大規模連想語データベースの構築

テリー・ジョイス
(「大規模知識資源-21 世紀 COE」東京工業大学)
key words：連想語　データベース　語彙連想マップ

連想は、人間の認知の基本的な原則の一つであり(Deese, 1965; Cramer, 1968)、連想語－単語と概念の間にある関係・結合・構造－に着目する研究は、認知のメカニズムへの貴重な洞察をもたらす可能性がある。例えば、Nelson, McEvoy, & Schreiber (1998)が構築した大規模英語連想語データベースを利用して、Steyvers & Tenenbaum (2005)は自然な意味ネットワークの特徴を抽出した。Steyvers らは、その Nelson らのデータベースと、WordNet, Roget の類義語辞典に基づいたネットワークをグラフ理論的な分析した結果、密なクラスタ群をハブ的なノードが結合し、リンクの分布がべき乗となることが自然な意味ネットワークの特徴であることを示唆している。

多くの認知科学分野では包括的な連想語データベースが必要であるため、Joyce (2005a)は日本語の基本語彙における自由連想語の大規模連想語データベースの構築し、そのデータを利用して語彙連想マップの作成を始めた。最初の調査対象コーパスとして、日本語基本語彙の 5,000 項目の漢字と単語を選択した後、Joyce (2005b)は、第一回目のデータ収集として約 1,000 名の回答者に質問紙を配り、回答を得た。本研究では、第二回目のデータ収集を行うとともに、Web を用いた質問フォーマットを公開のため準備を続けることである。

方法

回答者：大学生 300 名。

対象項目：Joyce (2005b)は、5,000 項目の語彙コーパスから 2,000 項目をランダムに選択して、第一回目のデータ収集とし、質問紙の調査を行った。本研究は、第二回目のデータ収集として、残りの 3,000 項目を対象とした。

質問用紙：各項目に対して 10 回答ずつを得るように、3,000 項目を 30 リスト（各リストを 100 項目）に分けた。リスト内の連想がないように各リストを調べた後、各リストの 10 名の回答者に対して個人の質問紙での提示順序の並べ替えを行った。質問紙では、「印刷されている文字を見て、一番最初に思い浮かんだ日本語の単語を 1 つ、下線部に書いてください。意味的に関係がある単語なら何でもけっこうです。」という教示を与えた。

結果

大規模連想語データベースを構築するために、Joyce (2005b)が行った第一回目のデータ収集で得られた約 100,000 弱の解答に加えて、本研究で行った調査から約 30,000 の解答を得た。

考察

大規模連想語データベースをより効率的に構築するために、Web を用いた質問フォーマットも準備した。その過程における問題点は、質問紙リスト内の連想をできる限り統制しながら、多数の回答者に対して各質問紙を自動的に作成できる方法を工夫することであった。その問題を解決するために、調査対象コーパスの各項目を、読み、表記、構成要素漢字、意味の分類、質問紙回答に関する情報でコード化した。

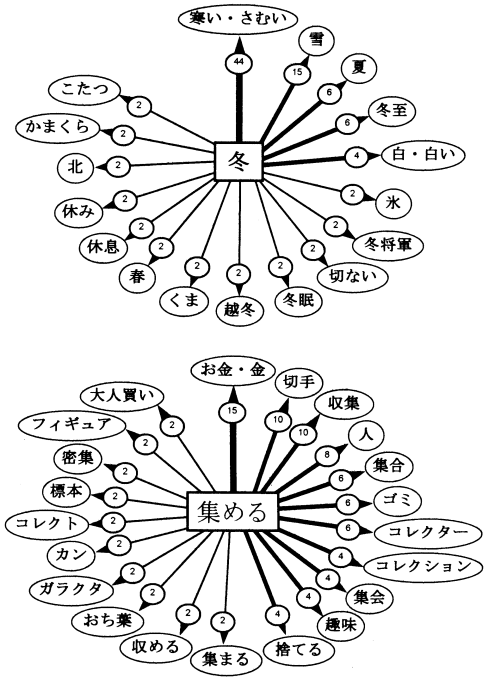

図 1.　「冬」と「集める」に対して連想語の集合

本研究で構築している大規模連想語データベースは、認知心理学実験のための有用な資料となりうる。連想語データベースの他の応用としては、語彙の重要な特徴と語彙どうしの接続性を捉えることのできる語彙連想マップ（意味ネットワークの一種）として利用することも考えられる。図 1 は、「冬」と「集める」に対しての連想語の集合である。「冬」という名詞には修飾語である「寒い」という形容詞との間に強い関係が見られる。一方で、「集める」という動詞に対して強く連想された単語には、その目的語と思われるものが多い。

この連想語データベースに、逆の連想と集合間の連想も加えていくことにより完全な語彙連想マップができる。また、この語彙連想マップによって、レンマ・ユニット・モデル (Joyce, 1999, 2002, 2004)における意味表象部分のモデル化が可能となる。更なる応用は、ユーザ・フレンドリな検索方法および見出し語の追加・補足による辞書編集と包括的な漢字データベースに基づく日本語学習システムなどがある。

(Terry JOYCE)

本研究は、文部科学省 21 世紀 COE プログラム「大規模知識資源の体系化と活用基盤構築」の一部である。