係り受け解析を用いた複合語の自動分割法†

宮崎 正弘††

漢字，かな，英数字などの各種の文字で構成され，一般語の他に固有名詞も含んだ一般的な複合語に対する新しい自動分割法（係り受け解析法）を提案する．本解析法は次の三つの部分から構成される．第一は，すべての可能な分割パターンを効率よく生成する部分，第二は，複合語を構成する単語が意味的にどのように結合しているかを解析するための係り受け解析部分，そして最後は，係り受け解析結果のなかから最適な分割パターンを選択する部分である．新聞記事に含まれる複合語の分割に適用した実験結果によれば，本手法は従来の最長一致法や分割数最小法に比べて精度よく，複合語を分割できることがわかった．

1. まえがき

通常の日本語文は，漢字，ひらがな，カタカナ，英数字など用いられる字種が多く，単語単位に分ち書きされず単語の区切りが明確でない．さらに，漢字は造語力が強く連続した漢字列で構成された複合語が頻繁に現れる．このような漢字かな混りの日本語文をコンピュータで解析するには，まず文章中の単語を正しく認定し，他の単語と分離する自動分ち書きが基本的な技術の一つとして重要である．

このうち，複合語に関しては，語数が多いこと，次次に造語が生まれることなどにより，これをすべて辞書に登録することは不可能なため，複合語を辞書にある基本的な単語の組合せに意味的にも正しく分割することは自動分ち書きにおける重要な課題の一つとなっている．ここで，ひらがな列は付属語，活用語尾，形式名詞，ひらがな書きされることの多い自立語など比較的限定された単語より構成され，単語間の文法的接続関係による単語接続検定が有効である．これに対し漢字列で構成された複合語は，名詞，接辞など限定された品詞の単語より構成されるので単語間の文法的接続関係による単語接続検定は有効でない．また，固有名詞まで含めると数十万に及ぶ膨大な単語が，その構成要素となりうる．さらに，カタカナ語，英数字なども漢字単語といっしょに複合語を構成する場合がある．上記の理由により，漢字かな混り文の自動分ち書きにおいて，複合語の自動分割はひらがな列の分割に比べむずかしい点が多く，本分野の最も困難な問題となっている．

† Automatic Segmentation Method for Compound Words Using Semantic Dependent Relationships between Words by MASAHIRO MIYAZAKI (Yokosuka Electrical Communication Laboratory, N. T. T.).

†† 日本電信電話公社横須賀電気通信研究所

従来，複合語の自動分割についてはいくつかの研究が行われている．長尾らは漢字列で構成された複合語を2字漢字による辞書引きにより分割し，分割にあいまいさのある部分列を1文字漢字の接頭語，接尾語となる確率をもとに，3～5文字漢字列の代表的な分割パターンを組み合わせて分割する方法を提案している[1]．しかし，この方法では，3文字以上の単語や固有名詞に対する配慮がなされておらず，またカタカナ語や英数字を含む複合語を取り扱うことができない．池田らは長尾らの方法に改良を加え，日本語キーワード向きの複合語分割法を提案している[2]．しかし，この方法でも数詞に対する配慮はなされておらず，固有名詞などの特殊語は辞書引きにより，最初に優先的に複合語より抽出し，他の部分とは分離しているため，この優先抽出による分割失敗が生じうる．とくに姓，名，地名，企業名など各種の固有名詞が多数辞書に収録されると，かなりの数の1文字漢字が固有名詞になること等により，本来一般語の一部となるべき漢字までが固有名詞として優先的に抽出されてしまうなど分割失敗が多数生じる．したがって，きわめて限定された固有名詞しか使われない分野にしかこの方法は適用できない．

中井らは JICST のキーワードより抽出した基本的な単語を収録した語基辞書により，語基で切断可能なすべての分割パターンを求め，より長い語基で切断された分割パターンを選ぶという手法で漢字，カタカナ，英数字等を含む複合語を分割する方法を提案している[3]．しかし，この方法では最も長い語基で切断された分割パターンが複数ある場合には，分割は一意に定まらない．また，科学技術文献を主たる対象としていることから，固有名詞に対する配慮はされていない．

この他に，コンピュータによる日本語文の解析によ

く用いられている最長一致法[4]を複合語分割に適用した例も報告されている[5),6)]．しかし，最長一致法では最初に辞書とマッチした文字列が優先的に単語として抽出されるが，その根拠は必ずしも明らかでない．したがって「輸出国政府」のように複数の分割パターン（輸出/国政/府，輸出/国/政府）がある場合，正しい分割パターン（輸出/国/政府）が最初から生成されずに分割誤りとなるなどの問題がある．また，辞書にない未知語を含むため，分割パターンの生成に失敗した場合の扱いも問題となる．

一方，複合語よりすべての単語候補を抽出し，文法的接続関係を考慮してこれらを接続し，すべての分割パターンを生成する総当り法では，複数の分割パターンがある場合，最長一致法のような問題が生じない．また，未知語を含む場合でも，その位置を推定することができる．しかし，複数の分割パターンから，いかにして正しい分割パターンを選択したらよいかという新たな問題が生じる．吉村らは，べた書き日本語文に対する文節数最小法による単語分割法を提案している[7)]．ここで，複合語を一つの文，複合語内の単語を文節とみなして，複合語の分割に文節数最小法を適用することが考えられる（以下，これを分割数最小法と呼ぶ）．従来，複合語分割に分割数最小法を適用した場合の分割精度などについての定量的な評価については発表されていない．しかし，分割数最小法では単語の意味的結合関係を考慮していないため，名詞，接辞など限定された品詞の単語より構成される複合語に対し適用した場合には，限界があることが予想される．

これに対して本論文では，複合語の解析に単語間の意味的接続関係（係り受け規則）を導入し，漢字，カタカナ，英数字，ひらがな等各種の文字で構成され，一般語のほかに固有名詞も含んだ一般的な複合語を，従来の方法に比べて精度よく分割することができる新しい自動分割法（係り受け解析法）を提案する．

本方法ではまず，総当り法によりすべての分割パターンを効率的に生成する．

次に，各分割パターンにおいて単語間の係り受け解析を行い，複合語の構造を解析するとともに，一般語，固有名詞，接辞，数詞などの単語種別を明らかにする．ここで，係り受け規則は数詞，固有名詞，接辞，用言性名詞（サ変動詞化，形容動詞化する名詞），用言化しない一般語に関するもの 14 種を作成した．複数の分割パターンが生成された場合には，分割数，係り受け数等を基に最適な分割パターンを選択する．

1日分の新聞記事を用いた実験によれば，係り受け解析法は，従来の最長一致法や分割数最小法に比べ，複合語の分割失敗件数をそれぞれ 1/15，1/7 に減らすことができ，辞書を充実することにより 99.8% の分割精度を達成できることがわかった．

2. 分割パターンの作成

本論文において対象とする複合語は，単語辞書にある単語（原則として短単位で収録されている）が複数結合した語であり，全体として文法的に一つの品詞としての働きをするものである．ここで用言の複合語は2単語の結合が大部分であるのに対し，名詞の複合語は，多数の単語が結合して，長い複合語を構成している場合が多い．したがって，本論文では複合語分割を行う上で問題となる名詞を主たる対象とする．なお，ここでは，実際の文中にはよく出現する「昨日山」のように本来複合名詞とはいえない「副詞的名詞＋名詞」，「一斉射撃」や「我が国」のように「副詞/連体詞＋名詞」の形で複合名詞的に使われるもの等も複合名詞に含めて取り扱うこととする．

2.1 単語候補の抽出

分割の対象となる複合語（文字数：M）を $\{k_m\}=k_1k_2\cdots k_m\cdots k_M$ とおく．k_m $(m=1\sim M)$ から始まる l 文字の部分列 $K_{m,l}=k_mk_{m+1}\cdots k_{m+l-1}$ $(l=1\sim M-m+1)$ で単語辞書引きを行い，$\{k_m\}$ に含まれるすべての単語候補を抽出する．

ここで，無限の単語が生成される数詞，種々の外来語，擬音語・擬態語等を表すカタカナ語，種々の略語等を表わす英字略語，通常，漢字書きされるひらがな語をすべて辞書に収録することは辞書の容量の点から困難である．したがって，辞書引きにより文字列全体が一つの単語候補として抽出されていないカタカナ列，英字列，ひらがな列は，その文字列全体を単語候補（未知語名詞）として抽出する．また，数詞列については数詞と同形異語となる単語［七五三（しちごさん），一（はじめ）…］，数詞＋助数詞と同形異語になる単語［八戸（はちのへ）…］があることを考慮して，無条件に数詞列全体（数詞，および数詞にはさまれた小数点，カンマ等で構成される）を単語候補（数詞）として抽出する．

なお，無効な分割パターンの生成を抑止するため，単語ごとに当該単語が複合語の先頭（前要素），中間（中間要素），最後尾（後要素）に位置することができるか否かを表わす情報（複合語要素フラグ）を単語辞

書に収録し，複合語要素フラグに合致しない単語候補を，分割パターンの生成前に除去することとした．図1に単語候補の抽出例を示す．

2.2 分割パターンの生成

2.1節で抽出した単語候補のうち，複合語の最終構成要素とならないものの集合を $\{s_i\}$，複合語の最終構成要素となるものの集合を $\{s_j\}$ とする．各 s_j について s_j の前方に文法的に接続可能な単語を $\{s_i\}$ から選び，前方に次々に連鎖していくことによりすべての可能な分割パターン* を生成する．なお，2.1節で生成した未知語名詞に完全に包含される単語連鎖が生成された場合，当該未知語名詞を上記の単語連鎖で置換する．

ここで，分割パターン数が爆発的に増加することを抑止するため，以下のような工夫を行った．

① 単語間の文法的接続関係をチェックするため，各単語に文法カテゴリーを付与する．ここで同一の文法カテゴリーを有する同形異語は一つのグループにまとめ，単語連鎖において一つの単語として扱う．

② 複合語の品詞が名詞の場合，一般名詞（用言性名詞，連用形名詞などを含む），固有名詞，数詞＋(助数詞)，時詞（副詞的に使われる時を表す名詞）には同一の文法カテゴリー（名詞）を付与する．なお，名詞を同形異語にもつ体言型接辞（単語＋接辞全体が名詞となるもの，助数詞を含む）において，名詞，および，接辞の両方が文法的接続条件を満たす場合，当該接辞は名詞と同一のグループにまとめる．

③ 接頭辞，接尾辞がともに文法的接続条件を満たす場合，当該接辞（助数詞を含む）の文法カテゴリーを接頭/接尾辞として一つのグループにまとめる．

④ 他の長い単語候補* に完全に包含される単語連鎖は原則として生成しない．たとえば「物理学」という単語候補*がある場合，「物理学」に完全に包含される「物理/学」という単語連鎖は生成しない．なお，本規則により有効な分割パターンが生成されず，分割エラーとなる場合がある．たとえば「東進」（とうしん）というサ変動詞化する名詞により，「東/進」（あずま/すすむ）という姓＋名の分割パターンが生成されず分割失敗となる場合がある．そこで，上記のような分割失敗を生じる可能性のある単語**を調べ，当該

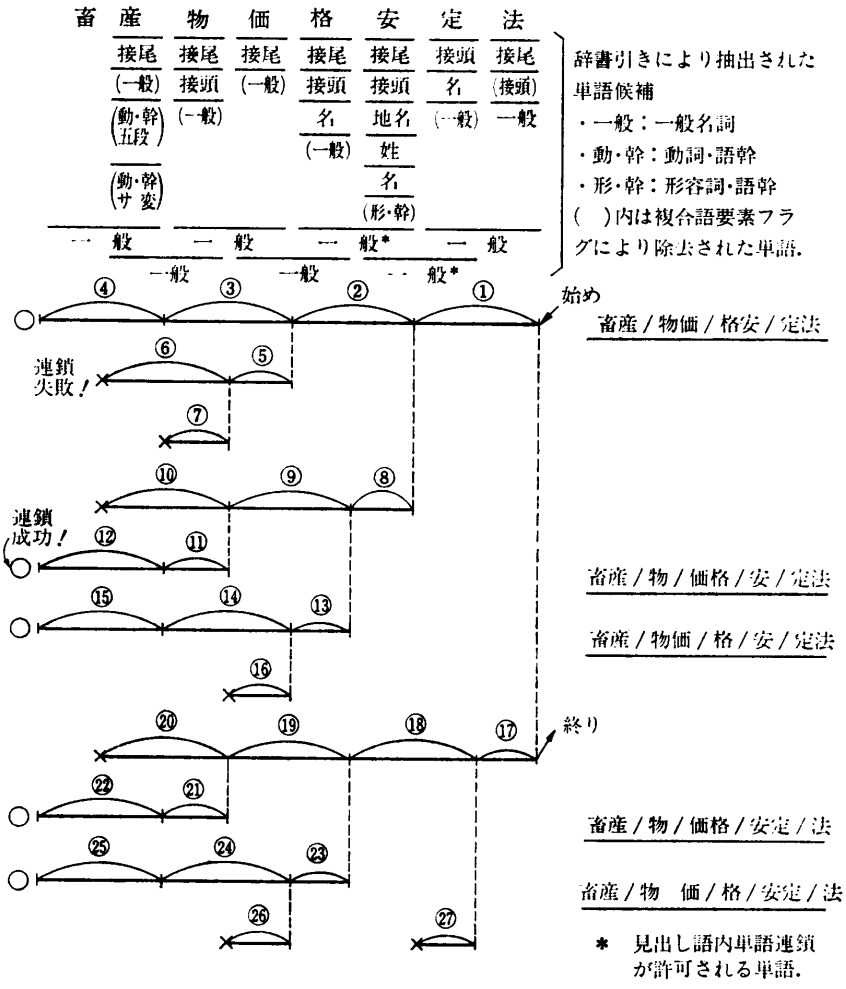

図1 分割パターンの生成例
Fig. 1 An example of word segmentation pattern generation.

* 複合語の先頭まで到達した単語連鎖を連鎖成功とし，一つの分割パターンとする．

* 連鎖成功した単語連鎖の構成要素となる場合に限る．

** 単語内に姓＋名，接尾＋接頭，助数詞＋接辞・名詞を含むもの等がある．

単語に完全に包含された単語連鎖の生成を許可する旨のフラグ（見出し語内単語連鎖フラグ）を設定した．

もし，分割パターンの生成にすべて失敗した場合には，単語連鎖に失敗した部分の前方に1文字以上の長さをもつ未知語名詞を仮定し，再度連鎖を試みる．以下，単語連鎖が複合語の先頭に到達するものが少なくても一つ得られるまで上記の処理を繰り返す．

図1に分割パターンの生成例を示す．

3. 係り受け解析

複合語を構成する単語が意味的にどのように結合しているかを解析し，一般語，固有名詞などの単語種別を明らかにするために，2章で生成された各分割パターンにおいて単語間の意味的接続関係（係り受け）の解析を行う．複合語を構成する単語を先頭から，適当な付属語等を補いながら読み換えることにより通常の文に変換できる* ことに着目して，係り受け解析では，通常の文解析における係り受けと同様に，以下の原則に従うこととした．

① 前方の単語から後方の単語に係る．
② 単語の係り先は一つに限る．
③ 複数の単語を一つの単語が受けてよい．
④ 係り受けの非交差性を守る．

以下，複合語の構造を解析するための各種係り受け規則，および係り受け解析法について述べる．

3.1 数詞に関する係り受け規則

数詞は単独で用いられるよりも，助数詞，接辞と結合して用いられることが多い．助数詞には，「約」，「第」などのように数詞の直前に位置する前置助数詞と，「本」，「回」など数詞の直後に位置する後置助数詞がある．接辞には，「強」，「未満」などのように後置助数詞の直後に位置する接辞（助数詞承接型接辞と呼ぶ）や，「以上」，「未満」など数詞の直後に位置する接辞（数詞承接型接辞と呼ぶ）がある．以上，まとめて表1の1～4のような四つの係り受け規則を作成した．係り受け解析は複合語の前方から数詞を探索することにより行う．

3.2 固有名詞に関する係り受け規則

地名，人名，企業名などを表す固有名詞はその種類，語数も多い．ここでは，このような固有名詞を意味によって図2のように分類し，各固有名詞に固有名詞属性を付与した．なお，単語辞書の容量をできる限り小さくするため，固有名詞の辞書への収録は原則として短単位とし，本来固有名詞と考えられるもののみを収録した．これにより，たとえば「京都府」，「京都市」，「京都駅」，「京都大学」，「京都銀行」などは「京

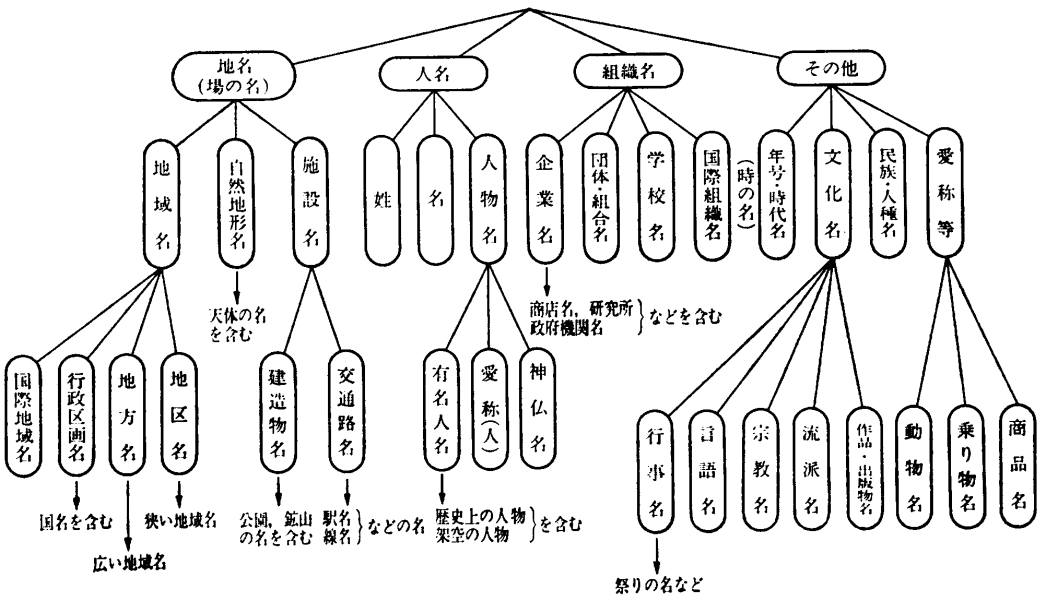

図2 固有名詞の意味分類
Fig. 2 Classification of proper noun semantics.

* たとえば，「情報処理システム」は「情報を処理するシステム」と変換できる．

表1 係り受け規則（その1）
Table 1 Rules of semantic dependent relationships between words (No. 1).

No.	係り受けの型		例
1	前置助数詞－数詞		約 10 ，第 八 回
2	数詞－後置助数詞		二 本 ，50 パーセント
3	後置助数詞－助数詞承接型接辞		50 kg 強 ， 数 % 台
4	数詞－数詞承接型接辞		100 未満 ， 10 以下
5	固有名詞－固有名詞承接語	地名	東京 駅 ，関東 平野
		人名	平野副 社長 ，一 郎 君
		組織名	三 井 信 託 銀 行，上 智 大 学
		その他の固有名詞	明治 時代 ， アイヌ 人
6	役職承接型接辞－役職		美濃部前 都 知事 ，元 総 裁
7	姓－名		加 藤 一 二 三
8	包含関係のある地名の連接		神奈川県 横須賀市 武
9	接頭辞＋単語		超 短 波 ，極 超 短 波
10	単語＋接尾辞		近代 化 ， 大型 機 用
11	一般語－一般語		一 日 朝 ，県 立 高 校，我 が 国
12〜14	用言性名詞－単語		表2参照

都」という固有名詞に「府」，「市」，「駅」などの接辞や「大学」，「銀行」などの一般語が結合して長単位の固有名詞を形成しているものとみなし，固有名詞としては地名「京都」のみを辞書に収録し，行政区画名（府，市），交通路名（駅）などを固有名詞属性として付与した．

固有名詞は単独で用いられるより，複合語内で特定の単語と共起する場合が多い．たとえば，敬称（氏，様など）は人名（姓，名），役職（社長，教授など）は人名（姓）や組織名（企業名，学校名など）*，行政区画を表す接辞（県，市など）は地名（行政区画名），銀行，電気などの単語は組織名，地名と共起する．また，姓と名の連接，包含関係のある地名（行政区画名）の連接（例：東京都千代田区丸の内）など固有名詞同士の共起現象もある．ここでは複合語内で固有名詞と共起する一般語や接辞（固有名詞承接語と呼ぶ）を約2000語収録し，どのような固有名詞と共起するか，また共起する固有名詞は当該固有名詞承接語に対して複合語内でどのような位置にあるか等を分析し，その結果を係り受け情報として辞書に収録した．

以上の共起現象から表1の5〜8のような四つの係り受け規則を作成した．係り受け解析は，複合語の前方から固有名詞承接語を探索し，探索された固有名詞承接語の係り受け情報をもとに，当該固有名詞承接語と共起する固有名詞を探索することにより行う．ここで，役職のように複数の固有名詞と共起するものについては，係り受け情報として固有名詞の探索順序を収録し，その情報をもとに固有名詞を探索する*．姓，名，地名が探索された場合，当該固有名詞と共起する固有名詞（名，姓，包含関係のある地名）を探索する（固有名詞承接語を伴わずに固有名詞同士が共起する場合には，別途このような係り受けをチェックする）．連接する固有名詞と固有名詞承接語の係り受けが成立し，固有名詞承接語に固有名詞属性が付与されている場合，固有名詞＋固有名詞承接語を新たに固有名詞（後部単語に付与された固有名詞属性をもつ）とする．また，姓＋名も新たに固有名詞（人名）とする．

ここで，固有名詞が一般語や接辞と同形異語となる場合，以下のような例外処理を行う．

① 探索された固有名詞承接語が固有名詞と同形異語の場合，固有名詞承接語と仮定して係り受け解析を行う．以後の処理において，矛盾が生じ当該単語を固有名詞と認定した場合，固有名詞承接語と仮定して行った係り受けを無効とする．**図3**の①に例を示す．「時代」を固有名詞承接語と仮定し，「江戸時代」を固有名詞（時代名）と認定する．しかし，「時代」の直後

* 新，前，元，現などの特定の接頭辞（役職承接型接辞と呼ぶ）も役職と共起する．

* 役職承接型接辞の探索も同様に行う．

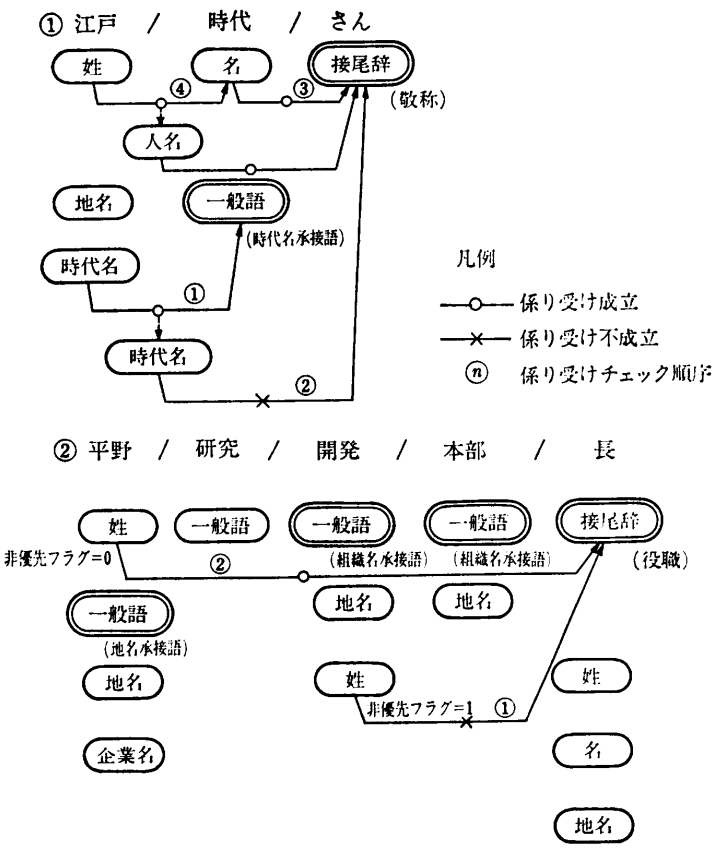

図 3 同形異語を持つ場合の係り受け処理の例
Fig. 3 Examples of semantic dependent relationships between words process for compound word with identical characters.

にある固有名詞承接語「さん」は直前に姓，名などを要求し，時代名が直前に位置することはない．したがって，「時代」を名（ときよ）と認定し，名と共起する姓を探索し「江戸」を姓と認定し直す．一般に複合語全体の意味を規定する上で複合語の後方の単語ほど重要であるから，係り受け解析で矛盾が生じた場合，後方の固有名詞承接語による係り受けを優先することは妥当である．

② 探索された固有名詞が一般語や接辞と同形異語の場合，固有名詞の非優先フラグ* か“0”ならば固有名詞と認定する．しかし，非優先フラグが“1”ならば当該単語と共起する固有名詞がない限り固有名詞と認定しない．図3の②に例を示す．探索された姓「開発」は非優先フラグが“1”で，共起する固有名詞がないため，一般語とみなし，さらに前方にある姓を探索する．探索された「平野（ひらの）」は一般語と同形異語であるが非優先フラグが“0”のため姓と認定される．

* 一般語，固有名詞，接辞，数詞＋(助数詞)間の同形異語などにおいて，使用頻度のきわめて低い単語である旨を示すフラグ．未知語名詞にも付与される．

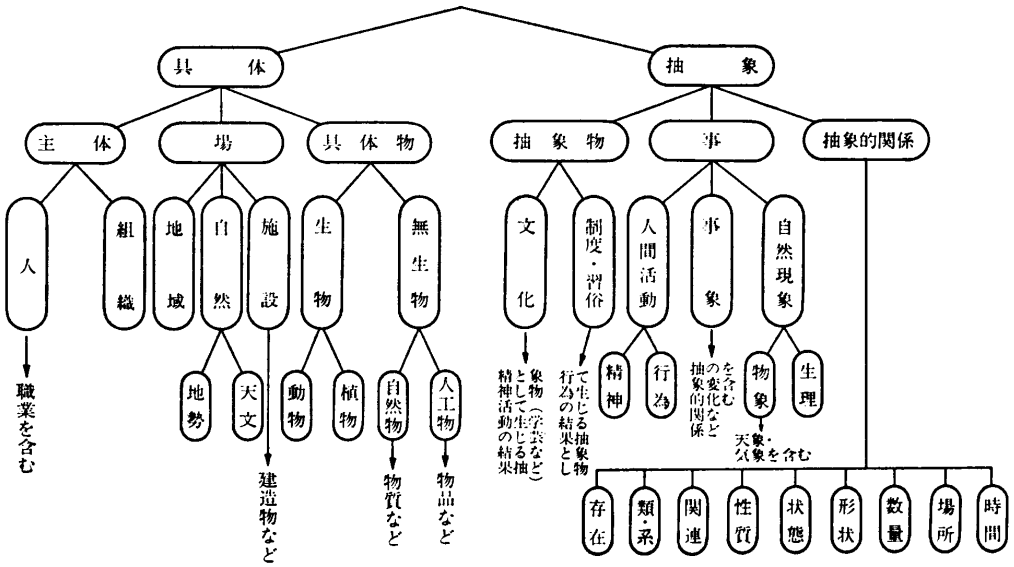

図 4 一般名詞の意味分類
Fig. 4 Classification of common noun semantics.

3.3 接辞に関する係り受け規則

接辞は必ず他の単語と結合して用いられる．ここでは，3.1～3.2 節において，係り受け関係の成立しなかった接辞を対象に，表1の9～10のような接辞の結合対象となる単語（接辞承接語と呼ぶ）との係り受けをチェックする．ここで，一般名詞を意味によって図4のように分類し，国研の分類語彙表[8]等をもとに，各一般名詞に一般名詞属性を付与した．接辞が特定の一般名詞属性の単語を接辞承接語とする場合，逆に特定の一般名詞属性の単語を接辞承接語としない場合には，その情報を接辞の係り受け情報として辞書に収録し，接辞と接辞承接語が意味的に結合しうるかをチェックする．ここで連接する接辞承接語と接辞の係り受けが成立する場合，接辞＋接辞承接語を新たに単語（後部単語に付与された一般名詞属性をもつ）とする．なお，接辞が接頭辞/接尾辞の両方になりうる場合で，上記の係り受けにより判定できない場合には，使用頻度の大きいほうを選択する．

3.4 一般語の係り受け規則

数詞，固有名詞，接辞を含まず，一般語がいくつも結合した単語列については，3.3 節までの方法では，単語間の意味的結合関係を解析できない．ここでは，2語の単語で構成された複合語における単語間の意味的結合に関する従来の研究[9)~13)]を発展させ，多数の一般語同士が結合して複合語を構成する場合，サ変動詞化する名詞（サ変動詞型名詞），形容動詞化する名詞（形容動詞型名詞）が，形は名詞のまま，意味的には表2に示すように用言化して二つの名詞を結合することに着目して，単語間の意味的結合関係を解析する．ここで，接辞承接語＋接尾辞（化，視），接辞承接語＋接尾辞（的，性，型，用，風，など）* は，それぞれサ変動詞化，形容動詞化するのでサ変動詞型名詞，形容動詞型名詞として扱う．また，「国際」，「積極」，「民主」，「自動」，「個別」，「先進」などのように通常，「的」などの接尾辞と結合して形容動詞化するが，複合語においては接尾辞と結合しない形で表2の14①のように形容動詞的に使われる名詞や，否定の接頭辞（無，不，未，非）が接続することにより全体を形容動詞化する名詞（不安定，不合理など）なども形容動詞型名詞として扱う．

係り受け解析は，複合語の後方からサ変動詞型名詞，形容動詞型名詞を探索し，辞書に収録された以下の係り受け情報をもとに，表2の 12～14 の係り受けをチェックすることにより行う．

表 2 係り受け規則（その2）
Table 2 Rules of semanic dependent relationships between words (No. 2).

No.	係り受けの型		係り受けの条件		例
			γ_iの位置等	γ_iのα or βに対する格関係等	
12	サ変動詞型名詞（α）	①αするγ_1 ②αするためのγ_1 ③αすることによって生じたγ_1	αの後方の単語*（αの直近の単語より順次，係り受けをチェックする．）	①の場合No.13と同じ．（γ_2と一致する格は除く．） ②～③の場合，γ_1は抽象的な種々の名詞**をとりうる．	予想最高気温（対象） する を 受験資格 するための 勤務成績 することによって生じた
13		γ_2（が／を／にと／へ／で／より／まかよ／で／について／にのために…）αする	αの直前の単語*（No.12の係り受けチェックの前に本係り受けをチェックする．）	・動作主 ・対象 ・道具／手段／材料 ・起点／目標 ・目的 ・原因／理由／結果 ・共同 ・時間／場所／数量	政府発表（動作主） が する 情報処理（対象） を する 観光旅行（目的） のために する 水力発電（手段） で する
14	形容動詞型名詞（β）	①βなγ_1	βの後方の単語*（βの直近の単語より順次，係り受けをチェックする）	②の場合と同じ	特別扱い な が 専門的知識 な が
		②γ_2がβな	βの直前の単語*（①が成立しない場合のみ本係り受けをチェックする．）	主格	効果十分 が な 人気絶頂 が な

* 接頭辞＋単語は一つの単語とみなす．
** 形容動詞型名詞を除く．
* 「急性」，「新型」，「洋風」などの名詞もこの型に含める．

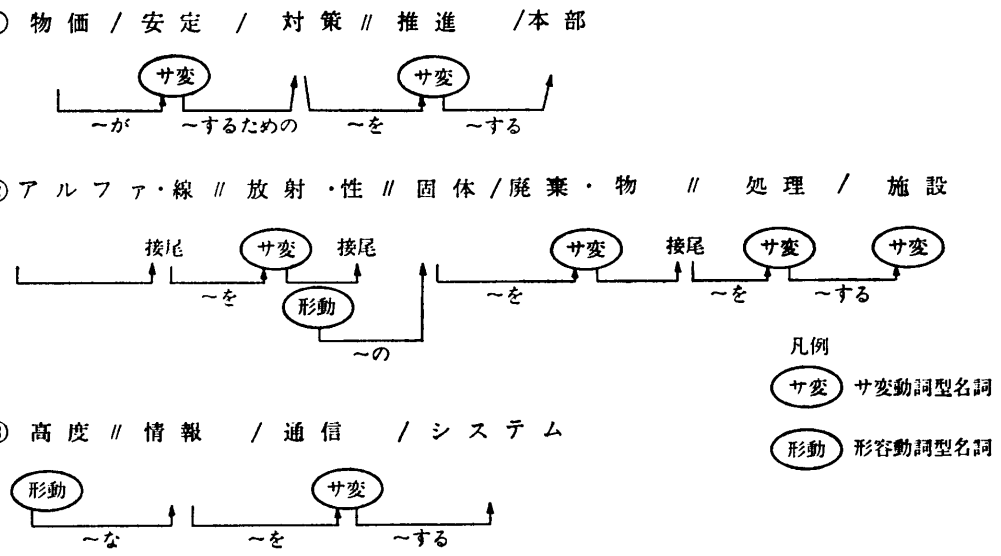

図 5 係り受け処理の例
Fig. 5 Examples of semantic dependent relationships between words process.

① サ変動詞型名詞：修飾語 γ_2 がとりうる格，および各格がとりうる一般名詞属性．

② 形容動詞型名詞：修飾語 γ_2 がとりうる一般名詞属性．

係り受け解析の例を**図 5** に示す．

用言化しない一般語同志 γ_1, γ_2 は，格助詞「の」を介して「γ_1 **の** γ_2」の形で結合するが，その結合は，通常，上記で述べた用言性名詞の場合ほど明確でない．ここでは，直前，または直後の特定の単語と意味的に強く結合する場合，前方/後方承接語の一般名詞属性を係り受け情報として辞書に収録して，表 1 の 11 のような用言化しない一般語同士の係り受けをチェックする．

4. 分割パターンの選択

複数の分割パターンが生成された場合には，最適な分割パターンを選択する必要がある．ここで，分割パターン j における分割数を α_j，2 単語間の係り受け数を β_j，非優先フラグ＝"1" の単語数を γ_j とする．α_j が最小のものを選択するのが分割数最小法であり，その有効性は文

表 3 分割パターンの選択例
Table 3 Examples of optimal word segmentation pattern selection.

例 No.	分割パターン No.	分割パターン	δ_j	最適解
例1	1	畜産 / 物価 / 格安 / 定法 (形動)	3	
	2	畜産・物 / 価格・安 / 定法 (接尾辞) (接尾辞)	3	
	3	畜産 / 物価 / 格 / 安 / 定法	5	
	4	畜産・物 / 価格 / 安定・法 (接尾辞) (サ変) (接尾辞)	2	○
	5	畜産 / 物価 / 格 / 安定・法 (サ変) (接尾辞)	3	
例2	1	森永 / 前・日銀 / 総裁 (姓) (役職承接型・接辞) (企業名) (役職)	1	○
	2	森永 / 前日 / 銀 / 総裁 (姓) (役職)	3	
例3	1	区立 / 江戸 / 川野 / 球場 (地名) (地名) (地名承接語)	2	
	2	区立 / 江戸川 / 野球・場 (地名) (地名承接語)	1	○
例4	1	藤沢 / 工場・内 (地名) (地名承接語) (接尾辞)	1	○
	2	藤沢 / 工 / 場内 (姓) (名)	2	

凡例 ×→ 係り受け失敗

献 7) ですでに報告されている．また，β_j が大きいほど係り受けにより複合語を構成する単語が意味的に強く結合されていると考えられる．したがって，$\delta_j = \alpha_j - \beta_j$ とおくと，α_j が小さいほど，β_j が大きいほど δ_j は小さくなり，δ_j が最小のもののなかに最適な分割パターンが存在すると考えられる．なお，δ_j により解が一意に定まらない場合には，単語の使用頻度による選択を行う．ここで，γ_j が大きいほど，使用頻度の低い単語を多く含んでいるため，γ_j が最小のものを選択すればよい．

上記の議論より，以下の手順で分割パターンの絞り込みを行い，最適な分割パターンを一意に定める．

① δ_j が最小の分割パターンを選択する．

② γ_j が最小の分割パターンを選択する．

③ 複合語を構成する単語の出現確率の積が最大な分割パターンを選択する．

分割パターンの選択例を**表 3** に示す．最長一致法では例 1，例 2 がエラーとなり，分割数最小法では例 1 がエラーとなり，例 2～例 4 は分割パターンが一意に定まらないが，本手法ではすべて正しい解が得られる．

5. 評　価

係り受け解析法の有効性を確認するため，適当に選んだ新聞記事 1 日分に出現する複合語* (5422 件) を対象に，係り受け解析法，分割数最小法，最長一致法による複合語分割の実験を行った．実験には，一般語や接辞約 8 万語，固有名詞約 20 万語，専門用語 (時事用語など) 約 2 万語の合計 30 万語を収録した単語辞書を用いた．単語の使用頻度は 90 日分の新聞記事

表 4 実験結果
Table 4 Result of experimentation.

手法	分割失敗件数*	分割失敗例/( ) 内は正解パターン
係り受け解析法	11 件 (1.6%)	他の方法でも失敗するもの (9件) 最大/手(最/大手)，ダウ/最高/値(ダウ/最/高値)，主戦/場(主/戦場)，米国/内/シェア(米/国内/シェア)，長期/間/貯蔵(長/期間/貯蔵)，欧州理事/会議/長(欧州/理事/会/議長)，EC/外相/理事/会議/長(EC/外相/理事/会/議長)，戦域/核・削減(戦域・核/削減)，資金・金/十五億/円(資金/金・十五億/円)
		最長一致法では正解が得られるもの (2件) 最/盛期(最盛/期)，他/国民(他国/民)
分割数最小法	74 件 (10.6%)	分割数最小のもの以外に正解があるもの (49 件) 大阪/大卒/二人(大阪/大/卒/二人)，七十/年代(七十/年/代)，輸出/補助/金的(輸出/補助/金/的)，十/年越(十/年/越)，セルビア/共和/国内(セルビア/共和/国/内)，五/割増(五/割/増)，三十一/日夜(三十一/日/夜)，坑内/員/九百/人台(坑内/員/九百/人/台)
		分割数最小のものが複数あり，分割パターンの選択を誤ったもの(22 件) 来年/度(来/年度)，各国/営/企業/体(各/国営/企業/体)，前年/度/比/伸び/率(前/年度/比/伸び/率)，ハヤカワ/氏/上院/選出/馬(ハヤカワ/氏/上院/選/出馬)，旭光/学/工業(旭/光学/工業)，第/二次/大/戦後(第/二次/大戦/後)
		品詞の認定を誤ったもの (3件) 石油・高/価格/時代(石油/高・価格/時代)
最長一致法	170 件 (24.4%)	再出/国(再/出国)，新制/度(新/制度)，初会/合(初/会合)，高利/子(高/利子)，両地/域(両/地域)，手作/業(手/作業)，主幹/事(主/幹事)，大成/功(大/成功)，高水/準(高/水準)，小委/員/会(小/委員/会)，旧帝/大系(旧/帝大/系)，米国/防/総省(米/国防/総省)，都議/会/各党(都/議会/各党)，新品/種/開発(新/品種/開発)，わが/国政/府(わが/国/政府)，生産/性向/上(生産/性/向上)，調理/師団/体(調理/師/団体)，三原/則(三/原則)，高価/格/エネルギー(高/価格/エネルギー)，丸一/年/後(丸/一/年/後)，三十一/日発/表(三十一/日/発表)，モスクワ/三十一/日時/事(モスクワ/三十一/日/時事)，五十六/年上/期(五十六/年/上期)，事実/上野/放し(事実/上/野放し)，工業/技術/院/総務/部会/計/課長(工業/技術/院/総務/部/会計/課長)

* ( ) 内は分割パターンが複数生じたものの中での分割失敗率

* 正しい分割パターンに含まれる単語を未知語（かな，英字の未知語名詞を除く）として含まないすべての名詞．

の語彙統計の結果を用いた．なお，分割数最小法において分割数が最小のものが複数ある場合には，以下の手順で分割パターンの絞り込みを行い，分割パターンを一意に定めた．

① 数詞と後置助数詞の連接しているものを選択する．

②～③ 5の②～③を適用する．

また，最長一致法，分割数最小法において，同形異語が存在する場合には，文法的接続条件を満たすもののなかから，数詞，接辞，一般語/固有名詞の順に単語を選択し，そのなかで使用頻度の最大の単語を選んだ．

実験結果を表4に示す．全入力データの 12.8% にあたる696件で複数の分割パターンが生成された．実験により以下のことが明らかとなった．

（1） 係り受け解析法は従来の最長一致法や分割数最小法に比べ複合語を精度よく分割でき，辞書を充実することにより 99.8% の分割精度が達成できる．

（2） 分割数最小法は，最長一致法より精度はよいが，分割パターンが複数生じたもののなかで 58% にあたる402件で分割数が最小のものが複数生じ，分割数最小のものの中に正解が含まれない場合は7%であった．

なお，本実験では名詞のうち数詞は独立の品詞とし，接辞は接頭辞と接尾辞を別品詞とし，単語分割パターンと各単語の品詞が正しいものを正解とした．

6. むすび

漢字，カタカナ，英数字，ひらがな等の各種の文字で構成され，一般語の他に固有名詞を含んだ一般的な複合語に対する新しい自動分割法（係り受け解析法）を提案した．

そこで，分割パターンの生成においては，無効な分割パターンの生成を抑止し，効率的に分割パターンを生成する方法を提案した．次に，複合語における単語間の意味的結合関係を解析するための 14 の係り受け規則を提案し，それらを用いた係り受け解析により一般語，固有名詞，接辞，数詞などの単語を認定する方法について述べた．さらに，複数の分割パターンのなかから最適な分割パターンを選択する方法を提案した．

新聞記事を用いた実験によれば，係り受け解析法の精度は，従来の最長一致法に比べ 15 倍，分割数最小法に比べ7倍に向上することがわかった．

本論文で提案した手法は，現在試作中の，漢字かな混り文の日本語文を音声に変換するための日本文音声変換システム[14]に組み込まれているが，本手法は漢字かな混り文の自動分ち書きを行うための基本的な技術として，機械翻訳など幅広い分野での応用が可能である．

謝辞 本研究にあたり，新明解国語辞典の使用を承諾してくださった三省堂の各位に深謝します．

参考文献

1) 長尾，辻井，山上，建部：国語辞書の記憶と日本語文の自動分割，情報処理，Vol. 19, No. 6, pp. 514-521 (1978).
2) 池田，杉山，中村：部分一致検索における複合語分割法，56年度前期情処全大，No. 4L-1 (1981).
3) 中井，岡野，佐藤，中瀬，長田，古賀，石川：日本語における語基と構文についてⅢ，情処全大，No. 5E-9 (1978).
4) 牧野，木澤：べた書き文の仮名漢字変換システムとその同音語処理，情報処理，Vol. 22, No. 1, pp. 59-67 (1981).
5) 野村，森：漢字かな変換システムの試作，信学論，Vol. J 66-D, No. 7, pp. 789-795 (1983).
6) 田中，森，糸賀：語基の接続規則に基づく漢字仮名変換方式の提案，57年度後期情処全大，No. 2K-8 (1982).
7) 吉村，日高，吉田：文節数最小法を用いたべた書き日本語文の形態素解析，情処論，Vol. 24, No. 1, pp. 40-46 (1983).
8) 国立国語研究所：分類語彙表，秀英出版，東京 (1964).
9) 野村：三字漢語の構造，国立国語研究所報告 51, pp. 37-62 (1974).
10) 野村：四字漢語の構造，国立国語研究所報告 54, pp. 36-80 (1975).
11) 野村：接辞性字音語基の性格，国立国語研究所報告 61, pp. 102-138 (1978).
12) 野村：造語法，岩波講座日本語 9，pp. 245-284, 岩波書店，東京 (1977).
13) 田中，水谷，吉田：語と語の関係について，情報処理学会，自然言語処理研究会，No. 41-4 (1984).
14) Miyazaki, M., Gotō, S., Ooyama, Y. and Shirai, S.: Linguistic Processing in a Japanese-Text-to-Speech System, Proc. of ICTP '83, pp. 315-320 (1983).

（昭和59年2月27日受付）
（昭和59年5月15日採録）