自然言語処理 118－18
（1997．3．22）

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

岡本 潤† 内山 清子‡ 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

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。小学生の学習基本語彙中の名詞を刺激語として連想実験を実施し、様々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。本研究で、オンライン連想実験システムにより得た実験データは、多量のデータのため、中には表記のゆれ、入力ミス、勘違いなどが含まれており、それらを効率的に修正し、データを集計するシステムを作成した。したがって、今まで多くの労力を必要としていた連想実験を効率良く実行し、データを収集することができた。最後に、集計されたデータをもとに概念辞書構築にむけて考察を行なう。

On-line association experiment system and building concept dictionary for basic vocabulary in elementary school

Jun OKAMOTO† *Kiyoko UCHIYAMA*‡ *Shun ISHIZAKI*†

†Faculty of Environment Informantion,Keio University
‡Graduate School of Media and Governance, Keio University

On-line association experiment system was built for extracting semantic information and contextual information from human memory. We carried out the association experiment about basic nouns in elementary school textbooks, and collected various conceptual information. Collecting many experimental data will provide us general knowledge on the basic vocabulary. The row data by the association experiment included various errors such as mistaking of typing and misunderstanding by subjects. We collected conceptual data associated by human subjects efficiently by using the error correction system.We describe the building of concept dictionary based on the data.

1 はじめに

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、実際に役立つシステムを構築する上で十分ではない。即ち、システムによる理解の深さにはあきらかに限界がある。[1]

このような人間との大きなへだたりを埋めるために、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。

従来からの大規模な知識ベースの開発では自然言語理解のための知識ベースの開発がＥＤＲで約40万語の概念について行なわれた。[2] しかし実用に耐えるようなレベルの一般的な常識の蓄積、言語の意味の記述までには至っていない。[3]

こうしたなか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。

今回概念辞書を構築するにあたって、小学生の学習基本語彙の名詞について連想実験を実施し、その結果を概念辞書に反映させることを目指していく。

2 連想実験システムの構築

連想実験システムはすべてキャンパスネットワークのオンラインシステム上で稼働できるように作成した。これによって被験者は都合のいい時に好きなだけ時間をかけて実験を行なうことが可能となる。また、実験結果はすべて電子テキストで得られるため、集計·統計処理などを直接コンピュータで行なうことができた。それにともない概念辞書構築が効率化されている。

2.1 連想実験の概要

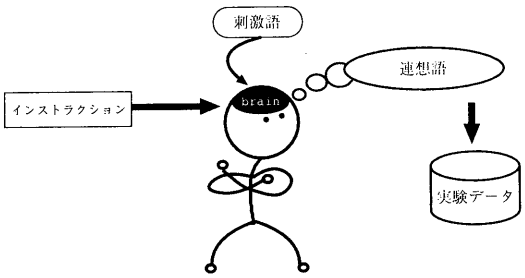

図 1: 連想実験の流れ

◎手続き

光村図書「語彙指導の方法」[4] から小学校の学習基本語彙の名詞約 1400 個の中から、「母」,「母親」などの同義語や多義語のある名詞を除いて約 800 個の名詞を得た。次にその中からランダムに 100 語を選択し、10 語ずつ 10 個のグループに分けて刺激語とする。今回は 100 語のうち、50 語について実験を終了したので結果を報告する。各刺激語に対して被験者 10 人に選択した名詞を課題ごとに提示し (図 2)、その刺激語から連想する単語をかな漢字変換システム (kinput2) を用いて入力させる。課題は「上位概念」、「下位概念」、「部分・材料」、「属性」、「類義語」、「動作概念」、「動作環境」の 7 つである。(表 2)

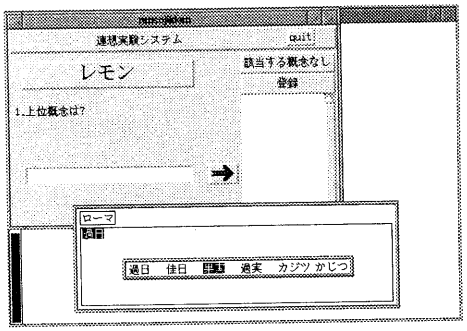

図 2: 連想実験の実験画面

実験システムでは、図 2 で示すように連想語の入力はすべてローマ字入力、カナ漢字入力で行なわれる。単語を入力したら入力欄の右矢印を押して、登録 BOX に登録する。単語入力が終了したら、登録ボタンを押して次の課題に進む。刺激語 10 語に対してそれぞれ 7 つの課題を終了すると実験は終り、実験データは実験者のもとに送られる。

◎提示刺激

表 1 は今回行なった連想実験での刺激語である。意味分類によって、共通の意味を持つ単語 (類語) がまとめられる。[7]

表 1: 連想実験の刺激語一覧

刺激語	意味分類	刺激語	意味分類
視覚	感覚	地下	上下
たんす	家具	学生	専門的技術的職業
海岸	海, 島	産業	業
恐怖	苦悩, 悲哀, 恐れ, 怒り	ふろ	家具
列	束, 片, 列, 枠	辞書	辞書, 目録, 暦
組織	内容, 組織	手帳	ふだ, 帳
読者	作者, 仕手, 主役, 観客	糊	文具
校舎	家屋	ため息	表情
研究	研究, 実験, 調査, 検査	地図	表, 図, 譜, 式
共通語	言語	作文	創作, 著述
直観	感覚	はだ	皮, 毛髪, 羽毛
親友	友, なじみ	距離	長所, 広狭
敵	相手, 仲間	勇気	自信, 名誉, 勇気
生涯	永久, 一生	旅	旅, 行楽
順序	順序	機械	機械
照明	灯火	迷信	信仰, 宗教
井戸	その他の土木施設	孫	子, 子孫
座席	範囲, 席, 跡	まくら	寝具
旗	標識	海	海, 島
岩	鉱物	徒歩	足の動作
外国	国	鍵	農工具
課長	相対的地位	頬	頭, 目鼻, 顔
熱	熱	折り紙	紙
小説	芸術, 文芸	広場	地類
親子	家族	のど	頭, 目鼻, 顔

◎課題

表 2: 7つの課題の説明

上位概念	上位概念とはその単語の意味を含む概念のことです。例えば「さかな」の上位概念には「生物」や「食物」が考えられます。
下位概念	下位概念とはその単語に含まれるより細かい概念のことです。例えば「さかな」の下位概念には「さんま」や「深海魚」が考えられます。
部分材料	部分材料とはその単語を構成する、または部分をあらわす語です。例えば「さかな」の部分材料には「うろこ」や「ひれ」が考えられます。
属性	属性とはその単語の持つ特徴です。例えば「さかな」の属性には「えら呼吸する」「冷たい」「ぬるぬる」が考えられます。
類義語	類語とはその単語と意味が似ている語です。例えば「さかな」の類義語には「魚類」が考えられます。
動作概念	動作概念とはその単語がする、あるいはされる動作です。例えば「さかな」の動作環境には「泳ぐ」、「釣る」、「焼く」が考えられます。
動作環境	動作環境とはその単語が用いる、あるいは用いられる環境です。例えば「さかな」の動作環境には「海」、「魚屋」、「食卓」が考えられます。

◎結果

表 3 は刺激語「辞書」に対して、上位概念ならば「書物」「本」「文献」という順番に連想し、各々の数字は連想時間である。

表 3: 刺激語「辞書」における
一人の被験者の実験結果の例

上位概念	{書物 7} {本 12} {文献 18}
下位概念	{英語辞典 6} {国語辞典 12} {漢和辞典 19}
部分・材料	{見出し語 18} {語釈文 33} {ページ 38} {表紙 44}
属性	{難しい 6} {わかりやすい 11} {楽しい 16}
類義語	{辞典 8} {事典 17}
動作概念	{読む 5} {調べる 11} {引く 15}{探す 19}{買う 29}
動作環境	{図書館 6} {本屋 27}

2.2 データ集計法の概要

被験者から送られてきた実験データには刺激語に対する課題に明らかにふさわしくない記述、かな漢字入力のために生じる漢字とひらがなの表記の揺れ、また送り仮名などの違いが少なからず見受けらる。

たとえば、「海」の属性に「広い」と記述する被験者と「ひろい」とする被験者がいる。また、送り仮名の違いとして「気持ちいい」「気持いい」などが出てきた場合、ある基準を用いて表現を統一しなければならない。

次のようなシステムによって実験データの修正がスムーズに行なえるようになり、連想実験によるデータ収集と集計が効率化された。

2.2.1 実験データ形式の修正

本システムでは、コンピュータ上で自動的に集計・統計処理を行なうので、実験結果が指定されたフォーマットで収集されなければならない。しかし、時に人為的なミスにより指定されたフォーマット以外で得られる場合がある。

この過程では、実験データチェックツールにより被験者ごとに実験データファイルのどこにエラーがあるか容易に発見することができる。実験データファイルのフォーマットエラーは結果リストに表示され、それを参照しながら元ファイルの形式を修正する。

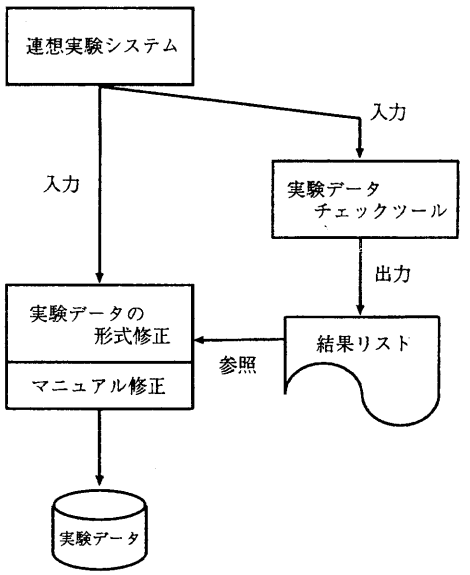

図 2: 実験データの形式修正フロー

2.2.2 実験データ内容の修正

ここでは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。具体的にはチェックリストとひらがなカタカナリストを生成する。

ひらがなカタカナリストは、連想語において「食べる」と「食べる」のように、見た目ではわかりづらい“へ”“べ”“ぺ”のひらがなとカタカナの違いをチェックするために使用する。

生成されたチェックリストを、ひらがなカタカナリストを参照して修正する。また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、あるいは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。

修正されたチェックリストは実験データへの変換ツールで課題移動編集リストを作成し、また実験データ(修正済み)、固有名詞、不使用語に分割して格納する。

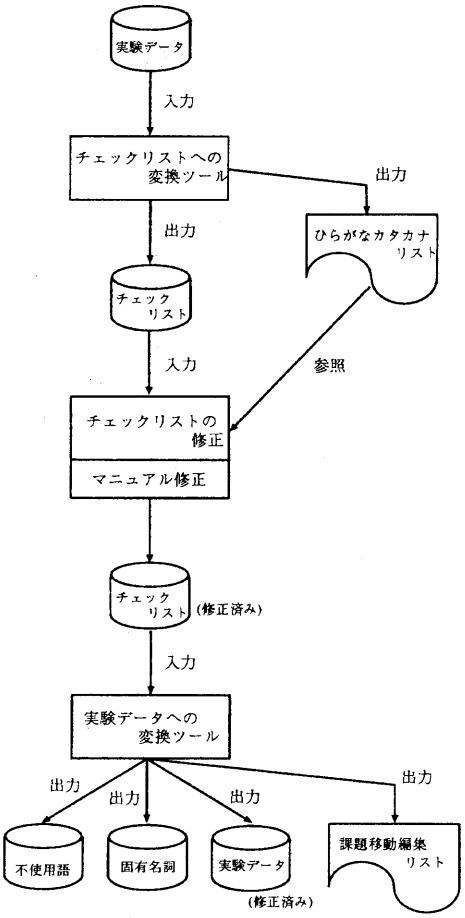

図 3: 実験データの修正フロー

2.2.3 刺激語毎の集計データの作成

実験毎の集計ツールと刺激語毎の集計ツールにより実験の刺激語毎に対して各課題毎にどのような連想語が何人の被験者によって連想されたか集計するプログラムである。

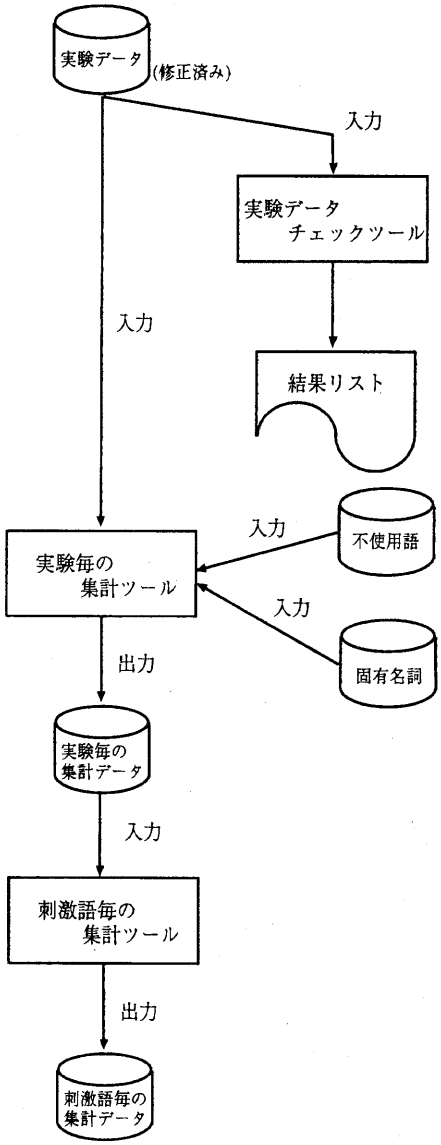

図 4: 刺激語毎の集計データ作成フロー

3 連想実験の結果

表 4: 刺激語ごとの連想個数の例

刺激語	井戸	岩	海	親子	折り紙	ため息	直観	勇気	外国	距離	機械	生涯	恐怖
上位概念	3	5	8	10	9	6	3	4	3	3	5	8	3
下位概念	5	10	13	10	9	0	1	0	15	7	21	1	4
部分材料	11	10	26	15	11	17	8	18	16	8	34	19	19
属性	15	11	19	20	19	34	19	30	18	30	29	29	27
類義語	5	2	8	10	5	4	8	17	9	13	7	7	13
動作概念	12	14	15	15	16	7	15	14	19	17	30	15	21
動作環境	10	17	10	9	15	26	17	21	12	18	20	13	19
合計	61	69	99	89	84	94	71	104	92	96	146	92	106

刺激語として提示される名詞には連想しにくい名詞がある。「ため息」「直観」「勇気」「恐怖」「距離」「生涯」などの抽象名詞は主に上位概念・下位概念・部分材料が課題として与えられた時連想しにくい。それは「岩」「海」などの具体名詞に比べて物体として容易にイメージできないからだと考えられる。

その反面、抽象名詞は状況的な事柄をあらわす言葉を連想しやすく、また同時に状況を説明する形容詞なども連想しやすいため属性・動作環境の連想頻度が大きかった。

上位概念、下位概念という課題は被験者にとって困難な場合がある。上位概念・下位概念から他の課題に移したのは課題移動を行なった全体の7割近くであり、上位概念・下位概念として被験者が連想した項目における誤りの多くは部分・材料や属性であった。また、「岩」(岩石のうち大きなもの) と定義が決まっているものについて、「岩石」は「岩」上位概念であるし、「石」(岩石のうち、割合に小さいもの。砂ほど小さくなく岩ほど大きくない) は「岩」の部分材料である。しかし、「岩」の上位概念にも下位概念にも「岩石」や「石」は連想語としてあらわれてきた。連想実験では被験者の知識体系や感じ方をそのまま反映してしまうので、学術的用語の扱いに注意しなければならない。

3.1 基本語彙の概念空間の広がり

異なり語数とは7つの課題において刺激語は違っていても同じ連想記述が現れる。例えば、刺激語「作文」,「小説」の動作概念として「書く」が連想された場合、延べ数を数えるのではなく1つと数える。

"表 5: 連想語語数" の表から連想語の異なり語総数は特に動作概念に顕著に現れ、連想語延べ数に比べて減少している。これは、刺激語から連想される動詞はある程度重なりがあって概念空間は小さくなっていくことを表している。本研究では50語の刺激語について連想実験を行なったが、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、もしくは連想しやすい、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。

表 5: 連想語彙数

課題	連想 単語数	課題別 異なり語数	全体 異なり語数
上位概念	306	240	
下位概念	543	543	
部分材料	920	920	
属性	1001	778	
類義語	394	390	
動作概念	960	683	
動作環境	898	627	
合計	5022	4181	3575

4 概念辞書構築へむけて

表 6 は「海」という刺激語に関して、課題ごとに連想された概念の連想順位と連想頻度を実験データを集計して算出したものである。

表 6: 海に関する連想語と連想順位・連想頻度

課題	連想語	連想順位	連想頻度
上位概念	地理	1.00	0.10
	地球	1.20	0.50
下位概念	地中海	4.00	0.10
	太平洋	1.50	0.60
部分材料	海流	4.00	0.10
	大陸棚	3.00	0.20
	水	1.13	0.80
属性	しょっぱい	4.00	0.20
	深い	3.33	0.60
	広い	1.44	0.90
類義語	母	1.00	0.10
	海洋	1.00	0.20
動作概念	潜る	4.00	0.10
	航海する	2.00	0.20
	荒れる	1.00	0.40
	泳ぐ	1.20	0.50
動作環境	港	3.00	0.10
	地図	1.50	0.20
	地球	1.00	0.50
used-in	海岸	部分材料	
	産業	動作環境	
	地図	部分材料	
	はだ	動作環境	

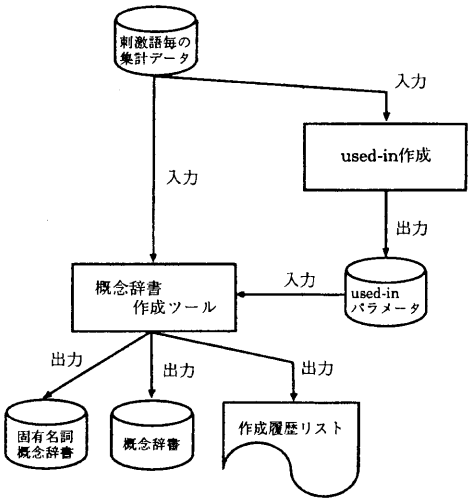

図 5: 概念辞書作成フロー

概念辞書作成のために、刺激語毎に集計された結果をもとに used-in パラメータを作る。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 へのリンクをはることができる。

概念辞書作成ツールに刺激語毎の集計データと used-in パラメータを入力することにより概念辞書が作成される。

5 今後の課題

本研究では表 2 のような課題についてのインストラクションを与えて被験者に良く読んでもらってから実験を行なったが、実験者が被験者につきっきりで実験を行なってもらうわけではないので、インストラクションを熟読しないで実験を行なう被験者が少数であるがいると思われる。

上位概念や下位概念の意味を正しく理解してもらうために実験中でも必要に応じて課題の説明 (表 2) を呼び出して見ることができるが、被験者にとってインストラクションを詳しく読まなくても容易に連想できるようなインターフェースを構築することが望ましいと思われる。したがって実験画面を以下のようにして実験を行なうことを予定している。

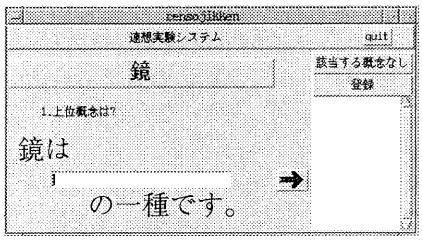

図 6: 新しい連想実験画面

謝辞

本研究は DENO の「メディア情報と言語情報の統合と学習技術の基礎研究」と科研費の「言語の状況依存性の認知モデルと文脈理解システムの研究」の助成を受けました。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] 板橋秀一, 知識・知能と情報 —脳のはたらきと情報処理—, 近代科学社,1992.

[2] 日本電子化辞書研究所,EDR 電子化辞書使用説明書,1993.

[3] 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.

[4] 甲斐睦朗 松川利広, 語彙指導の方法, 光村図書,1996.

[5] 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.

[6] 岩波国語辞典, 岩波書店,1994.

[7] 国立国語研究所, 分類語彙表,1993.