自 然 言 語 処 理　130－11
（1999．3．5）

# 概念辞書の構築と概念空間の定量化

## ～連想実験による概念空間の抽出～

岡本潤†，石崎俊§

§慶應義塾大学　環境情報学部
†慶應義塾大学　政策・メディア研究科

小学生の学習基本語彙を刺激語として連想実験を行い、得られた連想語といくつかのパラメータから、線形計画法を用い刺激語と連想語の距離を定量化することによって概念辞書を構築する。次に概念辞書から身近で分かりやすい語を選択して、その上位概念／下位概念を二次元空間に配置し、概念間の繋がりを調べる。配置した概念の中には、上位下位関係で双方向にリンクのある概念対がいくつか見受けられたので、これについて考察を行う。また、概念辞書には「属性」「動作」「環境」という項目も記述されており、これらの情報から「食物」「動植物」の二つの文脈をとることができる語について考察する。

# Construction of electronic concept dictionary and quantification of concept space

## --- Extraction of concept space by association experiment ---

Jun OKAMOTO†, Shun ISHIZAKI§

§Faculty of Environment Information, Keio University
†Graduate School of Media and Governance, Keio University

In this paper, distances between concepts were calculated using a linear programming method, which combines two parameters, frequency and sequential order of the associated concept. Using the quantified distances a concept dictionary was built by organizing the stimulus concepts and associated concepts with a hierarchical structure. All of the associated concepts were connected to the stimulus concepts with the distances.

Next, the following three themes was discussed. Firstly, arranging concepts, those are basic nouns in elementary school textbooks, in the second dimension. Secondly, connecting to concepts each other with the distance. Finally, stimulus concepts, witch have some contexts - "food", "animal" and "plants" -, using attribute, action and situation describing in this electronic concept dictionary.

Effective applications of the concept dictionary to natural language processing were discussed.

## 1. はじめに

人間が書いたり、話したりする自然言語をコンピュータで処理するためには、言語学的情報に基づいて構文解析や表層的意味解析を行なうだけでは十分ではない。われわれが言語理解に用いている一般的な知識、当該分野の背景的知識などの必要な知識(記憶)を整理し、自然言語処理技術として利用可能な形にモデル化することが重要になってくる[4]。

また人間は文章を読む時や、会話を行う時、ある単語をもとにさまざまなイメージや概念を連想している。連想によって関連付けられた概念のネットワークをもとに目、耳から入った情報と結び付けて理解していく。そこで、一般性のある自然言語理解のためには、現実の世界で成り立つ知識を構造化した知識ベースが必要であり、そのためには人間がどのように言葉を理解しているかを調べる必要があると考えている。

本研究では連想実験を行い現実の世界で人間が利用している知識を概念辞書として構造化し、概念空間を定量化する。従来の電子化辞書などでは、距離が定量化されておらず、木構造の粒度に依存したアドホックなものであったため本研究の意義は大きいものと思われる。

## 2. 研究の概要

連想実験から得られたさまざまなパラメータを用い、線形計画法によって刺激語と連想語の距離を定量化した。構築した概念辞書からいくつか語を選択し、その語の上位／下位概念を二次元空間に配置することで概念間の繋がりを調べた。また、文脈によって観点が変わる語について考察する。

## 3. 連想実験システム環境の概要

連想実験システム環境は、「連想実験システム」と「データ集計システム」の２つから構成されている。このシステムを用いることで、被験者に対して連想実験を行い実験データを収集し、蓄積されたデータをデータ集計システムの個々のプログラムによって効率よく修正作業を行い概念辞書を構築していく[4]。

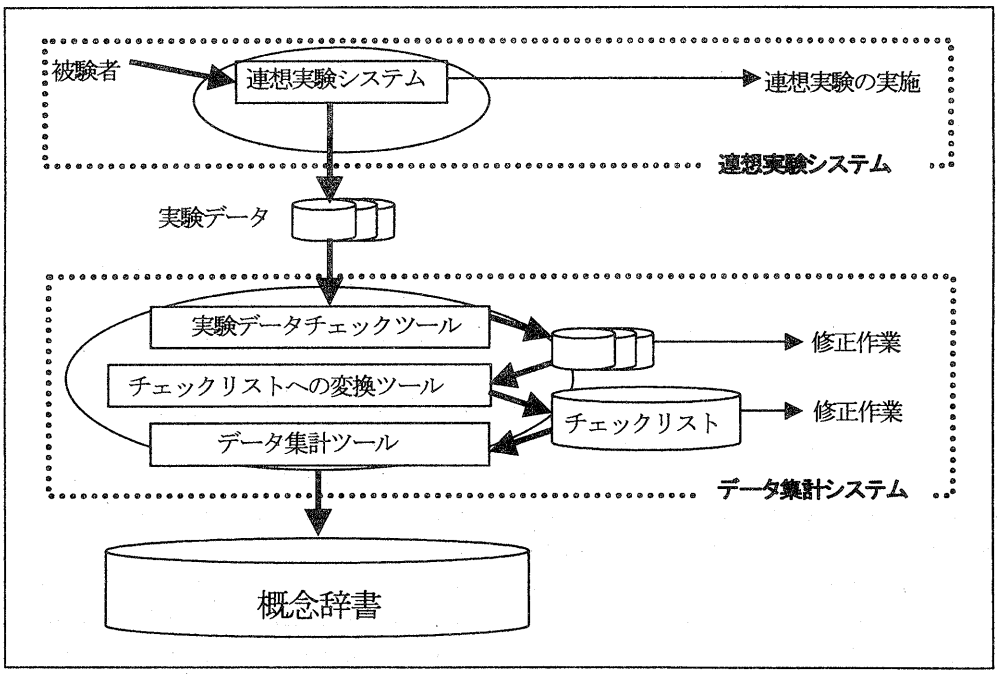


図１　概念辞書構築の流れ

### 3. 1 連想実験の実施

人間が一つの概念から関連する概念を連想する仕組みを調べるために連想実験を行った。これにより人間の語彙記憶に対応した概念情報を抽出し、概念辞書に反映させることを目指している。

概念辞書における概念体系を明らかにするためには上位・下位概念という情報が必要になってくる。本研究では、上位・下位概念の情報の他に、概念の持つ部分材料に関する情報や、その概念の特徴をあらわす概念（属性）、概念がどのような動作を伴って普段の日常生活で用いられているか、どのような環境（状況）において用いられるかという情報を概念辞書に記述するため７つの課題を設定した。

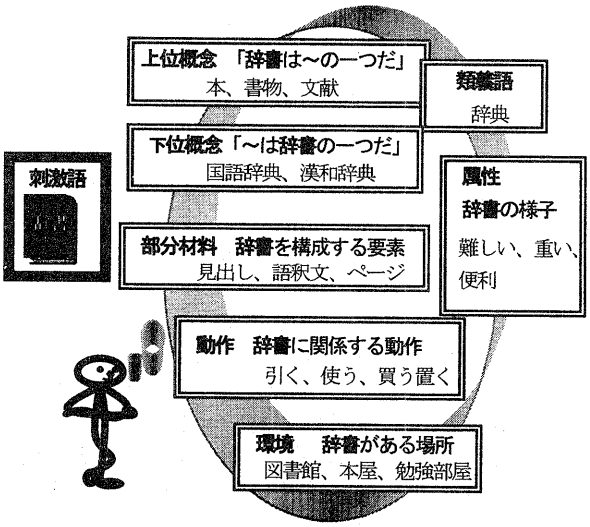


図２　連想実験

表１　連想実験の手続き

| | |
|---|---|
| 刺激語 | 光村図書「語彙指導の方法」の名詞、約 1400 語から同義語を省く 800 語＋任意選択 200 語 |
| 課題 | 上位概念・下位概念・部分材料・属性・類義語・動作・環境 |
| 被験者 | SFC の学生 3, 4 年生及び大学院生 |
| 被験者数 | 各刺激語に対して 10 人ずつで行う。（各被験者には 5 語ずつランダムに提示） |

言語理解のためには、現実世界で成り立つ知識を構造化した知識ベースが必要であり、そのためには人間がどのように言葉を理解しているかを調べる必要がある。この意味で、連想によって得られた情報に基づいて概念体系を構築することによって、人間の語彙記憶のメカニズムを取り入れた概念辞書へと発展させることができれば情報検索、文書要約、観点の違いを取り入れた概念の類似度計算など、数多くの応用が見込めると考えている。

## 4. 連想実験の集計結果

図３は連想実験の結果を集計して各課題ごとの連想語数と異なり語数をグラフ化したものである。

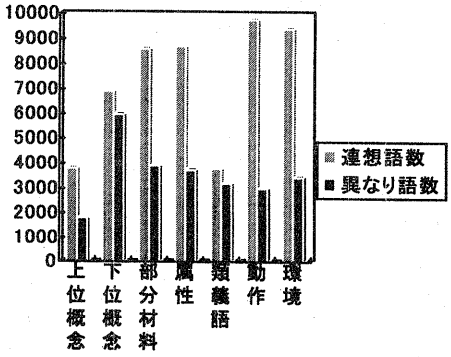


図３　課題語との連想語数と異なり語数

連想語数とは連想された語の合計数のことである。異なり語数とは刺激語が違っていても同じ語が連想される場合、同じ単語として一つと数えるものである。例えば、刺激語「小説」「作文」の動作として「書く」が連想された場合、連想語数は「２」と数え、異なり語数では「１」と数える。

上位概念、部分材料、属性、動作、環境は異なり語数の数は大幅に減少している。これによって、同じ語がいくつもの刺激語から連想されていることがわかる。

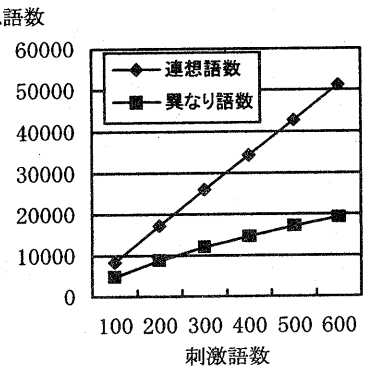

図4　連想語数と異なり語数の推移

図4では、連想語数は刺激語数が増えていくにつれて一次関数的に増加していく。しかし、連想語異なり語数に関してはある上限の値に近づくかのように、曲線を描いて増加していく。

これは、刺激語から連想される「上位概念」「部分材料」「属性」「動詞」「環境」はある程度決まった概念を連想していることによるものである。本研究では約600語の刺激語について集計を行なったが、今後、連想実験を行なっていくにつれて、ユニークな連想語総数（連想語異なり語数）のおおよその上限が明らかになっていくのではないかと考えられる。それによって、われわれ人間が普段使っている大体の語彙数が予測できるのではないだろうか。今後膨大な量の知識データを扱う場合、基本語彙での連想語彙数の収束の値や、連想しやすい語、いいかえれば日常生活で普段使っている語彙はどのようなものがあるのかというデータは有用なものとなり得るだろう。

## 5.　線形計画法による距離の計算式の決定

本研究では概念間の距離（D）の式を次の一次式とする。

$$D=\alpha\times T+\beta\times S+\gamma\times\frac{1}{F}\cdots(1)$$

$$T=\frac{1}{n}\sum_{i=1}^{n}t_i\times\frac{1}{60}$$

$$S=\frac{1}{n}\sum_{i=1}^{n}s_i$$

$$F=\frac{n}{N}$$

$t_i$ ＝　被験者が連想に要した時間

$s_i$ ＝　被験者が連想した語彙の順位

$n$ ＝　被験者の合計数

$N$ ＝　連想人数

距離の式の係数α、β、γの値を求めるために、

$\alpha\geqq 0$、$\beta\geqq 0$、$\gamma\geqq 0$

とし、線形計画法を用いて下記の形の問題を解いていき、係数α、β、γを決める[5]。

最小化　$Z=c_1\times\alpha+c_2\times\beta+c_3\times\gamma$

条件　$a_{11}\times\alpha+a_{12}\times\beta+a_{13}\times\gamma=D_1$

$a_{21}\times\alpha+a_{22}\times\beta+a_{23}\times\gamma=D_2$

$\alpha\geq 0,\beta\geq 0,\gamma\geq 0$

まず、目的関数のC1、C2、C3は、C1≧C2≧C3とする。これは連想頻度、連想順位、連想時間の順で信頼性が高いからである。これによってα、β、γの係数は$\alpha\leqq\beta\leqq\gamma$が期待される。

次に、条件として以下の場合を考える。

刺激語と連想語の距離がD1になる場合を、「連想時間が短く」「一番初めに連想された語」「被験者全員が連想」した時と仮定する。距離がD2になる場合を「連想時間がある程度長く」「連想順位が大きい」「全被験者のうち一人だけが連想した語」の時と仮定する。

以上の条件下でパラメータを変化させながらシンプレックス法を用いてα、β、γの最適解を求める。得られた最適解と、実験結果からT（連想時間），S（連想順位），F（連想頻度）を用いて、身近で分かりやすいと思われる刺激語と連想語の距離を計算し、空間的に配置してみてもっとも妥当な最適解を採用する。

以上より、

目的関数の係数　$(c_1,c_2,c_3)=(10,5,1)$

条件式の係数と値　$(a_{11},a_{12},c_{13},D_1)=(0.1,1.0,1.0,1.1)$

$(a_{21},a_{22},a_{23},D_2)=(1.0,4.0,10.0,9.0)$

の時、$\alpha=0$、$\beta=0.33$、$\gamma=0.77$となり、概念間の距離は、

$$D(\text{概念間の距離}) = 0.33 \times S + 0.77 \times \frac{1}{F}$$

となった。

オンライン実験システムでは実験者が被験者を間近で観察していないため、正確な時間を測定することは難しい。また、連想時間にはキーボードの入力時間も含まれているため、被験者のキーボード操作の熟達度がT（連想時間）に著しく影響し、あまり信頼できる値とはいえない。そのためα=0となるのは、妥当であると考えられる。

## 6. 概念辞書の構築

図1のデータ集計システムによって作成された刺激語毎の集計データとused-inパラメータを概念辞書作成ツールに入力することにより概念辞書が作成される。固有名詞は概念辞書とは別に「固有名詞辞書」が構築される。

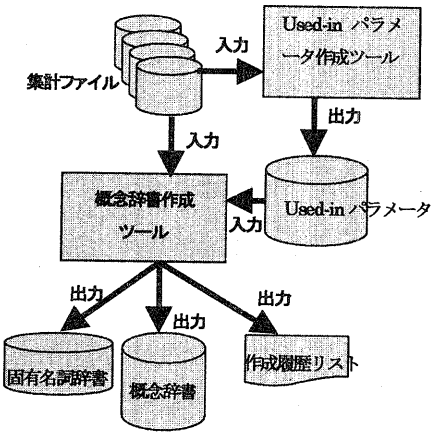


図5　概念辞書の構築

used-inパラメータとは刺激語Aが他の刺激語Bの連想語として現れた場合、概念Aの概念記述の中に刺激語Bと課題を記述するものである。これによりAからBへのリンクを張ることができる。

構築された概念辞書は以下のようになっている。

```
(いす
  (上位概念
    (家具                1.292)
    (座る物  …)
      :
    (連想語  …))
  (下位概念
    (ロッキングチェア  2.134)
    (座椅子  …)
              :
    (連想語  …))
              :
  (used-in
    (家具  下位概念)
      :
    (ロッキングチェア  上位概念)))
(入れ物
    :
  (used-in
    (かご  上位概念)
      :
    (ふくろ  上位概念)))
    :
```

図6　概念辞書の記述フォーマット

図6では「いす」の上位概念として「家具」が連想されており、「いす」と「家具」の概念間の距離は1.292である。「上位概念」の他に「下位概念」「部分材料」「属性」「類義語」「動作」「環境」などの課題語とに連想語が刺激語「いす」との距離と共に記述してある。used-inでは、「いす」は「家具」の下位概念として連想されていることを示す。

## 7. 二次元での概念の配置

辞書に記述してある単語のうち身近で分かりやすい語を選択して、実際に二次元空間で概念を配置し、その繋がりを調べる。選択した語は、連想実験で被験者に提示した刺激語の中で比較的連想しやすく連想語数が多い刺激語が主である。この刺激語を中心として、上位概念、下位概念にどのような語が連想されているのかを調べた。これらの語は私たちにとって日常馴染みのある単語であろう。また、刺激語から連想された語を刺激語として連想実験を行なっているものもあり、密な語彙ネットワークが出来上がると考えられる。

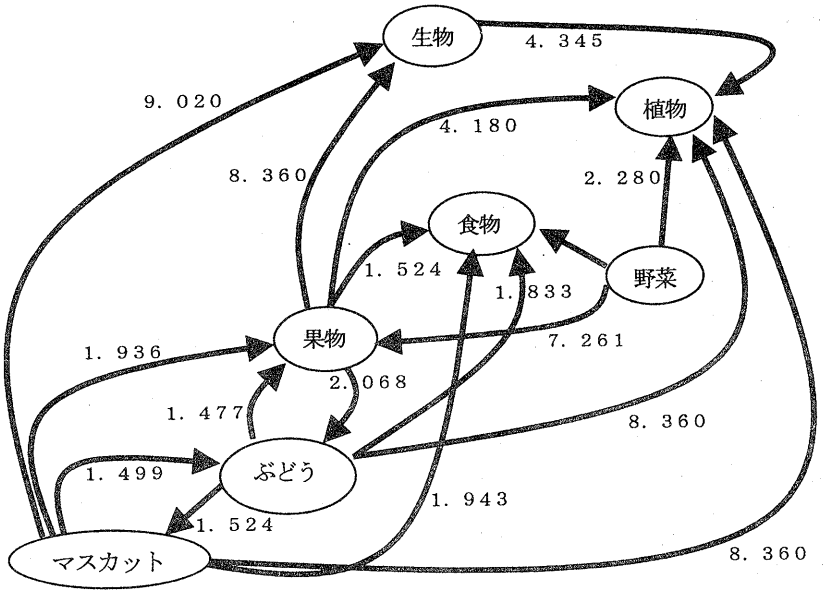


図7　果物を中心とした概念の配置

図7では、「ぶどう」の上位概念として「果物」「植物」を連想されている。概念間の距離は「ぶどう」「果物」の間が 1.477、「ぶどう」「植物」の間が 8.360 である。「ぶどう」の連想語である「果物」「植物」はどちらとも刺激語として連想実験を行っているので、何らかの連想語に繋がっており、「果物」は上位概念として「植物」を、下位概念として「ぶどう」を連想している。「ぶどう」「マスカット」と「果物」「食物」の概念間の距離は短く、「ぶどう」から「植物」「生物」までの距離は大きくなっている。

これは「ぶどう」「マスカット」という語は日常生活において食卓の上や果物屋という状況において用いられ、「食物」として取り扱う機会が多いためと思われる。このように上位概念、下位概念を連想する時、刺激語に関するエピソード的な記憶をもとに想起されると考えられる。

## 8.　双方向のリンクがある概念対

図7の「マスカット」「ぶどう」「果物」のように上位・下位概念の双方向で連想される場合がある。600 語の刺激語のうち双方向にリンクのある概念対は 161 組みあった。

上位から下位、下位から上位までの距離が共に短い場合や、下位から上位までの距離は短いが、上位から下位までの距離は長いなどの場合がある。

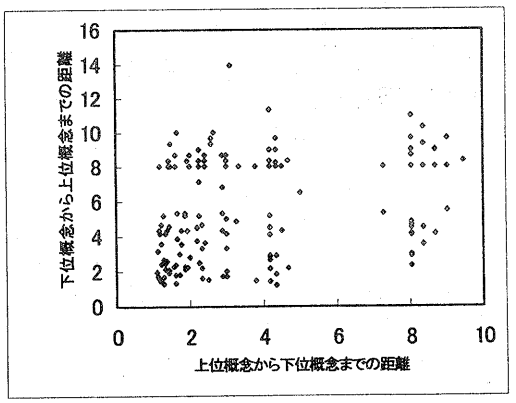


図8　双方向にリンクのある概念対の分布図

表2　双方向にリンクのある概念対での概念間の距離の平均と分散

| 下位から上位までの距離 | | 上位から下位までの距離 | |
|---|---|---|---|
| 平均 | 3.994 | 平均 | 5.651 |
| 分散 | 6.960 | 分散 | 8.494 |

上位概念を連想する時のほうが、下位概念を連想する時よりも平均の距離が短い。これは、双方向にリンクがある概念対で身近で日常的な語であっても上位概念のほうが連想しやすいことを示していると考えられる。「桜」を例にとると、「桜」上位概念は「植物」「木」などを真っ先にあげることができる上、連想できる数は限られてくる。逆に、「木」の下位概念としてすぐに「桜」でてくるとは限らない。被験者の生活環境や経験などによって「木」の下位概念は変わってくるだろう。被験者が連想する語は「松」「杉」であるかもしれない。連想の順位や頻度で概念の距離を計算しているので、連想順位の影響が大きく出ていると考えられる。

表8　双方向にリンクがある概念対の例

| 下位概念 → 上位概念 | 距離 | 上位概念 → 下位概念 | 距離 |
|---|---|---|---|
| 家 → 建物 | 1.723 | 建物 → 家 | 1.807 |
| 車 → 乗り物 | 1.292 | 乗り物 → 車 | 1.292 |
| 鉄 → 金属 | 1.613 | 金属 → 鉄 | 1.332 |
| 台所 → 部屋 | 4.180 | 部屋 → 台所 | 4.107 |
| 給食 → 食事 | 1.668 | 食事 → 給食 | 10.010 |
| ﾏｽｶｯﾄ → 果物 | 1.936 | 果物 → ﾏｽｶｯﾄ | 8.030 |
| 鏡 → 家具 | 2.667 | 家具 → 鏡 | 10.010 |
| 絵本 → 本 | 1.186 | 本 → 絵本 | 8.030 |
| 雪 → 天気 | 8.360 | 天気 → 雪 | 4.180 |
| 文字 → 記号 | 8.030 | 記号 → 文字 | 4.180 |
| 火 → 道具 | 8.030 | 道具 → 火 | 8.030 |
| 台風 → 災害 | 8.690 | 災害 → 台風 | 9.020 |

双方向のリンクの距離が共に短いものには「鉄」「金属」などのように上位概念・下位概念として共に連想しやすいものと考えられる。これらは上位／下位関係としてお互いに典型的な例と考えることができる。

一方「マスカット」と「果実」、「絵本」と「本」、「鏡」と「家具」、「給食」と「食事」などは、双方向のリンクの距離が共に短いものにくらべて上位概念から下位概念がより連想しにくいものとなっていると考えられる。たとえば、「鏡」の上位概念は「家具」であると連想した人が多く、「家具」の連想順位も高い。つまり、すぐに連想され、概念間の距離も短いのに対して、「家具」の下位概念は「鏡」であると連想した人は少なく、「たんす」「いす」などの概念の後に「鏡」が連想されており、連想した人も少ない。これによって「鏡」「家具」間の距離が長くなっている。つまり、「家具」の下位概念として「鏡」は典型的な例ではないといえる。

## 9. 概念の観点の違い

構築した概念辞書には食物の下位概念として「魚」「鳥」「マスカット」「ぶどう」「豆」「野菜」という見出し語がある。これらの語には「食物」としての観点と「動植物」としての観点などが考えられる。概念辞書で「環境」の項目は「刺激語がどのような環境（状況）で用いられるか」をたずねたものである。たとえば「魚」には「海」「魚屋」「食卓」「スーパー」「池」「寿司屋」「水」「まな板」「生け簀」「居酒屋」「レストラン」「東京湾」「台所」「太平洋」「テーブル」「自然」「沼」「舟」「市場」が連想された。この場合「海」「池」「水」「東京湾」「太平洋」「自然」「沼」は自然の下位概念にあたり、「魚」が生息している場所というとらえ方ができる。一方「魚屋」「食卓」「スーパー」「寿司屋」「まな板」「生け簀」「居酒屋」「レストラン」「台所」「テーブル」「舟」「市場」は魚が商品、食物として存在するような状況と考えられる。

「魚屋」「スーパー」「台所」など人間が関与する場所・状況には「買う」「食べる」「調理する」などの動詞が共起されやすいと予想される。また、「食物」を刺激語として連想される形容詞（属性）には「おいしい」など味覚に関する形容詞の連想語が多く、「うれしい」などの心情語も連想されている[1]。概念を観点の違いで捉えるには、上位・下位関係の他に、状況（環境）、動詞（動作）、形容詞（属性）との繋がりを調べる必要がある。これによって文脈解析など高次の自然言語処理システムが望めるのではないだろうか。

## 10. 概念辞書の応用

コンピュータの普及に伴ってネットワークが拡大し、電子化されたテキストが世の中にあふれ、

さまざまな情報をネットを通じて入手する機会が増えてきている。それと同時に本当に自分が必要な情報を効率よく検索することが難しくなってきた。これは従来のキーワードマッチングでは、キーワードの意味を理解して検索をしていないためノイズが入ること、サーチャーなどのデータベースの特徴を理解した人でないと検索に効率のよいキーワードの与え方が分からないなどが問題であると考えられる[3]。

そこで、キーワードから連想する言葉も検索語として入力したり、キーワードと一緒にそのキーワードがどのような観点で用いられているものかを入力することによってユーザーが望む検索結果に導くことができると考えられる。また、システム側で関係のありそうなキーワードをユーザーに提示し、その中から選びながら、的を絞った検索方法も考えられる。

「魚」は「生物」「食べ物」など観点の違いによって連想語にも違いが出てくる。つまり、文の構成素としての「魚」という単語に繋がりやすいと思われる動詞、「魚」の属性がある程度類推できる。「魚」について「生物」という観点ならば「泳ぐ」「食べる」等の身体的動作、「ヌルヌル」「冷たい」等の形状を示すような「属性」、「環境」としては生息する場所などが連想されうる。「食べ物」という観点では、「食べる」「買う」等の「魚」を食べ物や商品として扱う場合の動詞、「熱い」「美味しい」などの属性、「環境」では商品として扱う場所が連想されると考えられる。このような観点の違いを考慮に入れた概念辞書検索システムを構築することで、前後の文脈による「コソア」の指示詞の同定や、「あれ[魚]、食べた？」という談話中の「魚」と「食べる」の関係（「魚」が動作主か、人間が動作主か）の推定などに応用できると考えている。

今回構築した概念辞書は記述されている語彙数がまだ少なく網羅性という面で大きな課題が残っているが、連想実験での刺激語を増やしつつ辞書の整備をしていきたいと考えている。

## 謝辞

連想実験の被験者の皆様に心より感謝いたします。また、適切な助言と実験を手伝って下さった慶應義塾大学石崎研究室の皆様に、また実験データの修正を手伝ってくださった概念辞書班の皆様に感謝いたします。

## 参考文献

[1]安藤まや・石崎俊,名詞・形容詞の共起関係の定量的考察～名詞基本語彙の連想実験から～,言語処理学会第４回年次大会,B2-5,1998.

[2]内山清子,岡本潤,石崎俊,概念辞書における日本語と英語の語彙空間の違いについて,電子情報通信学会　言語理解とコミュニケーション研究会,NLC97-2,1997.

[3]大熊智子,認知実験に基づく概念辞書の構築と検索,情報処理学会報告自然言語処理　112-18,1996.

[4]岡本潤・内山清子・石崎俊,オンライン連想実験システムと学習基本語彙の概念辞書化,情報処理学会報告　自然言語処理 118-18,1997.

[5]岡本潤・石崎俊,連想実験に基づく概念間の距離の計算方法と概念辞書の構築～学習基本語彙による距離空間の定量化～,言語処理学会第４回年次大会,B2-4,1998.

[6]甲斐睦朗 松川利広,語彙指導の方法,光村図書,1996.

[7]高野陽太郎,認知心理学２～記憶,1995.

[8]Tulving,E. ,Elements of Episodic Memory,Oxford University Press,1983.
（太田信夫 訳 1985,『タルヴィングの記憶理論』,教育出版）.