単語の関連性判別の分析 —類義性, 対義性, 連想性—

笠原 要 †　　稲子 希望 †　　金杉 友子 ††　　永森 千晴 ††
加藤 恒昭 †††

† 日本電信電話 (株) NTT コミュニケーション科学基礎研究所
†† NTT アドバンステクノロジ　　††† 東京大学大学院総合文化研究科

1　はじめに

本稿では基本的な 200 の単語に対する類義語，対義語，連想語の取得を目的とした被験者への 2 段階のアンケート結果について報告し，関連性判別の性質を考察する．

人間は、刺激として与えられた単語に対して様々な種類の関連性を持つ単語を挙げることができる．例えば「母」という刺激語に対して，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，反対の意味である対義語として「父」や「子」等，刺激語から思い浮かべる語である連想語として「家庭」,「肝っ玉」等を挙げることができる．他にも上位，下位，部分，属性等の様々な関連性があると考えられる．関連語自体については，言語学での意味論の研究などで検討されている．また，このような関連性を判別する人間の認知機能については長期記憶や概念構造等の研究として心理学で取り上げられている．

現状の処理技術では，テキストデータを文字列や単語列としてのみ捉えて処理を行う．そのために，人間が判断すれば関連性があると思われるデータ同士であっても，同じ文字や単語を含まないデータであれば関連がないと判断されてしまう．従って処理結果に欠落が生じる場合があるため，より高度な処理が要求されている．多数の単語に対する関連性の判別をコンピュータで再現し利用することは，1 つのアプローチとして期待されている．

情報検索では，ユーザの質問に適合する文書として質問中の単語を含むものを出力する．そのために，質問と同じ単語は含まないが関連性のある単語を含む文書や内容として関連性がある文書を出力できず，検索の再現性を下げてしまう．そこで，検索文書集合や国語辞典等のテキストコーパスを用いて単語の潜在的な関連性を表現し，人間の感覚に沿った検索を実現するための検討が行われている [Deerwester 90, Schütze 95, 熊本 99]．また，コンピュータを用いてテキストデータを内容に基づいて分割するテキストセグメンテーションの研究でも，テキストを単語列とみなした単純な処理よりも単語の関連性を考慮した処理の方が精度が高いことが報告されている [別所 01]．

さらに，ネットワークを用いたサービスモデルやシステムとして注目されている，動的な B2B を指向する Web Services[丸山 01]，メタ情報の記述を可能とする Semantic Web[荻野 01]，P2P の意味情報ネットワークを指向する Semantic Information-Oriented Network，SION[星合 01] では，組織やコミュニティ内の用語間の関連性を記述したオントロジーの存在を前提としている．現在のオントロジーの構築やオントロジー間の比較は手作業によって行われており，作業コストの削減やシステムの規模を拡張するための方法として，コンピュータによる関連性判別技術は有望と考えられる．

様々な種類の関連性の判別技術を評価し改良するためには，コンピュータによる判別処理結果がどの程度適合しているかを判断する必要がある．評価方法の 1 つとして，処理結果の適合性を直接人間が判断することが挙げられる．しかし，方式を改良する度に人間による評価が必要な点が問題である．もう 1 つの方法としては，あらかじめ人間の関連性判別結果を収集しておき，コンピュータによる判別結果を比較評価する方法である．そのためには，方式の有意性の議論できる規模の単語に対する常識的で網羅的な関連語の基準データが必要である．このような基準データとしては，各種辞典や言語学，心理学等の関連する研究分野で作成されたデータベースが有望である．

国語辞典には多数の見出し語に対して類義語や対義語が掲載されており，これらは辞書学者が事例の収集に基づいて判断した関連語が挙げられている．しかし，全ての見出し語に対して関連語が定義されてはいないので，網羅性

†笠原 要 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 e-mail: kaname@cslab.kecl.ntt.co.jp

には欠けている．また，類語辞典には上位性，下位性，同位性の関連性が網羅的に記述されている．しかし各辞典ごとに分類の方針や実際の設定が異なっており，統一的な見解はない点が問題である．例えば類語国語辞典 [大野 90] では，3 段階の階層による 1000 分類で分類を設定しているのに対し，日本語語彙大系 [池原 97] では，深さが一定ではない階層による 3000 分類を設定している．そのために，様々な類語を比較する研究も行われている [岡本 01]. さらに，国語辞典，類語辞典の両者とも辞書学者や研究者の思考に基づいて関連性が定義されており，一般的な成人が行う関連性の常識的な判別とは一致していない場合も考えられる．

一方，心理学や国語学の研究では，記憶の研究を目的として心理実験やアンケート調査によって基準データが作成されている (例えば，[Deese 65, 梅本 69, 国研 65, 岡本 01])．例えば梅本は，210 語の刺激語に対して大学生 1000 人に自由連想を行わせ，連想基準表を作成している [梅本 69]．岡本らは，大学生 50 名を対象に，100 の名詞からなる刺激語それぞれに対し，連想制限として「上位」,「下位」,「部分・材料」,「属性」,「類義語」,「動作」,「環境」という 7 種類の関連性を与えて得られた概念カテゴリー基準を元にして，概念間の距離の定式化等を試みている [岡本 01]．一方 [国研 65] では，類義語として考えられる単語対について，語の形態や文法機能の差異等の様々な側面から類義語として適しているかをアンケート調査している．

心理学と国語学での単語の関連性の調査を比較すると，関連性判別には 2 つの側面があることが推測される．1 つは，刺激語に関連語を思い浮かべることに関わる人間の機能 (関連語想起機能) であり，もう 1 つは，刺激語と対象となる 1 つの語が提示された時に，対象となる語が関連語として適合しているかを判別することに関わる機能 (適合性判断機能) である．もしも刺激語に対して人間が想起する単語は，適合性が高い順であるならば，関連語想起機能は，刺激語に対する個々の単語との適合性を判断する機能に基づいていると考えられる．しかしながらこのような推測を裏付けるような規模の大きな比較調査は行われて来なかった．コンピュータ上での単語の関連性判別を再現するために基準データとして大規模な関連性判別データベースを作成するためにはまず，関連語想起機能と適合性判断機能の関係を明らかにすることが必要である．両者が同じ機能に基づいているならば，刺激語に対して思い浮かぶ (想起する) 関連語の調査結果のみで関連語の基準データは十分である．一方異なる機能に基づく場合，コンピュータによる関連性の判別の適用先によっては，想起する関連語の調査とともに，刺激語に対して対象とする全ての語彙との関連性の適合性調査が必要かも知れない．

そこで本稿では，200 語の基本的な単語について，2 種類の機能に関わるアンケート調査を行い，両者の機能について分析を行った結果を報告する．関連性の種類としては，検討の第一歩として類義性，対義性，連想性を取り上げた. 2 章では，アンケート調査の方法について説明し，3 章で，調査結果の分析結果について紹介する．さらに 4 章では，関連語の判断における被験者ごとの差異や，関連語想起機能と適合性判断機能の関連について考察する．

2　関連語の収集方法

本章では，まず刺激語の選定について説明し，次に，自由回答アンケート調査について説明する．最後に，自由回答アンケート調査で得られた刺激語に対して関連語の適合性を調査する適合性アンケート調査の方法について説明する．

調査の全体的な方針としては，被験者の言語感覚を実験者の考え方に誘導しないように試みた．また，調査結果を集計する際には，単語の表記揺れや誤字等の問題が必ず発生するが，その解消についてもできるだけ客観的な処理を試みた．

2.1　刺激語の選定

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．今回は検討の第一歩として，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を行った．

日本語語彙特性 [天野 00] を用い，約 8 万語より 200 語を刺激語として選定した．上記文献では約 8 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 含まれている.

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 関連性の説明

今回は関連語として, 類義語, 対義語, 連想語の収集を試みた. 関連語のアンケートを行う際には, 被験者に個々の種類の関連語とはどのようなものであるかを説明する必要がある. ここで問題となるのは, 関連語の定義は明確にはできない点である. 例えば類義語の場合, 任意の単語について似ている度合いがどの程度であれば類義語であるかという境界を広義として説明はできない. さらに, その境界は個人によってどの程度異なっているかも定かではない. 国語辞典には類義語の説明が記述されているが, その内容や例は, 辞典によって多少の相違がある. 被験者自身が考えている関連語の定義をできるだけゆがめないように, 収集に際し特定の考え方に依存しない関連語の説明を行った (表 2). また, 同様の理由で説明には関連語の例を補足として用いなかった.

連想語は, 類義語や対義語と共通することが多いと予想される. そのために, 同一の被験者に対して連想語と類義語, 対義語のアンケートを同時に行うと両者の依存性に基づく結果が生ずる恐れがある. そこで, 連想語に関する 2 種類のアンケート調査は, 類義語と対義語の調査を行った被験者と別の被験者を用いて行った.

2.3 自由回答アンケート調査

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) を被験者とした関連語の自由回答アンケート調査を行った. アンケートの指示書に基づいて, 被験者は刺激語が記載された所定の用紙に, 指定した 1 つの関連性 (類義/対義/連想) において刺激語ごとに頭に思い浮かんだ関連語を記入する. これを 200 語それぞれに対して行わせた.

表 2: アンケート調査における被験者への関連語の説明

関連語	説明
類義語	刺激語と似たような意味を持った単語
対義語	刺激語と反対の意味や対となるような意味を持った単語
連想語	刺激語から思いつく単語, 連想する単語

人間の認知的連想構造を調査する手段として自由回答アンケートを用いる研究 [Deese 65] では，1 刺激語に対して被験者に回答させる単語を 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関連語を複数記入するアンケート調査を実施した．

2.4 自由回答アンケートの集計

調査結果について，刺激語に対して被験者が一致して回答した関連語（類義語/対義語/連想語）の回答者数を集計した．結果の集計においては，表記が全く同じ回答語のみをまとめて，その回答者数を単語の出現頻度とすることを考えていたが，得られた結果 (例えば，表 3) を見ると次のことが明らかとなった．つまり，単語が持つ意味やそれと関連する概念を研究するという立場に立った場合，意味論的，語用論的にほぼ同じとみなして構わないと判断される単語 (例えば，表 3 中の「月の輪熊」「月の輪グマ」「ツキノワグマ」) が異なる複数の表記を持っており，それらを別の単語と扱うよりは，同じ単語の表記揺れであるとしてまとめ上げて回答者数を議論した方がより適切であると感じられたのである．

表 3: 刺激語「熊」に対する類義語の自由回答アンケート結果 (一部)

パンダ (18), 白熊 (12), 白態 (1), 白クマ (1), しろくま (1), ベアー (10), ベア (9), ベア (bear)(1), 月の輪熊 (1), 月の輪グマ (1), ツキノワグマ (1), テディベア (1), テディーベア (1), テディ・ベア (1), 動物 (3), ヒグマ (3)
(括弧内は，回答した被験者数)

しかしながら，回答語表記より同一の単語の表記揺れであるかどうかを適切に判定する規則は存在しないために，客観的に表記揺れを解消することは容易ではない．そこで 44 の規則を順次設定し，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 01] を参考とされたい．

2.5 適合性アンケート調査

表記揺れを解消した自由回答アンケート結果で得られた関連語と刺激語の対を 76 名の被験者に提示して，関連語として適合であるかを判断する適合性アンケート調査を行った．

自由回答アンケート調査で得られた関連語の異なり語数を表 4 に挙げる．

表 4: 自由回答アンケートで得られた関連語数 (異なり語数)

	自由回答アンケート結果		適合性アンケート対象			
関連語	合計	回答者数 1	頻度 2 以上	頻度 1	頻度 0	合計
類義語	7,988	5,079	2,909	600	600	4,109
対義語	6,514	4,285	2,229	596	600	3,425
連想語	25,633	17,418	8,215	1,000	1,000	10,215

基本的には，得られた関連語全てについて適合性の調査を行う必要があるが，関連語数が非常に多いので，すべてを調べることは困難である．また，回答数が 1 の関連語，すなわち，100 人の被験者中で 1 人のみが回答した関連語は表 4 から分かるように非常に多い．そこで，回答者数が 2 以上，すなわち，100 人の被験者の内で 2 人以上が共通して回答した語が関連語として適当であると仮定して，調査の対象とした．上記の仮定が適切であるかを検証するために，回答頻度 1 の関連語の約 1 割を無作為に選択して分析の対象として加えた．また，今回の自由回答アンケート調査で全く回答されなかった単語が関連語としてどの程度の適合性を保有するかを調査するために，3

種の関連性に対して回答語として挙げられた全語彙中の出現頻度が 5 以上の単語中で，注目する刺激語と関連性において頻度が 0 の単語を，アンケート語数の 1 割程度無作為に抽出して調査の対象とした．

上記の内容の関連語，類義語 4,109，対義語 3,425，連想語 10,215 と刺激語とを対にした回答用紙を作成した．76 名の被験者に提示し，指定した関連性にある語として適当であるかを判定させた．判定は，適合している/いないの二択式とした．ただし，知らない語や単語に誤字が含まれていると思われる場合，判定の対象外とした．単語対の提示の順番が調査結果に影響しないように，4 種類の提示順のアンケート用紙を作成し，19 名ずつ 4 グループの被験者に配布した．被験者が記入したアンケート結果については郵送で収集し，ワープロ投入して集計した．

アンケート調査を集計した結果の一例として，刺激語「きつね」に対する自由回答アンケートでの回答語と回答者数，適合性アンケートでの「きつね」に対する回答語の適合者数を表 5 に挙げる．

表 5: 刺激語「きつね」に対する調査結果 (括弧内は “回答者数:適合者数”)

類義語	たぬき (21:11), フォックス (21:57), あぶらあげ (7:25), 稲荷 (5:33), あげ (4:23), キタキツネ (4:31), おいなりさん (3:28), 狼 (3:7), こんこん (3:16), 犬 (3:9), 鹿 (2:6), 化ける (2:16), 詐欺師 (2:10), コヨーテ (2:11), 動物 (1:16), うそつき (1:8), 手ぶくろ (1:2)
対義語	たぬき (80:33), 犬 (3:1), 虎 (2:4), 素直 (1:1), ネズミ (1:0), ぶた (1:2)
連想語	うどん (38:70), たぬき (26:72), 北海道 (15:63), あぶらあげ (11:73), 黄色 (9:61), 化ける (9:72), いなり (8:69), だます (7:68), 北国 (7:68), 神社 (6:63), しっぽ (6:66), おいなりさん (6:70), 山 (6:49), ねこ (5:27), ごんぎつね (5:70), 毛皮 (5:62), 化かす (5:72), コンコン (5:72), 動物 (5:69), 雪 (4:46), 赤い (4:50), エキノコックス (4:23), かわいい (4:47), 北きつね (4:69), 赤いきつね (4:62), そば (4:63), 病気 (3:8), えりまき (3:49), 目 (3:67), 手袋 (3:24), ずるがしこい (3:69), 北の国から (2:61) きつね目 (2:70), 神様 (2:40), 銀 (2:40), あげ (2:65), さぎ (2:24), 葉っぱ (2:44), つり目 (2:70), フォックス (2:71), きつね色 (2:71), ゴン (2:54), 素早い (2:49), 冬 (2:50), ずるい (2:67), 北 (2:55), どんべい (2:64), きつねうどん (2:69), 鋭い (1:35), 面 (1:33), かむ (1:17), こんがり (1:51), みどり (1:17)

3 分析結果

上記の調査結果に基づく分析で明らかとなった結果について紹介する．

3.1 被験者ごとの回答の揺れ

自由回答アンケートで，被験者の回答結果にどの程度のばらつきがあるかを分析した．表 6 は，刺激語に対する回答数の平均及び，被験者 (100 名) と刺激語 (200 語) それぞれを変数とした場合の 95%信頼区間を記載したものである．回答語数の平均が大きいほど信頼区間は大きくなっている．例外は連想語の回答数であり，3 種類の関連語の中で平均が最も大きいにも関わらず，被験者を変数とした時の平均の信頼区間が他の関連語の場合よりも小さくなっている．この結果は，被験者への指示が影響していると推測される．記載する関連語数の上限には制限を設けなかったが，アンケートの説明中に “最低 3 語以上の関連語の記入を目標とする” と指示をした．そのため被験者は，ある刺激語に対してはもっと多くの関連語を記載できるにも関わらず，それまでの刺激語あたりの記載数の平均を無意識に推測して，平均 3 語程度となるように記載した恐れがある．

表 6: 自由回答アンケートでの刺激語に対する回答語数

	類義語	対義語	連想語
平均	1.4096	1.0993	3.2206
95%信頼区間 (被験者)	0.0434	0.0382	0.0218
95%信頼区間 (刺激語)	0.0863	0.0625	0.0951

次に，適合性アンケート結果における 76 名の被験者の回答の揺れについて説明する．自由回答アンケートで得られた関連語の内で，回答者数が 2 名以上の単語と回答者数が 1 の単語の一部，それに人工的に作成した頻度 0 語

を合わせたものを刺激語と併せて表示した際に関連語であると判断した割合 (適合率) を表 7 に示す．一般的には適合性判断の件数が増えるにつれて被験者間の判断の揺れが大きくなると予想されるが，件数が他 2 種の関連性の倍以上もある連想語は，95%信頼区間の値が対義語よりやや大きい程度で類義語より小さくなっている．このことは，連想語の適合性は，他の関連語の適合性に比べて被験者の判断が一致していることを示唆する．

上記のような，被験者の回答語数や適合判断の比較のみで被験者ごとの関連性判別の変動の傾向を推測するのでは十分ではなく，刺激語ごと，あるいは適合性判断の単位となる刺激語-関連語ごとに被験者を比較することが必要である．今後は，このような詳細な分析を行う予定である．

表 7: 適合性アンケートでの被験者の判断の揺れ

	類義語	対義語	連想語
質問数	4,109	3,425	10,215
適合率 (%) 平均	37.9354	21.7107	73.2838
95%信頼区間 (被験者)	4.5395	3.1388	3.5813

3.2 関連語の想起と適合性判断の相関

刺激語が与えられた際の関連語想起機能と，関連語としてふさわしいかを判断する適合性判断機能の相関について，３種類の関連性それぞれに対して検討する．ここ では，自由回答アンケートでの回答者数を前者の，適合性アンケートでの適合者数を後者の機能の指標とする．

2.5 章で述べた通り，適合性判断の調査は，自由回答アンケート調査で得られた関連語の内で，回答者数が 1 の単語の一部を除き人工的に作成した頻度 0 語を数百から千語追加した，類義語 4,109 語，対義語 3,425 語，連想語 10,215 語で行った．その場合の個々の関連語においての自由連想回答者数と，その関連語の適合判断者数の相関係数は，それぞれ，0.465，0.418，0.294 であった．この数値のみより推定すると，類義語と対義語は関連語としての思い浮かび易さと，関連語としてのふさわしさは中程度の正の相関があるが，連想語については相関があまりないと推測される．

次に，回答者数の同じ関連語について，適合者数を平均化して対応関係を調べた結果を図 1, 2,3 に示す．個々の点が，回答者数に対する適合者数の平均を表し，縦のバーは，平均の 95%の信頼性区間を表す．類義語と対義語については，これらの図は，相関係数から得られる推測を支持している．すなわち，多くの人が共通して思い浮かぶような関連語である程，関連語としてふさわしいと思う傾向がある．また，類義語は，回答者数が半数を越すと適合者数はほぼ 6-7 割に達しており，単調増加の傾向は少なくなる．対義語の場合も，半数を越すと適合者数は大きくなるが，その適合者数の変動は大きい．

連想語は，自由回答アンケートでの回答者数が低い語であっても，高い適合者数となっていることが，相関係数を他の種類の関連語よりも下げている原因であることがわかる．100 人中ただ 1 人のみが回答した連想語についても，76 名中の約 6 割が連想語としてふさわしいと考えていることになる．つまり連想語の場合は他の関連語と異なり，あまり思い浮かばないような関連語でも，提示されれば妥当であると判断しやすい傾向があることを示している．

図 4 は，上記の推測を裏付ける実験結果である．人工的に作成した頻度 0 の自由連想回答の関連語に対する適合者数の相対頻度の分布を表している．個々の単語について，刺激語と関連性があるかを実験者で判断することは行っていないため，偶然に関連語が含まれる可能性がある．類義語や対義語については，8 人未満が適合していると判断しているにとどまり，被験者によっては関連していると思う語もあるが，半数以上が適合していると判断する語は存在しない．それに対して連想語では，半数以上の被験者が適合していると判断した語が多くあり，類義や対義よりも多くの単語に対して関連性があると人間が思う傾向にあることがわかる．

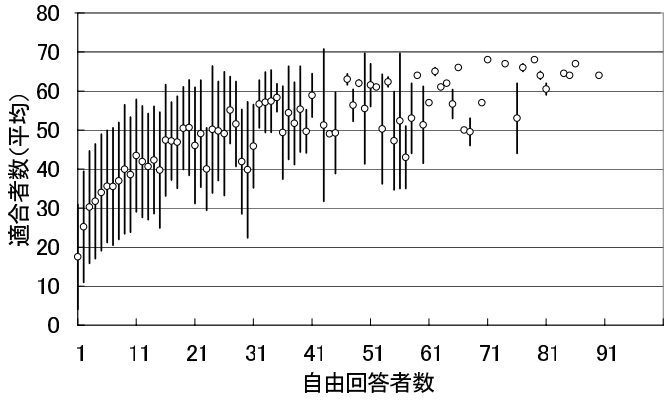

図 1: 類義語の回答者数と適合者数の関係

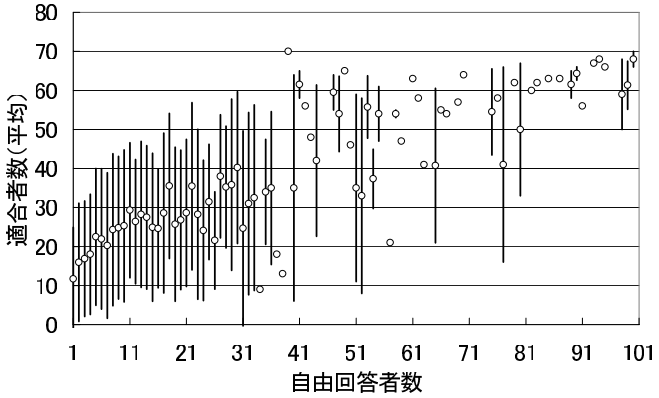

図 2: 対義語の回答者数と適合者数の関係

4　考察

本章では，分析結果より推測される関連語想起機能と関連語判断機能のつながりについて考察する．そして，どのように人間の関連性判別の基準データを作成するか，また，実験の規模をどのようにスケールアップするかについて述べる．

4.1　関連語想起機能と関連語適合性判断機能

関連語の想起や適合性判断の被験者ごとの一致性と関連語の妥当性が対応しているならば，前章の分析結果では，類義語や対義語の想起のしやすさと適合性の判断の度合は相関が高いことになる．想起と適合性判断の機能では，記憶中の同じ語彙を対象として，関連性があるかどうかの判定を同一の機能で行っている可能性が示唆される．

それに対して連想性では，1 人の被験者のみが回答した回答語に対しても多くの人が適合であると回答している．この原因の 1 つとしては，想起する機能と適合性を判断する機能が異なっていることが考えられる．想起は単純に記憶中の連想語のリストを参照するのみであるのに対して，適合性の判断では刺激語との因果関係を推論するならば，機能が異なっていると言えよう．回答者がいなかった単語に対しても連想の適合性が高いことは，また，1 つの刺激語に対して適合していると判断される単語が非常に多いことを示しいる．記憶の構造の推測という側面から見た場合，今回明らかになった適合性に関する刺激語と連想語の密なつながりに対する解釈としては，単語同士が直接的に密に結合している場合と，多段に結合しているが直接的にはそれほど多くは結合していない場合の 2 種類が考えられる．これを明らかにするためには， 想起性や適合性以外の尺度を設定した調査が必要となるであろう．

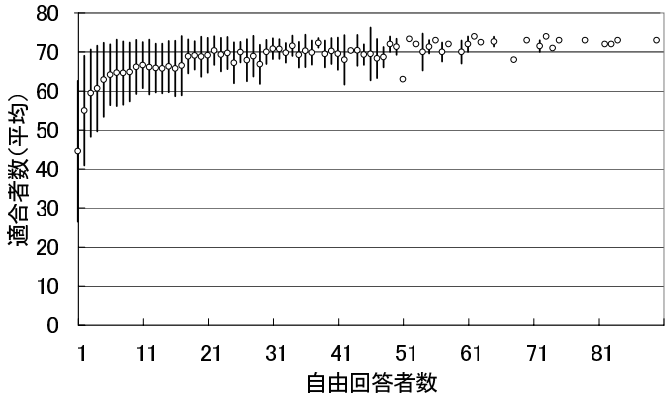

図 3: 連想語の回答者数と適合者数の関係

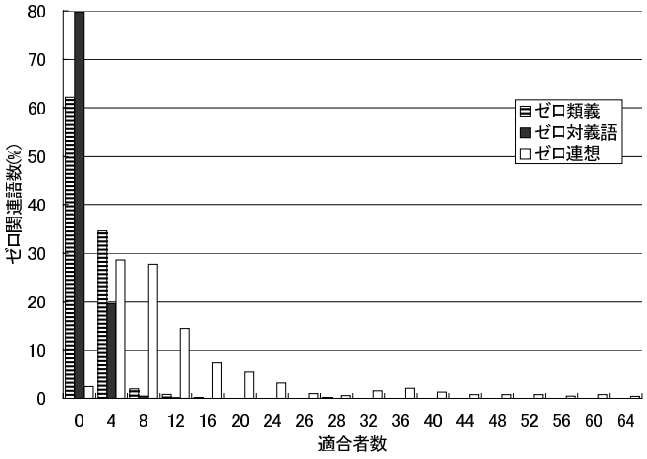

図 4: 回答者数 0 の関連語の適合者数分布

4.2　2 つの調査結果の例外的な関連性

3.2 章では，マクロ的な分析を行ったが，それだけではわからないミクロ的な視点の分析を行う．

図 5,6,7 は，関連語の自由回答アンケートでの回答者数と，適合性アンケートでの適合者数のバブルプロットである．自由連想回答者数の多い／少ない，そして，適合者数の多い／少ないという特徴によって，関連語を大きく 4 群にわけることができる．その中でも自由連想回答者数，適合者数がともに少ない群は，そもそも関連語としてふさわしくないので検討から除外される．残された 3 群に関して考察する．

[1] 自由連想回答者数:多 - 適合者数:多

表 8 の例の通り，関連語であることが常識的に妥当な単語が現れている．しかし，これらの関係は国語辞典や類語辞典には必ずしも記載されていない．例えば「やかましい」に対する類義語「うるさい」を日本の代表的な国語辞典 3 冊 [金田一 88, 新村 92, 松村 92] で参照すると，文献 [金田一 88] のみが類義語として記載されており，多数の人が共通して考える類義語の全ては国語辞典中に記載されていないことを示唆している．

また類義語では,「フリー」に対する「自由」のように，外来語と翻訳された日本語の対応が多いことが特徴的である．外来語を言語に直してその訳語を検索することは容易であるので，この場合には自動的な類義語の取得が期待できる．

[2] 自由連想回答者数:多 - 適合者数:少

表 8: 自由連想回答者数:多 - 適合者数:多の関連語

刺激語	類義語	回答-適合	刺激語	対義語	回答-適合	刺激語	連想語	回答-適合
フリー	自由	90-67	入口	出口	100-72	口座	銀行	90-75
やかましい	うるさい	86-70	ノンフィクション	フィクション	100-70	ペット	犬	84-75
ステーション	駅	84-68	遅い	速い	100-69	写る	写真	82-74
パワー	力	85-67	中古車	新車	100-66	オアシス	砂漠	79-75

多くの被験者が共通して思い浮かんだが，刺激語と対にして提示した場合は適合しないと判断された関連語が現れる領域である．図 5,6,7 を見ると，対義の場合のみ現れている．

表 9 は，典型的な対義語の例である．対義のみこの群に単語が現れる特異性として，以下の 2 点が挙げられる．対義語を思い浮かべることの難しい刺激語が多く，被験者は無理に回答した恐れがある点である．例えば,「熊」や「ホテル」のような単語は関連語を思い浮かべることは難しいと思われる．そのために，強いて思い付く対義語は回答者の共通性が高いが，改めて適合性を問うと，妥当とは思わないという理由が考えられる．

もう一つは，刺激語と対義語の関係は複雑であるためと考えられる.「生」と「死」のように一方の否定が他方になっている対義語は定義が明確である．しかし「白い」に対して「黒い」や「赤い」のように否定関係ではない対義語や,「国立」,「公立」,「市立」のような対立関係が無い対比的な関係もある.「豚肉」に対する「牛肉」では，想起する時の定義と適合性を判定する時に適用する定義が大きく異なっていたために，この群に含まれるような結果となったと思われる．

表 9: 自由連想回答者数:多 - 適合者数:少の関連語

刺激語	対義語	回答-適合
きつね	たぬき	80 - 34
豚肉	牛肉	77 - 16
教科書	ノート	65 - 24
水	氷	65 - 20

[3] 自由連想回答者数:少 - 適合者数:多

基本的には，思い付きにくいが，提示されれば適合している関連語を含む群である．この群と，自由連想回答者数も適合者数も低い関連語の群を分離することが重要である．そのための方法の 1 つとしては，自由回答アンケート調査の被験者数を増やすことにある．単純に想起の確率が低いだけならば，これにより，回答者数を高めることができる．

一方,「夕日」に対する類義語「沈みゆく太陽」のように，被験者数を増やしても，おそらく一致する回答者はいないような関連語もある．これは指示書に「単語」を記載するよう指示したにも関わらず，名詞節が記載された，いわば回答の制約に違反する例である．

自由連想の回答者が意識的あるいは無意識的に回答の制約を感じほとんど挙げなかったが，一度提示されれば関連語として適合する単語はこれだけではない．例えば,「緊張」に対する連想語「緊張感」は，' 感' のみの文字列の差異であり，連想語としてそもそも思い浮かべないが，上記と同様に，提示されれば適合すると多くの被験者は判断する．極端なケースが，刺激語に対して同じ語を回答するケースである．連想の記憶構造の研究 [Deese 65] では，刺激語自身は決して回答しないが，暗黙の回答語として考慮すべきと主張している.「緊張」に対する「緊張感」は，暗黙の回答語に近い関係にあると言えよう．このような回答は，単純に被験者数を増やしても単純には回答者数は増加しないと思われる．そのために，類義語や連想語の適合性をコンピュータで再現する場合には,「感」,「風」,「的」のような意味を大きく変化させない接尾辞が結合し得る刺激語対して接尾辞が付与した語は，無条件で関連語とする機能が必要である．

表 10: 自由連想回答者数:少 - 適合者数:多の関連語

刺激語	類義語	回答-適合	刺激語	対義語	回答-適合	刺激語	連想語	回答-適合
入口	はいり口	1-65	安全	デンジャラス	1-62	焼きそば	ソース焼きソバ	1-76
夕日	沈みゆく太陽	1-62	やかましい	物静かな	1-61	収入	サラリー	1-75
広げる	広くする	1-60	深夜	真昼間	1-60	緊張	緊張感	1-74
正しさ	公正さ	1-59	答え	クエスチョン	1-59	考える	思索	1-73
姿	身なり	1-58	やかましい	寡黙	1-57	答	返答	1-73

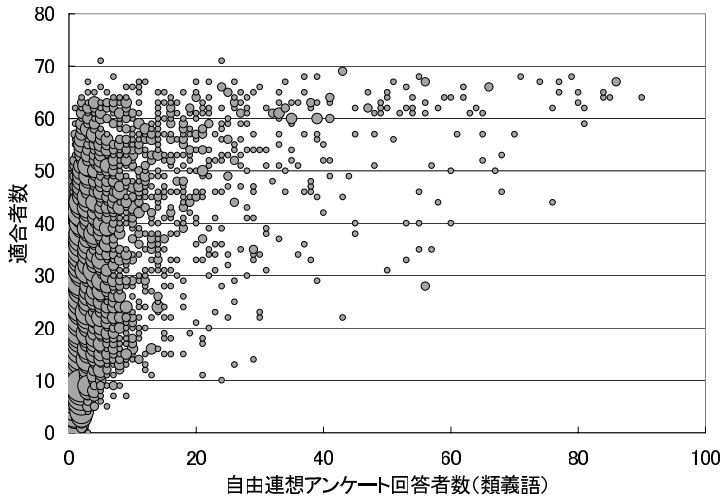

図 5: 類義語の回答者数と適合者数の関係

5 おわりに

本稿では，単語の関連性の工学的再現の研究を行うための基準データ作成を目的とした，単語の関連性の 2 種類のアンケート調査の方法と，その分析結果について紹介し，考察を行った．

類義性と対義性に関しては，刺激語に対して被験者が思い浮かべる関連語の傾向と，刺激語と関連語を提示した時の適合性の傾向に相関があり，想起と適合性判断は記憶中の同じ語彙や判断機能に基づくものであると推測される．

また，連想性では，どのような刺激語に対しても関連語として適合しやすいことが明らかとなった．連想の因果関係のような自由回答者数や適合者数以外の尺度による調査が必要と思われる．

今後は，これらの分析結果をいかして基準データを決定し，それを用いて国語辞書やコーパスを用いた単語のベクトル空間モデルによるシミュレーションの精度を向上させる予定である．また，単語の関連性判別に関わる人間の認知機能の解明を行うために，より多くの刺激語に対する基準データの作成を進める予定である．

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).

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

[天野 00] 天野, 近藤：日本語の語彙特性, NTT データベースシリーズ, 三省堂 (2000).

[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).

[梅本 69] 梅本尭夫 : 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).

[大野 90] 大野, 浜西：類語国語辞典, 角川書店, 第 4 版 (1990).

[岡本 01] 岡本, 石崎：概念間の距離の定式化と既存電子化辞書との比較, 自然言語処理, Vol. 8, No. 4, pp. 37–54 (2001).

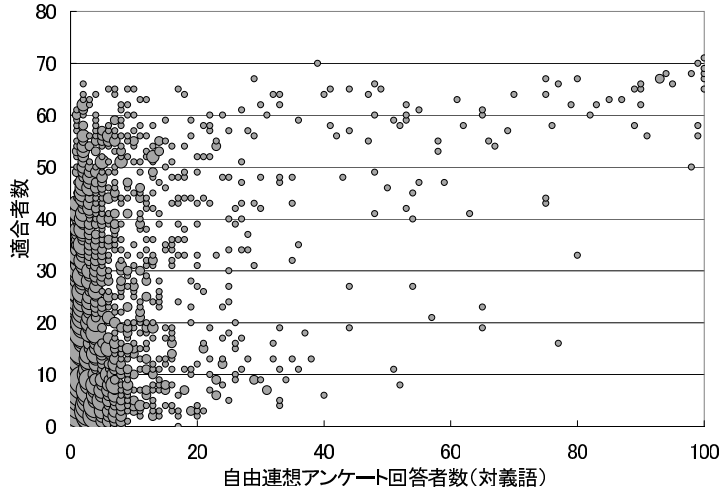

図 6: 対義語の回答者数と適合者数の関係

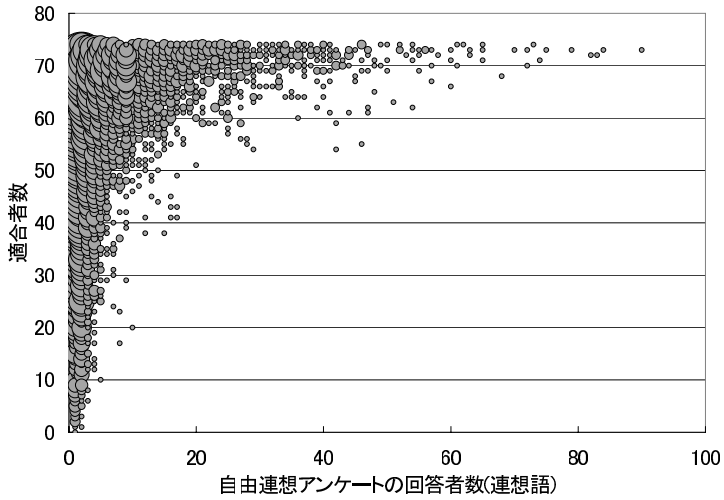

図 7: 連想語の回答者数と適合者数の関係

[荻野 01] 荻野達也 : セマンティック WEB の現状と課題, in *DBWeb2001*, pp. 71 – 77 (2001).

[笠原 01] 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).

[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).

[熊本 99] 熊本, 島田, 加藤 : 概念ベースの情報検索への適用—概念ベースを用いた検索特性の評価—, 情処研報, 第 SIG-ICS 115 巻, pp. 9 – 16 (1999).

[国研 65] 国立国語研究所 : 類義語の研究, 国立国語研究報告 28, 秀英出版 (1965).

[新村 92] 新村出 (編) : 広辞苑, 岩波書店 (1992).

[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).

[星合 01] 星合, 小柳, ビルゲ, 久保田, 柴田, 酒井 : 意味情報ネットワークアーキテクチャ, 電気情報通信学会論文誌, Vol. J84-B, No. 3, pp. 138 – 145 (2001).

[丸山 01] 丸山, 小坂, 浦本 : Web Services による動的な電子少取引の実現 - SOAP/WSDL/UDDI -, 情報処理学会誌, Vol. 42, No. 7, pp. 643 – 653 (2001).

[松村 92] 松村, 三省堂 (編) : 大辞林, 三省堂 (1992).