日本語解析システム「ささゆり」の諸機能

——構文解析，品詞解析と概念解析——

高　橋　　亘*

Functions in the Japanese Analysing System SASAYURI

——Syntax Analysis, Part of Speech Analysis and Concept Analysis——

Wataru Takahasi

Abstract : We have been developing a Japanese analysing system for several years. It was named SASAYURI after the parsing scheme of Japanese sentences in our system. The Japanese "sasayuri" means a kind of lily, which is a native of Japan. A Japanese sentence is understood in a manner similar to the leaves and a stem of a lily by our system.

The main feature of our system is represented by a decomposing scheme, which divides Japanese sentences into the perceptive collocations. A scheme with perceptive collocations provides a reasonable prescription for our concept analysis, but it must be afraid of a disturbance. A kind of modification relation, in which a short verb phrase modifies a noun phrase, is often considerd as a perceptive collocation. But such a collocation disturbs for the system to find much bigger modification relation in many cases. We should ommit such collocations out of the candidates of perceptive collocations in the first step. We must remedy the modification relation after the understanding of complex sentence sturucture. The parsing method to shear a complex sentense of some complete verb phrases which modefy the noun phrases, provides a structure which resembles a lily with a stem and some leaves.

In order to decompose a huge number of sentences into the perceptive collocations, we should possess in the system a copious number of perceptive collocations, which are desired to be gained from some machine learning system. Our system SASAYURI contains already a machine learning function, with which the system is able to pick up some collocations for the candidates of the perceptive collocations from some text document autmatically.

The more accurate method for the machine learning needs the more efficient function to estimate the parts of speech for the words in the text sentences. In this article, the basic scheme of SASAYURI in the concept analysis is discussed in an comprehensive manner. Especially, a function of syntax analysis, a function of part of speech analysis and our basic sheme of concept analysis are discussed.

Key words：日本語解析システム Japanese analysing system　知覚連語 perceptive collocation　意味解析 semantic analysis　概念解析 concept analysis　品詞解析 parts of speech analysis　機械学習 machine learning　自然言語解析 natural language analysis

*関西福祉科学大学社会福祉学部　教授

1　はじめに

この論文は，先に“日本語解析システム「ささゆり」の品詞解析機能と概念解析”と題する論文[1)]として，『Proceedings 2005 M Technology Association of Japan』に公表した内容に，理論の基礎付けと例証を加筆し，その妥当性と整合性を強調するものである．

我々が日本語解析システムの初期の構想を初めて公表したのは 1999 年の夏のことであった．[2~4)] 当初は日本語文の切断に関して，M 言語の大域変数の階層構造をアルゴリズムに組み込んだ方法の有効性を示唆するだけの素朴なものであったが，開発が進むにつれ，我々のシステムは人間の意味知覚と密接に関連していることが解ってきた．[5~10)]つまり，我々の日本語切断の方法は漢字の読みを決定する程度の局所的コンテキストを反映した連語で日本語文を切断していくものであるが，漢字の読みを決定する問題は人の言語知覚の問題や日本語文の意味解析の問題と密接に関連しているのである．[11, 12)]

言語学者のソシュールは，言語記号である単語が，記号内容である概念と不分離の関係にあることを強調したが，この概念がどのような意味内容を持っているのか，ということについてはあまり問題にしなかった．しかし，単語と対応する概念というものは，それほど明確な意味内容を持っているものではない．単語の持つ意味内容は，通常考えられているよりずっと豊富なものである．ソシュールによって指摘されたように，語が結合すると，語の結合規則が単語間の類推規則を与えるので，元々恣意的な記号と概念の関係にある種の規定性を与える．「花」という単語の持つ概念はそれほどはっきりしたものではないが，「薔薇の花」のような連語はかなりはっきりした意味内容を持ち，さらに「白い薔薇の花」となると，人間の明確な意味知覚と対応するようになる．このように知覚と直接対応するような連語を我々は知覚連語と読んだ．[11, 12)]

我々の日本語解析システムは知覚連語で日本語文を切断していくことを目指すものであるが，意味要素と密接に関連した対象である知覚連語を問題にするということは，意味要素とあまり関係がなく，文の文法的構成のみに寄与する，機能語を単離することになる．したがって，知覚連語を追求する問題は，逆に機能語が何であるかを明らかにするという問題でもある．

この論文の目的は，我々が近年開発してきた日本語解析システム「ささゆり」の概念解析の基本的なスキームを明確にすることと，システムが保持しているいくつかの機能の原理と例証を明示することにある．中でも，意味解析の基礎となる知覚連語と，機能語の文法的役割を，構文上の構造から如何にして明確に分別するのか（第 2 節と第 3 節），既存の形態素解析のシステムと比較して，我々のシステムが如何に働くのか（第 4 節），概念解析の背景となる，ベクトル空間の基礎理論が如何なるものであるか（第 5 節），について詳しく述べることを期している．

2　日本語解析システム「ささゆり」の構文理解

日本語の文のなかで，構文が最も簡単なものは単文と呼ばれ，単文は 4 つの種類に分類される．つまり，文尾に，助動詞の部分を除いて，名詞，形容詞，形容動詞，動詞がくるものをそれぞれ，名詞文，形容詞文，形容動詞文，動詞文とよぶ．単文が並列的に結合されている文，もう少し詳しくいうと，単文がそれぞれの間に修飾関係がなく，助詞や接続詞が間にはいって，列挙されていく文，（助詞，接続詞が入らず連用止めの単文が列挙されるものもある）これを重文という．最も複

雑な構文をもつものは複文であり，単文もしくは重文の骨格を持つ文の中の，名詞もしくは名詞句が他の単文もしくは重文によって修飾されているものである．

日本語文を知覚連語によって分解していくことを考えるとき，知覚連語の定義の仕方によっては上述の構文構造を著しく破壊してしまうことがあるので，知覚連語の定義には構文理解に関する注意が必要である．この節では日本語解析システム「ささゆり」が如何にして，複文構造を保持しながら知覚連語を切り取っていくのかということについて概観したい．

複文構造の中で，我々が注意しなければならないのは動詞文による修飾である．多くの文を見ていくときに，名詞文，形容詞文，形容動詞文が名詞もしくは名詞句を修飾するものは，比較的短い句を形成することが多い.[14]例をあげると，比較的長そうな句の例をえらんで，

じんじん端折りの頬冠りや、赤い腰巻の姉さんや、時には人間より顔の長い馬にまで逢う。【夏目漱石（草枕）】

というような文があげると，この文では「人間より顔の長い」という形容詞文が「馬」という名詞を修飾している．このような場合「人間より顔の長い馬」を一つの知覚連語として，定義すれば構文構造を壊すことなく，知覚連語に分解できる．

問題になるのは動詞文が名詞もしくは名詞句を修飾する場合である．というのはこのような動詞文は非常に短いものから大変長いものまであり，短いものの場合に「動詞」+「名詞」を知覚連語として定義すると，それが，もっと長い動詞文の構文理解を妨げることになるからである．

このため「ささゆり」では，原則として「動詞」+「名詞」の型の知覚連語は定義しない．ここで，動詞文による名詞（名詞句）修飾を含む文に対して「ささゆり」がどのような構文理解を示すのかを模式的に例示してみたい．例文として，「坊っちゃん」の文をあげる．

小学校に居る時分学校の二階から飛び降りて一週間ほど腰を抜かした事がある。【夏目漱石（坊っちゃん）】

「ささゆり」では，意味解析の際，この文をまず，次のように切断する．

小学校に居る　時分　≫　時分　学校の二階から飛び降り　て　≫　一週間　ほど腰を抜かした　事　≫　事がある。

ここで，知覚連語の切れ目にスペースが入っていることと，単文の切れ目に≫が入っていることに注意していただきたい．さらに，動詞文が名詞を修飾しているところで，次の単文の冒頭の名詞を引き出している点も重要である．次に，もとの文から動詞句が名詞句を修飾している部分を抜き取り，文を次のように分割する．

① 小学校に居る　時分
② 一週間　ほど　腰を抜かし　た　事
③ 時分　学校の二階から飛び降り　て　事がある。

①，②ではアンダーラインのあるところが動詞句によって修飾される名詞（名詞句）であり，①，②によって「時分」,「事」の 2 単語の意味が限定される．③の「時分」と「事」はこの意味と理解して，文全体の意味が決定される．このような文の分解は百合科の植物の茎と葉の関係に似ている．③が茎で①，②は葉である．このような構文理解のイメージを表象化したものが日本語解析システムの命名の由来である．この構文理解の構造は，理論物理の言葉を使っていえば，ファインマングラフのツリーダイアグラム（図 1）に似ている．

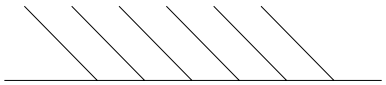

図 1　ファインマングラフのツリーダイアグラム

3　日本語解析システム「ささゆり」の品詞解析

第 2 節では，「ささゆり」の大まかな文の構造の理解について述べたが，このような構文理解のもとに，品詞列のパターンから知覚連語を特定し，知覚連語を学習する機能については既に昨年の論文で言及していることである．[13, 14] 品詞列を指標にして知覚連語の候補をリストするという目論見は，品詞解析が厳密であればあるほど，その効率が上昇することは論を待たない．しかしながら，形態素解析の機能の一部である品詞解析について，特に短い仮名綴りの形態素についての判断機能に脆弱性があることは多くの既成の形態素解析システム（日本語形態素解析システム JUMAN や日本語形態素解析システム『茶筅』）について認められることである．[15~17] この困難は助詞や助動詞の様な品詞について形態が等しく品詞的に多価であるものが多くあることからくる．つまりデフォルトで，ある品詞に特定していても実際に使用されるコンテキストでは別の品詞にシフトさせる必要が生じる．つまりコンテキストを判断して，その結果により再認する課程が必要なわけである．この再認課程が効率よく行われなければ，品詞解析を前提とした知覚連語学習システムは膨大な時間を要することになる．

我々はこのような問題について，ここで機能語接続行列（CMF；Connection Matrix for Function words）と称する方法を提示したい．これは簡単に言えば，先行詞の活用語尾と後続詞の組の整合性によって品詞の再認判断を行う一般的方法であるといえる．

先行詞の活用語尾を判断する関数は，動詞の場合を例に挙げると，次のようなものである．

S Z＝$$^NWFGVTL（WORD,.GYOU,.DAN）．　（1）

ここで関数値 Z は動詞語尾で 1，それ以外は 0 を返す．つまり 1 が返ると動詞の語尾の可能性があることになる．第一引数 WORD は現在問題にしている語の文中のそのままの形であり，第二引数 GYOU は語尾の行，第三引数 DAN は語尾の段を返す．DAN の値は次のようである．

あ段→1,　　い段→2,　　う段→3,
え段→4,　　お段→5,　　イ音便→6,
撥音便→7,　　促音便→8,
ワ行五段ウ音便→31,
カ変の「来」→25.

したがって，「る」語尾の場合 GYOU＝“ラ”，DAN＝3 が返るが，この応答では，ラ行五段の終止形や，上一，下一，カ変，サ変の終止形と連体形の判断が後に続かなければならない．またワ行五段の「い」語尾とカ行，ガ行の音便語尾「い」の区別も必要である．語尾の「行」と「段」を決定するようなアルゴリズムはプログラミングによって差のあるものではないが，ここに見られるような多価問題を解決するには M 言語のような階層型データベース言語には一分の利がある．これらの区別のために変化形と品詞典型の関係をあらかじめ大域変数に登録しておき，この値があるかどうかで絞り込みを行う．たとえば，

^NWPRO（“書い”,“カ五”）＝“書く”，　（2）

のようなデータがあると，ここにデータがあることで，ワ行五段の「い」語尾でなく，カ行五段のイ音便の「い」語尾であることが判断できる．

後続詞の形態判断については，次のような関数を用意する．

S Z＝$$^NWFSEQC(HPS, LWD, LDCWD)．　(3)

ここで，HPS は予想される先行詞の品詞，LWD は後続する機能辞の文中での形態，LDCWD は後続する語が連語である場合に，その単語分解した語列を与え，これにより先頭に機能辞が来ていないかどうかの情報を与えるものである．関数値は後続する機能辞の原形に応じて次のような値を返す．（ナ行，マ行，ガ行以外の五段活用動詞の場合）

ぬ→未 1,	ず→未 11,	ない→未 2,
せる→未 3,	れる→未 31,	う→未 4,
たい→用 1,	ます→用 11,	つつ→用 12,
そうだ→用 13,	ながら→用 14,	た→用 2,
て→用 21,	たり→用 22,（助詞）	、→用 23
たり→用 24,（助動詞）	ん→終 1,	し→終 2,
ば→仮.		

以上のような関数により，先行詞の語尾と後続詞の整合性を判断することで，デフォルトで与えられている品詞判断を，より正確なものに再定義することが可能である．このような整合性の判断のためのデータリストは M 言語では容易に作成出来よう．

先行詞と後続詞の整合性を判断する問題は，時には二項問題から三項問題に拡張されなければならない．例えば，通常の学校文法で準体言助詞と呼ばれる「の」は，先行する動詞句，形容詞句，形容動詞句などの修飾を受けて，引き続くの部分の中で主語として振る舞うことがある．このような助詞は，単独では主語になれないにも拘わらず，形態素解析の上では形式名詞として扱えあれることが多い．しかし主語として振る舞っているかどうかの判断には，「の」に後続するものが何であるかを見る必要がある．つまり，「の」が〔形式名詞〕と判断されるべきかどうかの判定は，本来，三項問題なのである．「ささゆり」は，このような，先行詞，中央詞，後続詞の三次元をもつ行列による再認機能を保持している．

4　日本語解析システム「ささゆり」の品詞解析機能の実効性

我々はここで，いくつかの例文を上げ，これらに対し我々の品詞解析機能が如何に働くかを示してみたい．判断の的確性を検証するため，既存の二つの形態素解析，JUMAN と『茶筅』の結果を挙げて，比較してみたい．

まず，最初の例文は，

> しかしある日、突然、すべてが変わってしまった。【ブルック・ニューマン作，五木寛之訳，リサ・ダーク絵（リトルターン)】

である．これに対し，JUMAN は，

> しかし〔接続詞〕 ある〔動詞〕 日〔名詞〕 、〔読点〕 突然〔副詞〕 、〔読点〕 すべて〔副詞〕 が〔助詞〕 変わって〔動詞〕 しまった〔接尾辞〕 。〔句点〕

と切断し，各部分の品詞判断は〔　〕内のようである．（この後の議論でも同様の表記法を用いることにする）同じ文に対し，『茶筅』での切断と品詞判断は，

しかし〔接続詞〕 ある〔連体詞〕 日〔名詞〕 、〔記号〕 突然〔副詞〕 、〔記号〕 すべて〔名詞〕 が〔助詞〕 変わっ〔動詞〕 て〔助詞〕 しまっ〔動詞〕 た〔助動詞〕 。〔記号〕

のようである．JUMAN は，「ある〔連体詞〕」と「すべて〔名詞〕」について，品詞を誤認しており，形態素解析にしては，「て〔助詞〕」と「た〔助動詞〕」を動詞語尾に含めてしまっている点が不完全である．この意味では，この文に関する限り，『茶筅』の方が幾分進化しているといえる．

我々の「ささゆり」では，同じ文を，次のような切断と品詞判断をする．

しかし〔接続〕 ある日〔連語＊名詞〕 、〔記号〕 突然〔副詞〕 、〔記号〕 すべてが変わってしまっ〔連語＊ワ五〕 た〔助動＊タ〕 。〔記号〕

知覚連語による切断のため，二つの連語の部分があるが，それぞれ，内容把握は次のようであり，確かである．

ある日＝ある　日〔連体　名詞〕

すべてが変わってしまう＝すべてが　変わっ　て　しまう〔名詞　助詞　ラ五　助詞　ワ五〕

第二の例文は，

ぼくがそれまで知っていたことは、もはやそこにはなく、ぼくの前にあるのはまったく未知の世界だった。【ブルック・ニューマン作，五木寛之訳，リサ・ダーク絵（リトルターン）】

であるが，JUMAN と『茶筅』の結果を順にリストする．

- ぼく〔名詞〕 が〔助詞〕 それ〔指示詞〕 まで〔助詞〕 知っていた〔動詞〕 こと〔名詞〕 は〔助詞〕 、〔読点〕 もはや〔副詞〕 そこ〔指示詞〕 に〔助詞〕 は〔助詞〕 なく〔形容詞〕 、〔読点〕 ぼく〔名詞〕 の〔助詞〕 前〔名詞〕 に〔助詞〕 ある〔動詞〕 の〔形式名詞〕 は〔助詞〕

  まったく〔副詞〕 未知の〔ナノ形容詞〕 世界〔名詞〕 だった〔判定詞〕 。〔句点〕 ［JUMAN の結果］
- ぼく〔名詞－代名詞〕 が〔助詞〕 それ〔名詞－代名詞〕 まで〔助詞〕 知っ〔動詞〕 て〔助詞〕 い〔動詞〕 た〔助動詞〕 こと〔名詞－非自立〕 は〔助詞〕 、〔記号〕 もはや〔副詞〕 そこ〔名詞－代名詞〕 に〔助詞〕 は〔助詞〕 なく〔形容詞〕 、〔記号〕 ぼく〔名詞－代名詞〕 の〔助詞〕 前〔名詞－副詞可能〕 に〔助詞〕 ある〔動詞〕 の〔名詞－非自立〕 は〔助詞〕 まったく〔副詞〕 未知〔名詞－一般〕 の〔助詞〕 世界〔名詞－一般〕 だっ〔助動詞〕 た〔助動詞〕 。〔記号〕 ［『茶筅』の結果］

この例でも，JUMAN　で未分化に理解されていたものが，『茶筅』で分化し，正確に判断されているのが解る．ただ「名詞－非自立」とあるのは形式名詞のことだろうか？いずれにしても，形態素解析としては，『茶筅』で一応の判断が出来るものと思われる．

同じ例文に対し，我々の「ささゆり」の判断を示したいが，知覚連語の機械学習をはさんで，判断がどのように変化するかを見てみたい．

- ぼく〔代名〕 が〔助詞〕 それまで〔連語＊副詞〕 知ってい〔連語＊上一〕 た〔助動＊タ〕 こと〔形名〕≫ こと〔形名〕 は〔助詞〕 、〔記号〕 もはや〔副詞〕 そこ〔代名〕 に〔助詞〕 は〔助詞〕 なく〔形容〕 、〔記号〕 ぼくの前〔連語＊名詞〕 に〔助詞〕≫ ある〔ラ五〕 の〔形名〕≫ の〔形名〕 は〔助詞〕 まったく〔副詞〕 未知の世界〔連語＊名詞〕 だっ〔助動＊形動〕 た〔助動＊タ〕 。〔記号〕［学習前］

  ぼく〔代名〕 が〔助詞〕 それまで〔連語＊副詞〕 知ってい〔連語＊上一〕 た〔助動＊タ〕 こ

と〔形名〕≫　こと〔形名〕　は〔助詞〕　、〔記号〕　もはや〔副詞〕　そこ〔代名〕　に〔助詞〕　は〔助詞〕　なく〔形容〕　、〔記号〕　ぼくの前にある〔連語＊ラ五〕　の〔形名〕≫　の〔形名〕　は〔助詞〕　まったく〔副詞〕　未知の世界〔連語＊名詞〕　だっ〔助動＊形動〕　た〔助動＊タ〕　。〔記号〕［学習後］

複文構造があるため，第 2 節で述べたように，≫が二カ所ずつはいって，次の区間の形式名詞を引き出しているが，例外が一つあって，「僕のまえ　に」の後に入っている≫は「僕のまえに」が副詞句であることを示すためのものである．学習前の判断から

ぼくの前にある＝ぼく　の　前　に　ある〔代名　助詞　名詞　助詞　ラ五〕

を候補として提示し，これを登録すると，学習後のような切断に変化する．前節でも述べたように，この連語に後続する「の」は，通常の学校文法では準体言助詞と呼ばれて，助詞に分類されるが，複文の構造をはっきりさせるためには形式名詞の理解が合理的である．

これまでの例では品詞判断に限って言えば，形態素解析システム『茶筅』と日本語解析システム「ささゆり」の間で認識率にさほどの相違があるようには思われない．「ささゆり」の方が構文認識機能や知覚連語の学習機能を持っている点が違うぐらいである．しかし，次の例になると，品詞判断についても「ささゆり」の有効性が明らかになる．最後の例文は，次のようなものである．

「証明に、美しい、美しくないの区別なんてあるんですか」【小川洋子（博士の愛した数式）】

前の例文と同様，JUMAN と『茶筅』の結果を順にリストする．

- 「〔括弧始〕　証明〔名詞〕　に〔助詞〕　、〔読点〕　美しい〔形容詞〕　、〔読点〕　美しく〔形容詞〕　ない〔接尾辞〕　の〔形式名詞〕　区別〔名詞〕　なんて〔助詞〕　ある〔動詞〕　んです〔助動詞〕　か〔助詞〕　」〔括弧終〕　［JUMAN の結果］
- 「〔記号〕　証明〔名詞〕　に〔助詞〕　、〔記号〕　美しい〔形容詞〕　、〔記号〕　美しく〔形容詞〕　ない〔助動詞〕　の〔名詞－非自立〕　区別〔名詞〕　なんて〔助詞〕　ある〔動詞〕　ん〔名詞－非自立〕　です〔助動詞〕　か〔助詞〕　」〔記号〕　［『茶筅』の結果］

この例では，どちらの形態素解析システムでも，「ない」「の」の判断に問題がある．まず，「ない」について言えば，〔接尾辞〕　は規定仕切れていないし，〔助動詞〕は誤認である．言語学的な立場の違いがあるのかもしれないが，形容詞について，打ち消しの意味を表す「ない」は通常の学校文法では，〔補助形容詞〕もしくは〔形式形容詞〕と判断される．形容詞の連用形に助動詞がつくのは不自然である．さらに「の区別」の接続を考えると「の」が〔形式名詞〕であると判断するのは誤認である．

同じ例に対し，三項判断機能をもつ「ささゆり」の判断は，次のようである．

「〔記号〕　証明〔名詞〕　に〔助詞〕　、〔記号〕　美しい〔形容〕　、〔記号〕　美しくない〔連語＊形容〕　の〔助詞〕　区別〔名詞〕　なんて〔助詞〕　ある〔ラ五〕　ん〔形名〕≫　ん〔形名〕　です〔助動＊形動〕　か〔助詞〕　」〔記号〕

ここで「美しくない」はかって学習されたもので，この連語の品詞認識は〔形容　補助形容〕である．

5　日本語解析システム「ささゆり」の意味解析のスキーム

我々の意味解析の基本的な基盤は「単語は本来多義なものであり，語が結合することによって意味が限定され，意味的に純粋な状態を実現する」という認識である．

ソシュールは記号の恣意性を強調したが，言語記号が全く恣意的なものであれば意味が生じようがない．しかし，ソシュール自身も恣意性というのみならず，無契性という言葉をもって，言語記号の本性を述べた．この無契性という言葉が言語学的に意味のないものとして，否定的にとらえる人たちもいるが，意味論からすれば無契性こそが，言語に意味が生成される基礎である．つまり，ソシュールの言葉を借りて言えば「結合語は，被結合語間の関係性を類推する規則を与えるので，単語ほどには無契的でない」と言える．このことは，裏返してみると，ほぼ無数に使用されている結合語の被結合語間の関係性を記号論的に読み解くことが意味構造を規定すると言ってよい．我々は既に，このような視点から語の結合が如何にして意味を規定するのかということを，ベクトル空間の言葉を用いて議論した.[11, 21]同じことを知覚連語という言葉を使って表現すれば，知覚連語こそ，人が用いた言語による意味の実現化であり，知覚連語によって意味要素を規定し，その被結合語間の関係性を読み解くことが単語の持つ意味（これは多義になる）を規定する方法である．

ウィトゲンシュタインは「言語使用が言葉の意味を規定する」といったが,[18]彼の認識はある言葉の使われ方がその言葉の意味を規定するということであり，我々の認識とは少し異なっている．我々の認識を少し哲学的に言えば「人々が，文章においてどのような連語を使用したか，と言うことが言語の意味構造を決定する」ということになる．

もう少し具体的な議論に移って,「ささゆり」の意味解析の基礎となっている理論について述べる．我々の理論で意味空間の基礎を与えるものは，語が十分結合し，意味的に純粋な状態を提示する知覚連語である．このような知覚連語をもとに意味要素を決定する．例を挙げると第 1 節で上げた知覚連語「白い薔薇の花」が提示する意味要素を［白バラ花］で表現すればこのような知覚連語に対して，意味空間の基底ベクトル

$$|\text{白バラ花}\rangle$$

を用意する．ここで注意しなければならないことは，このようにして知覚連語にふられる意味要素は単に知覚連語の独立性をラベルしているだけであって，それ以上の役割は持っていない．したがって，ここでの意味要素の名付け方の是非の議論は全く意味がない．このことは議論が進むにつれて自ずから明らかになる．

このような基底ベクトルによって意味空間の正規直交基底を構成していくのが我々の意味空間の構成の仕方である．つまり十分に長い知覚連語を一つとり，これによって意味要素を一つ提起し，この意味要素に対して一つのベクトル用意する．新しい意味要素が今までに提起された意味要素と独立であれば，用意するベクトルを今までの基底ベクトルに直交するように定義し，これを新しい基底ベクトルの一つとする．このようにして独立な意味を提起する知覚連語とそれに対応する基底ベクトルを定義していけば，基底ベクトルは次第に増えていくが，知覚連語のすべてが新しい意味要素を提起するわけではない．たとえば,「赤い薔薇の花」という知覚連語と「真紅の薔薇の花」という知覚連語は，色彩の微妙な違いを別にすれば，ほぼ同義と見て良い．もし色彩の微妙な違いを問題にするとすれば異なる意味要素を提起する．つまり，意味的差異をどの程度細かに区別するかによってかによって意味空間の定義の仕方は異なる．いま，二つの知覚連語が示す意味が同義であるとするなら，これらの知覚連語は同義という規則によって一つの同値類に属し．共通の基底ベクトルを提起する．

$$\text{「赤い薔薇の花」} \equiv \text{「真紅の薔薇の花」} \Rightarrow |\text{赤バラ花}\rangle. \qquad (4)$$

つまり独立な意味要素を提起する十分に長い知覚連語を，順次十分な個数とることによって意味空間の正規直交基底を構成できるとするのが，我々の基本的な仮説であり，スキームである．

我々は語が結合することによって意味を限定していく課程を表現するために，一つの語に一つの線形演算子を対応させる．独立な意味要素を提起する知覚連語と意味空間の基底ベクトルを一対一に対応させる方式では，知覚連語を表現する演算子は対角行列で表現される．しかし，知覚連語によって提起されていく独立な意味要素は単に知覚連語の独立性をラベルしているだけであって，我々が日常的に概念を分類するための指標としている意味要素とは必ずしも一致しない．したがって，実用的な意味で日常的な概念の分類を優先させるためには，日常的な概念分類の意味要素に対応した新たな基底ベクトルを導入することによって，基底ベクトルを主軸変換する必要がある．このような主軸変換を行えば，もはや知覚連語が対角行列で表されるという保証はない．二種類の意味要素を区別する必要があるので，今後，知覚連語と一対一に対応する意味要素を知覚意味要素，日常的な概念分類の意味要素を概念意味要素と呼ぶことにする．

原理的な問題を考えるときには，知覚連語を対角化する基底を用いるほうが便利である．このようにするとき，“白い薔薇の花”という知覚連語は，［白バラ花］を知覚意味要素として，次のような線形演算子 $W_{\text{白い薔薇の花}}$ で表される．

$$W_{\text{白い薔薇の花}} = P, \tag{5}$$

$$P \equiv |\,\text{白バラ花}\rangle\langle\text{白バラ花}\,|. \tag{6}$$

正規直交基底の定義から，演算子 P は射影演算子である．

$$P^2 = P. \tag{7}$$

知覚連語の構成語である“白い”，“薔薇”，“花”の各語も線形演算子と考え，これらの語はその多義性に応じて，いくつかの意味要素に対する基底演算子で構成される射影演算子の和で表されるとする．これらが掛け合わさって上述の射影演算子 P 残すためには，P 自身を成分として持たなければならない．かくして，3 つの構成語に対する線形演算子は演算子 P を成分に持つことになる．以上の様な方法で構成語に対する線形演算子を定義していくと

$$W_{\text{白い}} W_{\text{薔薇}} W_{\text{花}}$$

のような積で，基底ベクトルの直交関係により，他の成分は落とされ，

$$W_{\text{白い薔薇の花}} = W_{\text{白い}} W_{\text{薔薇}} W_{\text{花}}, \tag{8}$$

のような，関係が成立するようになる．このようにして，語の結合により意味要素が限定される関係を線形演算子で表現できる．

文が知覚連語分解されていれば，文の意味要素を線形演算子 S で表すと，

$$S = \sum_{\text{文を構成する知覚連語}} W_{\text{知覚連語}}, \tag{9}$$

の様に表される．複文のような場合であっても，動詞句＋名詞［名詞句］が一つの知覚連語と考え，この意味要素を名詞［名詞句］に託すという方法をとれば，このような分解は可能である．もし意味空間をベクトルで表現したければ，文の意味ベクトルを $|V_S\rangle$ として，

$$|V_S\rangle = S|N\rangle, \tag{10}$$

のように定義すればよい．ここで $|N\rangle$ は

$$|N\rangle = \sum_{\text{すべての知覚意味要素}} |\,\text{知覚意味要素}\rangle, \tag{11}$$

である．基底ベクトル｜知覚意味要素〉の正規直交性を考慮すると，

$$|V_S\rangle = \sum_{\text{文を構成する知覚意味要素}} |\,\text{知覚意味要素}\rangle, \tag{12}$$

ように計算される．

ここで，先に述べた概念意味要素を用いた表現を見てみたい．概念意味要素による基底ベクトル｜概念意味要素〉の完全性を用いると，｜知覚意味要素〉は次のように展開される．

$$|\text{知覚意味要素}\rangle = \sum_{\text{すべての概念意味要素}} |\text{概念意味要素}\rangle\langle\text{概念意味要素}|\text{知覚意味要素}\rangle. \quad (13)$$

この式の右辺が知覚連語が保持する意味内容であり，この式が概念意味要素の言葉で表現された知覚意味要素の定義式であるとを考えれば，知覚意味要素の名前は単に知覚連語との対応関係を指定するためにふられたラベルであることが明瞭である．

式（12）と式（13）は，概念意味要素による概念解析のアルゴリズムを示している．まず，ある知覚連語が依存する概念意味要素の集合を ｛概念意味要素 1，概念意味要素 2，･･･｝ として，次のような大域変数 ^NCDIC を定義する．

^NCDIC（“知覚連語”, “変化形態”）＝“概念意味要素 1/概念意味要素　2/･･･”. (14)

ここで変化形態とは，連語と単語の区別や語尾変化に対応する品詞名による指標である．これをもとに知覚連語の変化形態の一つ一つと概念意味要素の相関を決定する大域変数 ^NWCAMP を定義する．

^NWCAMP（“知覚連語の変化形態”, “概念意味要素”）=〈概念意味要素｜知覚意味要素〉. (15)

この大域変数の値をもとに意味ベクトル｜V_s〉を決定する方式は既に明らかである．

6　まとめと展望

日本語解析システム「ささゆり」の構文理解，品詞解析と意味解析のスキームについて述べた．

「ささゆり」の構文理解の仕方は，重文，複文などの複雑な構文，中でも複文の構造を分解してして，動詞文が名詞（名詞句）を修飾する部分を百合の葉のように枝分かれ構造として捉える．これがこのシステムが「ささゆり」と名付けられたゆえんである．

我々のシステムは日本語文を構文構造を壊さないようにして，知覚連語と機能語に分解していく．したがって，多価性の強い機能語に対しても品詞解析が有効に働く．我々は品詞解析に機能語接続行列（CMF；Connection Matrix for Function words）と称する方法を提示し，先行する活用語と後続する機能語の整合性から，先行詞と機能語の品詞を再認識させる方法を考案した．我々の機能語接続行列の方法は，二項判断を行うのみではなく，必要があれば，先行詞，中央詞，後続詞の三項の整合性を判断する三次元行列に拡張される．さらに，このような複雑なパターン参照の問題に対して，我々の方法では M 言語の大域変数をうまく参照していくことによって，高速処理と判断の正確さを保証するアルゴリズムを実現している．

「ささゆり」の意味解析の方法は，知覚連語によって純粋な意味状態が構成され，知覚連語によって意味空間が対角化されるという原理によって，意味空間を構成していく点に意味解析の大きな特徴を持っている．さらに「ささゆり」の構文認識の方法は，複文などの複雑な構文に対しても，知覚連語による意味空間の対角化スキームを支障なく推し進めることを保証する．

最後に，我々は知覚連語による意味ベクトル空間の原理的な規定を与えるとともに，現実的な概念意味要素による表現との関係を明確にした．

References

1）高橋　亘，“日本語解析システム「ささゆり」の品詞解析機能と概念解析”，『Proceedings 2005 M Technol-

ogy Association of Japan』，9～14（2005）．
2）高橋　亘，“大域変数の階層構造と日本語文切断のアルゴリズム”，『Proceedings ’99 M Technology Association of Japan』，7–1～7–4（1999）；『MUMPS』22, 29～36（2002）
3）高橋　亘，“音声的ユニバーサル・インターフェイスと日本語解析”，『電子情報通信学会技術研究報告』WIT 99–1～22［福祉情報工学］，第二種研究会資料　Vol. 99 No. 1, 59–64（1999）
4）高橋　亘，“日本語文切断のアルゴリズムと　M 言語の大域変数の階層構造”，『情報科学研究』（関西学院大学情報メディア教育センター），No. 14（1999）
5）高橋　亘，“M 言語によるコンテキスト判断機能持つ人工知能と TTS インターフェイス，”『Proceedings 2000 M Technology Association of Japan』，55～58（2000）
6）高橋　亘，“ユニバーサル・インターフェイスにおけるコンテキストに依存する漢字の読み分けと人の言語知覚，”『電子情報通信学会技術研究報告』WIT 00–26～38［福祉情報工学］，第二種研究会資料　Vol. 00 No. 3, 31–36（2000）．
7）高橋　亘，“脳と言葉（2）－コンピュータによる言語解析が示唆するもの－，”『関西福祉科学大学紀要』No. 4，（2001）．
8）長谷川直子，清藤秀樹，高橋　亘，“日本語における失文法失語と言語知覚の単位”，『電子情報通信学会技術研究報告』SP 2001–76，WIT 2001–30（2001–10）［音声・福祉情報工学］，Vol. 101 No. 352，23–30（2001）．
9）長谷川直子，清藤秀樹，高橋　亘，“日本語における失文法失語と言語知覚の階層構造”，『関西福祉科学大学紀要』No. 5, 75–89（2002）．
10）高橋　亘，“言語知覚の単位を考慮した M 言語による日本語解析機能”，『Proceedings 2002 M Technology Association of Japan』，37～42（2002）．
11）高橋　亘，渡邊大樹，“M 言語による概念カテゴリー解析機能”，『Proceedings 2003 M Technology Association of Japan』，29～32（2003）．
12）高橋　亘，渡邊大樹，“コンピュータによる概念解析の方法”，『関西福祉科学大学紀要』，Vol. 7, 59～81（2004）．
13）高橋　亘，“M 言語による意味解析システムの学習機能”，『Proceedings 2004 M Technology Association of Japan』，49～52（2004）．
14）高橋　亘，“概念解析における学習機能”，『関西福祉科学大学紀要』，Vol. 8, 17～26（2005）．
15）黒橋禎夫，長尾　眞，“日本語形態素解析システム　JUMAN version 3.61”，京都大学大学院情報学研究科（1998）．
16）黒橋禎夫，河原大輔，“日本語形態素解析システム　JUMAN version 4.0”，東京大学大学院情報情報理工学系研究科（2003）．
17）松本裕治，北内　啓，山下達雄，平野善隆，松田　寛，高岡一馬，浅原正幸，“形態素解析システム『茶筅』version 2.3.3　使用説明書”，奈良先端科学技術大学院大学・情報科学研究科・自然言語処理学講座（2003）．
18）ウィトゲンシュタイン，『哲学探究』，ウィトゲンシュタイン全集　8，大修館書店（1976）東京．