【総合福祉科学研究】 創刊号 091–100 頁

【論文】

日本語解析システム「ささゆり」における日本語文簡易化の方法と知覚連語間の意味的距離

高橋 亘

Semantic Distances between the Perceptive Collocations and
Reduction Methods for Japanese Sentences with the Japanese Analysis System SASAYURI

Wataru Takahasi

01
2010年3月
総合福祉科学研究
Journal of Comprehensive Welfare Sciences

【論文】

日本語解析システム「ささゆり」における日本語文簡易化の方法と知覚連語間の意味的距離

高橋　亘 *

Semantic Distances between the Perceptive Collocations and Reduction Methods for Japanese Sentences with the Japanese Analysis System SASAYURI

Wataru Takahasi

要　旨

日本語解析システム「ささゆり」における二つの日本語文簡易化の技術が議論される．どちらの技術も知覚連語の意味的距離が問題になる．この論文で，新しい知覚連語間の意味的距離が定義される．指定された意味要素の組から意味的距離がある範囲内にある知覚連語の高速検索を実現するために，知覚連語 - 意味要素について相関関数と逆相関関数が定義された．

日本語文簡易化の技術の第一は，形式名詞が連体修飾をうける複文の単文化の技術である．複文を捌き，形式名詞の意味を推定する方法はこれまでの我々の研究で明らかであるから，この論文では，文脈によって限定された形式名詞の意味に最も近い日常語の検索技術が議論される．

日本語文簡易化の技術の第二は，難解語の言い換えに関する技術である．共通の単語を含む知覚連語の同値類と意味的に距離の近い知覚連語の同値類との双方の特性を活用した言い換え表現を提供する技術である．特に意味的に距離の近い知覚連語の同値類の活用はキーワード検索では検索不可能な異表現で同義な知覚連語の検索を可能にさせる．

Abstract

Two reduction methods for Japanese sentences with the Japanese analysis system SASAYURI are discussed. Both methods are closely related to the semantic distances between the perceptive collocations. In the present article a new definition of semantic distances between perceptive collocations is given. We need to retrieve the perceptive collocations within a certain range of semantic distances from a set of semantic elements. In order to quickly retrieve the perceptive collocations, we defined the correlation functions and their inverses that correlate the perceptive collocations to the semantic elements.

The first reduction method for Japanese sentences applies to complex sentences in which a pseudo-noun is modified with a verb phrase. As discussed in our previous articles, we have already established the method for decomposing a complex sentence into a triplet, made up of a verbal modifier, a modified noun and the remaining skeleton of the sentence. In our previous theory, we can specify the meaning of the pseudo-noun in the complex sentence from the context. A method for replacing the pseudo-noun with a

* 関西福祉科学大学　社会福祉学部　教授

noun having a concrete meaning is presented in this article.

The second reduction method for Japanese sentences applies to the methods for rephrasing sentences containing difficult words into sentences that are easy to understand. We have two equivalence classes of perceptive collocation. One class is of the perceptive collocations that comprise a special word, and the other class is of the perceptive collocations that carry the same meaning. In order to rephrase such sentences, two classes with specific features work together. Chiefly, our method for identifying the perceptive collocations with the same meaning presents a new method, different from that for retrieval by key words. And it presents a technique for retrieving sentences that use different representations but have the same meaning.

● ● ○ **Key words**　日本語解析システム「ささゆり」Japanese analysis system SASAYURI ／知覚連語 perceptive collocations ／知覚連語の言語学 linguistics of perceptive collocations ／形成規則 formative grammar ／意味解析 semantic analysis ／意味的距離 semantic distance ／日本語文の簡易化 reduction of Japanese sentences ／コミュニケーション支援 communication support ／ M 言語 M language

1．はじめに

近年我々は，日本語解析システム「ささゆり」の基礎をあたえる知覚連語の言語学の構築を推し進めてきた.[1] この新しい言語学の基本原理をまとめると次のようになる.

知覚連語の言語学は，通常の言語において単語が明確な意味を保持せず，連語を形成するときに意味が純粋化されるという原理を基礎にとって，明確な意味を伝達すること，つまり明確な知覚を誘発する連語を形成することこそが言語習得の原動力となることに基礎を置いている．知覚連語とはこの明確な知覚を誘発する連語のことである.

知覚連語は二つの側面を有する．その一つは知覚連語の形成過程に由来するものである．知覚連語の言語学によれば，知覚連語はその形成規則にしたがって形成され，形成規則は知覚連語の構成要素の範疇列によって定義されるのであるから，知覚連語は最終的に語彙範疇の列によって構成される．もう一つは，十分な長さをもつ知覚連語は純粋な意味に対応するのであるから，より少ない意味要素によってラベルされることに由来するものである. 知覚連語の言語学によれば，単語は通常多くの意味要素を保持しているが，単語が連語を形成し，連語が長くなるにつれて，連語の形成語間の相互規定のために連語が保持する意味要素はきわめて少なくなる．つまりうまく構成された連語は純粋な意味を保持するようになる．この原理が語の結合を単に連語と呼ばずに知覚連語と呼ぶ根拠である.

知覚連語の形成規則はある種の語の結合を意図的に禁止している．その一つは連体修飾をする動詞文と被修飾名詞の結合の禁止である．一般に動詞文による連体修飾の構造をもつ複文では，修飾する動詞文が長くも短くもなりうるので，短い動詞文と被修飾名詞の結合を連語名詞（連名詞）と把握してしまうと長い動詞文の修飾関係が把握できなくなるためである．我々の日本語解析システムでは動詞文による名詞の修飾を知覚連語の形成規則から排除することによって，期せずして複文の構文解析の機能を保持することになった.

このような複文の構文解析では修飾する動詞文（修飾子）と修飾される名詞（接合名詞）との対応関係と接合名詞を含んで後続する知覚連語（後続子）の構成関係の二つの関係が接合名詞の意味を限定する.[2] この接合名詞の意味限定の原理は，形式化した名詞，つまり形式名詞についても同様である．一般に複文のような構造を持つ文は聾者にとって分かりにくいものとされている．複文の要としての接合名詞に形式化したものが入ると，さらに理解を困難にする．我々は聾者に対する情報保障のために，形式名詞を含む複文を単文化する問題に我々の意味限定の方法を適用することを試みた .[3] このような目的を解決するためには，意味を限定された形式名詞を限定された意味に最も近い，意味の明確な，日常的名詞に置き換える必要があ

る．ここに語（知覚連語や単語）が保有する意味要素の集合の間の意味的距離を規定しなければならないという問題が浮上するのである．この論文の第一の主題はこのような意味要素の集合の間の意味的距離の測度の問題である．

知覚連語の形成規則のいま一つの知覚連語形成の禁止則は，機能性の高い機能語が端に来るような結合の禁止である．構文が長くなると語の結合は単に意味を規定し合うだけでなく，異質な意味の間の関係性を規定するようになることがある．このような関係性を把握するためには，機能性の高い機能語を知覚連語に内包させるべきではない．このような構文については，知覚連語間の接続関係によって，我々が高次の総合関係と称した指標付き直和ベクトル空間として把握するべきである．指標付き直和ベクトル空間の問題は，機会を改めて議論したい．

この論文で問題にする第二の主題は知覚連語間の同値類の問題である．先に述べた知覚連語の二つの側面は二種の同値類の存在を示唆する．その一つは知覚連語を構成する範疇の共通性がもたらす同値性であり，共通な範疇を保持する知覚連語が一つの同値類を構成する．このような同値類は，いわば共通のキーワードを保持する知覚連語の探索技術と直接的に関わるものである．このような知覚連語の同値類を，今後，共通語同値類と呼ぶことにする．いま一つは共通の意味要素を保持する知覚連語の同値類であり，意味的距離の近い知覚連語が一つの同値類を構成する．これは意味空間の構成が前提であり，知覚連語の構成によって始めて定義が可能になるものである．このような同値類は，共通のキーワードを保持しない知覚連語の探索技術を提起するものであり，情報検索の新しい技術である．このような知覚連語の同値類を，同義性同値類と呼ぶことにする．

二種の同値類，つまり共通語同値類と同義性同値類は，双方の利便性が相まって情報検索に寄与するものであるが，コミュニケーション支援の基本的方法を提示する．コミュニケーションは，人物 A が自分の記号化規則を用いて，自らの意図を言語記号によるメッセージを構成することに始まる．メッセージが他者に伝わると，メッセージを受け取った人物 B は，これをその人の記号化規則を逆に用いて復号し，メッセージ構成者の意図を推し量る．一般的に記号化規則は人ごとに異なっているから，復号された意図はメッセージ構成者の元のままではない．そこで，B に，B が理解した A の意図を，B の記号化規則にしたがってメッセージ化してもらうことをしてみると B の構成したメッセージによって，A は自分の意図がどのように相手に伝わったのかを知ることが出来る．この時，多くの場合，B の理解した A の意図はしばしばいくらかのずれをともなって伝達されることが観察させる．コミュニケーションが正確に成り立つためには，少なくとも数回のメッセージのやり取りが必要な所以である．

やり取りされるメッセージの中に A と B のどちらかが，知らないか，もしくは分かりにくい言葉（記号）があると，このような対話は特異的になる．特異的な対話では，対話が進行するためには，一方が理解できない言葉を含むメッセージ（これは共通の言葉を含む知覚連語の集合，つまり共通語同値類を構成する）を，その言葉を理解出来る側から，同じ意味を持つ，他の言葉で表現された，別のメッセージ（これは意味的に近い知覚連語の集合，つまり同義性同値類を構成する）を作成する必要性が生じる．つまり，未知の語を既知の語を用いた言い換えの集合を構成することで，未知の語を含むメッセージの理解が促進されることが期待されるのである．

このような未知の言葉を含む文の言い換えの技術は，障害があるために，ある種の言葉に系統的に分かりにくい言葉が集中する時にも有効な方法を提供する．聾者に分かりにくい日本語文は（1）語彙自体が分かりにくいもの，（2）構文が難解なもの，（3）日本語と日本手話で補助化する単語が異なるもの，などが挙げられる．[1] 先に言及した複文は（2）の例であり，形式名詞は（1）と（2）にまたがっている．（3）の問題は，いまは触れないにしても，（1）の語彙として分かりにくいものには機能語を筆頭に，補助化した内容語や抽象名詞，オノマトペ（擬音語・擬態語）などが系統的に分かりにくいとされている．こうした言葉には上述の言い換えが必要なのである．

第 2 節では，知覚連語間の意味的距離の測定方法が，第 3 節では，2008 年以来取り組んできた形式名詞を含む日本語文の簡易化の問題が取り扱われる．[3] これには構文の簡易化と形式名詞の意味推定の技術，さらに限定された意味要素の集合に最も近い日常語を見つけ

る技術などが総て投入される．この節で述べられることは，その前半は，宮地との共同研究として 2009 年の本学紀要に公表された内容の要説，後半は，宮地の修士論文として発展させられた内容を筆者の観点から補遺するものである．第 4 節では，知覚連語の二種の同値類，共通語同値類と同義性同値類の一般論が述べられる．また，この技術の適用例である，聾者の情報保障のためのオノマトペの言い換え技術が，津村との共同論文として本誌に公表される．[4]

2．知覚連語間の意味的距離と知覚連語 - 意味要素相関関数

一つの知覚連語にはその知覚連語が表現している意味要素の組が対応する．二つの知覚連語があって，それぞれの知覚連語が意味要素の組 A，B を持っていたとする．このとき，二つの知覚連語の間の意味的距離をどのように定義すればよいのだろうか？A，B の要素数がそれぞれ n_A 個，n_B 個であったとし，A，B 共通のものが $n_{A\cap B}$ 個あったとすれば，共通性のない意味要素の個数は $n_A + n_B - 2n_{A\cap B}$ であり，この要素数が二つの知覚連語の意味的隔たりの程度を表していることに異論はなかろう．これに対して共通の意味要素の個数 $n_{A\cap B}$ 個は二つの知覚連語の意味的共通性の程度を表していることもまた然りである．しかしこれらの二種の程度を表している量はどちらも個数という物理的には意味のよく分からない量を単位として測定するものであることも事実である．数学や物理学では物事の特性を示す数量を定義する際にしばしば単位系によらない数量として無次元数を定義する．これに習えば，今の場合，知覚連語の間の意味的距離を次の式で定義するのが妥当であると考えられる．

$$d_{AB} = \frac{n_A + n_B - 2n_{A\cap B}}{n_{A\cap B}}.$$

この定義では，A と B が集合として等しい場合に意味的距離が 0 となり，A と B に共通する意味要素がないときに無限大となる．（もちろんコンピュータで無限大の数量を扱うことは出来ないので，情報科学的には意味的距離の最大数をコンピュータで扱いうる最大数によって制限することになる）また，この定義は，隔たりの程度と共通性の程度の比であることから，隔たりの程度の大きいものに対して値が大きく，共通性の程度が大きいものに対して値が小さくなるという特性も知覚連語の間の意味的距離の性質をよく保持していると考えられる．

ここで，我々は知覚連語間の意味的距離の測定を必要とするいくつかのアルゴリズムについて触れておきたい．日本語解析システム「ささゆり」においては，知覚連語が一つ存在するとそれが保持する意味要素の集合は，システムが保有する日本語概念辞書に対応する大域変数 ^NCDIC によって一意的に決定される．大域変数 ^NCDIC は，

^NCDIC（Col,PS）=SemSet

の形を持っており，ここで大域変数をラベルする Col は知覚連語，PS は知覚連語の連語範疇である．大域変数の値である SemSet は知覚連語が保有する総ての意味要素をセパレータ “/” を挟んで接合したものである．したがって，知覚連語とその連語範疇が与えられるとその意味要素の集合が一意的に与えられるので，二つの知覚連語とその連語範疇が与えられれば，これらの間の意味的距離を測定する方法は意味的距離の定義から自明である．

ここで我々が問題にしたいのは，一つの知覚連語かあるいは一つの意味要素の集合があるときに，いずれの場合も意味要素の集合が一つ決まるわけであるが，与えられている知覚連語をのぞいて，この集合に最も近いか，もしくは意味的距離がある基準の範囲内にあるかの知覚連語を見つけ出すアルゴリズムである．現在，日本語解析システム「ささゆり」が知覚連語として学習完了しているものは単語も含めて 700,000 語を超えている．このような膨大なデータを総てなめ尽くして検索するのに必要な時間は，4 GHz を超える処理速度を持つコンピュータであっても，数秒がかかることになる．したがって，検索効率を上げるには，あらかじめ知覚連語と意味要素の相関，逆相関の関係を記憶しているデータを，M 言語の大域変数として，

^NWCSAMP（Col,Sem），

^NWCSIAMP（Sem,Col）

のようなものを定義しておく方法が考えられる．ここで Col は知覚連語，Sem は個々の意味要素である．このような大域変数を ^NCDIC から生成する方法は半ば自明であるが，この操作に必要な時間は数 10 秒と

推定される．しかし，一度定義しておけば，ある知覚連語を指定して，その知覚連語が依存する意味要素を総てリストするときには相関関数 ^NWCSAMP を用い，逆に意味要素を指定して，これを含む知覚連語の総てをリストするときには逆相関関数 ^NWCSIAMP を用いて，それぞれ第二階層を手繰れば，どちらの検索もほとんど一瞬に完了する．

いま我々が問題にしているのは，指定された意味要素の組 S に最も近い知覚連語を探すことであるから，アルゴリズムの第一段階は，意味要素の組 S の個々の意味要素に相関を持つ知覚連語をリストしてこれらを候補とすることである．このとき，一般には意味要素の組 S の複数の要素を保持する知覚連語も存在する可能性があるわけであるから，単純に候補をリストしていく手法では，重複して候補を挙げる可能性がある．このような知覚連語に対し一度候補に挙がった知覚連語を重複してリストしない仕掛けが必要である．

アルゴリズムの第二段階は，こうして，リストされた知覚連語の一つ一つと指定された意味要素の組 S との意味的距離を測定し，その最短のものを見つけることである．先に二つの意味要素の組の意味的距離を求める方法は確立しているので，この段階のアルゴリズムはほぼ自明である．

3．形式名詞の意味推定と形式名詞を含む日本語文の簡易化

一般的に連体修飾のある複文は聾者に分かりにくいとされているが，被修飾名詞が形式名詞になると，形式名詞の意味が単語のみでは推定し難いために，理解が一層困難になる．我々は，2008 年以来，形式名詞が連体修飾されている複文の機械的簡易化の技術について開発を続けてきた．[3]

これまでの我々の方法を要約すると次のようになる．日本語解析システム「ささゆり」によれば，動詞文による連体修飾のある複文は，連体修飾する動詞文（修飾子）と修飾される名詞（接合名詞），接合名詞が体言の役割を果たす骨格文に分解される．[1] 被修飾名詞が接合名詞と呼ばれるのは，これが修飾子と骨格文の接合部分を形成するからである．修飾子は，それ自体が知覚連語であり，接合名詞は，それが体言として働く骨格文の，特に接合名詞とそれに後続する部分の中で知覚連語（後続子）を形成する．

接合名詞の意味は一般に二つの関係によって限定される．二つの関係とは，

(1) 修飾子と接合名詞の対応関係，

(2) 接合名詞を含む知覚連語（後続子）の形成関係，

の二つである．接合名詞の意味は，前者の関係によって，修飾子の保持する意味要素の集合と接合名詞が元々保持していた意味要素の集合の積集合に限定され，後者によって，後続子の保持する意味要素の集合と限定前に接合名詞が保持していた意味要素の集合の積集合に限定される．したがって，二つの関係によって接合名詞の意味は，接合名詞が元々保持していた意味要素の集合と修飾子の保持する意味要素の集合，後続子の保持する意味要素の集合の三つの集合の積集合に限定される．[2]

形式名詞の意味推定の方法は，このような接合名詞の一般的な意味限定の方法を形式名詞に適用するものである．先の論文 [3] で詳説したように，我々の立場では形式名詞とは，「連体修飾を受け，骨格文で体言として振る舞う」ということによって始めて明確な意味を持つ名詞であると言うことが出来る．そして，意味の上から言えば，形式名詞は「大変多くの意味を保持してしまったが故に単独では意味を特定できないもの」と言うことが出来る．形式名詞は接合名詞で，かつ形式化したものであるから，一般の接合名詞の意味限定の方法によって意味が限定される．

以上がこれまで我々が開発に際して明らかに為し得た内容の要約であるが，ここで，この論文の主題の一つである，形式名詞を含む複文を単文化する最終段階の議論に移りたい．日本語解析システム「ささゆり」によれば，複文を修飾子と接合名詞の組と骨格文に分解する方法は既に確立しているのであるから，形式名詞を含む日本語文の簡易化は，上述の方法で意味を限定し，限定された意味要素の組に最も近い日常的名詞を見つける方法が確立すれば，その大半のアルゴリズムが確立することになる．いま我々が必要とするのは指定された意味要素の組に最も近い日常的名詞を見つける方法であるが，この方法は第 2 節で既に確立している．指定された意味要素の集合に最も近い知覚連語を，連語範疇を名詞に限定して，第 2 節のアルゴリズムを適用するだけのことである．

以上で，限定された意味要素の集合に最も近い意味を持つ日常名詞を見つける方法のアルゴリズムが確立したことになるが，我々はここで，我々の方法で簡易化される日本語文の実例を挙げて，本当に分かりやすい日本語文が提供されるのかどうかを検証しておくことには意味がある．ここで先の論文で引き合いに出したいくつかの例について，我々の方法を適用した結果を例示しておきたい．まずは形式名詞の特性を示す例文として，作為的に構成され，よく引き合いに出されるものを挙げてみよう．

［例 1］“ 彼がいうことは信用できない。”

このような例文に我々の方法を適用すると

（1）彼がいう こと［事柄 / 内容 / 言葉 ; 話］

< こと >（1）は信用できない 。

のような解析結果が出る．一行目が修飾子と接合名詞の対応関係，二行目が骨格文を表している．[ ] 内の “;” の前が限定された意味要素であり，後が形式名詞の代わりに用いるべき日常的名詞ということになる．解析結果から，［例 1］の単文分解した簡易化文は，

“ 彼がいう。その話は信用できない。”

ということになる．つまり修飾子と接合名詞の対応関係で形式名詞を捨象して言い切る．そして，骨格文で，限定された形式名詞の意味を表現する日常的名詞を用い，これに指示的連体詞を付加して形式名詞と置き換えた文を後続させたものが，元の文を簡易化した言い換え文ということになる．

同様の趣向をもつ，短い例文を 4 例あげておきたい．

［例 2］“ 毎朝牛乳を飲むことにしている。”

（1）毎朝牛乳を飲む こと［当為 / 習慣 ; 習慣］

< こと >（1）にしている 。

［例 3］“ ホッチキスとは紙を留めるのに使うものです。”

（1）紙を留める のに 使う もの［道具 / 学習 ; 学用品］

ホッチキス と は >> < もの >（1）です 。

［例 4］“ 昔、あそこへよく行ったものでした。”

（1）あそこへよく行っ た もの［経験 / 習慣 ; 習慣］

昔 、>> < もの >（1）でし た 。

［例 5］“ 人間は死ぬものです。”

（1）死ぬ もの［事態 / 必然 / 運命 ; 定め］

人間 は >> < もの >（1）です 。

［例 4］に関しては，使用される名詞 “ 習慣 ” にかかる連体詞は “ その ” ではなく，“ そんな ” の方が自然である．使用するべき自然な連体詞を決定する方法はこれからの問題であると考えられる .

上に挙げた例は，例文として作為的に構成されたものであるが，実際に文献に使用された例文についても我々の方法は有効であり，このような例を 2 例挙げておく．

［例 6］“ 太陽の引力とみあうだけの反対方向の力が惑星に対して働かねばならない。”【P. C. W. デイヴィス著，松田卓也，二間瀬敏史訳（ブラックホールと宇宙の崩壊）】

（1）太陽の引力とみあう だけ［平衡 / 程度 / 計量 ; 同程度］

< だけ >（1）の反対方向の力が惑星に対して働かね ば ならない 。

［例 7］“ このうち 3 羽は、去年巣立ったばかりの若い鳥だった。”【記者不詳（朝日新聞サイエンス動物）】

（1）去年巣立ったばかり［時期 / 状態 / 状況 / 直後 ; 直後］

このうち 3 羽 は 、>> < ばかり >(1) の若い鳥 だった 。

いずれも，機械的推定の妥当性が観察される．

4．共通の単語を含む知覚連語と意味的に距離の近い知覚連語

この節では，知覚連語の二種の同値性，構成要素として共通の単語を含むという意味での知覚連語の同値性（共通語同値類）と，意味的に距離が近いという意味での知覚連語の同値性（同義性同値類）とについて知覚連語の検索技術について述べ，その後で，これらの同値性を用いた難解語の言い換え技術について考察したい．

同義性同値類に関連する検索技術は，ある単語を含む特定の知覚連語に意味的に近い知覚連語をリストする技術である．これは二段階の操作で完了する．第一段階は，与えられた知覚連語の保有する意味要素の集合を取得することであるが，この集合は第 2 節で既に述べられている大域変数 ^NCDIC の検索で直接的に与えられる．第二段階は，与えられた意味要素の集合

に最も意味的に近い知覚連語の一群を検索することであるが，この第二段階は，第 2 節の後半で既に述べられている．

共通語同値類に関連する検索技術は，共通の単語を含む知覚連語の高速検索の技術である．日本語解析システム「ささゆり」によれば，知覚連語はその形成規則にしたがって機械学習される．知覚連語形成規則は範疇列の数語のパターンについて定義されているから，機械学習された知覚連語はその構成要素の列と構成要素の品詞の列が，ある系統性をもって配列されている．これらの配列は知覚連語の辞書に相当する大域変数 ^NWDIC の値の一部として記憶されている．この大域変数は，

^NWDIC（Col,PS,No）
=YOMI_HT_ADD_HT_STRING_
HT_PARSING_HT_SGN

のような構造を持った大域変数である．ここで大域変数をラベルする Col は知覚連語，PS は知覚連語の連語範疇，No は作品の特性などを指定する指標（通常は 1 に固定）である．大域変数の値は HT（アスキー番号 9 番の制御コード ; Horizontal Tabulation）で区切られた個々の特性であり，STRING が単語列，PARSING が品詞列である．（単語列と品詞列はそれぞれ半角スペース“ ”で区切られている）第 2 節で議論した大域変数 ^NCDIC は大域変数 ^NWDIC と双対的に生成されるものであるから，大域変数 ^NWDIC もまた総数 700,000 個を超えた数量を持っている．したがって，構成要素としてある単語を保持する知覚連語を総て見つけ出すという今の我々の課題は，ある意味要素を保持する知覚連語を総て見つけ出すという第 2 節の課題と酷似している．少し事情が異なる点は，STRING がスペース区切りの単語列であり，個々の単語は文中の形として活用形を保持しているから，これらの原型をデータベースに問い合わせて確認する必要があることである．このような確認作業を含めると，検索に要する時間は数 10 分に達する．したがって，検索効率を上昇させる必要性は先の例よりもはるかに大きい．我々はここでもまた，予め知覚連語と構成要素の相関と逆相関の関係を記憶しているデータを，M 言語の大域変数として，

^NWCWAMP（Col,Word），
^NWCWIAMP（Word,Col）

のようなものを定義しておくことが得策である．議論を繰り返すことを避けて結論を述べれば，ある知覚連語を指定して，その知覚連語が依存する構成要素を総てリストするときには相関関数 ^NWCWAMP を用い，逆に構成要素を指定して，これを含む知覚連語を網羅するときには逆相関関数 ^NWCWIAMP を用いて，それぞれ第二階層を手繰れば，どちらの検索もほとんど一瞬に完了する．

以上の議論で既に明らかなように，特定の単語を構成要素として含む知覚連語をリストする問題は大域変数 ^NWCWIAMP（Word,Col）の定義と検索の問題に集約される．

以上で二種の同値性に関連する検索技術の議論は完了するが，残された問題は，二種の同値性を用いた難解語の言い換え技術である．我々の置かれた状況は次のようなものである．ある人にとって難解な言葉が与えられたとすると，当該の言葉を含む知覚連語の集合 A が存在する．ここで，集合 A に属する知覚連語の一つ一つを X_i で表すことにする．一般に，一つの言葉はいくつも意味を持っているから，集合 A は言葉の使われ方，つまり知覚連語の形成のされ方にしたがって，X_i の保持する意味が規定される．いま，第 2 節で考察した知覚連語どうしの意味的距離を考えると，一般的に，ある知覚連語に意味的に近い知覚連語の集合という同値類が考えられる．意味的に近いという同値性はある単語を含んでいるか，いないかという同値性とは独立したものであるから，この同値類に含まれる知覚連語には，指定された単語を含んでいるものと含んでいないものが存在する．問題となる一つの言葉を決めると，これを含む知覚連語の集合 A に属する個々の知覚連語 X_i に対して，意味的に近い知覚連語の集合 B_i が存在する．つまり，当該の言葉を含む知覚連語の集合は，その個々の要素に意味的に同値な知覚連語の集合によって類別される．（図 1）ここに知覚連語の集合 B_i の一つ一つの要素には当該の言葉を含んでいるものといないものが存在する．つまりある言葉が難解であるとする人にこの言葉の意味を理解することを支援するためには，この人にこの言葉に対する言い換え表現が必要なわけであるが，いま述べた知覚連語の類別に登場する各々の意味的同値の集合における当該の言葉を含むものと含まないものとの照合体系こそ，我々が必要とする言い換え表現である．つまり

我々に必要な言い換え技術は，当該の言葉が含まれる知覚連語の意味を同定し，これと意味的に近く，当該の言葉を含まない知覚連語を検索する技術である．この技術のアルゴリズムはこの節の前半の議論で既に明らかであろう．

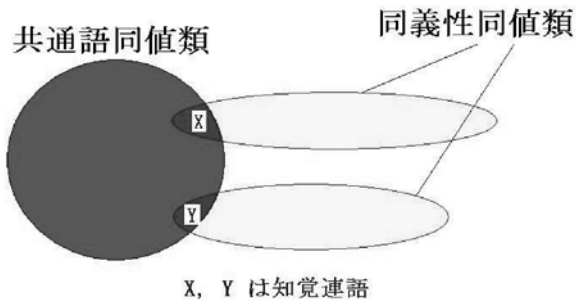

図 1 共通語同値類と同義性同値類

我々の技術は，様々な障害によってコミュニケーションに困難さが生じている個々の状況に応じて，有効な方法を提示するものと期待される．この技術の適用例として，聾者の情報保障のためのオノマトペの言い換え技術は，津村との共同論文として本誌に公表される．[4]

5．まとめと展望

我々は，日本語解析システム「ささゆり」における二つの日本語文簡易化の技術について議論した．二つの新しい技術は知覚連語間の意味的距離と密接に関連するものであった．

我々は最初に知覚連語間の意味的距離を定義することから議論をはじめた．一つの知覚連語には一組の意味要素の組が対応するから，知覚連語間の意味的距離は二組の意味要素の組の間の距離でもある．我々は二組の意味要素の組の間の意味的距離を二組に共通しない意味要素の個数と二組に共通する意味要素の個数の比として定義した．我々は，意味要素の組の間の意味的距離を用いて，意味的距離がある意味要素の組から指定された範囲にはいる知覚連語を検索するアルゴリズムを考察した．これには，知覚連語と意味要素の相関関数と逆相関関数を記憶する M 言語の大域変数を定義しておくことで検索効率を大幅に上昇させることが判った．

知覚連語間の意味的距離の定義は，まず，形式名詞を含む日本語文の簡易化の技術に適用された．我々は既に，先の論文[2,3]で，形式名詞を含む日本語文を修飾子，接合名詞，骨格文に分解する方法と文中における形式名詞の意味を推定する方法とを確立させていたが，これに，指定された意味要素の組に最も近い日常名詞を検索する技術が追加されることによって，形式名詞を含む日本語文の単文化の技術を確立することになった．

我々は，さらに知覚連語間の意味的距離の定義を意味的に距離の近い知覚連語を一つの同値類（同義性同値類）として扱う技術に適用した．知覚連語がその形成規則によって学習されるものであることから今一つの知覚連語の同値類，つまり共通の単語を含む知覚連語という意味での同値類（共通語同値類）の存在が予見される．二つの同値類は知覚連語について二種類の検索技術を提供する．共通語同値類の検索はいわばキーワードによる検索であるから，旧来からよく知られたものである．我々は共通語同値類の検索の効率を上げるために，知覚連語と共通語の相関関数と逆相関関数を記憶する M 言語の大域変数を定義しておく手法を提唱した．同義性同値類に属する知覚連語の検索技術は新しい検索技術であり，同じ意味内容を異なる表現の仕方で表現している知覚連語の検索技術である．

コミュニケーション支援の立場からみれば，知覚連語間の意味的距離の定義が適用された前述の二技術は日本語文の簡易化表現を与える技術として捉えられる．複文は聾者にとって分かりにくいものとされているが，複文の要となる接合名詞に形式名詞が来ると難解さはさらに増加する．こうした問題に対し我々の技術は，問題の日本語文を，意味の具体的な日常名詞を用いた単文に機械的に分解する手法を与える．

二種の同値類を組み合わせた検索技術は，一般的に，二者のコミュニケーションで一方が知らない言葉をもう一方が言い換えによって意趣を伝えるコミュニケーションの手法を与える．この技術は障害などによって系統的な言葉が分かりにくい場合の情報提供の方法としても重要である．たとえば，オノマトペは系統的に聾者に分かりにくいとされるが，我々の方法はオノマトペを文脈に従って別の言葉で表現する方法を与える．

この論文では，知覚連語の共通語同値類と同義性同

値類の検索技術についての原理的考察を行ったが，こうした技術を活用するには適切なインターフェイスが必要である．

共通語同値類については，検索する専門用語を入力して当該の用語を含む知覚連語と連語範疇，特性，概念辞書に登録された意味要素などをグリッド形式に表示するものが考えられる．このようなインターフェイスには，リストする知覚連語の個数を絞り込むための補助キーを入力するテキストボックスや知覚連語が保持すべき意味要素を概念辞書に追加登録するボタンなどが必要である．

同義性同値類については課題の文と専門分野，意味的近さの境界距離などを入力し，課題文が含む専門用語や意味要素の組を表示し，課題文と意味的距離が境界距離以内にある同義分をリストするものが考えられる．課題の文が，多くの専門用語の解説文に見られるような複文の場合にはこれを構文解析して，修飾子と接合名詞の対応関係から推定される接合名詞の意味要素を表示するテキストボックスも欲しい．

現時点でこのような二種のインターフェイスを実現する基礎技術はそろっているのであるから，我々は近い将来こうしたインターフェイスの活用を含めた議論を公表すべきである．

同義性同値類についてのインターフェイスは先に述べたコミュニケーション支援の観点からも重要であるが，障害者を支援するエディターや電子図書館などに組み込んでインライン化を図るなどの工夫をすれば，実用性は一気に高まる．こうした進展も近い将来の課題である．

引用文献

[1] 高橋 亘,『コミュニケーション支援の情報科学』，現代図書（相模原, 2007, 4 月）.
高橋 亘，“日本語解析システム「ささゆり」の言語学”,『Proceedings 2007 M Technology Association of Japan』, 14 ～ 18（2007）.
高橋 亘，“日本語解析システム「ささゆり」の基礎を与える言語学”,『関西福祉科学大学紀要』，Vol. 11，41 ～ 48（2008）.

[2] 高橋 亘，“M 言語による日本語解析システム「ささゆり」の意味解析 --- 連体修飾のある日本語文の意味解析 ---”,『Mumps』, Vol. 24（2008）27 ～ 33.
高橋 亘，“日本語解析システム「ささゆり」における連体修飾のある日本語文の意味解析”,『関西福祉科学大学紀要』, Vol. 12，21 ～ 30（2009）.

[3] 宮地絵美，高橋 亘，“M 言語による聾者のための日本語簡易化機能 --- 連体修飾のある日本語文の単文化と形式名詞の意味推定 ---”,『Mumps』, Vol. 24，35 ～ 40（2008）.
高橋 亘，宮地絵美，“聾者のための日本語簡易化法 --- 連体修飾のある日本語文の単文化と形式名詞の意味推定 ---”,『関西福祉科学大学紀要』, Vol. 12，31 ～ 39（2009）.

[4] 高橋 亘，津村雅稔，“オノマトペを含む日本語文の代替表現機能 --- 聾者のための情報保障の技術 ---”,『総合福祉科学研究』, Vol. 1（2010）115 ～ 122.