

柔軟な意味解析のための概念空間の定量化

石崎 俊[†]

田中茂範[‡]

今井むつみ[§]

慶應義塾大学環境情報学部

自然な発話状況における対話では、従来の文法に基づいて判断するとおかしいが、意味は自然でよく分かる文が多く見られる。そのような対話文では意味を主体にした解析手法が望ましい。しかし、従来の意味解析手法や辞書の機能では「固い」ため、そのような意味解析を実行するには不十分なことが多い。そこで、本研究では、従来よりも柔軟な意味解析機構をもつモデルの構築を目標とし、それに適した意味解析法と、概念辞書における距離空間の導入について基礎的な検討を行う。概念辞書において概念間の距離を、従来のように階層構造をたどるのではなく、多次元尺度法を用いて定量的に距離を測ることによって距離空間を構成することを検討する。次に、状態を表す概念を指標としたSD法を用いて意味空間における距離を計算し、動的な意味の変化について基礎的な検討を行う。

Introducing Metric Space In Concept Dictionary for Flexible Semantic Analysis

ISHIZAKI Shun, TANAKA Shigenori and IMAI Mutsumi

Faculty of Environmental Information

Keio University

In spontaneous conversation, we have many ungrammatical but semantically natural sentences. Analysis based on semantics or context seems appropriate for such conversational texts. The traditional methods for semantic analysis or structures of concept dictionaries are not enough to treat with such texts. This paper discusses, therefore, a computational model which has a flexible mechanism for semantic analysis and discusses functions and a structure which concept dictionaries should have. For example, using cognitive psychological experiments, MDS method, a metric space among concepts can be organized. Then, SD method using concepts of state as parameters introduces a metric space where dynamic semantic analysis can be formalized. This paper discusses possibilities of these new approaches.

[†]ishizaki@sfc.keio.ac.jp

[‡]stanaka@sfc.keio.ac.jp

[§]imai@sfc.keio.ac.jp

1 はじめに

自然な発話状況における対話を観察すると、従来の文法に基づいて判断すると、いわゆる非文だが、意味は自然でよく分かる文が多く見られる。そのような対話文では構文解析を第一に考えるよりも意味を主体にした解析手法が望ましいと考えられる。従って、対話モデルにおける処理単位としては、文などの構文情報を中心としたものよりも、意味を主体としたまとまり(チャンク)を中心として情報を交換するというモデルを考える[田中 1994]。

しかし、従来の意味解析手法や辞書の機能では「固すぎて」、そのような対話文において意味を主体として解析を進めるためには不十分なことが多い。そこで、このようなモデルに適した意味解析法や概念辞書が持つべき構造と機能に関して基礎的な検討を行う。

問題は、従来の意味解析などで、意味的な距離や概念間の距離が十分に定義されておらず、アドホックな方法で処理を済ませていたことである。そこで、本論文では概念辞書における概念間の距離を従来のように階層構造をたどるだけでなく、多次元尺度法を用いて定量的に距離を測ることによって距離空間を構成することを検討する。次に、状態を表す概念を指標としたSD法を用いて意味空間における距離を計算し、比喩文を例として意味空間における意味の動的な変化の考察、情緒的情報の導入などに基づいて、自然言語の意味を柔軟に解析するための基礎について述べる[石崎 1994]。

2 概念体系と距離空間

2.1 概念辞書

自然言語の解析・理解を進めるための情報のうち、日本語や英語に依存した情報はそれぞれの単語辞書に記述する。すなわち、文法情報や言語の表層の格情報、単語と概念を結び付ける情報は日本語や英語の単語辞書に記述する[石崎 1989]。

一方、特定の言語への依存性の少ない一般的な意味情報は概念辞書にまとめて記述すると仮定する。概念辞書には、意味解析を行うための基礎となる意味情報が記述されており、一つ一つの概念には曖昧性がなく、一通りの意味を持つと仮定する。概念は

上位概念、下位概念などによって体系化されている[Ishizaki 1990]。

一般に、上位概念を一つだけに制約すれば概念体系は木構造になるが、概念はいろいろな視点によってながめられ解釈されるので[石崎 1987]、多くの概念を一つの木構造で表しきることはむずかしい。そこで、複数の上位概念を許したネットワーク構造をとる方が便利である。

ところで、そのような構造の概念体系において、概念の間の距離はどのように定義できるだろうか。概念から概念までの枝(リンク)の数、もう少し詳しくして枝に距離を定量的に与えてその和をとるなどの手法が考えられる。しかし、概念体系の粒度の一樣性を確保することが難しいことや、概念の上位、下位関係の決定が恣意的で不安定なことなどが原因で、天下り的に概念空間に距離を与えることは困難である。

2.2 MDS法

そこで、心理学で従来から行われている多次元尺度法(Multidimensional Scaling, MDS法[Kruskal 1989])を用いて概念構造に距離の尺度を導入することを考える。MDS法は対象間の類似度(similarity)を被験者に判定させ、そのデータにモデルとして単調関数を仮定して(ノンメトリック)反復計算によって当てはめを行う(SYSTATを使用)。モデルと観測データの適合度を表す指標としてストレス(stress)が用いられる。ストレスが小さいほど誤差が少なく表現されていることになる。ストレスの減少の様子からMDS法の次元数を選ぶこともできるが、データ構造を視覚的にとらえるためには2ないし3次元程度が望ましい。

データとして、コンピュータ用語における比喩表現を多く収拾している「テクノバブル」を選び、次のような比喩表現を例文として用いることとした。

- コンピュータ技術はワインのように成熟する時間が無い。
- 地位の説明が暗号のように難解だ。
- 言葉の意味がゴム粘土みたいに姿を変える。
- ここはコンクリートと交通渋滞の海だ。
- ニューズレターはテクノバブルの金脈である。

- 少数の人がコンピュータを奴隷のままにする。など…

コンピュータに関するいろいろな表現には、今までにない新しい物事を説明するために比喻表現が多く使われるが、たとえる語 (vehicle) としては、われわれ人間の身近な表現が多い。

- 人間に関するもの 開拓者 上司 高僧 音痴 胎児 など…
- 自然に関するもの 海 泥沼 金脈 など…

また、たとえられる語 (topic) としては、次のようなものが多い。

- コンピュータ関連 テクノロジー コンピュータ技術 2進数 コンピュータリテラシー ツール など…
- 情報・語関連 語意 文章 情報 ニュースレター など…

このような概念の中から topic と vehicle を合計 66 個選び、学生 8 人によるカテゴリ分類の予備実験を行った。

図 1 と図 2 に MDS 法で得られたそれぞれ 1-2 軸、2-3 軸の概念の配置を示す。ストレスは 2 次元で 1.0245、3 次元で 0.0758、4 次元で 0.0529 であったので、本実験では 3 次元空間を選択した。図 2 では、大分類として、人間、人工物、情報および自然という 3 つのカテゴリに分類できそうである。図 1 では、情報および自然のうちから情報が分離されていることがわかる。また、それらのカテゴリは、ほぼきれいに分かれている場合もあれば境界が曖昧な場合もあり、これは生のデータを扱う場合に常に生ずる現象である。

被験者にカテゴリ分類させて得た上記のデータにクラスター分析を行えば、よく知られているように木構造のクラスターが得られる。たとえば最大距離法でクラスターを構成してみると図 3 が得られた。この場合はたまたま、上記 4 分類が大きなカテゴリとして分類されている。クラスター分析はいくつかある手法によって異なる結果がでる。また、測定値のわずかな差によっても結果が異なる。例えば、図 1 と図 2 でカテゴリの境界にある多くのデータはクラスター化の時に大きく変動する可能性が高い。さらに、クラスター解析の結果は対象を 1 列に並べて表示するが、この順序に意味はなく、となりあった対象が常に類似度が大きいわけではない。

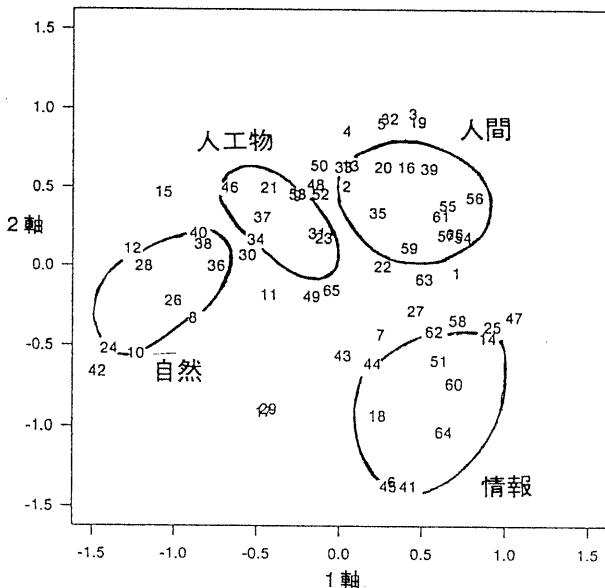


図 1: 多次元尺度法 (1-2 軸)

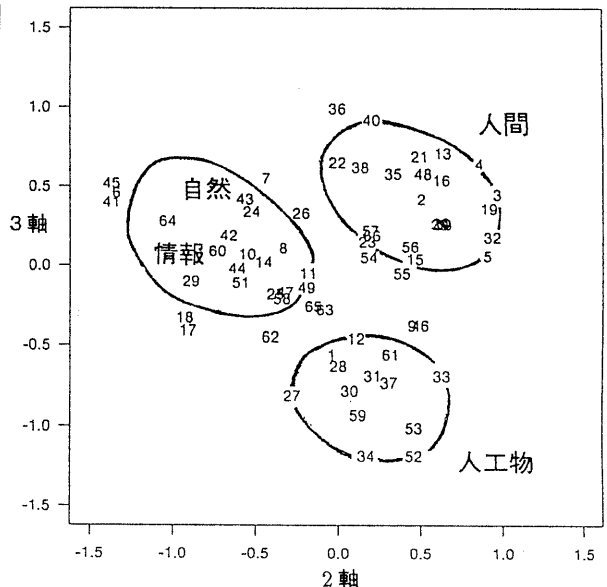


図 2: 多次元尺度法 (2-3 軸)

表 1: MDS 法による概念間の距離の例 (図 1, 図 2 に対応)

1. コンピュータ			55. コンピュータ・ユーザ		25. 情報		10. 海	
順位	距離	概念	距離	概念	距離	概念	距離	概念
1	0.341	コンピュータ社会	0.239	コンピュータの天才	0.144	テクノロジー	0.282	湿地
2	0.371	初期のコンピュータ	0.241	ハッカー	0.214	2進数	0.315	泥沼
3	0.407	コンピュータ技術	0.303	パソコン・ユーザ	0.216	コンピュータ・リテラシー	0.415	金脈
4	0.422	ツール	0.330	コンピュータの素人	0.389	テクノバブル	0.479	宇宙
5	0.461	コンピュータ・リテラシー	0.401	専門家	0.443	コンピュータ技術	0.847	流路
	
64	2.165	エイリアン	2.114	海	2.252	流路	2.277	コンピュータの天才
65	2.335	泥沼	2.277	泥沼	2.382	泥沼	2.297	発明者
66	2.378	湿地	2.378	湿地	2.435	湿地	2.314	テクノロジー

このクラスター分析と MDS 法による表現とは一つの対応関係があるが、MDS 法では、すべての概念間の距離が定量的に計算できる点、および次元数を選べる点に特長がある。また、MDS 法の 3 次元表示から概念の階層構造を抽出するとき、いろいろな視点から階層を構造化することができると思われる。これはちょうど、概念階層に視点を導入して分類の基準を表し、従来の概念構造よりも精密な情報を記述することに対応する [石崎 1987]。このような観点から、MDS 法による 3 次元空間はわれわれ人間が概念の階層構造をつくるときの基礎になる空間であり、その概念空間から出発して、与えられた視点に基づいた一つの概念階層が再構成できるものと考えられる。

表 1 には、ある概念と距離がもっとも近いいくつかの概念、もっとも遠いいくつかの概念を 4 つのカテゴリーから選んで示した。定量的に概念の距離が得られるようすがわかる。

3 意味空間と動的特徴

意味解析を行うとき、ある語に対する意味が、その語が述べられた状況によって制約され、正確に解析するには背景知識、領域的知識、常識などといわれるものが必要である。また、照応現象、情報の新旧などについては文脈に関する情報が特に重要である。

さらに、対話文を対象として意味を解析しようとするとき、古くは会話の作法が規範としてあり、また、上記のさまざまな情報以外にも話し手と聞き手がそれぞれ持つ知識を前提とした対話における情報の流れを考慮する必要がある。

従来のテキストの意味解析や文脈解析では特定の話題に関する知識をあらかじめ用意して語の意味範囲を限定し、ある程度深い理解を実現してきている [Ishizaki 1987]。しかし、少しでも対象を拡大したり、一般化しようとするとき、そのような知識をスキーマやスクリプトなどの枠組みで十分に用意するのも困難で、何らかのより一般的な意味解析手法が求められている。

3.1 SD 法

ここでは、心理学で従来から幅広く行われている SD 法 (Semantic Differential [岩下 1992]) を用

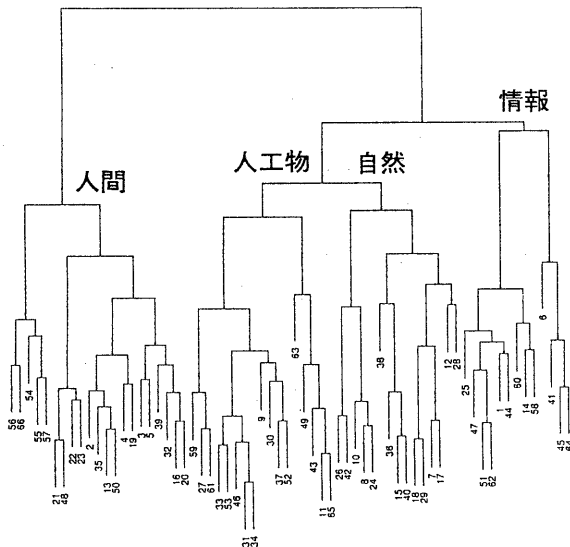


図 3: 概念のクラスター分析 (最長距離法)

いて、定量的な意味空間を自然言語処理に導入することを試みる。SD法では概念間の意味的な距離が定量的に計算できることから、概念辞書に書く情報の種類が豊富になり、解析に使える手法の柔軟性が期待できる。第2章で検討したMDS法では、概念そのものを被験者にカテゴリーへ分類させたのに比べて、SD法では、以下に述べるような状態を表す概念をパラメータとして意味的な距離を計量する。

具体的には、連想実験を行って、第2節で述べた概念に対して被験者にいろいろな状態を表す両極型の概念対を連想させる。たとえば、

大きい ⇔ 小さい 広い ⇔ 狭い
 速い ⇔ 遅い 長い ⇔ 短い
 多い ⇔ 少ない 難しい ⇔ 易しい
 精神的 ⇔ 物質的 基礎的 ⇔ 応用的
 連続的 ⇔ 不連続的

などであるが、この中には必ずしも形容詞に限らず、状態を表すさまざまな言語表現のペアの集合になっている。また、上記の両極型概念対の4段目以下は実験者が特に選んでつけ加えたものである。このようにして実験では、35個の両極型状態概念対を選び使用した。

第2章で使用した概念について、このような35個の両極型状態概念対をそれぞれ-3から3までの7段階に分けて、その概念の持つ特徴がどの段階にあるかを被験者に判断させる。第1回目の実験では、各概念を1つ1つ独立に被験者に示して評定させる。第2回目は1週間程度の間をおいて、比喻文の中における概念として評定させる。このようなデータを集めて因子分析することによって、意味空間内での概念の分布を計算する。本実験では、因子負荷量は主因子分析法によって反復計算した。

第1回目のデータからは、概念が独立なものとして意味空間中の配置が得られる。一方、第2回目の実験からは比喻文の中での文脈の影響を受けた位置が意味空間内で得られる。実験に用いた比喻文には以下のようなものがある。

コンピュータは奴隷だ。
 コンピュータは上司だ。
 コンピュータは生物だ。
 コンピュータはタワーだ。
 コンピュータ技術はワインだ。
 テクノバブルは侵入者だ。
 テクノロジーは胎児だ。
 知覚は扉だ。
 など…

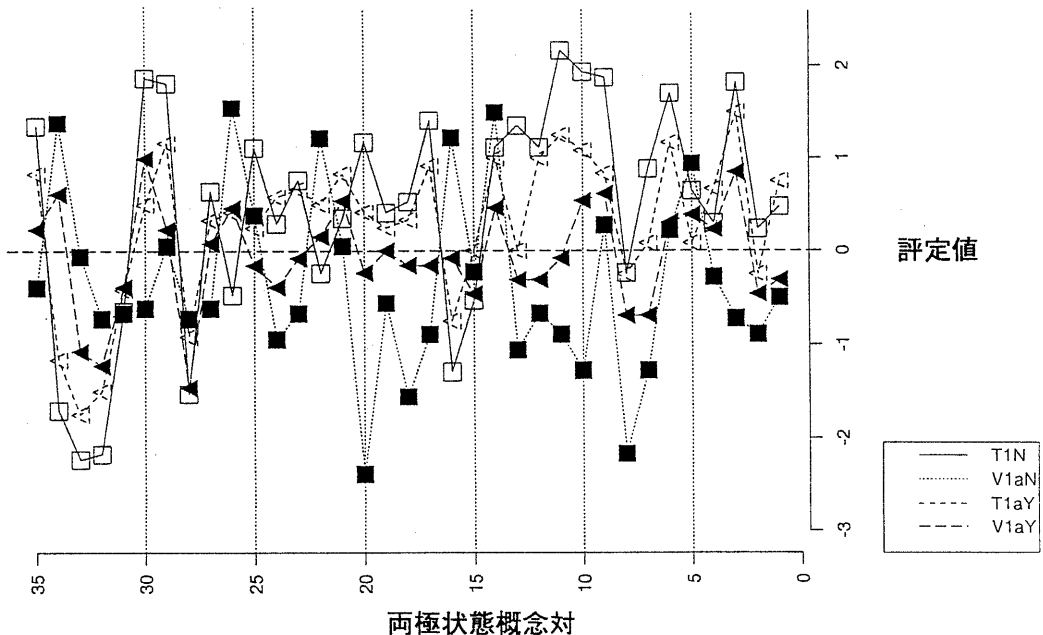


図4: SD法における諸特徴の変化

図4には「コンピュータ」と「奴隷」をそれぞれ独立に提示したとき、「コンピュータは奴隷だ」という文脈中で「コンピュータ」と「奴隷」を示したときの4通りの反応を状態概念対ごとに示している。8番目の状態概念対は「暗い-明るい」、10番目は「古い-新しい」、18番目は「汚い-きれい」、20番目は「貧しい-豊かな」などである。この図からもわかるように、概念によって変化するやすい両極型状態概念と変化するにくいものがあり、さらにその程度は概念の組み合わせにも依存する可能性がある。

この図で表されるデータは約60名の被験者を4つのグループに分けて実施し、被験者に関して平均化されたものである。その平均されたデータに因子分析を行い3次元で表示し、現在はデータを整理している段階である。比喩文という文脈でtopicが顕著に動く例、vehicleも同時に動く例、vehicleはほとんど動かない例、両方とも顕著には動かない例などが観察されている。これらの比喩文の解析、モデル化については[kusumi 1987]、[土井 1989]、[岩山 1992]で論じられている。

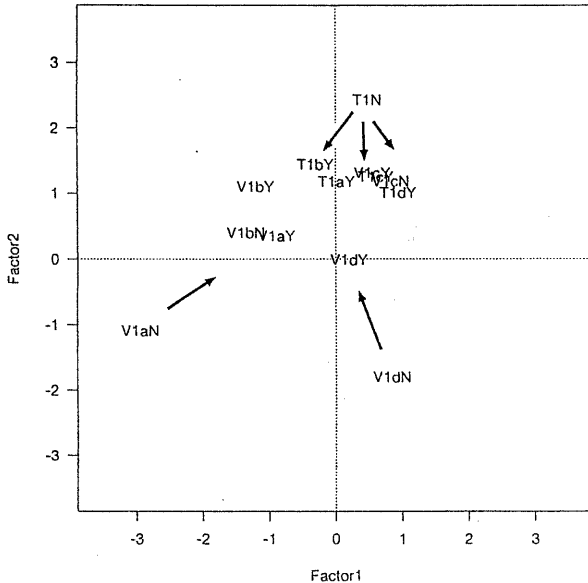


図5: SD法による概念の移動(第1-第2因子)

図5には、topicをコンピュータ(T1N)として被験者に独立(比喩文中でなく)に提示し、vehicleを奴隷(V1aN)、上司(V1bN)、タワー(V1cN)、生物(V1dN)という4種類として独立に提示したときのデータを第1因子と第2因子の空間で表示した。同図にはそれらが文脈の中に入ったときの位置を、topicに関してT1aY, T1bY, T1cY, T1dY、vehicleに関してV1aY, V1bY, V1cY, V1dYとして表示している。この図で、独立に提示したtopicであるコンピュータT1Nが、独立に提示したvehicleの奴隷V1aN、上司V1bN、生物V1dNの方向へ顕著に動いている。さらに、これらのvehicleもT1Nの方向へ移動している。しかし、タワーについてはtopicのみの動きがみられる。

3.2 概念記述への適用の可能性

我々が比喩文を聞いたり読んだりしたとき、含まれる語の意味を字義通りに解釈するか、比喩として解釈するかは何らかの意味的な制約の充足問題に帰着すると思われる。それは一種の選択制限であろう。その処理法としては、概念記述における制約として、字義通りの解釈の場合の条件を記述することが従来から考えられている。

一方、この選択制限に違反したものは比喩表現が無意味な表現に分類されると思われるが、実は、これら二つの表現に境界線を引くことは大変に難しい問題である。従って、この問題は今後の検討課題にしたい。現在の課題として比喩表現における意味の移動について、意味空間における定量的な評価と概念辞書における記述の有効性について検討したい。

たとえば、図4にあるように、概念ごとに顕著な特徴を示す両極概念対を見いだすことができる。それらの特徴を概念辞書中に記述し、比喩の理解に役立てることを考えている。比喩の理解は、類推の問題と関連が深く、概念の構造に関するマッピングをとらえることができる[Gentner 1983]。概念辞書における概念の構造記述に、そのような顕著な特徴を示す両極概念対を取り入れることによって、概念の意味の動的なモデルを作ることを検討している。

比喩における意味の種類には、カテゴリー的意味、情緒・感覚的意味、スクリプトの意味があるが[Kusumi 1987]、従来のSD法では、情緒・感覚的な意味に関して特に有効であるといわれている[岩下

1992]。これは、SD法の結果において、3つの軸を Evaluation, Activity, Potency として解釈するとき重要であることを意味する。しかし、そのほかの尺度で因子分析することを妨げるものではない。とくに、構造的なマッピングを考慮するときは、そのような制約よりも深い次元で考察しモデル化する必要がある。

4 まとめと課題

MDS法の使用では、3次元空間における概念の配置から任意の概念間の距離が計算できることを示した。このような概念空間に何らかの視点を与えて概念の階層構造を作れば、それは、概念の再構成過程としてとらえることができると考えられる。

一方、現在のMDS法の計算プロセスでは、分析したい概念をすべて実験の対象として扱う必要がある。概念数が増加すると1回の実験における被験者の負担は増大し、得られるデータの精度は必ずしもいいとはいえない。特定の領域や高度に専門的な領域に対象を限定するならば、この問題はある程度回避できそうであるが、一般性を持たせたり、常識的な概念などを対象とすると、大きな問題になる可能性がある。そこで現在は、複数の実験結果を統合するための方法について検討している。また、従来の人間が手作りした概念階層と、MDS法による概念空間を組み合わせることで実用的な概念空間を作ることについて検討している。

従来のSD法では、考慮している両極概念対をすべて実験の対象としなければならない。しかし、概念によっては顕著な特徴を持つ概念対が限定されるため、必ずしもすべての概念対を実験する必要はない。また、その概念を含む領域によって概念対の集合が限定できる可能性もある。このように、あらゆる両極概念対の母集団があるとしても、比喩の理解に必要なものは限定されているため、そのような母集団から関連する両極概念対を抽出する方法が重要になってくると考えられる。

一方、情緒・感覚的な特徴を持つ概念を扱うこのような手法は、自然言語における情緒的な情報の扱いに対する一つの方法を与えるので、情緒・感覚的な意味の導入と理解システムへつながると思われる。

本研究で述べるいくつかの実験については現在、多くのデータを整理したり、実施中であるため、本稿では、予備的な実験に関するデータや基本的な考え方について主に述べている。近いうちにまとまりの実験として報告する予定である。

謝辞

本論文をまとめるに当たって今井豊君、豊国潤君、中村隆英君、内山清子さんにデータの処理、図表示、 \LaTeX などでお世話になった。ここに感謝の意を表したい。また、この論文で行われた認知心理実験では玉村真樹子さん、野津恵さんをはじめ石崎研究室に所属する4年生、3年生の学生諸君に協力していただいたので特に感謝したい。

なお、本研究は文部省科学研究費(重点領域00245614)の援助を受けて行われたものである。また、本研究の一部は(財)人工知能研究振興財団の助成を受けて実施した。

参考文献

Ishizaki, S. et al., "Concept Dictionary as Interlingua," Bulletin of Electrotechnical Laboratory, Vol.54 No.4(1990)

Ishizaki, S. et al., "Natural Language Processing System with Deductive Learning Mechanism," in Nagao, M. ed., "Language and AI," Elsevier Science Publishers(1987)

石崎俊, 「言語理解の問題 — 文脈理解をめざした概念の体系と記述 —」 情報処理学会九州シンポジウム(1987年11月)

石崎俊, 他「文脈情報翻訳システム CON-TRAST」情報処理 30 巻 10 号 pp.1240-1249(1989)

石崎俊, 田中茂範, 「対話文理解のための解析手法と認知意味論モデルの研究」, in 音声・言語・概念の統合的処理による対話の理解と生成に関する研究, 重点領域研究成果報告書(1994年1月)

田中茂範, 石崎俊, 「日常言語における意味の生成: You know と I mean の役割」情報処理学会自然言語処理研究会資料 (1994年3月)

Kruskal, J. and Wish, M., "Multidimensional Scaling," Sage Publications(1989)

海保博之編著, 「心理・教育データの解析法 10講」福村出版 (1986)

岩下豊彦, 「SD法によるイメージの測定」川島書店 (1992)

Gentner, D., "Structure-mapping: A theoretical framework for analogy," Cognitive Science, 7, 155-170(1983)

Kusumi, T., "Effects of categorial Dissimilarity and Affective Similarity between Constituent Words on Metaphor Appreciation," J. of Psycholinguistic Research, Vol.16 No.6(1987)

土井晃一他, 「スベルベルの象徴解釈モデルに基づく隠喩の検出」情報処理学会論文誌 30 巻 10 号 (1989)

岩山真, 「計算機による比喩理解モデルに関する研究」東工大学位論文 (1992)

バリー著, 岩谷訳, 「テクノバブル」工学社 (1993)