論　文

日本語の報道記事を対象とする事象データ抽出システム

――コンピュータによる内容分析手法の概要と解析精度――

吉 田 文 彦

An Event Data Extraction System for Japanese News Articles : A General Description of the Content Analysis Software and Evaluation of Its Performance

Yoshida Fumihiko

Abstract

The author describes the organizational structure of a computer software developed by the author for extracting“event data” from Japanese news articles. The software, based on the result of morpheme analysis and syntax analysis, extracts event data, which is comprised of Actor, Action, and Target, from ordinary sentences found in such materials as newspaper articles. The author argues that the software should be of great use for various purposes not only for academic purposes but also for practical ones in business world. A detailed examination of the system's performance demonstrates that its accuracy in extracting event data from the original sentences seems to be comparable or in some situations may exceed the level usually expected for trained human coders. The author discusses further improvements to the software system, including an incorporation of a large scale electronic dictionary, an improvement of user interface, and a possible integration of the three separate software systems, the morpheme analysis system, the syntax analysis system, and the event data extraction system.

Ⅰ．はじめに

本稿は，日本語で書かれた報道記事の内容から，誰が誰に対し何をしたかという事象の根幹部分（以下それらの要素からなる文節のセットを事象データと呼ぶ）を抽出するためのコンピュータ・ソフトウェアの開発結果を報告するものである。

開発したシステムの構造，精度，そして今後の改善点などについて論じる前に，まずはそのようなソフトウェアの存在意義を主としてその用途の面から述べておくことにする。

大きく分けると2つの用途が考えられる。ひとつは研究のためのデータ生成である。もうひとつは，実務的な用途である。まず最初の用途であるが，そもそも筆者がこのシステムの開発に着手したのはマスメディア報道を研究するためのデータを生成することがその主目的であった。事象データを抽出するソフトウェアが存在すれば，CD-ROMやインターネット上のサイトから得られる電子化された報道記事を入力データとして用いることで，誰（あるいは，組織，国家）がどのような相手に対して，どのようなタイプの行動をとったかという事実関係を定型的な形でリストアップすることができることになる。しかも，それがコンピュータによって自動的に行えるのであれば，人手に頼る場合に比べて，圧倒的に速い速度で，しかもほとんどコストもかからずに精度の高いデータを入手できることになる。

そうして得られたデータは，さまざまな研究用途に用いることが可能である。たとえば，あるメディアの，ある問題に関する，ある期間の報道内容を対象に，そこから抽出された事象データに含まれる行動の主体（行為者），行動の対象となる国家，組織，人物，そして行為自体を何らかのタイポロジーによって分類し集計するならば，そのメディアの報道の特性を内容に即して評価するための基本的なデータが得られることになる。同様のことを他のメディアに対しても行えば，複数のメディアによるひとつの問題に関する報道内容の類似性や相違点などを実証的に比較することが可能となる。さらには，対象とする問題が国際的なイシューであれば，国内外の複数のメディアの報道内容をこのような方法で比較することもできる[1)]。以上のような意味で，本稿で報告するシステムはマスメディアの研究者にとっては新たな方法論上のツールとなり得るものである。

もともと事象データ（event data）というコンセプトは数量的な手法を用いて国際関係を研究する人々の間から生まれてきたものである[2)]。そのためか，マスメディアの研究者の間ではこれまで取り上げられることのなかったコンセプトでもある[3)]。国際関係論の分野では，それも特に米国の研究者にほぼ限定されるが，公刊されている文書（マスコミ報道や，特定地域に限定されたクロニクルなど）から事象データを抽出し，国際紛争などの研究の原データとして蓄積，分析することがかなり精力的に行われてきている[4)]。

理由は定かではないが，わが国ではこのような分野を専門とする研究者は今のところほとんど存在しないようにみうけられる。しかしながら，本稿で報告するシステムは事象データに依拠した国際関係の研究にも大いに貢献できるポテンシャルを持つものでもある。日本語による報道記事（あるいは放送原稿）などから，国際関係に関する事象データを定型的な形で抽出・蓄積する道を開くからである[5]。もちろんその用途は，必ずしも国際関係の研究に限定されるものではない。国内の政治やその他の社会現象に関しても利用可能である。

ここまでに述べてきたことからも既に明らかとは思われるが，国際関係や，国内の社会問題に関して，報道機関などにより公刊された文書から，日々生起する各種事象を定型的なデータとして蓄積することが可能となれば，そうした情報はさまざまな実務的な用途にも役立つものと思われる。世界各地で日々刻々と生起するさまざまな出来事を常時モニターしたり，あるいは過去のそのようなデータを頻繁に用いることが必要な業務内容であれば，本稿で報告するようなシステムは大いに役立ち得るはずである[6]。

以上のように，本稿で報告するシステムはさまざまな潜在的な用途を持つものと考えられるが，現在のところ市販のソフトウェア，シェアウェア，あるいはフリーソフトを見回しても，同様な機能を提供するソフトウェアは見当たらないのが実情である[7]。この意味で，本稿で報告する事象データ抽出システムの存在意義は十分に大きなものであると考えられる。

以下では，システムの基本構造，解析精度，今後の改善項目，などに分けて詳しく報告していくこととする。

II．システムの基本構造

本稿で報告する事象データ抽出ソフトウェアは，プログラムの形態としては単独の独立したものではあるが，実際に利用する場合には，これまでに筆者が開発してきた，形態素解析プログラム，構文解析プログラムとともに使用することを前提としてデザインされている[8]。より具体的には，ある報道記事（これは単独の記事であってもよいし，膨大な数の記事からなるファイルであってもよい）から事象データを抽出するには，まずそれを形態素解析プログラムに入力するところから始める。ついで，得られた形態素解析の結果を含む出力ファイルを構文解析プログラムに入力する。

構文解析プログラムは，構文解析の結果とともに，はじめに入力された形態素解析の結果も併せて出力するので，それをさらに事象データ抽出プログラムへの入力データとして用いるわけである[9]。実際の処理は以上のようなプロセスを経るわけであるが，ここでまず確認しておかなければならないことは，事象データの抽出にあたっては，形態素解析，構文解析の順で，いわゆる自然言語処理の基本的な手続きをまず行うということである。

形態素解析の段階では，どこからどこまでがひとつの語であるかを正確に識別するとともに，それらの品詞，活用形（用言の場合）を明らかにし，識別された語を文節の形にまとめる作業を行う。構文解析の段階では，それぞれの文章ごとに，文節間の係り受け分析を行い，どの文

節が他のどの文節に係っているのかという観点から文章の構造を解明する[10]。事象データの抽出段階では，以上 2 つの解析段階で得られた情報に基づき，誰が誰に対して何をしたのかを表す事象データを抽出するわけである。

本稿で報告するシステムでは，事象データの抽出過程は以下の 7 段階から構成される。

⑴助詞「が」,「は」,「も」で終わる文節を見つける（主格節の識別)。
⑵それらの係り先となる文節を識別する（述語節の識別)。
⑶ひとつの主格節が複数の述語節に係る場合，複数の事象に分割する（複数事象の識別あるいは並列構造の解析)。
⑷述語節の態，時制，否定などやモダリティを識別する（述語節の分析)。
⑸述語節に係る複数の文節から，時間，場所，対象物，相手，などにかかわる節を識別・判定する（格分析)。
⑹埋め込み文，発言内容を示す内容節などを識別する（埋め込み文，内容節の識別)。
⑺主格節の内容，行為の内容，行為の対象の存在などをもとに事象と非事象とに分類する（事象の識別)。

以下では，これら 7 つの段階について，それぞれ説明していくこととする。

１．主格節の識別

まず最初の，主格となる文節の識別であるが，助詞の「が」,「は」,「も」で終わる文節のみを取り上げるわけであるから，それ以外の助詞で終わるものは主格とはなりえないわけである。つまり，これら 3 つの助詞で終わる文節のみが，事象を形成する最初の要素である「行為者(Actor)」の候補とされるわけである[11]。したがってここではいくつかの問題が起きてくる。そのひとつは，人物，組織，国家など，いわゆる行為者とみなされるものの他にも，さまざまなことがらが拾い上げられるということである。しかしながら，それらは最後の段階⑺でふるい分けられるので，大きな問題ではない。

2 番目の問題は，日本語の文章では，主題が明示されていない場合がかなりあるということである。これは確かに事実ではあるが，この研究が解析の対象とする報道記事では，その頻度はこのような解析手法が無意味となるほどまでに多いわけではない。正確なデータは今のところ集めていないが，実際に報道記事を数多く調べた結果，そのような事例はむしろ例外的なものであり，根本的な問題は起きないのではないかと思われる。

3 番目の問題は，ひとつの述語節に複数の主格節が係る場合である。つまり，これは「象は鼻が長い」や「浜松はうなぎがうまい」といった有名な文例のように,「は」や「が」で終わる文節が同じ述語に係っているような場合，どちらを主格と見るかという問題である。確かにこの問題は一冊の本になってしまうほどの問題ではあるが[12]，ここで報告するシステムでは，助詞に先行する名詞の意味上の種類をもとに優先順位を与える方式をとり，問題への深入りを回避している。すなわち，名詞のタイプが個人，組織，国家などであれば，それら以外の名詞

を含む節よりも，主格である可能性が高いと判断する方式である[13]。

2．述語節の識別

主格となる文節が識別されれば，次は第2の段階として，述語節を識別することになるが，実際には既に構文解析によりどの文節が後続のどの文節に係るかが判明しているので，主格の節の係り先を見るだけで，述語となる節は自動的に見つかることになる。ただし，主格となる文節が，複数の述語の文節に直接あるいは間接的に係っているときは，話はやや複雑となる。この場合は次の第3段階で処理することとなる。

また，述語の文節が，2つ以上の文節から構成される場合や，述語の文節に動詞が含まれない場合もあり，それらの場合のための特別な処理もこの第2段階で行われる。すなわち，述語部分がどこで終わるかを識別しなければならないし，また報道記事によく現れるいわゆる体言止めを識別する必要もある。体言止めの場合は，実際に用いられる語自体は名詞の形であるが，機能的には動詞の働きをしているわけであるから，第3段階以降の処理においてそのことを処理プログラムが識別できるよう，述語部分が体言止めの形をとっているという情報を内部的に記録しておくことになる。

3．複数事象の識別および並列構造の解析

複数事象の識別については理解を容易にするために，実際の文例を用いて説明することにする[14]。

《例文1》外務省の林貞行事務次官は十五日夜，韓国の金太智駐日大使に電話し，韓国軍が竹島近海で行った演習に抗議した。

例文1では，林事務次官の行動に関しては，同次官が金大使に電話をしたということと，また同次官が金大使に抗議したという2つの事実が述べられている。実際の行動という側面から考えるなら，電話したということと，抗議したということとを必ずしも分離して考える必要はないかもしれないが，表層的な意味での文章の解析ということになれば，この例文からは「林次官が金大使に電話した」と，「林次官が金大使に抗議した」という2つの事象データが抽出されることになる。

また，この例文では，それらの事象はいずれも「十五日夜」に起きていることなので，どちらの事象に関しても，この時間を表す節を結び付けておく処置が必要となる。なお，この例文には「韓国軍が竹島近海で行った演習」という部分が含まれており，「韓国軍が演習を行った」という形で，さらに別の事象を取り出すことも可能である。この側面に関しては，6番目の段階のセクションで，埋め込み文との関連で論じる。

4．述語節の分析

以上の3段階の処理が終了したところで，第4の段階では述語部分の詳細な分析が行われる。

ここでは，主として述語節に含まれる助動詞の種類と活用の形態をもとに，⑴使役，⑵態，⑶推測・意志，⑷打ち消し，⑸希望，⑹時制，⑺断定，⑻様態・伝聞，⑼困難，⑽仮定，などの側面に関してそれぞれどのような表現形態が用いられているかを判断する。ここで得られた情報は，最終段階の事象のふるい分けの際に活用される。たとえば，時制に関する情報と，仮定に関する情報を用いれば，実際に起こったことと，仮定上の記述とを区別することが可能となり，実際に起きた事象とそうでないものとを区別する上で役立つ。

また，ここで得られる情報のかなりの部分はいわゆるモダリティに属するものであり，報道内容を質的に分析する上でも，きわめて有用な基礎データを提供することになる。たとえば，日本のマスメディアによる報道には，「……といわれる」とか「……のようだ」などという表現が用いられることがしばしばあるが，推測，伝聞などの表現形態をとる文章がどの程度一定の記事の中に見られるかという側面に関する分析もここから得られるデータを用いることで可能となる[15)]。

この第 4 段階では，この他，以下の 3 つの側面の分析も行う。そのひとつは，行動を規定する側面を持つ文章であるかどうかの判断である。たとえば，「……すべきである」という文章はその典型例である。また，対象となっている事象の候補が文章のどの部分に位置するものかに関する判断も行われる。ここでは，事象の候補とされている部分が，かぎカッコあるいは丸カッコの内部に位置するものであるかどうかが判断される。これは，事象の候補が何らかの発言の内容であるか，あるいは注釈の一部であるかどうかの判断材料となるものである。ただし，この面に関しては 6 番目の内容節の識別段階でより詳細な分析が行われることになる。

さらにこの第 4 段階では，述語部分に何らかの行為を示す記述が含まれている場合は，それが行動の対象となる相手にとって，明らかに敵対的・非友好的なものであるか，あるいは協調的・友好的なものであるかを判断し，「Negative」，「Positive」，「その他」のいずれかに分類する。その判断にあたっては，動詞および名詞の辞書データベースにあらかじめ記載してある行動内容の「正／負」の情報を用いている。この分析から得られるデータを用いれば，ある 2 国間の国際紛争に関する一連の報道記事などにおいて，自国が相手に対して行った行動と，相手の国が自国に対して行った行動がどのように報道されているかを，行動の内容に即して評価できることになる。

自国の行動に対しては最善の評価を行い，相手の行動に対しては，最悪の評価を行うという形での報道が行われているかどうかとか，相手の敵対的な行為に関する報道が，両国の行動全体の報道に占める割合はどうかという側面を，抽出された事象に含まれる行為のタイプ，内容から実証的に分析することも可能となる。より端的には，被害者意識に満ちた報道であるか，あるいはバランスのとれた報道であるかという側面をいわゆる評論ではない形で検証するための判断材料を提供できることになる。

5．格分析

この第 5 の段階では，行動を示す述語部分に係るさまざまな文節，あるいは節をその内容に即して分類する。分類にあたっては，現段階ではいわゆる 5 W 1 H の考え方を用いている。す

なわち，何時（When），どこで（Where），誰が（Who），誰に（Whom），何を（What），どのように（How）したかという一種のフレームによって，述語に係る文節や節を分類するという考え方である。しかしながら，ここで問題となるのは，本稿での中心概念となる事象データというコンセプトと，この５Ｗ１Ｈの考え方，さらには文法的な側面とに果たして整合性があるのかということである。

事象データは基本的には⑴行為者〈Actor〉，⑵行為自体〈Action〉，⑶行為の相手〈Target〉という３つの基本要素から成り立つ。これらは５Ｗ１Ｈの表現に直せば，「誰が（Who），誰に（Whom），何を（What）したか」という３つの部分に対応するわけであり，少なくとも事象データの考え方と５Ｗ１Ｈとの間には，一応の整合性はある。しかし，５Ｗ１Ｈと実際の日本語，あるいは文法的な考え方とを対比させると，多くの疑問が生じる。たとえば，どのように（How）という部分を取り上げるなら，それは手段を述べているものなのか，あるいは文法で言う副詞にあたるものなのかという疑問が生まれる。

また，５Ｗ１Ｈを用いるならば，行動の対象となる相手が人間や，組織，国家などではない場合には，実際の文章をその枠組みで捉えることはやや不自然でもある。また，これは部分的には日本語の特性によるものなのかもしれないが，行為の対象となる相手が実際の文章では明示されていない場合が無視できないくらいに多い。以上のようなことから，筆者自身も現段階では，事象データ，５Ｗ１Ｈ，日本語自体および文法上の各種概念を統一的に捉えることにはかなりの無理があるのではないかとの疑念を抱き始めているところではあるが，実際に作動するシステムを構築するという現実的な目的を考えると，よりアカデミックな考察は今後の課題とし，とりあえずは現実的な対処法を考えざるを得なかった。

そこで，本稿で報告するシステムでは，述語部分に係る文節および節の分類方式としては，５Ｗ１Ｈの考え方に基づきつつ，実際的な日本語の用法に対処できることを主眼に，以下の８項目を採用することにした。それらは，⑴時間，⑵場所，⑶行動の対象となるものやことがら，⑷行動の対象となる個人，組織，国家，⑸行動のさま，様相，⑹行動の関連事項，⑺言語行動の場合の発言内容，の７種類である。行動の主体となる行為者と実際の行為自体は，最初の３段階で，それぞれ主格の節と述語の節とによって識別済みであるので，この分類からは外してある。

また，この分類のうち，⑶を除く最初の５つは，基本的には５Ｗ１Ｈから得られたものである。⑹は何らかの行為が行われる場合，それがどのようなことについて，あるいはどのようなことに関連して行われたのかを意味する文節に対応するものである。⑺はいわゆる内容節と呼ばれる節に対応しており，「……と述べた」のような文章の「……と」の文節がそれにあたる。実際の分類にあたっては，これらの７タイプの他に，「どれにもあたらない」という項目を設け，各文節あるいは文節のグループからなる節をそれらのいずれかに分類する方式をとった。

実際の判定にあたっては，⑴各文節の最後の助詞のタイプ，⑵もし存在するなら文末のひとつ手前の助詞のタイプ，⑶助詞に先行する名詞の内容上のタイプ（やや大げさにいえば，意味素性マーカー），⑷述語節の動詞が言語行為を意味するものかどうか，⑸文節内に数字が存在するかどうか，などを含む多くの判断材料を用いた。ある文節が，助詞「で」で終わっており，

「で」の手前には地名を表す名詞が位置するならば，当然この文節は，場所を示すものと判断できるわけである。

また，ある文節に時間を表す名詞があり，しかもその文節に数字が含まれるならば，おそらくその文節は時間を表すものと判断できることになる。しかし，同じ助詞の「で」であっても，「竹島問題で，外相は……と述べた」という文章の場合は，「竹島問題で」という文節は上で示した分類方式では「行動の関連事項」に該当するものであって，場所を示すものではない。この例が示すように，ひとつひとつの文節が，上記の７つの分類のどれに該当するかを高い精度で判定することは容易なことではない。

そこで，本稿で報告するシステムでは，２つの独立した判定方式を用いて，最終的判断はその２つの方式から得られた結果を総合することによって決める形をとっている。どちらの方式でも，まず判定の対象となるひとつの文節に対し，８つのスコアを与えていくことになる。それぞれのスコアは，上記の７つの分類と，「その他」に対応するものである。判定方式のひとつは，判定に用いる判断材料に応じて，該当するスコアを積み上げていく方式である。

たとえば，「マニラのホテルで」という文節であれば，「で」という助詞があることから，場所である可能性と，関連事項である可能性が高いと判断し，場所のスコアと，関連事項のスコアを増やすわけである。また，「ホテル」という建造物を示す言葉があることから，場所のスコアをさらに増やすことになる。以上のようなスコアの積み重ね方式により，最大のスコアを得た項目が対象となっている文節の内容を示すものであろうと推測できるわけである。

ただし，実験の結果，この方式のみを単独で使用するよりも，他の方式と併用し両者の判定結果を総合したほうがさらに精度の高い判定が行えることが判明したため，本稿のシステムでは，もうひとつ別の判定方式を取り入れている。もうひとつの方式は，ファジー関係を用いたものである。コンピュータ・プログラム内に，35×８のマトリックスを用意し，35項目の事実のいずれかが文節の分析から見出された場合，それぞれの事実について，８つの分類が正しい確率はそれぞれどれくらいかを記述する方式をとっている。文節の分析から得られた事実と，このマトリックスの数値からファジー関係を計算することで，その文節が８つの分類項目のそれぞれである可能性が確率の形で導き出されるわけである[16)]。

２つの判定方式から得られた結果を総合するにあたっては次のような方式を用いた。スコアを積み上げる方式の場合も，ファジー関係を用いた方式の場合も，対象となるひとつの文節に関してそれぞれ８つの分類項目に対応する形で８つのスコアが存在するので，それぞれの分類項目に関して，スコア方式から得られた最終得点と，ファジー方式から得られた得点（実際は確率）とを掛け合わせていく。

この乗算により，その文節に関しては８つの分類項目のそれぞれに関して，最終的な判断に用いるスコアが得られることになる。最終的な判断は，それら８つのスコアのうち，最大値を得た分類項目を見つけることにより行われる。たとえば，場所に関する分類項目が最大のスコアを得たならば，その文節は場所を示すものだと判断されることになる。判定の精度に関しては，現在のところでは90パーセント弱というところであるが，このことに関してはさらに次のセクションで詳しく論じる。

6．埋め込み文，内容節の識別

ひとつの文章の中に，さらにもうひとつの文が組み込まれている場合，組み込まれているほうの文を埋め込み文という。

《例文2》池田行彦外相が「竹島は日本固有の領土」と発言したことに，韓国は猛反発している。

例文2では，「池田行彦外相が『竹島は日本固有の領土』と発言した」という部分が埋め込み文となる。内容節というのは，何らかの言語的行為によって発されたメッセージの内容を表した節である。例文2では，「『竹島は日本固有の領土』と」という部分がそれにあたる。第6の段階では，抽出された事象データの候補が，埋め込み文や内容節の一部であるかどうかを判断する。埋め込み文や内容節も含めて事象データを抽出するのか，それともそれらから抽出されるものは事象データに含めないのかということは，得られたデータをどのように利用するかによって決まってくることであるが，少なくともそれらを区別することができることが望ましい。

たとえば，実際に起こった出来事だけを報道内容から見つけたいという場合には，ある人物の発言内容に含まれている憶測や，一方的状況判断などが事象データに混入してくることが望ましくないと考えられることもあろう。その場合には，埋め込み文や内容節から得られた，事象データの候補は最終段階でふるい落とすことができるよう，あらかじめ各事象が埋め込み文や内容節から得られたものであるか否かを判断し，その結果を記録しておくことが望ましいわけである。

埋め込み文の識別が必要になるのは，ひとつのオリジナルな文章から，複数の事象データの候補が見つかった場合のみである。識別の方式としては，それぞれの事象データを構成する各要素に関して，それが同一の文章から取り出された他の事象データの構成要素，ならびにその事象データに付属する「場所」，「時間」などを表す節の一部に含まれるかどうかを，総あたり方式で調べていく。また，内容節の識別にあたっては，文節に先行するかぎカッコの存在，助詞「と」の存在，文節に後続する動詞あるいは名詞が言語行為に関連するものであるかどうか，などから判断する方式をとった。

7．事象の識別

以上のような方式でリストアップされた事象データの候補は，「何が何をした」という行動を表すタイプと，「何は何である」という描写，判断，評価などを表すタイプとに大別できる。本来の意味での事象データは当然ながら前者である。したがって事象データの選別にあたっては，前者のみを選ぶ必要がある[17]。これは述語にあたる部分の動詞の種類を調べれば容易に実行できる。明らかに行動を示すもののみを取り上げればよいからである。また，主格の部分を調べることにより，不必要な事象データの候補をふるい落とすこともできる。

たとえば，「交渉のデッドラインは目前に迫っている」という文章に関しては，主格は「デ

ッドライン」であって，何らかの個人，組織，国家などではない。何らかの具体的な行為者が主格となっている事象データを抽出することが目的なのであれば，このような事象データの候補は不要であり，ふるい落とすことが必要となる。この７番目の段階では以上のような方式により，最終的に取り出す事象データのみを選別するわけである。

以上の方式により識別した事象データは，大きく分けて２つの方式で出力される。現在は，ひとつひとつの事象データごとに詳しい分析結果とともに表形式で出力する方式と，リスト形式ですべての事象データを一括して出力する２つの方式を用いている。図１および図２は前者の方式による出力例である[18]。現時点では，いずれの場合も，ファイル形式で一括してハードディスクに出力する方式をとっており，分析結果を見るためには，いったんハードディスク上に書き出されたファイルをエディタなどのソフトウェアを用いて，画面上に呼び出すことになる。

III．解析精度と評価

事象データ抽出の精度を評価するにあたっては，訓練を受けた人間と同等の作業成果をコンピュータがあげることができるかどうかを主眼に評価することにした。そこで，まずテストデータとなる85の文章を用意した。これらはいずれも実際の新聞報道から得た文章である[19]。精度の判定にあたっては，まず筆者がこれらのテストデータから，事象データを抽出してみた。ついで，コンピュータにも同じ作業をやらせてみた。筆者が手作業で行った結果と，コンピュータによる抽出結果とが高い精度で一致するならば，事象データの抽出という作業に関しては，コンピュータに任せたほうがはるかに効率的だということになる。

事象抽出の段階は２つのプロセスに分けて考えることができる。最初のプロセスでは，助詞「が」,「は」,「も」で終わっている文節をまず見つけ，ついでその文節の係り先となっている文節を見つけて，それらから構成されるペアを事象データの「候補」として記録しておく。第２のプロセスでは，主題となる「が」,「は」,「も」で終わる文節の内容を調べるとともに，それらの係り先となる文節の内容を調べて，事象データの各候補が事象データとみなせるかどうかをチェックし，一定の条件を満たすもののみを残す作業を行う。したがって，解析精度も以上２つのプロセスごとに調べることができる。何度も精度を調べることは煩雑ではあるが，各段階での間違いを是正していく上では欠かせない作業といえる。そこで，本研究でも，以上の各プロセスに関して解析精度を調べた。

１．原文からの事象候補の抽出精度

原文から事象データの候補を抽出する段階では，テストに用いた85の文章のうち73文（85.9パーセント）に関しては，正確に事象データの候補が抽出できた。残りの12文（14.1パーセント）に関しては，少なくともひとつのエラーが確認された。以上は文章単位でみた精度であるが，テストに用いた85の文章からは全部で181の事象データの候補が抽出されたので，それらのひとつひとつが正しく抽出されているかという面から精度を判定することもできる。事象デ

図1　分析結果の出力例（その1）

【 55】[243]　事象番号(1/ 3)

この後、韓国政府は６５年４月と１２月、日本に「独島は韓国領土であり、日本の主張は考慮の対象にならない」との口上書を送った。

【ACTOR/SUBJECT】韓国政府は　　【ACTION】送った。

□□■□□□□□□□□□□□□□□■　≪事象基本要素のみ≫
■■■■■■■□□□□□□□□□□□■　≪識別済み文節グループ≫

（Ⅰ）５W1H

When　：６５年４月と１２月、
Where　：
What　：「独島は韓国領土であり、日本の主張は考慮の対象にならない」との口上書を
Whom　：日本に
How　：

（Ⅱ）関連事項：

（Ⅲ）内容節：

（Ⅳ）ACTION 分析結果

1.使役　　：
2.態　　：
3.推測・意志:
4.打ち消し　：
5.希望　　：
6.時制　　:過去形
7.断定　　：
8.様態･伝聞　：
9.困難　　：
10.行動内容　：
11.内容節　　：
12.行動規定　：
13.仮定　　：
14.体言止　　：

図2　分析結果の出力例（その2）

【 59】[247]　事象番号(1/ 2)

竹島（韓国名・独島）の領有権問題について、首相は「日本の立場は一貫している」と譲歩しない姿勢を示したうえで、漁業協定の見直し交渉を領土問題とは切り離して早急に開始するよう提案。

【ACTOR/SUBJECT】首相は　　【ACTION】示した

□□□□□□□□■□□□□□□□□□■□□□□□□□□ ≪事象基本要素のみ≫
■■■■■■■■■■■■■■■■■□■□□□□□□□□ ≪識別済み文節グループ≫

（Ⅰ）５W1H
When　：
Where　：
What　：譲歩しない姿勢を
Whom　：
How　：

（Ⅱ）関連事項：竹島（韓国名・独島）の領有権問題について、

（Ⅲ）内容節：「日本の立場は一貫している」と

（Ⅳ）ACTION 分析結果

1. 使役　　：
2. 態　　：
3. 推測・意志:
4. 打ち消し　：
5. 希望　　：
6. 時制　　:過去形
7. 断定　　：
8. 様態･伝聞　：
9. 困難　　：
10. 行動内容　：
11. 内容節　：
12. 行動規定　：
13. 仮定　　：
14. 体言止　：

ータの候補単位での抽出精度を見ると，抽出された181の事象データの候補のうち，正しく抽出されているものが164（90.6パーセント）あり，誤りを含んだものが17（9.4パーセント）であった。

一般に内容分析のコーディングにおいては80パーセント以上が許容可能な精度とみなされていることを考えると，以上の数値はまずまずの精度といってもよいだろう[20]。なお，この段階での誤りの多くは，いわゆる行為者が並列的に列挙してある以下のような文章から生じたものであった。

《例文３》交渉には日本側から外務省の加藤良三アジア局長，韓国側は金夏中アジア太平洋局長がそれぞれ代表として出席，次回はソウルで行うことに合意した。
《例文４》池田行彦外相，塚原俊平通産相ら，韓国側からは孔魯明外相，朴在潤通産相が同席した。

したがって，複数の行為者を含む文章からの事象データの抽出方式をさらに改善することで，さらなる精度の向上を目指すことが可能である。

２．事象データ選別の精度

次のプロセスである事象データ候補から事象データとみなせるものを取り出す段階では，⑴「が」，「は」，「も」で終わる文節に含まれる行為者は国，組織，個人のいずれかであること，⑵その係り先となっている文節は動詞を含むこと，⑶その動詞には希望，仮定，推測，伝聞などを意味する助動詞が含まれていないこと，などを条件とした。ここで行われる判断は，それぞれの候補に関して，それを事象データとみなすかみなさないかの二者択一である。そこで，ここでは筆者の判断とコンピュータ・プログラムによる判断とを比較してみた。その結果は表１に示されている。181の候補のうち両者の判断が一致したものが159（87.8パーセント）あり，不一致が22（12.2パーセント）であった。両者の不一致に関しては２つの種類がある。ひとつは，筆者は事象データとみなすべきだと考えたがコンピュータはそのように判断しなかったものである。もうひとつはその逆のケースである。まず，前者に関しどのようなものがそれにあたるかを見てみる。その約３分の１は何らかの交渉が開かれたことを伝える次のような文章であった。

《例文５》韓国との第１回交渉は８月13日に東京で行われた。

交渉開催を伝えるニュースは，例文５のようにほとんどの場合，受動態で書かれており，「が」，「は」，「も」で終わる文節に個人，組織，国家を表す語が用いられることがない。このため，これらは事象として識別されず振り落とされてしまうことになる。これらを事象として残すためには，受動態の原文を「（日本は）韓国と交渉した」という形に変換する手続きをプログラムに追加する必要がある。例文３が示すように，このタイプの文章には誰によってそれ

表 1　事象候補からの事象選択に関する精度

		コンピュータの判断		計
		事象	非事象	
筆者の判断	事象	109	14	123
	非事象	8	50	58
計		117	64	181

が行われたかということが明記されておらず，この変換はかなり困難であるが，今後の課題として取り組んでいきたいと考えている。

事象として残すべきではあるが，コンピュータプログラムによって振り落とされてしまったものとしては，上記のタイプのほかに，例文 6 のように「が」,「は」,「も」の手前に括弧が位置するものがあった。

《例文 6 》予算委では斉藤文雄氏（自民）が竹島問題の対応をただしたのに対し，…

4 つの事象候補がこれと同様の理由で事象とみなされなかったが，これはシステムの設計上，右側の括弧より手前の部分を助詞「が」,「は」,「も」と切り離し，別の文節として記録する方式を用いたことから生じた問題である。この問題に関しては，プログラムの修正は比較的容易なので，できるだけ早い時期に対処したいと考えている。

これらとは別に，表 1 に示されているように，筆者は事象とはみなせないと判断したがコンピュータは事象とみなしてしまった場合が 8 件ある。これらを詳しく見てみると，2 つのタイプに分けられる。ひとつは例文 7 のように予想や推測などを意味する動詞を含む事象の候補からなるグループである。

《例文 7 》韓国政府関係者は日本が日本近海で違法操業している韓国漁船などを取り締まるために二百カイリを宣言すると見ている。

例文 7 からは「韓国政府関係者は／見ている」，および「日本が／宣言する」という 2 つの事象の候補が得られるが，前者は何らかの具体的な行動を示しているわけではないし，後者は韓国側の推測の内容であり，実際に日本が行った行動ではない。したがってこれらを事象とみなすことには無理がある。今後は，動詞の内容に即して事象の判別が行えるようにシステムを改善していく必要があろう。

もうひとつのグループは，例文 8 のように，何らかの名詞（「機動訓練」）を修飾する節から

事象が誤って抽出されたケースからなる。

《例文 8 》韓国国防省は14日，竹島（韓国名は独島）近海で15日に海軍と空軍が参加した機動訓練を行う，と発表した。

例文 8 からは「韓国国防省は／発表した」，および「海軍と空軍が／参加した」という事象の候補が得られるが，後者に関しては予定であり実際に起きた事象ではない。このような候補が事象とみなされないようにするためには，いわゆる埋め込み文の動詞の時制をチェックする方法を何らかの形で組み込む必要があろう。

3．格分析の精度

ここまでは事象データの抽出に関しての精度を検討してきたが，以下では事象データの動詞部分に係る，時間，場所，行動の対象・相手などの識別を目的とした格分析の精度を検討する。格分析の結果は表 2 に示されている。表 2 は本来このように分類されるべきだと考えられる「正しい」分類を表側に，得点の積み上げとファジー関係とを用いて得られたコンピュータによる分類結果を表頭に示している。したがって，表 2 の左上から右下に至る対角線上に記載された数字が正しく分類された数となる。

なお，表 2 の最初の 6 項目は 5 W 1 H に対応したものであるが，最後の 3 項目は独自の観点から追加した分類項目である。7 番目の「About」は，ある行為がどのような事柄に関連して行われたかを示す「関連節（格)」とでも呼べる節（格）を示している。8 番目の「Content」はいわゆる内容節と呼ばれるもので，発言の内容などを示す節を示している。最後の「Drop」は以上のどれにも該当せず，分類不要（「その他」）とされた節を示している。

全体の識別精度は表 2 に示されているように89.0パーセントとなっており，約 1 割程度のエラーはあるものの，人手に頼って内容分析を行う場合に一般に容認されているエラーの許容範囲である 2 割前後と比べれば，かなりの高精度といえよう。しかしながら，個別に見てみると，「How」および「Drop」の 2 項目に関しては，80パーセント前後の識別精度となっており，さらなる改善が必要であろう。ただし，最も識別精度が低かった「How」に関しては，基本的には副詞を対象とするものであり，事象の抽出という大目的から見ると，その重要性は必ずしも高いとはいえない。

以上のように，事象の抽出および格分析の精度に関しては，内容分析を主目的とするソフトウェアとしてはまずまず期待のもてる結果が得られたものと考えられる。以上の分析結果から判断する限り，このまま実際のリサーチに応用しても，人手に頼るコーディングと同等あるいはそれ以上の精度でのデータの生成が可能だと考えられる。しかしながら，このセクションで検討した改善点のいくつかは比較的容易に実行できるものであり，まずはそれらについてシステムの改善を図りたいと考えている。

表 2　格分析の解析精度

		コンピュータによる分類									
		When	Where	What	Whom	How	About	Content	Drop	計	精度(%)
正しい分類	When	29	0	0	0	2	0	0	1	32	90.6
	Where	0	20	0	0	0	1	0	1	22	90.9
	What	0	1	68	2	0	0	1	1	73	93.2
	Whom	0	2	0	14	0	0	0	0	16	87.5
	How	0	1	1	0	15	0	1	1	19	78.9
	About	0	1	0	0	0	13	0	1	15	86.7
	Content	1	0	1	0	0	0	23	0	25	92.0
	Drop	0	0	0	1	0	2	2	20	25	80.0
	計	30	25	70	17	17	16	27	25	227	89.0

IV．今後の改善に向けて

前のセクションで述べたように，本稿で報告してきた事象データ抽出ソフトウェアは，解析精度の面からみれば，ほぼ実用化のレベルに達していると考えることができるが，本ソフトウェアは単独で使用するようにデザインされたものではなく，あくまでも形態素解析および構文解析システムと併せて利用することが前提となっている。したがって，その目的が研究用のデータ生成であれ，あるいは実務的な用途であれ，実際の用途に利用するためには，形態素解析と構文解析の部分を含めたシステム全体の解析精度や使いやすさを考えなければならない。この意味で，今後手をつけていかなければならないことは次の 4 点である：⑴システムの統合，⑵辞書の拡充，⑶構文解析システムの精度改善，⑷ユーザーインターフェースの改善。以下では，これらに関して今後の方針を述べることにする。

まず，システム全体の統合であるが，既に述べたように，現時点では形態素解析，構文解析，事象データ抽出の各部分がそれぞれ独立したプログラムとなっており，ひとつのプログラムの出力結果を次の段階のプログラムへの入力データとして使用する形となっている。したがって，利用者の立場からすれば使いにくいシステムであることは否定できない。したがって，いずれかの時点で，これら 3 つの部分を統合する必要があるとは考えている。また技術的にも，これらを統合することは比較的容易である。

しかしながら，各部分の精度をさらに高めるための改善努力を今後しばらくは続けていくとなると，当面はそれぞれを単独のプログラムの形にしておいたほうが開発の面からみれば取り組みやすいことも確かである。以上のような観点から，システムの統合はある程度先の話ということにして，当面は各部分の精度向上を最優先する考えである。

辞書の拡充に関しては，大きく分けて 2 通りの考え方ができる。ひとつは筆者が独自に構築したものをさらに拡充していくという考え方であり，もうひとつは公開されている既存のもの

を利用するという考え方である。前者の方式をとるなら，今後とも辞書データベースの構築を続けなければならないし，後者であれば，既存の辞書データベースの仕様に合わせて，開発済みのシステム自体をかなり大幅に修正しなければならなくなる。辞書構築とプログラム修正とに要する作業量を比較すると明らかに，後者のほうが容易であることは否定できない。以上の判断から，筆者は既に既存の辞書ファイルを入手し，それらを作成済みの形態素解析システムに組み込む作業を始めたところである[21)]。

構文解析システムの精度改善に関しては，既に触れたように[22)]，最新のニューラル・ネットワーク・モデルの利用も含め，構築済みの構文解析システムの見直しと精度向上の努力を続けているところである。また，筆者が開発した方式では現在のところ文法ルールが全く用いられていないが，ある程度のルールを導入した「ハイブリッド方式」の可能性も視野に入れた改善を志向している。

形態素解析，構文解析，事象データ抽出のいずれの部分に関しても，基本的には筆者自身が使用することを前提に開発を進めてきたため，いわゆるユーザーインターフェースに関しては，ほとんど考慮してこなかった。しかしながら，将来他のユーザーに利用していただくことも視野に入れるならば，ユーザーインターフェースを考えることが必要になってこよう。ただし，現時点では商品化を目的に開発を進めているわけではないので，筆者自身としては，この部分に関する優先度はあまり高くはおいていない。つまり，システムの根幹部分の精度を高めることこそが最優先事項である。ユーザーインターフェースを考慮したプログラム修正作業は，それが終了した時点での最後の作業段階と位置付けている。

今後は，目下進行中の構文解析システムの精度改善の作業を継続するとともに，構築済みのシステムを実際の応用研究に利用する実証作業を近く開始し，実証作業から得られるさまざまな知見をシステムの改善作業にフィードバックさせ，本システムの実用性を高めていきたいと考えている。

本稿で報告された研究の一部は2001年度文学部研究補助金による助成研究の成果に基づくものである。

註

1）本稿で報告するシステムは日本語のみを対象としているので，当然ながら国外の報道に関しては，まずはその内容を日本語に翻訳することが必要となる。

2）事象データを用いた研究方法を解説した初期の文献としては以下を参照。Edward E. Azar, Richard A. Brody, and Charles A. McClelland, International Events Interaction Analysis: Some Research Considerations, Sage Publications, 1972; Philip M. Burgess and Raymond W. Lawton, Indicators of International Behavior: An Assessment of Events Data Research, Sage Publications, 1972; John H. Sigler, John O. Field, and Murray L. Adelman, Applications of Events Data Analysis: Cases, Issues, and Programs in International Interaction, Sage Publications, 1972; Charles F. Hermann et al., CREON: A Foreign Events Data Set, Sage Publications, 1973.

事象データを本格的に研究に利用したのは数量的手法を用いる国際関係論の研究者であったが，

彼らに大きな影響を与えたのは心理学者であるチャールズ・オスグッドが提唱したEvaluative Assertion Analysisと呼ばれる分析手法であった。Charles Osgood, et al., "Evaluative Assertion Analysis," Litera (1956) 3, pp. 47-102.

3）筆者が調べた限りでは，日本のマス・コミュニケーション研究者による文献で「事象データ」に触れているものは見あたらない。

4）このあたりの事情に関しては以下を参照。Deborah J. Gerner et al., "Machine Coding of Event Data Using Regional and International Sources," International Studies Quarterly (1994) 38, pp.91-119.

5）マスメディアによる報道内容の研究データとして事象データを用いる場合と，国際関係の研究において事象データを用いる場合とでは前提が根本的に異なる。すなわち前者の場合は，報道内容から抽出される事象データを比較することにより，各種メディアの報道状況の比較を行うわけであるから，当然ながら，マスメディアは何らかの基準に基づいて，現実の世界から報道に値すると思われるものを取捨選択して報道しているとの前提に立っているわけである。これに対し国際関係の研究の基礎データとして事象データを用いる場合は，マスメディアの報道などを含む公刊された資料から事象データを抽出すれば，現実の世界で生起するほぼすべての出来事を網羅するデータが得られるとの前提に立つものである。このように両者の立場は大きく異なるが，本稿の目的は開発された事象データ抽出システム自体の報告であるため，ここでは両者の基本前提の相違を明確にしておくだけにとどめる。

6）あらかじめ定めておいたインターネット上のサイトを定期的に巡回して，そこに記述されている内容から事象データを抽出・蓄積していくようなシステムを構想することも可能である。また，情報収集にあたってはエージェント型の検索・探索方式をこれに採り入れることもできよう。

7）市販のテキスト・マイニング・ソフトは本稿で報告するソフトウェアの機能を一部実現しているかもしれないが，筆者はそれらのソフトウェアを実際に使用した経験がないので，この点に関しては適切な評価はできない。

8）形態素解析および構文解析のために筆者が開発したソフトウェアに関しては，以下の論文を参照。吉田文彦「内容分析のための日本語形態素解析システムの構築」『東海大学紀要文学部』（第68輯，1997年），49-59頁；吉田文彦「内容分析のためのニューラル・ネットワーク・モデルによる日本語構文解析システム構築の試み」『東海大学紀要文学部』（第70輯，1998年），39-49頁；吉田文彦「日本語の報道記事を対象とする構文解析システムの精度改善の試みと事象データ抽出システムの試作結果」『東海大学紀要文学部』（第74輯，2000年），59-72頁。

9）ここで述べたように，現在のところは，形態素解析，構文解析，事象データ抽出の各部分が，それぞれ独立したソフトウェアによって処理される形になっているが，これらを統合することに関しては特に技術的な問題はない。ユーザーにとっての使いやすさを考えれば，当然統合することが必要であろう。現在のような形になっているのは開発上の便宜を考えたからである。この点に関しては，さらに最後のセクションでも論じる。

10）日本語の構文解析に関しては既にさまざまな研究成果があるが，筆者はニューラル・ネットワーク・モデルを利用することで，文法ルールを全く使用しない新たな方式を採用した。その結果，実用化に一応足る程度の成果は得られているが，さらなる解析精度の向上を目指し，現在ニューラル・ネットワーク・モデルの研究者との共同研究を進めているところである。Ryotaro Kamimura and Fumihiko Yoshida, "Analysis of Complex Systems by Teacher-Forced Learning". 9th International Conference on Information Processing（ICONIP）にて発表（シンガポール，2002年11月）。

11）文節の最後の助詞が「が」，「は」，「も」のいずれかであっても，そのひとつ手前に別の助詞がある場合は，「のが」，「のは」，「のも」を除いて，主格としては取り上げていない。

12）三上　章『象は鼻が長い（増補版）』くろしお出版，1964年。

13）名詞に関しては，意味に応じて，11種類のタイプからなる分類方式を用いている。形態素解析に用いる名詞用のデータベースには，各エントリーがどのタイプに属するかを記述してあり，形態素分析ではその情報も出力ファイルに含まれている。したがって，事象データ抽出の際にもその情報が利用できるようになっている。

14）以下で用いる例文はすべてCD-ROM版の毎日新聞（1996年）より得たものである。

15）モダリティに着目した報道記事の分析に関しては以下を参照。藤田真文，「新聞報道における論評の表明――モダリティ概念によるテクスト分析」，鶴木眞（編）『客観報道:もう一つのジャーナリズム』成文堂，1999年，93－125頁。

16）ファジー関係に関しては，以下を参照した。電気学会（編）『あいまいとファジィ――その計測と制御』オーム社，1991年。

17）「何は何である」というタイプの記述部分も定型的な形で取り出すことができるわけである。したがって，これらの部分に関しても分析用のデータを出力することは当然可能である。

18）図1の例文からは3つの事象が，図2の例文からは2つの事象が抽出されたが，これらの図ではそれぞれ最初に抽出された事象のみが示されている。

19）解析精度の検証に用いた文章はいずれもCD-ROM版の毎日新聞（1996年）より得たものである。なお，これらはすべて日本と韓国との間での「竹島」の領有権問題に関する記事から選ばれたものである。

20）内容分析におけるデータの信頼度（再現性）に関しては，どの程度までを許容範囲とみなせるかという一般的な基準が存在するわけではない。しかしながら，多くの場合80パーセント程度の数値がひとつの目安とみなされているようである。

21）公開されている辞書としてはたとえば奈良先端科学技術大学で開発されたものがよく知られているが，筆者は新世代コンピュータ技術開発機構（ICOT）により開発された辞書ファイル群を用いることにし，形態素解析システムの修正に着手している。後者は，無償で提供されており，しかも内容の改変や配布に関しても何らの制約も課されていないためである。

22）註10参照。