自然言語処理 70－5
（1989. 1. 20）

複合語の解析による語の上位－下位関係の自動構築

原田隆史（慶應義塾大学大学院），細野公男（慶應義塾大学文学部），田村俊作（慶應義塾大学文学部），高柳敏子（独協大学経済学部），後藤智範（愛知淑徳大学文学部），岸田和明（慶應義塾大学大学院），坂田亮子（慶應義塾大学文学部）

複合名詞の造語特性に着目した解析方法を開発し，語の上位－下位関係の構築を試みた。

構築手順は以下の通りである。1)実際の文献から複合語を自動的に切り出す，2)基本辞書とのマッチングによって構成要素に分割する，3)接頭辞，接尾辞の処理を行う，4)各構成要素に付与された品詞情報に基づいて上位－下位関係を構築する。

情報処理学会論文誌 100 件およびＪＩＣＳＴ科学技術文献速報 11,200 件の抄録を対象として，上位－下位関係の構築実験を行った結果，得られた上位－下位関係のうち93.5％は人手による結果と一致し，複合語の解析から上位－下位関係を構築することの妥当性が示された。

Developing an automatic construction method of the BT-NT relations from Japanese compound nouns.

Takashi HARADA[1], Kimio HOSONO[2], Shunsaku TAMURA[2], Toshiko TAKAYANAGI[3], Tomonori GOTOH[4], Kazuaki KISHIDA[1], Ryoko SAKATA[2]

1) Graduate School of Library and Information Science, Keio University.,
2-15-45, Mita, Minato-ku, Tokyo, 108 Japan.

2) School of Library and Information Science, Keio University,
2-15-45, Mita, Minato-ku, Tokyo, 108 Japan.

3) Faculty of Economics, Dokkyo University.,
600, Sakae-machi, Soka-shi, Saitama, 340 Japan.

4) School of Library and Information Science, Aichi syukutoku University
9, Katahira, Nagakute, Nagakute-cho, Aichi-gun, Aichi, 480-11 Japan.

Based on the fact that a lot of Japanese compound nouns consist of modificands and their modifiers, we developed an automatic construction method of the BT-NT relations from Japanese compound nouns.

The procedure we used is as follows : 1) extract compound nouns from a text, 2) decompose these nouns into their elements with reference to the special word dictionary, 3) examine special characters which are considered to be functioned as prefixes or suffixes, 4) construct the BT-NT relations from the elements processed according to the predetermined criteria.

As the result of the experiments, more than 90% of the NT-BT relations constructed automatically conformed to those by human beings.

１．はじめに

現在，情報検索あるいは人工知能，自然言語処理の分野で，シソーラスや意味解析用辞書の重要性が広く認識されている。シソーラス・意味解析用辞書を整備する際に，語間の関係，特に上位－下位関係を明らかにすることは，きわめて重要であると考えられてきた。そして，語間の上位－下位関係の構築には多大な労力を必要とし人手による上位－下位関係の構築には多くの困難が伴うため，さまざまな自動化が試みられてきた。語間の上位－下位関係を自動的に構築しようとする試みの一例として，共出現頻度に基づく統計学的手法があげられるが，上位－下位関係の構築が不十分であり，限界が指摘されている[1)]。

それに代わる方法としては，複合語の造語特性に着目し，複合名詞の解析によって語の上位－下位関係を構築することが考えられる。すなわち，以下の手法によって，上位－下位関係を自動構築するものである。

(1) 実際の文献から複合語を切り出す

(2) 複合語を構成要素（語基）に分割する

(3) 分割された語基を複合語の造語特性に従って関係づける

そこで，我々は，複合語の解析による上位－下位関係の抽出法を理論的に検討し，さらに複合語の解析によって上位－下位関係を構築することの可能性を探るために，実際に自動構築システムの一部を構築して実験を行った。今回は，この手法についての理論的背景と実験結果について報告する。

２．概念の上位－下位関係

1970年にUNESCOから発行された"Guidelines for the Establishment and Development of Monolingual Thesauri"[2)]を参考に，個々の用語間の関係を分類すると，以下の３つに分けることができる。

(1) 同義関係：例えば，通俗名と科学名の関係がこれにあたる。真の意味での同義語に加え，索引作業や検索を行う上で区別する必要がないと考えられる疑似同義語も同義語に含めて考える。

(2) 階層関係：用語同士が上位－下位の関係にある場合である。たとえば，上位語がクラスを表現し，下位語がそのメンバーを示すような場合が階層関係に該当する。シソーラス中で，階層関係は一般にＢＴ，ＮＴで示される。語の階層関係は，さらに以下の３つに分けることができる。

(a) 包含関係：この関係はクラスとそのメンバー，あるいは類と種との間のつながりを明らかにするものである。この関係においては上位語も下位語も同一の基本的概念タイプ（物，行為，特質etc.）に属している。

例）金属（原料），非鉄金属（原料）　－　包含関係にある

金属（原料のクラス），鋳造（行為）－　包含関係にはない

(b) 階層的全体－部分関係：部分名が，どんな文脈においても，それを所有している全体名に含まれる場合の関係。全体名は上位語，部分名は下位語としての役割を果たすことによって階層構造が作られる。たとえば，地名などがあげられる。

例）アジア－日本－関東地方－東京都－港区

(c) 例示関係：普通名詞で表現された，物や事柄についての一般的なカテゴリーと，そのカテゴリーに属する具体的な例を示す語との間の関係。個々の例は固有名の形態をとる。

例）客船－クイーン・エリザベスII世

(3) 関連関係：その他，同義関係でも階層関係にある訳でもないが，索引や検索を行う場合に，つながりを明確にさせるべきであると考えられる場合に設定される。通常の全体－部分関係は，この範疇に含まれる。

本研究では，これらの関係のうち(2)(a)包含関係に着目し，複合語の解析から自動的に上位－下位関係を構築するための手法について検討する。

ＪＩＣＳＴシソーラス見られる上位－下位関係のうち包含関係の占める割合は，電気工学分野では，約45％であった。これは，複合語の解析による手法を使用することによって上位－下位関係の約半分を自動構築できる可能性を示しているといえよう。

(2)(a)の包含関係が成立するには，上位概念も下位概念も同一の基本的概念タイプ（物，行為，特質など）に属し，さらには下位概念の内包は上位概念の内包を含み，上位概念の属性を継承していることが必要である。この包含関係の特質は，概念のラベルである語に次のように反映される。例えば，「システム－情報システム－経営情報システム－集中型経営情報システム」という上位－下位関係は以下の構造を持っている。

システム
情報　システム
経営　情報　システム
集中型　経営　情報　システム

この例では，「システム」という概念のラベルである「システム」という語が，下位の概念のラベルに継承され，さらに下位の語は上位の語を限定・修飾する語を前の部分に伴うようにして形成されている。したがって，複合語の解析を行い，複合語と構成要素との組み合わせで，語間の上位－下位関係を決定することが可能であると考えられる。

しかし，日本語における複合語は，このような修飾関係にある語の組み合わせのみで作られるわけではない。複合語を構成する独立的要素間の結合パターンについては，国立国語研究所の野村[3]らの包括的な研究があり，以下の５つのパターンに分類できることが明らかとなっている。

(1) 修飾関係１：前部分の語基が後部分の述語成分の状態・程度などを詳しくする関係。
(2) 修飾関係２：前部分の語基の性質や状態を，後部分の語基が修飾したり限定したりするもの。
(3) 補足関係　：述語に相当する成分の内容を，語基が［ガ・ニ・ヲ・ト・デ］などの関係で補足するもの。
(4) 並列関係：類似の意味の語基を並べたもの。
(5) 対立関係：意味の対立する語基を並べたもの。

上述の「集中型経営情報システム」という語は，(1)の関係である。５つのカテゴリーのうち，(1)に示す関係の語については，前部分の語基が後部分の語基を修飾したり限定したりする係受け関係が成立しているために，文字列を分割することによって語間の上位－下位関係を自動的に構築することが可能となる。

しかし，「温度測定」という語は，「測定」という語が，対象である「温度」と結びついて複合語を構成しており，(3)補則関係の例である。また，「調査研究法」の上位語は「研究法」でも「法（法

律）」でもない。これは，調査法と研究法という並列関係にある語を組み合わせて作られた複合語の例である。このように，(2)～(5)の場合には上位－下位関係を構築できる係受け関係が成立しないため，自動構築は無理である。そこで，実際の複合語が持つ構造について，どのような場合に上位－下位関係が表現されるのかを明かにする必要がある。

3．分析プロセス

3．1　複合名詞の定義

本研究で対象とする複合語は，特に複合名詞に限定し，本来単独の用法を持ち得る語を２つ以上結合して，新たに１語としての意味・機能を持つようになったものと定義する。なお，複合語を構成する要素の一方が，非独立的要素（接辞）であるものは，これに含めない。また，1)外来語などのカタカナ・アルファベットで表現される語については，そのひとまとまりを１つの独立的要素として扱う，2)２字熟語は，２字１組で１つの独立的要素として扱い，それ以上に分割しないこととした。

3．2　分析方法

複合語の解析によって，上位－下位関係を決定することができるかどうかは，複合語の各構成要素が持つ品詞情報を利用して決定できると考えられる。たとえば，①「自然言語」は，相言である［自然］が後行要素の［言語］を修飾しているため［言語］を上位語と考えることができる。しかし，②「自然破壊」は，体言である［自然］は破壊の対象であって，修飾語とはみなしがたいので，前述の修飾関係１を満足しえないと考えられる。したがって，［破壊］を上位語にすることには問題がある。このように，後行要素が（体言類）である場合には，上位－下位関係を抽出できるのに対して，後行要素が（用言類）である場合には，上位－下位関係を抽出することはできないと考えられる。

そこで本研究では，以下の規則に基づいて，複合語の分析による上位－下位関係の自動抽出を試みた。実際の処理手順を図１に示す。

(1) 複合語の後行要素として体言類が来る場合には，複合語を後行要素である体言の下位語であると判断する（例；電磁石－コイル）。

(2) 複合語の後行要素として相言類が来る場合には，上位－下位関係は構築出来ないものとする（例；利用－可能）。

(3) 複合語の後行要素として用言類（サ変動詞）が来る場合には上位－下位関係は構築出来ないものとする（例：エネルギー－計測）。ただし，最後行要素の前にくる要素が体言であった場合には，複合語全体がその体言と最後行要素である用言（サ変動詞）が結びついた語の下位語であると判断する。
（例；「パルス－磁気－エネルギー－計測」は，「計測」の下位語とはならないが，「エネルギー－計測」の下位語となる）

(4) 体言（または用言）－相言－体言の場合には，はじめの体言（または用言）－相言を，１つの構成要素としてまとめられるものとして処理する（例；情報－処理的－アプローチ）。（したがって，「情報－処理的－アプローチ」は，「アプローチ」の下位語となるが，「処理的アプローチ」の下位語とはしない）

4．上位－下位関係の構築実験

4．1　分析対象

分析対象としては，情報処理学会論文誌1986年４月～1987年３月号に掲載された，情報処理分野に属する100論文の抄録中に含まれる語512語，および，ＪＩＣＳＴ科学技術文献速報1988年４月～５月

分の11,200論文の抄録中に含まれる11,834語とした。これらの語はいずれも異なり語数である。分野を限定したのは，各主題分野によって複合語の造語特性に違いがあると思われるためである。

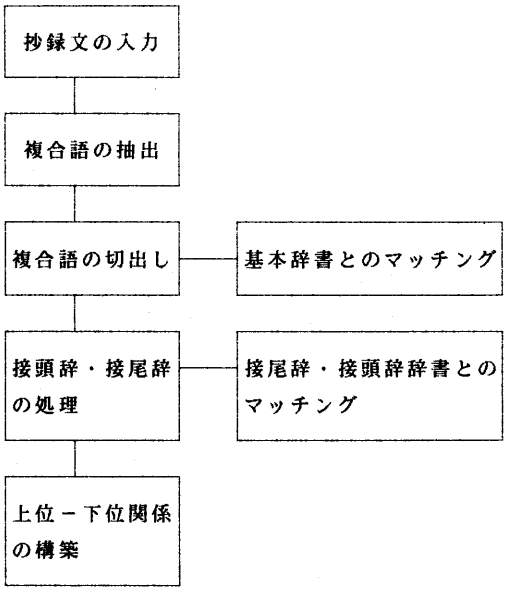

図１．上位－下位関係構築の処理手順

４．２　複合語の抽出

情報処理学会論文誌およびＪＩＣＳＴテープ中の抄録を対象に，「ひらがな」および「記号（。，，・等）」を区切り記号として，漢字およびカタカナ，アルファベット列のみからなる部分を抽出した。この方法を用いた場合，「新システムの研究及び開発を行った」（「新システム」「研究及」「開発」が抽出される）や「研究上の諸問題」（「研究上」「諸問題」が抽出される）ような文中の「及」，「上」といった漢字が複合語の構成要素として含まれるが，これらの字については最終的には接辞辞書と照合して取り除いた。（なお，前処理としてパーザーを利用することも考えられる。）

４．３　複合語の語基への分割法

複合語の語基への分割は，基本辞書に登録されている熟語が含まれているかどうかを調べ，含まれている語を構成要素の候補とする方法で行った。その際，何通りかの解析が可能であった場合には，複合語を構成する文字のうち，基本辞書中の語とマッチした文字数が最大となるような解析法を採用した。

例えば，基本辞書中に，「円筒」「中空」「陽極」がそれぞれ含まれていた場合，「円筒中空陽極」は「円筒－中空－陽極」のように分割した。また，「核融合」「融合」「電磁石」「磁石」「コイル」が基本辞書中に含まれていた場合には，「核融合用電磁石コイル」は基本辞書中の語とマッチした文字数が最大となるように，「核融合－用－電磁石－コイル」のように分割した。なお，「多重度」（「多重」「重度」），「複合体」（「複合」「合体」）のように３文字の単語が２文字の単語２つから構成されている場合には，３文字全体で１つの単語（名詞）であると判断した。さらに，複合語の最後に「C14」のような数字，アルファベットを含んでいた場合には，これらをはずして処理した。

なお，複合語の構成要素と辞書とをマッチさせるのは，あくまでそれを参考に複合語を正しく分割するためであって，辞書にないからといって，その言葉を無視するものではない。

４．４　複合語を分割する際に使用する辞書

複合語を分割する際に使用する辞書としては，市販のワープロ用辞書を使用した。さらに，以下の工夫を加えた。

(1) 基本辞書中で，体言（名詞）と用言（動詞）の両方に分類されている語の品詞は用言と判断した。同様に，用言＞相言＞体言の順に品詞を判断した。

(2) 対象が電気工学分野の語であることを考え，この分野の専門語とはみなせない「器用」「体内」のような一般語を除いた。

４．５　接尾辞，接頭辞の処理

(1) 漢字１文字が分割された場合

複合語を分割する際に，漢字１文字が分割された場合には，その漢字が接頭・接尾辞であるかどうかの判断を行った。

(a) 漢字が接頭辞であった場合

後の語基とまとめて１つの語基であると判断した。その際，語基の品詞は後の語基に付与されたものとした。

例）「超／高速／通信」　->　「超高速／通信」

(b) 漢字が接尾辞であった場合

前の語基とまとめて１つの語基であると判断した。その際，語基の品詞は接尾辞に付与された品詞とした。体言とみなす接尾辞としては，「法」，「式」，「語」などが，相言とみなす接尾辞としては，「的」，「型」などがある。

例）「情報／処理／的／アプローチ」　->　「情報／処理的／アプローチ」

（この場合，語基「処理的」の品詞は，処理（サ変動詞）に付与された「用言」ではなく，「的」に付与された「相言」とした）

(c) 漢字が接頭・接尾辞でなかった場合

前の語基とまとめて１つの語基であると判断した。その際，語基の品詞は「体言」であると判断した。

例）「電子／計算機／操作／法」　->　「電子／計算機／操作法」

（この場合，語基「操作法」の品詞は，操作（サ変動詞）に付与された「用言」ではなく，「体言」とした）

(2) 基本辞書中にない２文字以上の語基が分割された場合

基本辞書中にない２文字以上の語基については，接尾・接頭辞を含むかどうかの判断を行った。すなわち，辞書中にない２字以上の語に対して，以下の処理を行った。

(a) ２字以上の語のうち，最後の漢字が接頭辞であった場合には，最後の漢字のみを分割して後の語と結びつけた。この場合，結合された語基の品詞は，後の語基のもつ品詞とした。

(b) ２字以上の語のうち，最初の漢字が接尾辞であった場合には，最初の漢字のみを分割して前の語と結びつけた。この場合，結合された語基の品詞は，接尾辞に付与された品詞とした。

(c) 接頭・接尾辞を含んでいない語については，「体言」とした。

なお，処理(a)，(b)の結果として漢字１文字のみが分割された場合には，さらに(1)の処理を行った。

(3) 複合語の最後の文字

複合語の最後の１文字が，「上」，「中」，「下」，「内」，「外」などの複合語の構成要素となる可能性の低い漢字であった場合には，その漢字を削除するものとした。たとえば，「投影／装置／上」の場合は，「上」を削除して「投影／装置」とした。ただし，最後の漢字が基本辞書中の語を構成する要素であった場合には，削除しないものとした。たとえば，「地盤／沈下」は「地盤沈」とはしない。

また，複合語の最初または最後の語基として，「現在」「過去」などの語基が分割された場合，これらの語基を削除するものとした。

5　分析結果

情報処理学会論文誌の100抄録中から抽出された512の異なり語のうち，80.3%にあたる411語について，上位－下位関係を自動的に決定することができた。これらの語のうち，人間の判断と一致した割合を複合語を構成する構成要素の数ごとに第１表に示す。また，ＪＩＣＳＴ科学技術文献速報の11,200抄録中から抽出された11,834の異なり語のうち，84.3%にあたる9,976語について，上位－下位関係を自動的に決定することができた。これらの語のうち，人間の判断と一致した割合を複合語を構成する構成要素の数ごとに，第２表に示す。

第１表　情報処理学会論文誌中上位－下位関係の構築が可能な複合語数

複合語の構成要素数	抽出された上位－下位関係の数	上位－下位関係が人間の判断と一致した場合	上位－下位関係が人間の判断と一致しない場合
２単位	299	293 (98.0%)	6(2.0%)
３単位	85	75 (88.2%)	10(11.8%)
４単位	18	14 (77.8%)	4(21.2%)
５単位	7	5 (71.4%)	2(28.6%)
６単位	2	2 (100.0%)	0(0.0%)
合計	411	389 (94.6%)	22(5%)

第２表　ＪＩＣＳＴ抄録中上位－下位関係の構築が可能な複合語数

複合語の構成要素数	抽出された上位－下位関係の数	上位－下位関係が人間の判断と一致した場合	上位－下位関係が人間の判断と一致しない場合
２単位	6132	5930 (96.7%)	202(3.3%)
３単位	2868	2581 (90.0%)	287(10.0%)
４単位	565	463 (81.9%)	102(18.1%)
５単位	242	196 (81.0%)	46(19.0%)
６単位	169	153 (90.5%)	16(9.5%)
合計	9976	9323 (93.5%)	653(6.5%)

これらの表に見られるように，抽出された上位－下位関係のうち，94.6%，93.5%については，人間の判断と同じ結果が得られた。

実際に上位－下位関係を構築できた例のうち，人間の判断と一致したものの例を以下に示す。

磁石　－　双極子磁石
絶縁体　－　磁気的絶縁体
装置　－　高精度加工装置　－　小型高精度加工装置
装置　－　多重極装置　－　静電多重極装置
導体　－　超伝導導体　－　中空超伝導導体

また，人間の判断と一致しなかったものの例を以下に示す。

・（人間の判断）　発生器　－　オゾン発生器　－　ガラス電極オゾン発生器
（システムの判断）
発生器　－　オゾン発生器　－　電極オゾン発生器　－　ガラス電極オゾン発生器

・（人間の判断）　磁石　－　高磁界高電流密度磁石　－　核融合研究用高磁界高電流密度磁石
（システムの判断）
磁石　－　密度磁石　－　高電流密度磁石　－　高磁界高電流密度磁石
－　研究用高磁界高電流密度磁石　－　核融合研究用高磁界高電流密度磁石

６．今後の展開

今回の分析で，複合語の解析からの上位－下位関係の構築の可能性がある程度示された。しかし，同時にいくつかの問題点も明らかになった。そこで，今後われわれは，以下の点について研究を進めていきたいと考えている。

①並列・対立関係の処理。これらの関係を持つ２つの語基は，なんらかの方法で自動的に１つの語基にまとめて，上位－下位関係構築の支障とならないようにしたい。
例：「研究－開発－機関」→「研究開発－機関」

②基本辞書や接辞辞書の充実とより効率の高いアルゴリズムの開発。

③自動的な「全体部分関係」「例示関係」の構築。このためには複合語の解析からではなく，構文の解析から情報を得ることが考えられる。

＜引用文献＞

1) Ghose, Amitabha and Dahwle, Anand S. Problem of Theasaurus Construction. Jounal of the American Society for Information Science. Vol.28, No.4, p211-217 (1977).
2) Guidelines for the Establishment and Development of Monolingual Thesauri. Second revised edition. United Nations Educational, Scientific and Cultual Organization, Paris. 1981.9, 64p.
3) 水谷静夫ほか著．文字・表記と語構成．東京，朝倉書店，1987．244p.