語彙の研究を考慮した専門分野コーパスの作成

岡田真穂　竹内孔一　吉岡真治　影浦峡　小山照夫
（国立情報学研究所）

1　はじめに

近年、自動専門用語抽出の研究をはじめとして、語彙の抽出や解析に絡んだ言語処理に対する関心が増している。しかし、従来のコーパスは構文解析を初めとする統語的な研究を想定したものが殆どであり、語彙の観点が余り重要視されていなかった。

特に、分かち書きされない日本語においては、語彙的観点のコーパスが存在しないことがかなり大きな問題となる。どこからどこまでが一つの語であるかということの認定が考え方やシステムによって大きく異なることと、殆どの形態素解析システムやコーパスは文以上の単位を処理することを主な目的としているため、語彙という観点からの一貫性にばらつきが見られることがその主な理由である。

こういった事情を背景に、我々は、専門用語処理や記述研究を始め、語彙的研究に関わる研究者が共通に利用できるようなコーパスの作成を進めてきた。このコーパスでは、語彙的単位という観点から、一貫性のあるタグ付けを目指している。本報告では、我々が作成しているコーパス（NTCIR言語タグ付きコーパス）の基本的な考え方と仕様の概略について述べる。

2　NTCIR言語タグ付きコーパス仕様の概要

本コーパスが想定しているテキスト母集団は、公開された媒体に書かれ、なおかつ安定した、専門分野依存のテキストである。実際には、我々は現在、人工知能分野と情報処理分野の2分野における抄録データを扱っている[1]。

2.1　基本的な考え方

本コーパスでは、語彙的研究のためのコーパスを作成するという目的上、語に関わる単位（単語／形態素）の認定に最も重点を置いている。

まず基本となるのが、文を構成する要素としての単語単位の認定である。この「単語」には、一形態素からのみ成る単純語と、複数形態素から成る複合語とがある。更に、単語として認定された語のうち、複数の要素からなる「複合語」については、形態素単位まで分割して記述する。本コーパスでは、このふたつのレベルで語の単位の認定を行なう。

ここで重要なのは、現れた語に対し、その時の状況に依存した形でのみ考えるのではなく、その語が本コーパスで関わる日本語の全体系の中でどのような存在であるかを常に考慮する、ということである。つまり、

1. 同じ語が異なる文中で用いられている場合
2. 異なる複合語の間で同じ語が要素として用いられている場合

のいずれの対比においても、一貫した単位の認定、品詞等の情報の付与が行なわれている必要がある。これらはそれぞれ1. 単語単位、2. 形態素単位の認定に関わる。また、同じ語が複合語の構成要素（形態素）として現れた場合と単純語（単語）として単独で現れた場合の対比においても、やはり一貫した単位の認定、品詞等の情報の付与が必要である。本コーパスで適用した仕様は、この点を中心に考慮したものになっている（文法及び品詞の定義自体は、益岡、田窪（[2]）及びその文法体系に基づく形態素解析プログラムJUMAN（[3]）にほぼ準じている）。

実際のデータの例を表1に記す。具体的には左から「出現語 読み 見出し語 品詞情報 語種情報」（複合語は「c 出現語 読み 見出し語 品詞情報」）という並びになっており、各項目を半角スペースで区切っている。

2.2　実際の分割基準

本コーパスでは、この後に述べる3つの観点（要素のレベル、語種、品詞）を総合的に判断して分割

[1]このうち人工知能分野の抄録データについては初期の仕様（[1]）に基づいた第一次版が完成しており、1999年に行なわれた「情報検索及び自動専門用語抽出に関するNTCIRワークショップ」の応用に提供、公開している。

```
更に さらに 更に CC W
c 強力な きょうりょくな 強力だ JNA_DArt
    強力 きょうりょく 強力 JNAst K
    な な だ TLJJNA_DArt W
c データ構造 データこうぞう データ構造 NN
    データ データ データ NN G
    構造 こうぞう 構造 NN K
を を を SCA W
持ち もち 持つ VTAry W
、 、 、 LRD
かつ かつ かつ CC W
c 曖昧さ あいまいさ 曖昧さ NN
    曖昧 あいまい 曖昧 JNAst K
    さ さ さ TLNPD W
も も も SSB W
c 取り扱えて とりあつかえて 取り扱える VArytate
    取り とり 取る VRAry W
    扱えて あつかえて 扱える VArytate W
、 、 、 LRD
c 完全性 かんぜんせい 完全性 NN
    完全 かんぜん 完全 JNAst K
    性 せい 性 TLNPD K
も も も SSB W
c 保証さ ほしょうさ 保証する VSHmi
    保証 ほしょう 保証 NS K
    さ さ する VSHmi W
れて れて れる TLVArytate W
いる いる いる TLVAbs W
。 。 。 LSD
```

表 1: タグ付きコーパスの例

基準を定めている。分割の目安という点に限れば、これらの観点の間に絶対的な優先順位はない。

2.2.1 要素のレベル

単語単位

単語単位には、前述のように、1単語1形態素から成る「単純語」(単純語の単位認定の基準は形態素単位の認定基準と同様) と、複数形態素で構成される「複合語」とがある。

単語単位の認定は、その語が内容語であるか機能語であるかによって扱いが変わってくる。基本的には、内容語であれば文構成要素として一つの働きを為す語のまとまりを一単語とし、機能語であれば一つの品詞として認められる最小の長さを優先する。以下に例を示す (下線の切れ目が単語の区切り位置、以下同様)。

例)
・本コーパス で は 、語彙的観点 が 考慮さ れて いる 。

この例では、「本コーパス」[2]「語彙的観点」「考慮さ (する)」といった内容語の単語単位が文構成要素の機能的単位と一致しているのに対し、機能語である助詞「で」「は」「が」や動詞性接尾辞「れて (れる)」は、機能のまとまりではなく、品詞的観点を優先しているために、文構成要素としての機能的単位とは必ずしも一致しない。

[2]この例は同格に相当する可能性があるが、現段階では複合語として扱っている。同格の扱いについては 2.3 節を参照。

形態素単位

形態素単位は、本コーパスにおいては実質的に最小かつ基本的な単位である。この認定の基準のうち内容語についてのものは、国立国語研究所のいわゆる M' 単位 ([4]) にほぼ基づいている。即ち、和語・外来語は意味を担った最小の言語単位 (1意味担体) を一形態素とし、漢語は漢字1字を1意味担体とみなして、その最小結合 (= 漢字2字) を原則的に1形態素とする。また機能語に類する語については、助詞列を認めていない、「-だ」型の助動詞をほぼ認めていないなど、比較的細かく区切るようになっている。以下に例を挙げる ("|"は形態素の区切り位置、以下同様)。

例)
・本 | コーパス で は 、語彙 | 的 | 観点 が 考慮 | さ れて いる 。
・テキスト に よる 手 | 続き の 記述 と 、……
・ＵＩ を 担当 | する クライアント | プログラム について 述べる 。

またこのレベルの分割では、品詞や文法的な機能の他に、語種という要素がかなり強く作用する (次項「語種」も参照のこと)。上の例では、「本コーパス」「考慮さ」「担当する」といった単語は異なる語種同士の複合語である。異なる語種の組み合わせで成る語は現実にも一語感が低いものが多く、語種による分割は、この点を比較的高い程度で反映しているといえる。

2.2.2 語種

本コーパスでは、語種を4つ (和語、漢語、外来語、その他) 設定している。このうち「その他」を除いた3つが、形態素単位の認定の最も明確な目安となる。語種を超えて高い膠着度を維持する語はパターンが少なく、形態素の切り出しが比較的容易なためである。特に、複合語内で同一の品詞に属する形態素が連続しているような場合、機能的・意味的観点からの判断よりも基準が明確で、なおかつ、現実の一語感と矛盾することも少ない。故に多くの場合、分割基準の中で最も優先的に考慮される観点である。以下に、典型的な組み合わせの語例を挙げておく。

組み合わせ	例
和語–漢語	縦 | 横
和語–外来語	子 | プロセス
漢語–和語	品詞 | 枠
漢語–外来語	下位 | レイヤ
外来語–和語	データ | 型
外来語–漢語	ＩＳＤＮ | 回線、データ | 領域

活用形	例
基本条件形	行なえば、(変換) すれば
意志形	行なおう、(変換) しよう
タ形	行なった、(変換) した
タ系連用テ形	行なって、(変換) して
タ形連用タリ形	行なったり、(変換) したり
タ系条件形	行なったらば、(変換) したらば

表 2: 活用形の定義 (一部)

しかし、漢語と和語の組み合わせについては、慣用性や歴史的経緯から分割するのが難しい、和漢入り交じった漢字熟語も存在する (「湯桶 (和 – 漢) /場所 (和 – 漢) /重箱 (漢 – 和)」など)。また、漢字による外来語及び和語の当て字表記 (「一日 (ついたち) /仕方 (しかた) /煙草 (タバコ)」など) も存在する。これらについては、語種を手掛かりにするよりも、前項「要素のレベル」で述べたような語構成要素の独立性の考慮を優先することになる。

2.2.3 品詞/活用の認定

品詞体系[3]は、特に用言系の語の扱いについて、いわゆる学校文法とは異なる。中でも顕著なのは、形容詞を除く用言の活用語尾、助動詞、接尾辞などの解釈と判定詞の存在である。形容詞を除く用言の最小活用部分の解釈の違いと判定詞を認めているところから、結果的に助動詞の範囲に相違が現れている。

まず、最小活用部分の範囲は、学校文法において助動詞や助詞とされるものの一部がこれにくるみ込まれていることがある (表 2 参照)。またできるだけ細分割するという方針上、学校文法で助動詞とされている「–だ」型の語 (「ようだ/そうだ」など) は、その殆どのケースにおいて「名詞 + 判定詞」と解釈し、それぞれを単独の単語として扱っている。これらの方針によって、助動詞と認定される語彙は、学校文法などに比してかなり限定されている。

体言系の語については、副詞/連体詞を長めにとっている点などで相違があるものの、品詞の種類やその定義自体はほぼ学校文法と同じである。

2.3 操作的な分割基準の適用の例外

以上のような観点からテキストを分割していくのだが、例外的に、これらの基準を操作的にのみ適用するのでは解決しないケースがある。助詞の省略と同格がその主なものであろう。

本プロジェクトで扱っているデータに限れば、助詞の省略は余り考慮する必要がない。論文の抄録というテキストの性質上、省略という現象そのものが起こりにくいし、起こったとしてもほぼパターンが決まっていて、それらは概ね「サ変動詞 + する」の形や複合名詞、副詞という形で表しても不都合がないことが多いからである。

一方、単語/形態素の同格については問題がある。同格は見かけ上、同一の (または機能上同一とみなすことのできる) 品詞が連続しているものであるから、品詞に基づく分割基準を操作的に適用すると、これらは複合語となるケースである。

しかし、同格は普通、並立する要素それぞれが文の構成要素として独立していると考えられ、この観点から判断すれば、複合語として扱わないのが適切といえる[4]。同格が起きているケースでは、文構成要素レベルでの認定を適切に行うためには、多分に文脈に依存した形で処理する必要がある。

2.4 補足：和語の扱いについて

以上のような分割基準は、本コーパスプロジェクトが開始された 1997 年からの約 4 年間で少しずつ練り上げられたものである。日常的な細部の変更に加え、2 度の大幅な方針変更を経て現在に至っている。

言語処理を考慮したコーパスは、理論的な要件を満たしつつ、一貫性を保持していなければならないのは言うまでもない。しかし一方で、語彙の研究において想定される全ての観点を取り入れることは現実的に不可能である。余りに細かい現象にまで踏み込むと、複雑になるばかりで、一貫性を保つことが

[3]厳密には、語彙的単位を形態素単位と単語単位に区別したことにより、品詞についての再検討が必要になってくると思われる。形態素に対する品詞相当のカテゴリとはどのようなものであるべきかは難しい問題であり、今後の検討課題としたい。

[4]ただし、並立する要素全てにかかるような形で一つの語が合成し、並立要素を呑み込んだ複合語を作る場合 (「教授・学習活動」など) は、この限りではない。また、同格的な成り立ちの複合動詞については現在検討中。

困難になる。初期の仕様 ([1]) は、実際の現象をできるだけ汲み取ろうとしたことで分割基準そのものが随分複雑になってしまい、大規模な処理を一貫して行なうには不向きなものになってしまっていた。

最も大きな問題だったのが和語の扱い方である。

初期の仕様では、分割に強いレベル（現在の仕様の区切りとほぼ同定義）と弱いレベルのふたつを設定していた。弱いレベルの区切りとは、例えば派生語（「重み／重さ」）や「和語名詞＋する」（「早寝する／ぶらぶらする」）など、その成り立ち自体は類型的であるものの、それぞれを全く別の形態素として扱うには不十分な語に対して用いていたものである。そしてこの「弱い区切り」を少し広めに解釈し、膠着度の強さを反映させるような形で、一般的な和語同士の複合語にも適用してきた。

しかし実際には、出現した状況によって各要素への対応が変わるため、それぞれの語の間では一貫性はあっても、弱い区切りによって分割された複合語と区切りの全くない単純語、または強い区切りを持つ複合語との間に理論的に明確な差のないケースが多々現れた。また、弱い区切りで分割されたそれぞれの要素が「語構成要素」としての単位区切りとして認められるパターンとそうでないパターンが混在してもいた。この弱いレベルの区切りは、それを膠着度（一語感）の反映として考える上では現実的な言語感に近かったものの、仕様全体から見るとかなりのばらつきがあった。

この問題に対して我々は、本コーパスで行なうのは「語構成要素」となり得る単位を認定することである、という点を明確にして、弱い区切りを解消[5]することで分割基準を単純化し、単位の長さをできるだけ細分化する方向で方針転換を行なった。それが、本稿で紹介した現在の仕様である。

細分割することで一貫性を保とうとする方法は、一方で、語源や表記といった点に考慮する必要性を増し、「現代語」の範囲の定義とも絡んでまだ議論の余地がある。しかし、現在の仕様を検証していく中での現実として、個別の状況による揺れが少なくなっているのは確かである。

3 おわりに

前節の終わりでも述べたように、本稿で紹介した分割基準の適用によって、一貫性を保つという目的はかなり高い程度でかなえられている。今後はこの分割基準をベースにして、より高度で且つ一貫したコーパスへと成長させていくことを目指す。なお現段階の仕様は第一次段階のもので、タグとして付与している情報も最小限でに留め、次段階の目標として、統語的及び複合語内の構造付与を行なうことなどを考えている。また対象とする言語データの拡張及び多言語への対応も可能なようにしていく予定である。

我々は、本プロジェクトで作成したコーパスを叩き台にして、語彙的応用を考慮したコーパスの在り方とその応用を巡る議論の場を作っていきたいと考えている。

[5]見直しの結果、一般的な区切りと同様に扱うことになったものと、区切りを完全に失くしたものとがある。

謝辞

本研究は、学術振興会の未来開拓学術研究推進事業による「高度分散情報資源活用のためのユービキタス情報システムに関する研究」の元で行なわれた。

また、本プロジェクトを進めて行く過程で多大な協力を頂いた、（株）日本電子化辞書研究所の荻野孝野氏、国立国語研究所の柏野和佳子氏、小椋秀樹氏、加藤安彦氏、山崎誠氏、仕様の立案からデータ検証にまで深く関わって頂いた波山万紀子氏、山口佳子氏に感謝の意を表する。

参考文献

[1] 吉岡真治、岡田真穂、影浦峡、小山照夫 (1999)「専門用語抽出・解析処理を考慮したコーパスの作成」『情報処理学会研究報告』Vol.99, No.20, pp. 41–48

[2] 益岡隆志、田窪行則 (1981)『改訂基礎日本語文法』くろしお出版

[3] 松本祐治、黒橋禎夫、山地治、妙木裕、長尾真 (1997)『日本語形態素解析システム JUMAN version 3.3』京都大学

[4] 野村雅昭、石井雅彦 (1988)『学術用語語基連接表』国立国語研究所

[5] 柏野和佳子、小椋秀樹、田中牧郎、加藤安彦 (2000)「話し言葉コーパスにおける単位切りと品詞付与の方法」『言語処理学会第6回年次大会発表論文集』pp. 99–102