連想実験に基づく概念間の距離の計算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～

岡本 潤 石崎 俊
{juno,ishizaki}@sfc.keio.ac.jp

慶應義塾大学 政策・メディア研究科

1 はじめに

意味情報と文脈情報を人間の記憶から直接取り出すことを目的としてオンライン連想実験システムを構築した。この実験では刺激語として小学生の学習基本語彙中の名詞を採用する。[6]

各刺激語ごとに被験者１０人に「上位概念」「下位概念」「部分材料」「属性」「類義語」「動詞」「環境」の７つの課題について連想させ、さまざまな概念情報を収集した。多数の被験者に対して連想実験を行ないデータを収集することで、人間の基本語彙に関する一般的な知識を得ることを図っている。[7]

概念辞書には概念情報の他に、連想時間、連想順位、連想頻度、概念間の距離という情報も付加し概念空間を構築する。

ここでは、概念間距離の定量化に関する一考察を行なう。現在、本研究では構築した概念辞書を可視化も試みている。

2 概念空間の距離計算

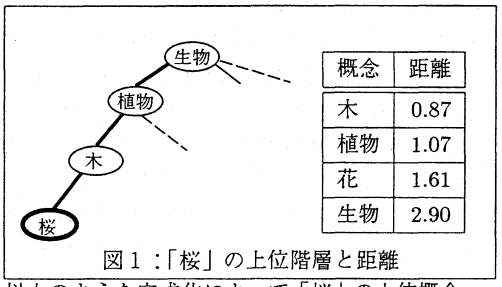
2.1 従来の計算方法

これまでの概念辞書での距離の計算はヒューリスティックに以下のような定式化を行なった。α, β, γについてはさまざまな値を用いて考察した結果、個々のデータの特徴を反映するもっとも妥当な値を採用した。[1]

$$D = \alpha \times T + \beta \times S + \gamma \times \frac{1}{F} \ldots (1)$$
$$\alpha = 0.1, \quad \beta = 0.3, \quad \gamma = 0.5$$

T(連想時間) $= \frac{1}{n}\sum_{i=1}^{n} t_i \times \frac{1}{60}$
S(連想順位) $= \frac{1}{n}\sum_{i=1}^{n} s_i$
F(連想頻度) $= \frac{n}{10}$

t_i = 被験者が連想に要した時間（秒）
s_i = 被験者が連想した語の順位
n = 連想人数...



概念	距離
木	0.87
植物	1.07
花	1.61
生物	2.90

図１：「桜」の上位階層と距離

以上のような定式化によって「桜」の上位概念の距離の上位４個の数値を見ると、「桜」の上位概念の階層構造に対応して距離が長くなっている。[1]

2.2 線形計画法による最適な式の決定

(1) 式で$\alpha, \beta, \gamma \geqq 0$として線形計画法を用いて１次式として表現する。

距離の式の係数α, β, γは、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。

$$\begin{array}{ll} \text{最小化} & Z = c_1 \times \alpha + c_2 \times \beta + c_3 \times \gamma \\ \text{条件} & a_{11} \times \alpha + a_{12} \times \beta + a_{13} \times \gamma = D_1 \\ & a_{21} \times \alpha + a_{22} \times \beta + a_{23} \times \gamma = D_2 \\ & \alpha, \beta, \gamma \geqq 0 \end{array}$$

まず、目的関数のc_1, c_2, c_3は$c_1 \geqq c_2 \geqq c_3$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。

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D_1になる場合を、「連想時間が短く」「一番最初に連想された語」「被験

者全員が連想」した時と仮定する。距離がD_2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。

$c_1, c_2, c_3, a_{12}, a_{12}, a_{13}, a_{21}, a_{22}, a_{23}, D_1, D_2$を少しずつ変化させて$\alpha, \beta, \gamma$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。

2.3 距離計算用の単語の選択

身近で分かりやすい単語として、比較的連想しやすいと思われる刺激語を選び、その連想語の中で連想頻度が高く、学習基本語彙の中に掲載されている単語を採用した。これらは私たちにとって日常馴染みのある単語であると考えられる。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ことになるので密な語彙ネットワークが出来上がる。以下はその例と実験結果の一部である。

表1：身近で分かりやすい単語の例

刺激語	課題	連想語	連想時間	連想順位	連想頻度
木	上位	植物	0.290	1.000	0.700
		生物	0.533	1.333	0.300
	下位	桜	0.800	3.333	0.300
桜	上位	植物	0.441	1.778	0.900
		木	0.258	1.500	0.400
		花	0.400	2.000	0.400
		生物	0.367	2.500	0.200
花	上位	植物	0.248	1.111	0.900
		生物	0.542	1.500	0.200
	下位	ばら	0.697	2.333	0.600
		桜	1.172	3.000	0.600

2.4 概念階層構造における距離関係の特徴

算出された数値をもとに、刺激語と連想語の上位概念と下位概念の位置関係を見てみる。

概念の階層構造を概念間の距離(A,B,C)に反映するように配置する。

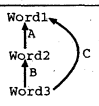


図2：概念間の階層構造とつながり

以上より、目的関数の$(c_1, c_2, c_3) = (10, 5, 1)$、条件式の$(a_{11}, a_{12}, a_{13}, D_1) = (0.1, 1.0, 1.0, 1.1)$, $(a_{21}, a_{22}, a_{23}, D_2) = (1.0, 4.0, 10.0, 9.0)$とした。この時、$\alpha = 0, \beta = 0.33, \gamma = 0.77$、概念間の距離$(D)$は

$$D = 0.33 \times S + 0.77 \times \frac{1}{F}$$

を得た。

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$\alpha = 0$となるのは、妥当であると考えられる。

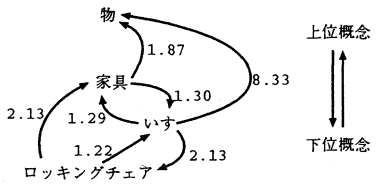


図3：「家具」「いす」「ロッキングチェア」の刺激語、連想語の配置

図3では「いす」→「家具」→「物」とたどった距離の合計よりも「いす」→「物」とたどった距離の方が長い。これは「物」が抽象的・包括的で、かなり上位層の概念が遠くにあるということが直観的に受け止められる。しかし、「ロッキングチェア」「いす」「家具」の場合や「桜」「木」「植物」、「桜」「花」「植物」の場合は他の概念を通って上位の概念に行く時の合計と直接上位の概念に行く時の距離とでは、直接上位の概念を連想した方が距離が短い。これは、どの語も具体的で身近な語であるため、お互いの概念距離が近いと考えられる。

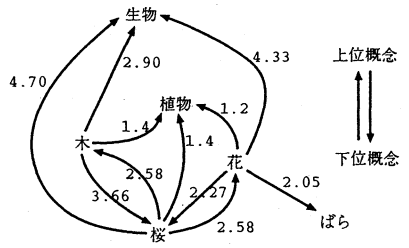


図4：「木」「桜」「花」の刺激語、連想語の配置

また、「桜」「木」「植物」、「桜」「花」「植物」に関しては「桜」の上位概念は「植物」であると連想する人が多く距離も近い、「花」という連想語は「桜」という「植物」そのものよりも桜の花の部分を想起させる場合が多いことを反映していると思われる。「木」という連想語は桜の幹や枝葉など桜の全体を想起していると考えられる。

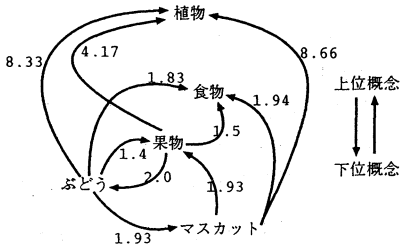


図5：「マスカット」「ぶどう」「果物」の刺激語、連想語の配置

「マスカット」「ぶどう」は「食物」「果物」として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るので概念間の距離は比較的近く、「植物」の概念間距離は遠い。これは「桜」の例とは異なり「ぶどう」「マスカット」を「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。

3 概念辞書の構築と可視化

3.1 データの修正作業と概念辞書の構築

まず、実験データ中の課題にふさわしくない連想記述、表記の揺れを修正する。課題にふさわしくない連想語は削除し不使用語とするか、7つの課題の他にさらに、「関連語」という項目を追加し、連想語としてふさわしい課題の場所へ移動する。

固有名詞は概念ではないので別のリストに収集するためにチェックしておく。表記の揺れの修正には「岩波国語辞典」の見出し語の表示を用いそれにしたがった。修正後、概念辞書作成ツールによって概念辞書が作成される。[7]

3.2 概念辞書の検索と可視化

刺激語から課題ごとの連想語を検索や、連想語から刺激語を検索する概念辞書検索システムを作成した。(図6)

検索された語は概念間の距離が短い順番にリスト化され課題ごとに表示される。また、概念辞書に記述してある距離を反映して空間的に語を表示する。今後、複数の検索語の語彙ネットワークを表示できるシステムにしていく。

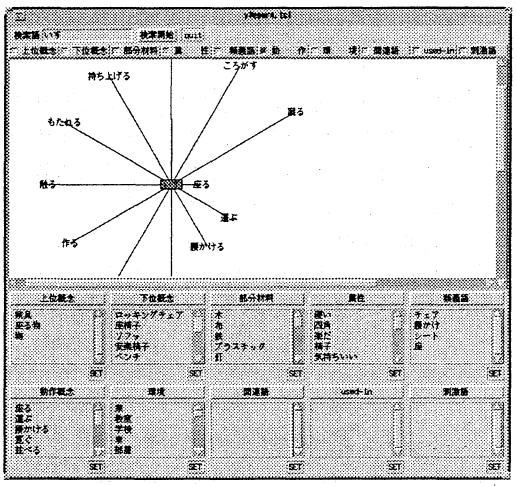


図6：概念辞書検索システム

謝辞

連想実験の被験者の皆様に感謝します。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感謝します。

参考文献

[1] 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.

[2] 岩波国語辞典, 岩波書店,1994.

[3] 国立国語研究所, 分類語彙表,1993.

[4] 日本電子化辞書研究所,EDR 電子化辞書使用説明書,1993.

[5] 牧野武則, 語彙の概念と知識について, 情報処理学会研究報告 自然言語処理 83-14,1991.

[6] 甲斐睦朗 松川利広, 語彙指導の方法, 光村図書,1996.

[7] 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.

[8] 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.