Vol.2014-IFAT-114 No.11
Vol.2014-DD-93 No.11
2014/3/29

ネイティブスピーカーの心的辞書構築のための連想語抽出法の提案

攪上 萌[1]　大野 一樹[2]　波多野 賢治[1]

概要：個人の持つ言語情報を内包する心的辞書の構築は，第二言語学習において重要とされる．心的辞書の構築には，語彙の増加に伴い辞書内の単語間の位置関係を決定していくことが必要である．本稿では単語間の複数の関係性を反映する連想に着目し，任意の単語に対する連想語を自動的に抽出する手法を提案する．単語間の連想強度を意味概念の類似度，文字列の類似度，共起頻度から数値化し，連想語の抽出を行う．この手法は，心的辞書内における単語間の関係を意識的に学習できるシステムの開発などに応用することができると考えられる．また，アンケート調査を介して，提案手法により構築された心的辞書の評価を行う．

キーワード：連想，単語類似度，心的辞書，第二言語学習

1. はじめに

個人の交流からビジネスまで，多様な場面において外国語習得の必要性が増してきている．一方で，日本のように公用語が唯一である国では，第二言語として外国語を自然と扱えるようになるための機会が多く用意されてはない．このように，幼少時から異文化に触れる機会が少ない国では，偶発的な学習のみで複数の言語を習得するのは困難である [1]．そのため，日本人の第二言語学習者は言語学習促進のために，意図的に語彙数を増やす必要性がある．

語彙獲得のための古典的な学習法として，単語帳を用いた学習法がある．単語帳は，単語とその訳，例文を一組として，カードあるいはノートに学習する単語を一つずつ書いて綴ったものである．学習者はこの単語と訳を交互に確認しながら，対応関係を学習することで，語彙獲得を行っていく．単語帳を用いた学習の利点としては，他の学習法と比較して学習開始時に必要な語彙力が求められておらず，また，学習コストも小さいため学習者にとって利用時の敷居が低い点が挙げられる．

しかし，従来の単語帳には学習する単語に対して知識の誤解を招くという問題がある．この問題は一つの単語に対して，母語訳を一対一で対応させて記憶を促すという単語帳のシステム的特徴に依存するものである．単語帳によっては一つの単語に対して複数の母語の意味を提示する場合もあるが，基本的には一つの単語に対して一つの訳語で学習するスタイルが取られている．これは一つの単語に対して複数の意味を提示しても学習が困難になることや，母語による直訳を利用した簡潔な学習が単語帳の容量や学習者の理解に適しているためであると考えられる．

しかしながら，母語を日本語とする日本人の第二言語学習者にとって，学習言語と日本語とでは語概念が完全には共有されていない．そのため，一つの単語に対して日本語による直訳を付与した単語帳を用いた学習では，厳密に正しい単語の意味を表すことができず，学習者に語概念の誤解を生じさせる可能性がある．したがって，母語を介した学習法は，結果的に後々にまで語の誤用 [2] を招く要因となりうる．

この問題に対して，本稿では言語習得において重要とされる概念である心的辞書の仕組みを考慮することで解決を図る．心的辞書とは，個人が知識として持つ単語の意味と品詞，音声情報などを含む言語情報のすべてを貯蓄する脳内で高度に体系化された動的な辞書のことを指す [3]．心的辞書の構築プロセスに着目することにより，既存の学習と比較して母語話者に近い言語運用を可能とする単語帳を用いた学習法の実現を目指す．

言語習得はこの心的辞書内に新しく覚えた単語を配置していく作業であると考えられる．そこで，学習には単語間の関係として心的辞書内における単語の位置情報が付与されていることが望ましいと考えた．本稿では，この位置情報を単語同士の連想関係として学習者に提供することを目

[1] 同志社大学文化情報学部
[2] 同志社大学大学院文化情報学研究科

的とし，任意の単語に対してその言語の母語話者にとって一般的な連想語を抽出するための抽出手法を提案する．

2. 関連研究

本節では，連想語の抽出手法を提案するにあたり，日本語母語話者の英語の学習過程と，連想関係の分類について述べる．また，連想語抽出時に利用される単語間類似度の計算手法を概説する．

2.1 第二言語学習者の英語学習過程

第二言語を習得する際，学習初期の段階では学習者は図 1 の左部のように，母国語の心的辞書 (L1) による翻訳に頼る．その後，学習が進み，第二言語の心的辞書 (L2) が構築され，習熟することで右部に示すように第二言語の語概念を理解できるようになる [4].

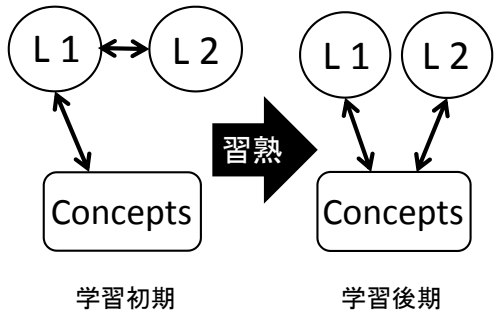

図 1 母語および第二言語の心的辞書と語概念の関係

英語を学習している日本語母語話者を対象に対して行った単語連想に関する比較実験が Yokokawa et al. によって行われている [4]．この比較実験によると，言語間で同じ意味を持つ単語から連想される日本語の連想語と，英語の連想語には質的な差があるという結果が示されている．その一方で，言語学習においては L2 のネットワーク構築には L1 のネットワーク構築と同じ枠組みが利用されていることが示唆されている．つまり，従来の学習法に則ると L2 のネットワークは L1 のネットワーク構築時に用いられる方法と同様の方法に基づいて構築されると考えられる．また，学習初期のように L2 と語概念の繋がりが希薄である場合，未知の単語を提示された際の意味の推測など，未熟な L2 だけで対応できない言語処理を L1 の翻訳に頼ることとなる．このことは結果的に語の誤った理解が行われる要因となり得る．

これらのことを踏まえて，語概念の誤解を防ぐためには，L2 の構築をより早く行うことで L1 に依存する時間を短くすることや， L2 の構築時に正しい単語の配置による正確な連想関係に基づいたネットワークを構築することが問題解決のための課題として考えられる．

2.2 連想関係の分類

ある単語（刺激語）を認識したとき，それに対して連想される単語のことをその語の連想語という．連想は，個々人の記憶および知識に基づいて生起する [5] ため，単語に関する連想は言語知識の貯蔵庫である心的辞書内の単語の配置に影響を受けていると考えられる．逆に，単語間の連想関係から心的辞書内の単語の配置を把握することができる．

心的辞書は概念同士の「類義関係」,「上位下位関係」,「近接関係」の三種類の関係によって構造化されている [3]．さらに，連想関係を八種類に分類した研究 [6] を受けて，秋山ら [7] は表 1 のようにこれらを対応させている．

表 1 概念間の関係の種類

大分類	小分類
類義関係	類義語
	反意語
	事例-事例
上位下位関係	全体-部分
	カテゴリ-事例
近接関係	熟語
	随伴
	統語

「類義関係」に分類される連想語は意味的な類似性によって連想される語であり，「宣伝」と「広告」のような「類義語」,「戦争」「平和」のような「反意語」,「色」の一種である「赤」「青」「緑」といった「事例–事例」に細かく分類される．「上位下位関係」は，類義関係と同様に連想語の意味定義に依存しているが，単語の属する概念の階層の深さが異なり，これに基づいた上下関係が存在する．この関係は「手」と「指」などの「全体–部分」関係,「動物」「犬」などの「カテゴリ–事例」関係として表される．「近接関係」に分類される連想語は，刺激語に対する「近接性」により連想される．近接性とは空間や時間など，ある側面で何らかの「近さ」があるということである．例えば,「やかん」という語から，やかんの中に入っている「お湯」という語を連想するのは空間的な近接性によるものである．この関係による連想語は，上記二つの関係と異なり，WordNet に代表されるような 一般的なシソーラス上には表現されていない [7]．また，表 1 では言及されていないが，そのほかの連想語として，発音やスペルの類似性を持つ語も連想語として存在する [8].

さらに他の連想関係として，刺激語と同じ品詞が連想される範列的反応と，刺激語と別の品詞が連想される統語的反応が Woodrow and Lowell [9] によって示されている．

しかし，これらはおよそ 7 歳で連想の反応は統語的なものから範列的なものに移行していくとされている．つまり，成熟した心的辞書は単語同士の範列的な繋がりが強まる性質を持っているといえる．

以上のような単語間の複数の関係性，連想関係によって心的辞書内の単語の配置が決定されることを考慮すると，これらの関係性を掌握することで，第二言語の語彙学習に心的辞書内の単語の位置情報を付与することができると考える．ゆえに，次の 2.3 節ではこの関係の計算手法に関する研究を関連研究として示す．

2.3 単語間類似度の計算手法

単語の連想の際に，連想語が生じる主な要因となるのは，単語間の意味的，音韻的，統語的な類似性である．その他に，個人的な経験や文化的背景等が連想に影響を及ぼす要因として考えられる．これは連想が意味記憶の影響を受けることに由来する [5]．しかし，本稿では一般的に妥当と判断されるような連想語を機械的に抽出することに焦点を当てる．そのため，一般化しにくい連想の要因に関しては除外するものとする．

単語間類似度の計算には，類似度計算の目的と，計算に利用される尺度に応じて多様な計算手法が存在する．以下では主に利用される類似度計算手法を概説する．

(a) シソーラスを用いた計算手法

シソーラスを利用した類似度計算手法では，単語間の意味的な類似度が計算できる．シソーラスとは語の概念辞書であり，語概念が階層的に内包されている．主要な英語シソーラスとしては WordNet[*1] や EDR 電子化辞書 [*2] がある．

シソーラスを用いて類似度計算を行った研究には，崔らの研究 [10] がある．単語が属する概念がシソーラスのどの階層に配置されているかによって，概念間の類似度を計算することが可能である．このように類似度を定義した距離を階層距離と呼ぶ．階層距離は，単語間に共通する上位概念の深さ，単語間のパスの長さ，共通概念の個数といった指標に基づいて計算される．

(b) コーパスを用いた計算手法

コーパスを用いた類似度計算手法では，目的に応じたコーパスから語の共起頻度を計算し，別の文脈内で同じ単語と共起している二単語を範列関係と捉えることで計算できる．この計算には，秋山ら [11] や橋高ら [12] の研究のように，ベクトル空間モデル [13] を用いることが多く見られる．ベクトル空間モデルは単語の意味的な特徴を多次元ベクトルとして表現するモデルである．ベクトルの各次元にはその単語と共起する単語の共起頻度を対応させることが多い．一般的に新聞記事や会話などのコーパスが使用されるが，辞典の語義文に基づいて単語概念の意味特徴を定義した概念ベースが用いられることもある．概念ベースに基づく類似度計算を行っている研究の例として高橋らの研究 [14] などが挙げられる．単語間類似度は二つのベクトルのなす角の余弦やユークリッド距離などを尺度として計算する．また，同じ文脈の中に出現しやすい統語的関係のある語に対してもコーパスの共起頻度から類似度を計算することができる．

(c) 単語の表層的特徴を用いた計算手法

単語の持つ意味や用法から類似度を計算する方法のほかに，単語そのものの表層的特徴，例えばスペルや発音の類似性に着目した方法がある．この方法では二単語間のスペルや発音記号の編集距離を類似度と捉えることができる．例えば “ZEBRA” という文字列から “ZERO” という文字列に変換を行う場合には，“ZEBRA” の “B” を削除する．次に “ZERA” の末尾の “A” を “O” に変えるという二回の操作が必要である．この操作にかかるコストを編集距離と呼び，ここでの編集距離は 2 となる．このように，編集距離とはある単語を別の単語に編集するために必要な基本操作数などを尺度とする距離である [15]．意味的な類似性に関わらず，語の表層の特徴のみを利用して距離を計算する．

ここで挙げた手法以外にも単語間の類似度の計算手法は様々であり，これらの手法は検索エンジンのクエリ変換や，テキスト解析を主な目的として研究，応用されている．

3. 提案手法

本稿では，前述した単語間類似度を利用し，英単語間の連想関係を計算することによって任意の単語に対する連想語を機械的に抽出する手法の提案を行う．第二言語学習の際に単語同士の連想関係を学習順序に反映させることで，学習言語における単語間の連想を学ぶことができ，第二言語の心的辞書の構築を支援することができると考える．連想データを入手する方法の一つとして，連想実験による連想語を格納するシソーラス EAT (Edinburgh Association Thesaurus)[*3] の利用が挙げられるが，対人実験によるデータの収集はコストが大きいため，含まれる連想語データに限りがあることが難点である．そのため，理論的に連想関係を定義することで，連想語の収録範囲の拡張が容易になり，実験で網羅できていない単語にも対応することが可能になる．

提案手法による連想語の抽出は以下の手順で行われる．

(1) 単語データの入力

[*1] http://wordnet.princeton.edu/wordnet/ 2013/12/20 閲覧
[*2] http://www2.nict.go.jp/out-promotion/techtransfer/EDR/J_index.html 2013/12/20 閲覧
[*3] http://www.eat.rl.ac.uk/2013/12/20

(2) 単語間類似度の計算
(3) 連想距離計算と連想語抽出

刺激語に対する連想語の抽出は，刺激語と連想語の二単語の連想距離の計算と，計算結果で連想語を降順に並べ替えることで距離の近いものから順に行う．まず連想距離を計算したい単語群を用意し，次に，用意した単語間の関係性として，複数の観点による類似度の数値化を行う．その後，各類似度を統合して連想距離とし，距離計算の結果に基づいて順序を付けることで連想語を抽出する．各処理内容については以下で詳述する．

3.1 単語データの準備

連想語として抽出対象となる単語の品詞は，単語自体が意味を持つ動詞, 名詞, 形容詞, 副詞である．助詞などの単語自体が意味を持たず，他の語に付随することによって役割を果たす付属語は連想されにくいため除外する．また本稿では使用ライブラリの関係上，今回は動詞と名詞に限って抽出を行った．

使用する英単語は，英単語を重要度でレベル分けしたJACET8000[16] から取得した．JACET8000 は，約 1 億語が収録されたイギリス英語コーパスである British National Corpus (BNC)[*4] を基に，アメリカ英語のデータも含むサブコーパスと合わせて語を選出した日本語母語話者向けの単語リストである．単語リスト中のレベル 1 からレベル 8 までの名詞 4547 語，動詞 990 語，計 5537 語が刺激語および連想語の候補となった．

本稿で提案する手法は単語データを品詞以外で制限することはないため，任意の単語データについて適用できる．

3.2 単語間関係の数値化

連想語を機械的に抽出するために，心的辞書内の単語の配置に関わっていると考えられる連想距離を計算する．この単語間連想距離を調べるために，第 2 章で触れた連想関係の分類を利用する．分類された単語間関係を，連想関係を構築する関係性の一部として捉える．つまり，これらの分類に当てはまる関係性の尺度を用いることで，複数の関係性の統合を経て連想距離を計算することができると考えた．

連想関係の一部を表す単語間関係として，まず単語の意味的な類似性が挙げられる．意味の類似性は，関係性の分類における「類義関係」および「上位下位関係」に当てはまる．これら二つの関係性の数値化は WordNet を用いた距離計算を行うことで決定する．概念間の距離を求める方法は幾つか存在し，特に WordNet 上の語の類似性を数値化する Java 言語のライブラリである ws4j (WordNet Similarity for Java)[*5] では，八種類の類似度を計算するメソッドが実装されている．本稿ではこの内，二単語の属する概念の最短経路の長さに基づいて類似度を測る LCH を意味類似度として用いた．ただし，特性上，品詞の異なる単語については直接距離を計算することができないため，よって，刺激語と同じ品詞の連想語の抽出を対象として距離の計算を行った．

次に挙げられるのが共起関係である．語の共起関係は関係性の分類における「近接関係」に当てはまる．近接関係を数値化するため，COCA (CORPUS OF CONTEMPORARY AMERICAN ENGLISH) [*6] に基づく共起頻度データを用いた．COCA は spoken, fiction, popular magazines, newspapers, academic journals の五つのジャンルから毎年同量のデータを収録している計 4 億語のアメリカ英語コーパスである．日本における英語学習テキストの多くはイギリス英語よりもアメリカ英語を標準としていることから，使用コーパスを選定した．本稿では，COCA が提供している共起頻度データの相互情報量 (MI) を共起頻度の指標として用いた．

更に，単語の音韻による連想をカバーするため，単語間の文字列の類似性を計算する．文字列の類似性は「音韻関係」を表すことができると考えられる．この関係は，文字列の編集距離を用いて数値化される．ここでは，一時的な文字列編集距離として Leven-shtein 距離を利用し，編集距離を算出する．音韻関係によって連想される語は，その他の関係によって連想される語に比べ少ない傾向にある [17]．しかし，本研究で提案する連想語抽出の手法は，より多くの関係性を含めることで連想関係の計算結果が実際の連想距離に近づくという立場を取るため，音韻関係も連想関係の一要素として計算に含める．ただし，音韻関係とその他の関係では，連想に及ぼす影響に差があると考えられる．

そこで，各関係性を統合して二単語間の連想距離を計算する際には，関係要素に連想への影響力を考慮した重みを付与し，パラメータを調整する．

3.3 連想距離計算と連想語抽出

上記で数値化した意味類似度，共起頻度，文字列編集距離を用いて，以下のように重み付けをした上で単語間の距離 $\mathbb{D}$ を計算する．意味類似度を α，共起頻度を β，文字列類似度を γ とした場合，単語間距離は以下のように計算される．

$$\mathbb{D} = 0.63 \cdot \alpha + 0.34 \cdot \beta + 0.03 \cdot \gamma$$

各数値に付与した重みは，母語における意味，共起，音韻の影響による連想語の割合がそれぞれ約 63 %，約 34 %，約 3 % であったという実験結果 [17] に基づいて設定した．

これら三つの関係性を統合した数値を利用することで，前節で述べた連想関係の分類で表される連想語を抽出する

[*4] http://www.natcorp.ox.ac.uk/ 2013/12/20 閲覧
[*5] http://code.google.com/p/ws4j/ 2013/12/20 閲覧
[*6] http://corpus.byu.edu/coca/ 2013/12/20 閲覧

ことができると考えた．上記の通り計算した連想距離は大きいほど連想しやすいことを示す指標であるので，この距離に基づき降順で並べた単語のうち，上位五語を連想語として抽出した．

4. 評価実験

本手法は第二言語学習において単語の心的辞書内の位置情報付与を行うために，学習言語の母語話者における一般的な連想語を抽出することを目指している．提案手法によって，実際に母語話者にとって一般的な連想語を抽出できるのかを，アンケート調査を介して評価した．本来であれば，本稿で学習言語として取り上げた英語の母語話者による評価を行うべきであるが，予備実験として日本人の男女を対象にアンケート調査を行った．回答者は 20 代の男性 5 人，女性 2 人の計 7 人であった．英語に対する知識の不足による誤差を軽減するため，TOEIC 700 点以上のスコア所有を回答の条件とした．アンケートは，刺激語と連想語一語ずつを一組として，刺激語に対する連想語の妥当性を判定するものであった．JACET8000 の単語リストに対して，レベル別の層別抽出によって選んだ 300 語の刺激語に対し，提案手法によって抽出した 5 語の連想語をそれぞれペアとし，1500 組の単語ペアを作成した．使用した単語の品詞の内訳が名詞 4547 語，動詞 990 語であったため，この割合に合わせて，サンプルの 300 語も名詞 246 語，動詞 54 語を抽出した．このペアを無作為に並べ替え，一人当たり 75 組で，品詞の内訳が同じになるよう配慮してアンケートの質問項目とした．調査結果を以下の表 2 に示す．

表 2 評価結果

項目	件数	割合
VALID	182	34.7 %
NOT VALID	284	54.1 %
NOT UNDERSTAND	59	11.2 %
計	525	100 %

全 525 組中，「妥当である（VALID)」と判断されたのは 182 件（34.7%）であった．「妥当でない（NOT VALID)」と判断された 284 組（54.1%）と比較すると少ない結果となった．また 7 人の中で“ VALID ”が“ NOT VALID ”を件数で上回ったのは 1 人のみであった．提示されたペアの内，どちらか一つでも意味を知らなかった件数は 59 件（11.2 %）であった．この結果から，本稿で提案した連想語抽出手法は適当ではない可能性がある．改善策としては，連想距離計算の際のパラメータ変更や，単語間の各関係性の数値化における手法の変更が考えられる．

ただし，このアンケートは前述の通り，本来英語母語話者を対象に行うべきである．心的辞書内の単語配置等の違いを考慮すれば，一概にこの結果が悪いと断じることはできない．以下の図 2 は英語母語話者 3 人（N1〜3）と上記評価実験に回答した日本人 3 人（J1〜3）のデータをグラフ化したものである．回答したアンケートは N1 と J1，N2 と J2，N3 と J3 で共通している．実施人数が少ないため，普遍的な評価とすることはできないが，英語母語話者群のほうが“ VALID ”が明らかに多くなっているのが読み取れる．今後，回答者を増やし統計的な手法での分析を行う等，一般化した結果を見出すことが望まれる．

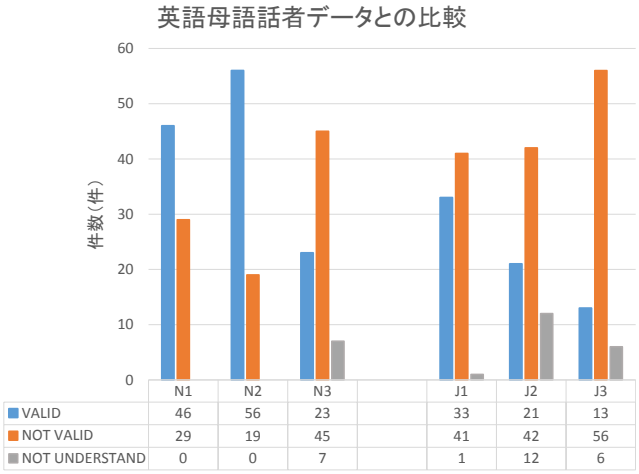

図 2 母語話者との比較

5. おわりに

本稿では第二言語の単語学習において，心的辞書の構築に直目し，学習言語の母語における一般的な連想語を抽出する手法の提案を行った．今回，手法の評価は英語が堪能な日本人へのアンケート調査によって行ったが，手法の目的を考慮すれば英語母語話者による評価を行うことが望ましいため，今後人数の増加とともにより詳しい分析を行っていく必要がある．評価結果の分析次第では，パラメータの変更等，連想距離の計算方法を再考し，より人間の連想に近い語を抽出できるように手法の改善をしていくことが期待できる．

参考文献

[1] 白畑知彦，若林茂則，村野井仁：詳説 第二言語習得研究 理論から研究法まで，研究社 (2010).
[2] 小山義徳：英単語学習方略が英語の文法・語法上のエラー生起に与える影響の検討，教育心理学研究，Vol. 57, No. 1, pp. 73–85 (2009).
[3] Aitchson, J.: *Words in the mind: An introduction to the mental lexicon*, Wiley (2003).
[4] Yokokawa, H., Yabuuchi, S., Kadota, s., Nakanishi, Y. and Noro, T.: Lexical Networks in L2 Mental Lexicon: Evidence from a Word–Association Task for Japanese EFL Learners, *Language Education and Technology*,

Vol. 39, pp. 21–39 (2002).
[5] 箱田裕司，都築誉史，川畑秀明，萩原　滋：認知心理学 Cognitive Psychology: Brain, Modeling and Evidence，有斐閣 (2010).
[6] 阿部純一，桃内佳雄，金子康朗，李　光五：人間の言語情報処理，東京：サイエンス社 (1994).
[7] 秋山哲史，内海　彰：概念間の関係に関する単語の意味空間の性質：コーパス, 構築手法, 文章単位による影響，認知科学，Vol. 17, No. 1, pp. 110–128 (2010).
[8] 池上嘉彦：意味の世界，NHK ブックス (1978).
[9] Woodrow, H. and Lowell, F.: Children's association frequency tables, *The Psychological Monographs*, Vol. 22, No. 5, pp. 1–110 (1916).
[10] 崔　進，小松英二，安原　宏：EDR 電子化辞書を用いた単語類似度計算法，情報処理学会研究報告自然言語処理, Vol. 1993, No. 1(1992-NL-093), pp. 1–6 (1993).
[11] 秋山哲史，内海　彰：ベクトル空間モデルに基づく単語の意味表現の性質，言語処理学会第 13 回年次大会発表論文集，pp. 1082–1085 (2007).
[12] 正薫橘高，将文萩原：単語ベクトル生成法と追加学習可能な言語処理ニューラルネットワーク，電子情報通信学会技術研究報告. NC, ニューロコンピューティング，Vol. 107, No. 542, pp. 391–396 (2008).
[13] Salton, G., Wong, A. and Yang, C.: A vector space model for automatic indexing, *Communications of the ACM*, Vol. 18, No. 11, pp. 613–620 (1975).
[14] 高橋良和，渡部広一，河岡　司：概念ベースを用いた記事間の意味的距離計算方式，電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション，Vol. 106, No. 517, pp. 19–24 (2007).
[15] 長尾　真，佐藤理史，池原　悟，中野　洋，黒橋禎夫：言語情報処理 (言語の科学 9)，岩波書店 (2004).
[16] 大学英語教育学会基本語改訂委員会 (編集委員会)：大学英語教育学会基本語リスト JACET List of 8000 Basic Words，大学英語教育学会 (2003).
[17] Tess, F. and Cristina, I.: Word Association in L1 and L2 : An Exploratory Study of Response Types, Response Times, and Interlingual Mediation, *Studies in Second Language Acquisition*, Vol. 33, pp. 373–398 (2011).