいずれも漢字部分と仮名部分とに分割されていた。

送り仮名が省略された語としては，「打合せ」「立入（検査)」「引下げ」などがあった。これらは「|打|合せ|」「|立|入|」「|引|下げ|」のように2単位に分割されていた。

このほか，外来語表記のゆれに起因するものもあった。例えば「ウェイト」「ユーザ」が「|ウェイ|ト|」「|ユー|ザ|」と分割されていた。なお，unidicには「ウエート」「ユーザー」のみが登録されていた。

以上のように，短単位境界に関する誤解析の主な原因は，正解の短単位や白書に出現した表記形がunidicに登録されていないことであった。今後，精度向上のために，未登録の語や表記をunidicに登録していく。ただし異表記を網羅的に登録することは，逆に精度を下げる可能性もある。各表記の頻度や出現分野の偏り等を見て，登録するか否かを判断する必要がある。そのためにも，今回のような分野を限定した解析結果の分析を継続的に行い，各分野の語や表記の出現傾向を明らかにしていきたい。

5. 終わりに

以上，本稿では，BCCWJの言語単位の設計方針，BCCWJで採用した長短2種類の言語単位のうち短単位の認定規定等について説明した。また，短単位解析のための自動解析用辞書の作成状況，その解析用辞書による白書データの解析精度及び誤解析の傾向についても報告した。

短単位の認定規定については，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，目標とする精度98%を達成するため，引き続き辞書の整備・拡充を図るとともに，学習用コーパスの整備，それに基づく学習等を進めていく予定である。

注

1 BCCWJの設計については，山崎誠（2007）を参照。
2 unidicについては，伝康晴（2006）を参照。

参 考 文 献

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（研究成果報告会）予稿集』

伝康晴（2006）「多様な目的に適した形態素解析システム用電子化辞書の開発」『特定領域「日本語コーパス」平成18年度全体会議予稿集』

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。また，unidicの拡充には，千葉大学 伝康晴氏の協力を得た。