関連語連想における単語親密度の影響

笠原 要[*1]　天野 成昭[*1]

[*1]日本電信電話 (株) NTT コミュニケーション科学基礎研究所

1. はじめに

人間は、質問として与えられた単語 (刺激) に対して様々な種類の関連性を持つ単語を思い浮かべて回答することができる。例えば、単語「母」に対して、被験者が思い浮かぶ単語を制限無く連想させた場合、「過程」、「肝っ玉」等が含まれるであろう。また、制限を与えて連想させた場合は、異なる単語が回答される。単語「母」に対して意味が似ている類義語として「母親」、「母さん」や「おっかさん」等、また、反対の意味を持つ対義語として「父」や「子」等が連想されるであろう。その他にも上位、下位、部分、属性等の様々な関連性にある単語も連想することができる。社会において人間が言語を通じて他者に意思を疎通させるためには、両者が言語が表す意味を共有していることが前提となる。それは、意味の基本単位である単語についても同じであり、ある単語から連想される関連語は、人によって多少の差異はあってもある程度は共通していると予想される。例えば梅本らは、1,000 名の大学生を被験者として 210 語を刺激とした自由連想の実験を行っており、回答された連想語は、平均して 2 割の被験者が一致して回答したと報告している [梅本 69]。

ヒトの単語連想の認知システムの特徴を明らかにすることができれば、単語やその意味がどのように記憶されており、また、どのように利用されているかについての心理学や認知科学研究への知見となるであろう。また、単語の連想をコンピュータを用いて工学的に再現することができれば、ヒトによる単語の関連性判断の代用として利用することが可能である。それによって、オントロジーの自動構築、情報検索や言語処理等のテキストに関する様々な情報処理へ単語の連想シミュレーションを応用し、処理の精度を向上させることができるであろう。実際に、国語辞典やテキストコーパスを用いて単語を多次元空間中に配置を行い、単語間の類似度を計算する単語の意味モデルが提案されている [Hindle 90, Deerwester 90, Schutze 92, 笠原 97]。また、それを用いた知的情報検索システム [Schutze 95]、意味に基づくテキストの分割方式 [別所 01] 等が提案されている。

ヒトの連想の認知システムの特徴を明らかにするため、[笠原 01a] では連想された関連語の性質が調査されている。これは、200 の基本語に対する自由連想および類義・対義の制限連想実験を 100 名の被験者を用いて行い、得られた関連語を刺激語と対にして別の 76 名の被験者により適合 / 不適合の判定を行った結果を分析した研究である。その結果、類義や対義の単語の制限連想実験では、一致して連想する被験者の割合と刺激語に対する関連語として適合と被験者が判断した割合に中程度の相関があることが判明した。類義した単語や反対の意味を持つ単語を思い浮かべる認知プロセスと提示された単語対が類義的、対義的な関係にあるかを判断する認知プロセスには、共通性があることが示唆される。それに対して自由連想の実験では、少数の被験者しか思い浮かべなかった単語であっても、適合実験では多くの被験者が適当な連想語であると判断していた。そのため、連想語の場合には思い浮かべることと、提示された刺激語との単語対が連想しているかを判断する認知プロセスが異る可能性が高いことが示唆されている。

上記の関連語の傾向についてさらに分析を行った結果、被験者が連想する単語の多くはなじみがある単語であり、また、多くの被験者が共通して連想する単語程、その単語親密度が高い傾向にあることが明らかとなった。単語親密度とは、単語から感じられるなじみの度合いの尺度である。本稿では、自由連想と類義語の制限連想において、回答される関連語がどのように単語親密度と関連があるかについてを分析した結果を報告する。また、この知見をコンピュータによる類義語の検索シミュレーションに適用し、単語親密度が一定値以上の単語の中から刺激語に類似した単語を出力することにより、類義語検索の精度を向上させることができたので、併せて報告する。

2. 連想実験

今回新たな分析を行う対象である、[笠原 01a] で行った連想実験の内容について説明する。

2.1 刺激語

選定する刺激語によって関連性の調査結果は大きく異なることが予想される。そのため、大多数の被験者が熟知しているような日常的な単語に関する調査が行われた。

日本語の語彙特性 [天野 00] を用い、約 7 万語よりなじみのある 200 語を刺激語として選定した。上記文献では約 7 万の語彙について、被験者を用いた心理実験を通して得られた単語親密度が付与されている。これは、刺激語が被験者にとってどの程度なじみがあると感じられるかを表した尺度であり、1 から 7 までの値をとる (1:なじみがない, 7:なじみがある)。名詞、動詞、形容詞、形容動詞で親密度が 5 以上の単語 25695 より 200 語を無作為に選出した。ただし、人称・時間・位置に関して相対的に言及する語 (例 : おととい (時間)、自分自身 (人称)、向こう (位置)) や、多義性のある固有名詞 (例 :「アポ

連絡先: 笠原 要　日本電信電話株式会社, NTT コミュニケーション科学基礎研究所　〒 619-0237 京都府相楽郡精華町光台 2-4 email: kaname@cslab.kecl.ntt.co.jp

ロ」(神/チョコ/ロケット)) は除外した。

選定した 200 語を表 1 に挙げる。品詞別では、名詞が 176、動詞が 16、形容詞が 4、形容動詞が 14 となった。これには「楽しみ」のような名詞と形容動詞の両方に入る単語が 11 含まれていた。

表 1: 刺激語リスト (200 語, 提示順)

熊, 建築, 消費, 飛行機, 食券, おかか, 街, 作戦, 踊る, 洋食, 馬, ハムスター, 春夏秋冬, タイガー, 検定, レベル, バーボン, ダイナミック, 引っ越し, ブラウス, 祭り, ボディーガード, 食卓, アルプス, 口座, 無神経, 着く, きつね, ミックス, 爆笑, サンデー, リスト, 美しさ, 関係, 安全, 新聞, シーツ, ツイン, 浮気, 広げる, 投げる, 健康, マッシュルーム, 裁判所, 楽しみ, 年賀, ロードショー, 工場, ワックス, ペット, 風呂場, 動く, 中古車, 感動, 夕日, 政治, 姿, パートナー, 診断, 友情, 運動会, ガーデン, 建てる, かけそば, 注射, 表面, 表情, 住所, マジック, 原作, 事務所, 企画, 入口, ファッションモデル, 恐れる, 写真, 外国, ディナー, 調査, スペース, やわらかい, 作る, やかましい, 氷, シーソー, 迷惑, 解放, 虫歯, 訓練, たらこ, 最適, アドバイザー, 村, 家族, 紙, 問題, 申し込む, ポリス, 集まり, 少年, 食物, 恋, ノンフィクション, おかず, サイレン, カセット, 教科書, ディスコ, チューリップ, 制服, セロリ, キング, ジャンプ, 圧力, 金庫, 大きさ, 戸籍, 中国, 遅い, 映画, フリー, ハート, 洗濯物, 角度, 鈍い, コック, 昔, 哲学, 深夜, 図書, 福祉, ビジネスマン, コットン, ネックレス, 夢, クッション, 選挙, ステーション, デザイン, ジャズ, 釣り, 保育, 運ぶ, 冷房, のこぎり, 対決, アロエ, 正しさ, 答, 多数決, 目立つ, 記号, チャーミング, ひょうたん, ボウリング, 雪ダルマ, 心配, 受話器, 地域, 悪口, パワー, 写る, シロップ, カボチャ, 化粧, バケツ, 冷やす, 七月, バナナ, 戦争, 夕食, 再会, 新車, タイミング, 恐い, めちゃくちゃ, 下げる, 姉妹, 窓, 完了, 焼きソバ, シンプル, 無意識, へそ, ぶた肉, 直線, オアシス, ひつじ, 注目, ボール, スペイン, 体調, 収入, 緊張, 四季, コマーシャル, 考える, スキップ, スーツケース, 倒す

2.2 実験

選定された 200 の刺激語それぞれについて、大学生 100 人 (男性 50 名、女性 50 名) を被験者とした自由連想実験を行った。被験者は刺激語が記載された所定の用紙に、刺激語から思い浮かぶ単語を制限なく記載させた。

人間の認知的連想構造を調査する手段とし自由連想実験を用いる研究 [Deese 65] では、1 刺激語に対して被験者に連想させる単語を 1 語に限定している。複数の単語を回答させると、回答語が直前の回答語に依存する恐れがあるためである。また、複数の刺激語に対して順番に自由回答アンケートを行うと、刺激語間で同様な依存関係が生じうる。そのため、1 名の被験者に対しては 1 つの刺激語を与え 1 つの回答語のみを収集するにとどめることが理想的である。しかし、上記の方法では多数の被験者を募る必要があり、多数の刺激語について関連語の基準データを作成する場合、調査コストが大きくなってしまう問題がある。我々は、コストの問題を考慮して、100 名の被験者それぞれに同一の 200 語の刺激語を提示して、1 刺激語あたり、30 秒以内に思い浮かぶ連想語を複数記入する実験を実施した。

思い浮かべる単語として類義語に制限し、それ以外は上記の自由連想実験と同じ条件で、別の 100 名の大学生を対象として類義語の制限連想の実験を行った。

それぞれの連想実験結果について、同じ刺激語に対して同一の関連語 (連想語・類義語) を何人の被験者が同時に連想したかについて集計した。連想された関連語の中には、同一の単語の表記ゆれと考えら得るものが多く存在した。例えば刺激語「熊」の類義の制限連想結果には、「月の輪熊」、「月の輪グマ」や「ツキノワグマ」出現した。結果の中には、このように比較的容易に同一の関連語の表記ゆれと判定できる場合から、別の単語として区別すべき場合まで様々存在する。日本語において、表記から同一単語であるかどうかを判定することや同一語の複数表記のどれを代表的な表記と決定するかについては統一的なな方法論は確立されていない。そこで、国語辞典を参照とし、44 の規則を独自に設定し、規則にのっとって表記揺れの単語を集計し関連語の回答者数を決定した。決定方法の詳細については、文献 [笠原 01b] を参考とされたい。

集計の結果、自由連想では、刺激語に対して平均 3.2 語の連想語が得られ、類義の制限連想では、平均 1.4 語が得られた。表 2 に連想された関連語の一例を示す。

表 2: 刺激語「街」に対する連想実験結果

類義語	回答者数	連想語	回答者数
町	53	ネオン	18
都市	28	街頭	14
都会	18	人	13
市街	9	街路樹	12
市	4	雑踏	8
...			
社会	1	ゲーム	1
買い物	1	清潔	1
街並み	1	温かい	1

3. 分析

実験結果における連想された関連語の傾向の分析結果を説明する。

まず、日本語の語彙特性 [天野 00] を用い、収集された関連語に単語親密度を対応付けた。[天野 00] は、約 7 万単語の表記約 9 万に対して、40 名の被験者よる心理実験で得られた単語親密度の評定値を含むデータベースである。類義語、連想語それぞれについて、その単語表記をデータベースの表記と比較した。その結果、下記の件数の関連語について単語親密度と対応づけられた。以下は、この限定した関連語に関する分析結果である。

表 3 は、心理実験によって得られた関連語が日本語語彙特性 (「DB」と呼ぶ) にどの程度含まれていたかを示すものである。類義、連想ともに収集した関連語の約 60%程度が含まれていた。また、関連語の単語親密度の値は DB での 7 万語の単語親密度平均に比べて 1 以上高くなっている。DB の評価対象は、国語辞典の見出し語である。辞書は、人間が語の意味を調べるために利用するものであるため、比較的難解な語や日常的にはあまり使われない単語も収録しておく必要がある。その

ため、そのような語彙の単語親密度平均に比べて、関連語の単語親密度の平均が高くなっていると推測される。

表 3: 連想実験による関連語の単語親密度

関連語	全数	DB 中の語数	単語親密度平均
類義語	7,988	4,655	5.75
連想語	25,633	16,810	6.01
(DB)	—	—	4.10

次に、制限連想実験において 200 語の刺激語に対して回答された類義語を共通して回答した人数 (最大 100 名) と、その類義語の単語親密度の相関を図 1 に示す。横軸は、刺激語に対して連想された類義語が 100 人中で何人が連想したかを示す。縦軸は、同一の回答者数の連想された類義語の単語親密度の平均値を表す。連想された類義語の単語親密度がおおむね高いことは先に述べたとおりであるが、さらに、共通に連想される類義語ほど単語親密度が高くなる傾向があることがわかる。このことは、自由連想の結果においても同様である (図 2)。

それぞれの図の直線は、回帰直線を表す。連想語に比べて類義語の回帰直線の勾配の方が急である。類義の制限連想の場合には、共通に連想される語である程その語の単語氏密度が高くなる傾向が強い。この実験では、刺激語の単語親密度が 5 以上に限定にされているため、あまりなじみが無い刺激語に対する連想の結果は明らかではなく、上記の傾向がどのような刺激語についてもあるかを検証するためには、さらなる調査が必要である。

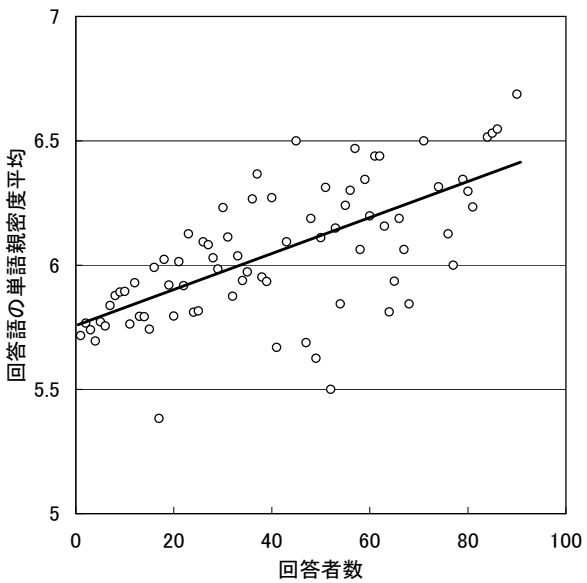

図 1: 連想された語の回答者数と単語親密度の関連 (類義語)

一方、連想される関連語の単語親密度は、刺激語の単語親密度の単語親密度には影響を受けない。図 3 は、200 語の刺激語の単語親密度に対して、連想された類義語の単語親密度の平均値との相関を表し、図 4 は、自由連想による単語の単語親密

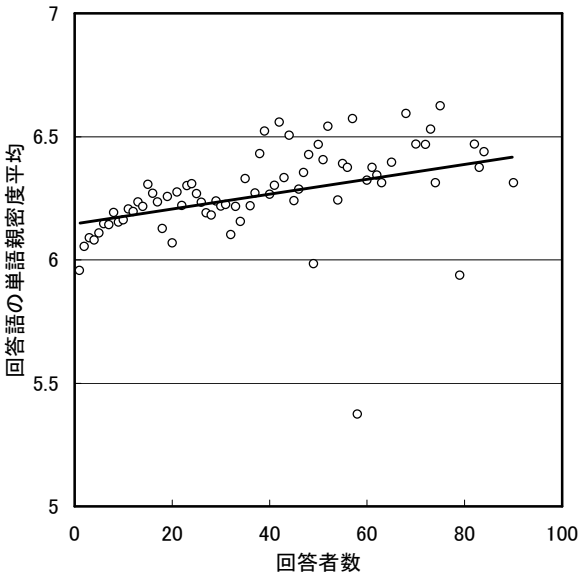

図 2: 連想された語の回答者数と単語親密度の関連 (連想語)

度平均との相関を表す。どちらの場合も、刺激語の単語親密度値は、関連語の単語親密度に影響を与えていない。図 2,4 の相関プロットに比べると、刺激語の単語親密度は、連想する単語の馴染みやすさには影響を与えていないと考えられる。

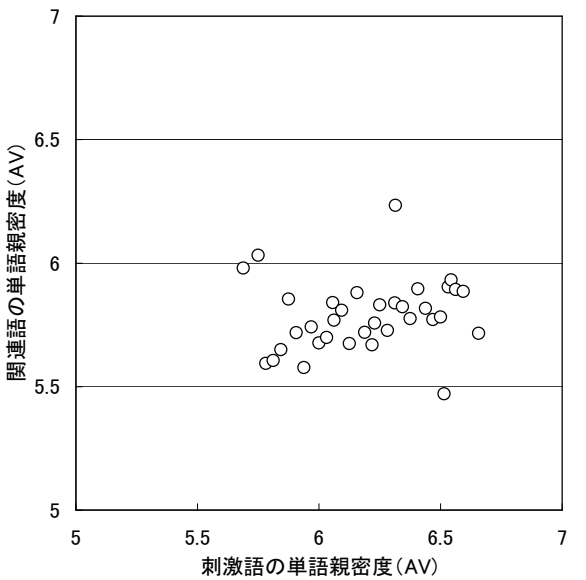

図 3: 刺激語と類義語の単語親密度の関連性

4. 類義語検索のコンピュータ・シミュレーション

多くの被験者が共通して回答する類義語や連想語には、単語親密度が高いことを工学的に再現することにより、コンピュータを用いた関連語検索を精度高く行うことが期待できる。具体

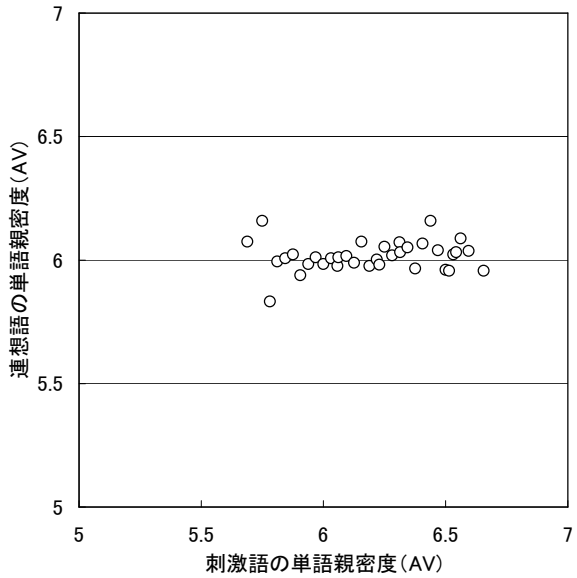

図 4: 刺激語と連想語の単語親密度の関連性

的には、単語親密度が低いなじみの無い単語を検索対象から除いて検索を行うことで実現できる。

ここでは、国語辞典を用いて単語を多次元空間に配置して単語間の類似度を計算する概念ベースの方式 [笠原 97] を用いて単語親密度の効果を検証する。

4.1 概念ベース

概念ベースとは、単語を属性の重みを表わす値 (≥ 0) を要素とするベクトル (「属性ベクトル」) で表現した知識ベースである。概念ベース中の単語 W_i $(i = 1, \cdots, n)$ の属性ベクトル $Word_i$ は以下の通りとなる。

$$Word_i = (v_{i1}, v_{i2}, \cdots, v_{in}) \tag{1}$$

属性として、概念ベース中の単語すべて (n 語) を用いており、概念ベース全体は、属性の重みを要素とする n 行 n 列の行列 (G_1) となる。また、n 語の属性を m 個のカテゴリーに分類するシソーラス (類語辞典) を用い、同じ分類に含まれる属性をカテゴリーに一般化する。

$$Word'_i = (v'_{i1}, \cdots, v'_{ik}, \cdots, v'_{im}) \qquad v'_{ik} = \sum_{l=1}^{n} v_{il} T(l,k) \tag{2}$$

$T(l,k)$ はシソーラスを表わす関数で、l 番目の属性が k 番目のカテゴリーに含まれる時は 1、それ以外は 0 を取る。また、シソーラスで属性を一般化した概念ベース全体は、n 行 m 列の行列 (G_2) となる。属性を一般化する方法にはこれ以外にも主成分分析や特異値分解を用いる統計的モデル [Schutze 92] やそれらを併用するモデル [笠原 02] がある。シソーラスを用いる方法は、250 次元以下では他のモデルよりも類似度計算の精度が低いが、G_0 がスパースな行列であるときに、その性質を保存しつつ次元を圧縮することができる特徴がある。カテゴリー数が十分高いシソーラスを用いれば、精度よく類似度の計算をすることができる [笠原 02]。

属性の重みは、国語辞典の見出し語に対する説明文中の単語の出現頻度に基づいて獲得する。獲得方法の詳細は、文献 [笠原 97] を参照されたい。なお、獲得された属性ベクトルそれぞれについて、個々の重みは正規化しておく ($\sum_{k=1}^{m} v_{ik}^2 = 1$)。

これを用いて、概念ベースに含まれる W_i、W_j $(1 \leq i, j \leq n)$ の類似度 sim $(0 \leq sim \leq 1)$ を対応する属性ベクトル $Word_i$、$Word_j$ のなす角度の余弦で表わす。

$$sim(Word_i, Word_j) = Word_i \cdot Word_j = \sum_{k=1}^{m} v_{ik} v_{jk} \tag{3}$$

現在は、国語辞典 [金田一 88] と、30 万語を 3000 カテゴリに分類したシソーラス [池原 97] を用い、約 9 万語の概念ベースが自動構築されている [永森 00]。

4.2 実験

上記概念ベースを用いて類義語検索のコンピュータ・シミュレーションの実験を行った。検索の方法は、刺激語の属性ベクトルに対して概念ベース中の全属性ベクトルとの類似度計算を行い、類似度の高い属性ベクトルと対応する単語を順次検索された類義語として出力する方法である。

200 語の刺激語に対して概念ベースを用いた類義語検索の出力について、先に述べた類義の制限連想結果から集計した基準の類義語と比較し、シミュレーションの評価値を計算した。基準の類義語としては、2 名以上が回答した単語について異なる被験者 73 名が刺激語に対する類義語として適当である / ないを判定し、半数以上の被験者が適当と判断した単語である。

図 5 は、0 から 1 までの 0.1 きざみの 11 点の再現率において 200 の検索結果の精度を平均した適合率-再現率曲線である。'all' のプロットは、概念ベースに含まれる 88,633 語全てを用いて検索を行った結果であり、'fam5' のプロットは、その内で単語親密度が 5 以上の 26,371 語を用いて検索を行った結果である。どのような再現率においても、親密度が高い単語のみで類義語検索を行った方が辞書のエントリーに含まれる全ての単語について類似度計算に基づく類義語検索を行うよりもはるかに精度良く類義語検索を行うことができることがわかった。

表 4 は、上記類義語検索結果の一部である。刺激語「街」に対して、88,633 語全てを用いて類義語検索をコンピュータで行った場合と、その中で単語親密度 5 以上の 26,317 語を対象として検索した結果の上位を表している。心理実験に基づく基準の類義語は、「都市」、「市街」、「都会」、「タウン」の 4 語であった。なじみのある単語に限定して検索した場合には、検索結果上位 10 語の内で正解が 3 語現われていた。

それに対して、全ての語彙を用いて検索を行った場合には、10 語中で正解がわずか 1 語しか現われていなかった。検索結果の中には、街や街中の意味を表す「街衢」が現われており、言語学的な観点での類義語と言えるがほとんど使われていないこのような単語が現われてしまう。この例からもわかるように、類義であるが馴染みが無い単語を出力しない工夫によって、類義語検索のシミュレーションの精度が向上する。

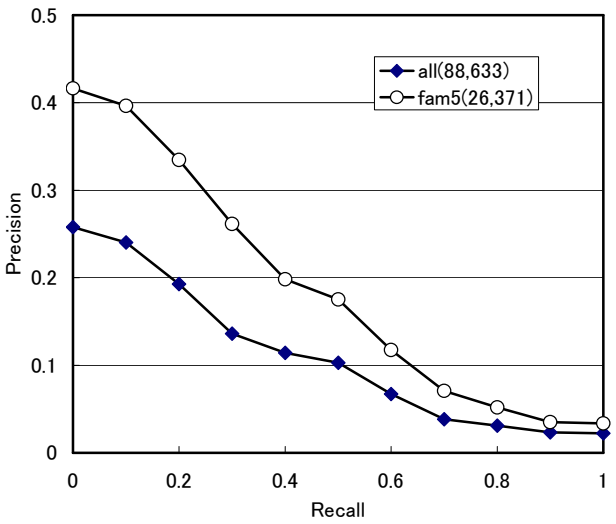

図 5: 類義語検索のシミュレーションの評価結果

表 4: コンピュータによる類義語検索例 (刺激語:「街」)

	語彙数=26,317		語彙数=88,633	
順位	検索語	類似度	検索語	類似度
1	市街	0.665	狭斜	0.758
2	都市	0.637	色里	0.732
3	ベッドタウン	0.596	花街	0.685
4	タウン	0.595	横町	0.679
5	田園都市	0.578	市街	0.665
6	近郊	0.557	街衢	0.658
7	大通り	0.556	遊郭	0.658
8	東海道	0.527	廃市	0.647
9	市	0.521	巷間	0.646
10	繁華街	0.520	洛内	0.643

5. おわりに

本稿では、ヒトが関連語を連想する際に、なじみの高い単語が多く、その傾向は単語の心的属性の一つである単語親密度で分析できることを報告した。自由連想、類義の制限連想のそれぞれにおいて、連想される単語は単語親密度が 5 以上の語が多く、さらに、共通して連想される割合と連想された単語の親密度の関係を回帰直線で説明することができることを述べた。さらに、国語辞典を用いて類義語検索をコンピュータで行うモデルにおいて、単語親密度が高い単語のみを検索の対象とした場合に検索の精度が上昇することを明らかにした。

参考文献

[Deerwester 90] Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R.: Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, Vol. 41, pp. 391 – 407 (1990).

[Deese 65] Deese, J. E.: *The Structure of Associations in Language and Thought*, The Johns Hopkins Press (1965).

[Hindle 90] Hindle, D.: Noun Classification from Predicate-Argument Structures, in *In Proceedings of ACL*, pp. 268–275 (1990).

[Schutze 92] Schutze, H.: Dimensions of Meaning, in *Proceedings of Supercomputing 92*, pp. 787–796 (1992).

[Schutze 95] Schutze, H. and Pedersen, J.: Information retrieval based on word senses, in *Fourth Annual Symp. on Document Analysis and Information Retrieval*, pp. 161–175 (1995).

[永森 00] 永森, 笠原, 松澤 : 概念ベース構築における表記と概念のマッピング手法, 人工知能学会全国大会, 第 14 巻, pp. 163 – 164 (2000).

[笠原 97] 笠原, 松澤, 石川 : 国語辞書を利用した日常語の類似性判別, 情報処理学会論文誌, Vol. 38, No. 7, pp. 1272–1284 (1997).

[笠原 01a] 笠原, 稲子, 金杉, 永森, 加藤 : 単語の関連性判別の分析–類義性，対義性，連想性–, 人工知能学会 ことば工学研究会資料, 第 SIG-LSE-A103-06 巻, pp. 17–27 (2001).

[笠原 01b] 笠原, 永森, 加藤 : 自由回答アンケートにおける単語の表記揺れとその解消, 人工知能学会ことば工学研究会資料, 第 SIG-LSE-A003-10 巻 (2001).

[笠原 02] 笠原, 金杉 : 国語辞典と英和辞典を用いた単語の類似性判別方式, 人工知能学会全国大会, 第 16 巻 (2002).

[金田一 88] 金田一, 池田 (編): 学研 国語大辞典 第二版, 学習研究社 (1988).

[池原 97] 池原, 宮崎, 白井, 横尾, 中岩, 小倉, 大山, 林 (編): 日本語語彙大系, 岩波書店 (1997).

[天野 00] 天野, 近藤 : 日本語の語彙特性, 第 1 巻 単語親密度, 三省堂 (2000).

[梅本 69] 梅本堯夫：連想基準表 -大学生 1000 人の自由連想による-, 東京大学出版 (1969).

[別所 01] 別所克人：単語の概念ベクトルを用いたテキストセグメーテンション, 情報処理学会論文誌, Vol. 42, No. 11 (2001).