言語の使い方をもとにした単語間の関係性の発展

橋本 敬
理化学研究所 国際フロンティア研究システム
情報処理グループ 情報表現研究チーム
〒351-01 埼玉県和光市広沢 2-1
takashi@irl.riken.go.jp
http://www.bip.riken.go.jp/irl/takashi/

Abstract

言語の意味は単語間の関係性で表現され、その関係性は言語の使い方から導き出されるという考えをもとに、構成論的な手法により単語間の関係性の発展を研究する。

1 Introduction

1.1 進化システムとしての言語

言語は進化システムとみなすことができる。言語は、起源の時期においては、簡単なシステムであったはずである。単語の数は少なく、統語構造も単純で、抽象的は概念を表す表現もほとんどなかったのではないだろうか。我々の言語は、語形成、文法化、表現の多様化という過程を通して複雑化、精緻化し、また多様化して来たと考えられる。そして、言語は現在も変化し続けている。例えば、ピジン語やクレオール語は観測可能なタイムスケールで複雑化しているし、日々新しい表現があらゆる言語に加わり続けている。すなわち、言語は常に変化するシステムと言うことができる。このような、常に変化し続ける言語というシステムを理解するために、言語を、単純なコミュニケーション・システムからの進化という面から研究することが重要だと考えられる。

言語の起源や進化には、複雑系に典型的に見られるような現象が様々なレベルで深く関連している。例えば、個々の言語使用者の間で共有される記号やルールの創発、組織化されたルールの集団運動、地域的あるいは社会的な方言が生じるクラスター形成から、複数の言語への分化という多様化、記号素、単語、節といった文法のレベルでの階層の形成、あるいは、様々な単語がカテゴリーをなすクラスター化やそのカテゴリーが階層的なシステムをつくるという現象である。また、言語システムは適応可能性と安定性の両方の性質を持たなくてなはらない。もし厳密に規定されすぎているならば、新しい多様な経験を記述することができなくなるし、不安定すぎるならば、コミュニケーションができなくなるであろう。

構成的手法は動的な複雑系を理解する良い手段である [1]。あるシステムを部分に分解し、個々の現象を分析的に記述するのではなく、システムをつくる要素のダイナミクスとその間の相互作用を設定し、そのシステムの大域的な振る舞いや複雑化の過程を、実際に提示することで理解しようとする。この手法は進化言語学の研究にも有用であろう。なぜなら、言語とは、分散した複数のエージェントのシステムにおいてエージェント間の相互作用を通して創発し、複雑化してきたものとみなされるからである。様々な言語現象の記述を行なう伝統的な言語学に対して、エー

ジェントの内的なダイナミクスとその間の相互作用を元にしたモデルを構成し、言語的な振る舞いとしての大域的な秩序の創発を観察するというやりかたで言語を捉えてみようというのである。しかし、大域的な秩序が創発するだけでは不十分であろう。言語はつねに変化をしていくシステムなので、モデルにおいても、大域的な振る舞いの創発だけではなくその発展を示すことが要求される。進化言語システムをモデル化する際の重要な点は、エージェント自体のダイナミクスを導入することであろう。この元になるダイナミク スを通してエージェント間の関係が変化し、大域的なレベルでのダイナミクスも現れえる。

この構成的手法を用いた例として、我々は形式文法を持つエージェントの言語ゲームによる言語進化のモデルを提唱した [2, 3]。その研究では、統語構造の進化と単語の使用方を共有する集団の形成が起きることが見い出された。そこで共有される共通の単語使用方は、個々の文法が進化することにより断続的に変化する。この報告では、上記モデルに意味論的な性質を付加することを考え、言語における意味構造の発展を議論したい。

1.2 言語の使い方をもとにした意味観

単語の意味とはなにかという問題は、言語学においてもっとも議論の行なわれるものであり、様々な主張がある。例えば、単語の意味とはそれが指示する外的な対象である、それらはいくつかの必要十分条件で表される、いくつかの特徴のベクトルである正規形式で表される [4]、あるいは、ある特性を持つか否かという二値特徴づけで表される、等々。我々は、ある単語の意味とは他の単語との関係で表され、言語の使い方という観点から議論されるべきであると考えている。すなわち、ある単語と他の単語との関係は様々な文の中での、その単語の使われ方から導き出されるべきだと主張しているのである。

単語間の関係には syntagmatic な関係と paradigmatic な関係の二種類がある。前者は、ある一つの文における単語の結び付き方で形成される。例えば、「私は本を読む」という文における「本」という語と「読む」という語の間に作られる関係は syntagmatic な関係という。後者は、二つの文において文法的に等価な単語の間の関係である。例えば、「私は本を読む」「私は雑誌を読む」という二つの文における「本」と「雑誌」という二語の間の関係が paradigmatic な関係と言われる。しかし、単語の使われ方を元にした関係の構築という立場からすると、paradigmatic な関係は syntagmatic な関係を通して理解することができる。上の例では、「私は本を読む」という文は「本」と「読む」の二語が関連付けられ、「私は雑誌を読む」という文は「雑誌」と「読む」の二語を関連付けている。そして「読む」という単語との関係を通して、「本」と「雑誌」という二語が関係を持つのである。これらの二つの文からだけでは、「私」という語も同様に関連付けられるが、単語は様々な異なる文の中で使われるので、それぞれの単語の使用方や頻度に応じてある単語の他の単語との関係の網目は異なる形態をなすのである。

この単語間の関係を数値化すると、単語間の関係を抽象的な空間 (我々はこの空間を語空間と呼ぶ) にマッピングすることが可能になる。本研究では、単語間の関係を文の中での使われ方の他の単語との類似性をもとに評価するが、具体的には Karov and Edelman により機械可読な辞書とコーパスにおける多義語の明確化のために提出されたアルゴリズム [5] を用いる。このアルゴリズムの重要な点は、単語と文の相互依存性である。すなわち、類似した単語は類似した文の中で使われるもので、類似した文とは類似した単語で構成されるものだと考える。彼らのアルゴリズムはコーパスと辞書という静的なものの中で単語間の関係を構築するが、我々は言語の使用を通して語空間の中にどのような構造が発展してくるかというダイナミクスに注目したい。

2 形式文法エージェントと単語間関係のモデル化[1]

2.1 エージェント、単語、文

生成文法 (書き換えルールのリスト) を持つエージェントがそれぞれの持つ文法に基づいて、記号列の生成・受理を行なう。文法は各書き換えルールの受理過程での使われ方に応じて変化する。

文と単語を以下のように定義する。文とはエージェントに与えられる、あるいはエージェントが生成する、0 と 1 の記号列である。そして、書き換えルールの右辺に含まれる、終端記号が一つ以上連続した部分を一つの単語と定義する。エージェントは文を語の連なりとして理解するが、文から語への分節は受理の仕方により異なる。例えば、文法として、$S \to A0B, A \to 10, B \to 11$ を持つエージェントは、"10011" という文を

$$10011 \overset{A \leftarrow 10}{\Longrightarrow} A011 \overset{B \leftarrow 11}{\Longrightarrow} A0B \overset{S \leftarrow A0B}{\Longrightarrow} S$$

という受理過程を経て、"10·0·11" という単語の列に分節して理解する。この受理過程の解析木は図 1(a) の様になる。

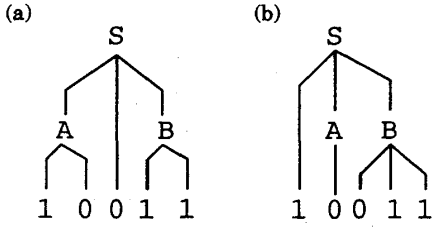


図 1: 解析木の例

一方、同じ文を、$S \to 1AB, A \to 0, B \to 011$ という文法を持つエージェントは "1·0·011" という単語列として受け取る。その時の受理過程は

$$10011 \overset{A \leftarrow 0}{\Longrightarrow} 1A011 \overset{B \leftarrow 011}{\Longrightarrow} 1AB \overset{S \leftarrow 1AB}{\Longrightarrow} S,$$

となり、解析木は図 1(b) である。すなわちエージェントの持つ文法により分節方法が異なる。

2.2 単語間の類似度の定義

単語間の関係を、単語の文中での使われ方の類似性の程度と定義し、Karov and Edelman の定義 [5] を修正したものを実際の計算に用いる。ここで単語と文の相互依存性が重要である。すなわち、類似した単語は類似した文の中で使われ、類似した文は類似した単語で構成される。単語の類似度の空間を語空間と呼ぶ。

単語間および文間の類似度を以下の式で定義する。

$$sim_{n+1}(w_i, w_j) = \begin{cases} \sum_{s \ni w_i} weight(s, w_i) aff_n(s, w_j) & \text{if } i \neq j \ , \\ 1.0 & \text{if } i = j \ , \end{cases} \tag{1}$$

[1] モデルの詳細は [6] を参照

$$sim_{n+1}(s_i, s_j) = \begin{cases} \sum_{w \in s_i} weight(w, s_i) aff_n(w, s_j) & \text{if } i \neq j \ , \\ 1.0 & \text{if } i = j \ . \end{cases} \tag{2}$$

関数 $aff_n(s, w)$、$aff_n(w, s)$ はそれぞれ、単語の文に対する親和度、および、文の単語に対する親和度であり、次の式で定義する。

$$aff_n(s, w) = \sum_{s' \ni w} weight(s', w) sim_n(s, s') \ , \tag{3}$$

$$aff_n(w, s) = \sum_{w' \in s} weight(w', s) sim_n(w, w') \ . \tag{4}$$

上記の4式で、添字 n はイテレーションの回数を表し、$w \in s$ はある文 s に含まれる単語を、$s \ni w$ はある単語 w を含む文を意味する。関数 $weight(s, w)$ と $weight(w, s)$ は正規化因子で、各々の単語と文の使用頻度や長さに関して重みづけを行なう。ある単語がよく使われると類似度等への影響が少ないが、よく出てくる文は影響が大きいとする。また、短い文の中の単語は長い文のなかの単語よりも重要であり、たくさんの文の中で使われる単語 (冠詞や be 動詞のようなもの) はたまにしか使われない単語よりも影響は少ないとみなす。これらの性質を反映させて、正規化因子は、

$$weight(s, w) = \frac{factor(s, w)}{\sum_{s' \ni w} factor(s', w)} \ , \tag{5}$$

$$factor(s, w) = \frac{p(s)}{\#(s, w)} \ , \tag{6}$$

$$weight(w, s) = \frac{factor(w, s)}{\sum_{w' \in s} factor(w', s)} \ , \tag{7}$$

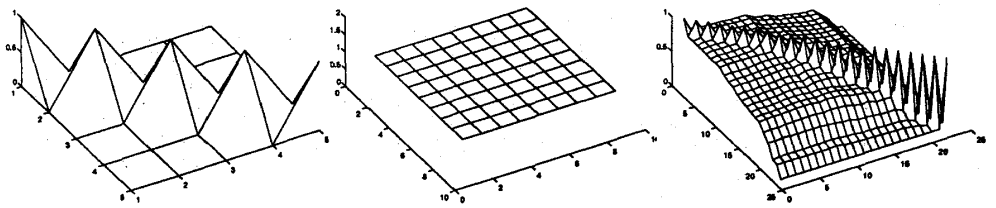
$$factor(w, s) = \frac{1}{p(w) lg(s)} \ . \tag{8}$$

で定義される。式 (6) と (8) で、$p(w)$ と $p(s)$ はそれぞれ、単語 w と文 s の出現確率で、$lg(s)$ は文 s の長さで、その文中に含まれる単語の数で定義される。また、$\#(s, w)$ は単語 w を含む文 s の出現回数である。

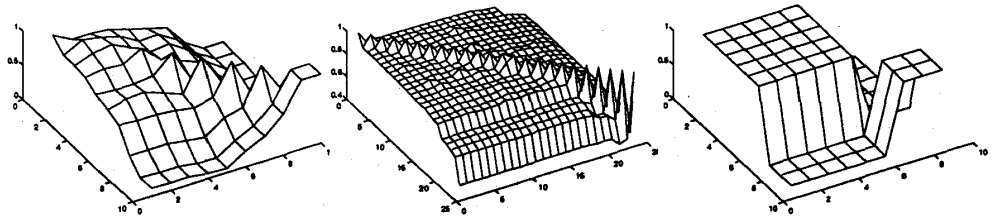
最初の計算ステップ ($n = 0$) において、単語それ自身に対する類似度 (単語間類似度の対角成分 $sim_0(w_i, w_i)$) は 1.0 に、それ以外は 0.0 に初期化する。これから $n = 0$ での単語-文親和度 (式 (4)) を計算する。そして、上記4式を (2) → (3) → (1) → (4) の順に計算する。

3 単語間関係の構造化

この章では、エージェントの持つ文法が理解過程での使われ方に応じて変化して行く場合のシミュレーションの結果を、語空間 (単語間類似度の空間) での構造に焦点を当てて見て行く。ランダムに生成された最大の記号数 8 の文をエージェントに与え、類似度の計算はエージェントが文の理解に成功するたびに行なう。文法の変異は 10 個の文与えるごとに行なう。



(a) solitary word (b) flat cluster (c) gradual cluster



(d) two-peak cluster (e) sub-clustering (f) two clusters

図 2: 語空間における構造の例。z 軸は語の類似度 (式 (1))。各単語は、構造が分かりやすい様に基準語を適当に選び、その語との類似度が降順になるように x 軸と y 軸に並べられている。語空間の構造によって 6 種類に分類される。(a) Solitary word. (b) Flat cluster. (c) Gradual cluster. (d) Two-peak cluster. (e) Sub-clustering. (f) Two clusters.

3.1 語空間の構造の分類

単語は語空間で、類似度クラスターを構成する。クラスターの構造を図 2に見られる 6 種類に分類する。

(a) soritaly word – 語が他のどれとも類似度を持たない場合

(b) flat cluster – クラスター内のすべての語がほとんど同じ類似度の場合

(c) gradual cluster – クラスター内の単語間の類似度が単語によって異なる場合。

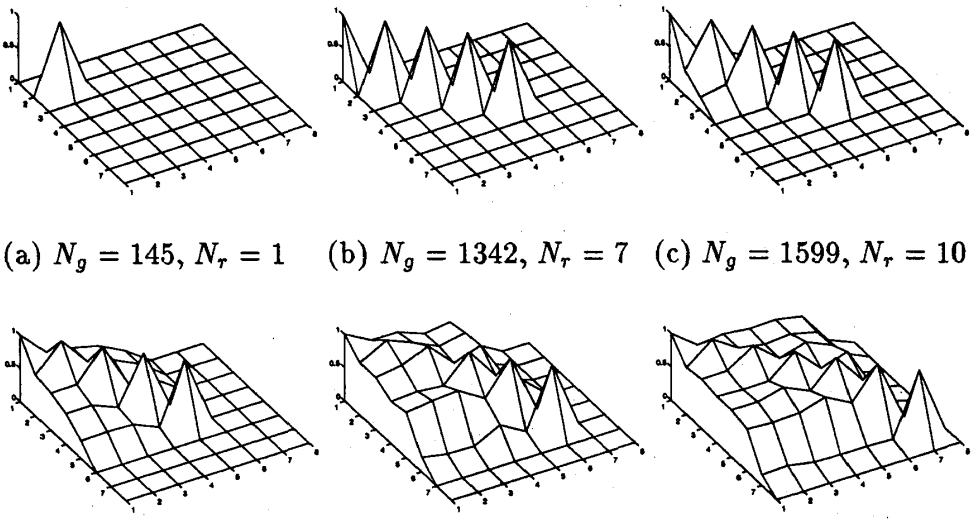
(d) two-peak cluster – 二つのピークがあるクラスター。

(e) sub-clustering – 階段的な構造になるクラスター。単語は一つのクラスター内で部分クラスターに分かれる。

(f) 複数クラスター – お互いに関連のない (類似性を持たない) 単語のクラスターが複数存在する。

3.2 語空間におけるダイナミクス

エージェントの文法をもっともシンプルなものからはじめた場合の発展のシナリオは、一般的に以下のようになる。まずはじめには、エージェントは一つの一語文を理解できるようになる。その後、複数の文を理解するようになるが、すべての文が一語文である。それゆえ、語空間にはいくつかの solitary word がある。やがて文をいくつかの単語の連なりに分節するようになり、単語間を関係付けはじめる。それにともない、単語は gradual cluster を形成するようになる。これらのクラスターが境界の拡大や二つのクラスターの結合といった過程を経て構造が変化して行く。この語空間での発展と並行して、文法は連続的構造から分岐構造をへて再帰構造へと発展する。



(a) $N_g = 145$, $N_r = 1$ (b) $N_g = 1342$, $N_r = 7$ (c) $N_g = 1599$, $N_r = 10$

(d) $N_g = 1641$, $N_r = 13$ (e) $N_g = 1747$, $N_r = 19$ (f) $N_g = 1853$, $N_r = 24$

図 3: 語空間の初期の発展。z 軸は単語間の類似度。単語は x 軸、y 軸に'0'、'1'、'01'、'101'、'10'、'11'、'011'、'00' の順に並べられている。各グラフの下の N_g と N_r は、それぞれ、エージェントに与えられた文の数、および、エージェントがそれまでに理解した文の数である。

このシナリオを二つの例を通して具体的に見てみよう。一つは初期の発展を見るために、もっとも簡単な文法からはじめたものの例をあげる。もう一方は、ある程度文を理解する能力をもった文法からはじめたもので、この例でクラスター構造の変化例示する。

3.2.1 初期の発展

書き換えルール $S \rightarrow 1$ だけを含む文法を持つエージェントが、ランダムに生成された文の理解を試みる。図 3に 語空間における発展を示す。最初に理解する文は、当然 "1" である。これは一語文なので語空間の発展は図 3(a) のようにたった一つの solitary word があるという状態から始まる。

エージェントに与えられる語の数 (N_g と表記) が 1580 まではすべての文 ("0"、"1"、"01"、"101"、"10") は一語文として理解される。よって 5 つの soliary word がある (図 3(b))。N_g が 1580 から 1900 で、エージェントが理解した単語の数 (N_r と表示) は急激に増加する。この期間のはじめにエージェントの文法は再帰構造を獲得している。再帰構造によりエージェントの理解できる文の集合は飛躍てきに大きくなる。この期間にエージェントは多くの新しい文を理解するが、それらは一語文ではなく、いくつかの単語の連なりとして理解するようになる。これらの文はそれまでの孤立した単語間を結びつけるものになる。例えば、N_g が 1599 ではじめて理解される "000111" という語は "0·0·01·1·1" に分節される。この文により、孤立語 '0'、'1' および'01' が関連付けされ、これらの単語は gradual cluster を形成する (図 3(c))。次に理解される文は "0·0·101·1·1·1" で、単語 '101' がこの gradual cluster に含まれるようになる。この様にして、$N_g = 1641$ で五つの孤立語が関連付けされる (図 3(d))。

新しい文の理解によってクラスターの境界が拡張される。$N_g = 1747$ で新しい単語 '11' が 文

"0·101·11" の理解によってクラスターに組み込まれる (図 3(e))。$N_g = 2853$ には文 "0·0·0·1·011·1" が境界を新しい語'011' にひろげる (図 3(f))。

3.2.2 クラスターの構造的変化

この章では、ある程度の大きさの文法[2]を持つエージェントの発展を例として、クラスターの結合、構造的変化、境界の拡張を見る。二つの two-peak cluster と一つの solitary word がある状態 (図 4(a)) が three-peak cluster を経て、一つのほとんど flat な cluster になる様子が、図 4(a)~(e) に示される。一つの文 "00·0101·1" が二つの two-peak cluster をつなぐ (図 4(b))。この文は新しい単語を含んではいないが、使い方が新しい。すなわち、これまで一つの文を構成していなかった単語とともに使われているのである。この結合部分を通してこれらのクラスターの関係はより強くなって行き (図 4(c))、関連付けられたクラスターは three-peak shape へと構造が変化する (図 4(d))。しかし、この three-peak の構造は長いあいだ続かず、境界を拡張しながらほとんど flat なクラスターになる (図 4(e))。

新しい文を理解することにより、このクラスターは大きくなって行く。また、solitary word がこのクラスターに含まれるようにもなる (図 4(f) と (g))。古い単語間の類似度はほぼ同じ値になるが、新しく理解されるようになった単語は、クラスターの境界で小さい類似度の値を持つ (図 4(h))。最終的にはこのクラスターはほとんど flat cluster になる (図 4(i))。

4 議論

4.1 カテゴリー化としてのクラスター形成

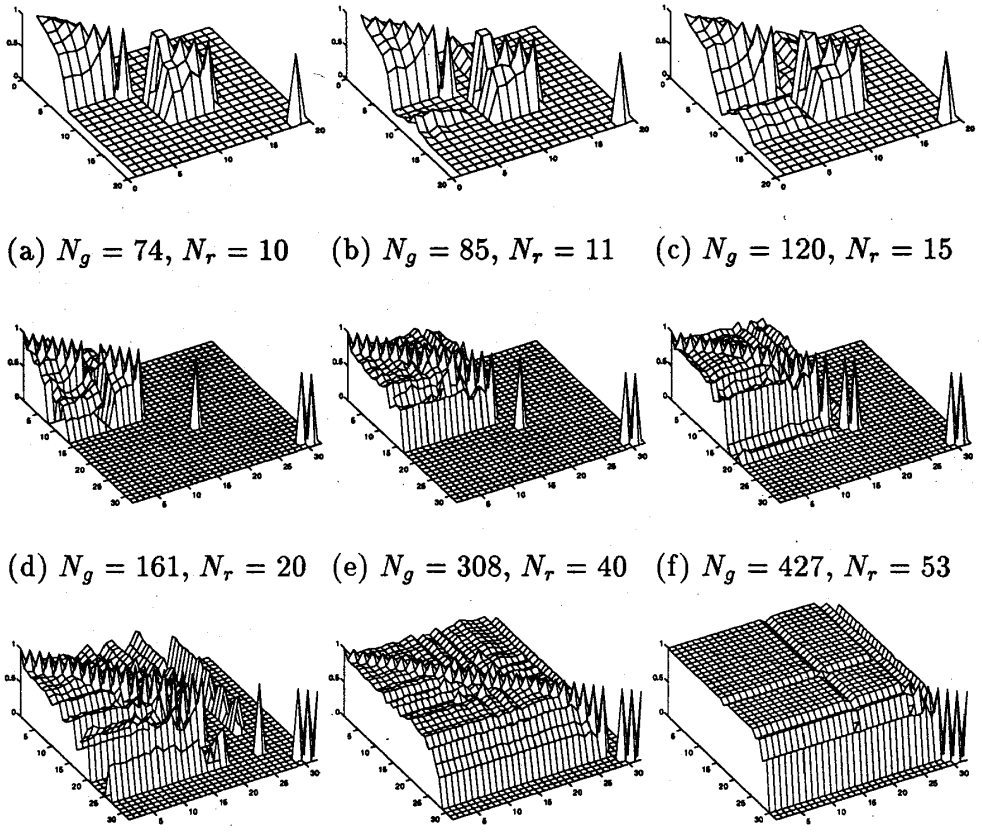
これまで、クラスターの構造とそのダイナミクスを見てきた。ここでは、このクラスター形成をエージェントによるカテゴリー化とみなした時の各構造について議論をする。ここでのクラスターとは、その中の単語は互いに強い関連性を持ち、クラスター外の単語とはほとんど関連を持たないというものなので、これをカテゴリーと結びつけるのは妥当であろう。

図 2(a) に示した solitary word は他のいかなる単語とも関連を持たない単語であるが、厳密にはこれはクラスターとはいえず、カテゴリーとはみなされないであろう。実際、たった一つのメンバーで構成されるカテゴリーというのは我々の知識システムにはない。しかしながら、発達のごく初期においてこのような形の単純な知識構造が存在するかもしれない。

すべての単語がほとんど同じ類似度をもつ flat cluster (図 2(b)) では、その境界、すなわちある単語がそのクラスターに属するかどうかが明確である。このタイプのクラスターは科学的な用語の様に、あるものがカテゴリーのメンバーか否かが必要十分条件で厳密に決定されているようなものに相当する。

一方、gradual cluster では 図 2(c) に見られるように類似度にピークがあり、単語によって大きいところから小さいところへと連続的に変化する。このピークはカテゴリーの中心的なメンバーで、この中心的メンバーと小さい類似度しかもたない単語は周縁的メンバーと見なされる。この構造は、プロトタイプ・カテゴリー [7, 8] と類似している。プロトタイプ・カテゴリーの理論では、あるメンバーがあるカテゴリーに属しているかどうかは段階的なものであり、周縁的なものがカテゴリーのメンバーかどうかはファジーなものであると考える。これらの点が、本研究での gradual cluster と類似する点である。

[2]この例では 33 個の書き換えルールの文法を初期に持つ



(a) $N_g = 74$, $N_r = 10$ (b) $N_g = 85$, $N_r = 11$ (c) $N_g = 120$, $N_r = 15$

(d) $N_g = 161$, $N_r = 20$ (e) $N_g = 308$, $N_r = 40$ (f) $N_g = 427$, $N_r = 53$

(g) $N_g = 500$, $N_r = 70$ (h) $N_g = 938$, $N_r = 200$ (i) $N_g = 1911$, $N_r = 700$

図 4: 語空間の構造変化の例。z 軸は語間の類似度。各単語はクラスタリングとそのダイナミクスが分かりやすい様に、x 軸、y 軸に並べられている。N_g と N_r はそれぞれそれまで与えられた文の数と理解した文の数を表す。

このプロトタイプ・カテゴリー的に考えると、two-peak cluster (図 2(d)) は多義的なカテゴリーとみなせる。すなわち、一つのカテゴリーにおいて中心的メンバーが二つあるのである。一つのピークを持つ構造の場合は、その中心メンバーとどの程度似ているかということで各単語を特徴付けることができるが、ピークが二つある場合には、一方の中心メンバーと似ているがもう一方とは似ていない単語や両方とある程度類似度を持つメンバーなどが存在する。

図 2(e) にあげたような sub-clustering を見せるクラスターは、一つのクラスターのなかで、さらに、類似度が高いグループと低いグループに 分かれているので、もっとも単純な階層的カテゴリー化をなしているものと解釈できる。実際のカテゴリーシステムは、このような階層的なカテゴリーが幾重にもなされているものであろう。

クラスターの構造の変化を §3.2.1と§3.2.2 でみた。そのダイナミクスは、境界の拡大、二つのクラスターの結合、孤立語の包含などがあった。 図 3(d) ~ (f) や図 4(e) ~ (h) に見られる境界の拡大の場合、 もともとの構造はそれほど大きな変化を見せず、周縁に新しい単語が組み込まれている。これは適応可能性と安定性を両立させるべきカテゴリーシステムの特性と一致し、プロトタイプカテゴリーの柔軟性 [8] とも類似している。

ある文が二つのクラスターを結合させる例が図 4(b) に示されているが、このクラスター間を結ぶ文の働きはメタファー的表現の働きに対応するものとみなせるだろうか。実際にメタファーは二つの意味領域を結びつける働きを持つ。しかし、メタファーは単に二つの領域を結びつけるだけではなく、元の意味領域の基礎的な論理構造を対象領域にマップすることで、大抵の場合に元の領域よりも抽象的で理解することが難しい対象領域の理解を行ないやすくするものだと考えられている [9]。 しかし我々の結果では単に二つのクラスターをつないでいるだけで、論理構造のマッピングなどは議論することはできないため、この対応関係を強く主張することはできない。

4.2 今後の発展

ここに報告したものは、ネットワーク的なコミュニケーションを通してエージェント内に構成される、あるいはエージェント間で共有される意味的構造の発展を研究するための基礎的な研究報告である。今後、多数のエージェント間のコミュニケーションへと発展させる予定であるが、その他にも以下のような問題を考えなくてはならない。

4.2.1 類似度の収束性

多くのクラスターは flat cluser へと近付きやすい。これは、われわれのモデルでの単語数やシンボルの数が現実の言語システムに比べて非常に少なすぎるということも一因であるが、我々の定義では単語間および文間の類似度は非減少関数になっており、計算のイテレーションを繰り返すと、 1.0 に収束することが大きな要因である。この収束性を避けるために、定義の修正、あるいは計算に対してなんらかの制限を考えなくてはならない。

4.2.2 言語外の要因

我々の研究で示した、クラスター化のダイナミクスは、カテゴリー化のダイナミクスの研究に関してなんらかの手がかりとなる可能性はある。しかし、この問題、特にプロトタイプカテゴリーの発展をより深く探求するためには、他のエージェントとのあるいは埋め込まれた環境との相互作用のみならず、言語使用能力と他の認知的身体的能力との相互作用を考慮しなくてはならない。

我々のモデルは、プロトタイプ・カテゴリーの様な認知言語学的な概念に関して語るには、構造主義的である。すなわち、エージェントの外部世界は存在せず、類似度やカテゴリー化を言語内の関係のみに立脚して議論している。認知言語学者はプロトタイプ・カテゴリーの形成には言語外的な要因が重要であると説く。例えば、Taylor は [8] において "Prototype effects ··· arise from an interaction of core meaning with non-linguistic factors like perception and world knowledge, and can thus be assigned to other components of the mind " と述べている。この様な言語外システムを我々のモデルにいかにして組み入れていくかは非常に大きな問題である。

参考文献

[1] Kaneko, K. and Tsuda, I., (1994), Constructive complexity and artificial reality: an introduction, *Physica*, **D75**, 1–10

[2] Hashimoto, T. and Ikegami, T., (1995), Evolution of Symbolic Grammar Systems, *Advances in Artificial Life*, F. Morán et al. (eds.), Springer, Belrin, 812–823

[3] Hashimoto, T. and Ikegami, T., (1996), Emergence of net-grammar in communicating agents, *BioSystems*, **38**,1–14

[4] Putnum, H., (1975), The meaning of 'meaning', in *Mind, language and reality*, Cambridge University Press, New York, NY

[5] Karov, Y. and Edelman, S., (1996), Similarity-based word sense disambiguation, Technical Report of Weizmann Institute, CS-TR 96-06

[6] Hashimoto, T., (1997), Usage Based Structuralization in a Space of Word Relationships, accepted for European Conference on Artificial Life 1997.
available from `http://www.bip.riken.go.jp/irl/takashi/papers/`.

[7] Lakoff, G., (1987), *Women, Fire, and Dangerous Things*, Chicago , The University of Chicago Press

[8] Taylor, J. R., (1995), *Linguistic Categorization – Prototypes in Linguistic Theory*, Oxford, Oxford University Press

[9] Lakoff, G and Johnson, M., (1980), *Metaphors We Live By*, Chicago, The University of Chicago Press