動詞連想概念辞書の構築とその特徴

寺岡 丈博† 岡本 潤‡ 石崎 俊†

†慶應義塾大学大学院政策・メディア研究科, {teraoka,ishizaki}@sfc.keio.ac.jp
‡慶應義塾大学 SFC 研究所, juno@sfc.keio.ac.jp

1. はじめに

コンピュータの自然言語処理技術は現在までに著しい発展を成し遂げてきたが，人間に近い言語意味理解を行えるまでの精度を求めるには困難性がある．なぜなら単語の品詞や文法などの言語学的な情報だけではコンピュータが人間のように言語を扱うには不十分であるからである．人間は言葉を話す或いは書く際にそれらの言語学的な情報だけでなく，言葉の背景にある膨大な情報を一般的な知識として利用している．即ち，コンピュータが人間と同様に文脈に基づいて言葉の意味を理解するためには知識とも言うべき複雑で膨大な情報を体系化したものが必要である．

そのアプローチとして単語に含まれている意味を表すための概念の収集と構造化が挙げられ，一例として連想概念辞書がある．連想概念辞書とは，人間の直感に基づいた大規模な連想実験のデータに基づいて定量化した連想距離を用いて言葉の背景にある情報を体系化したもので名詞を中心として構築されており[5]，重要文の抽出などに応用されている[6]．この連想概念辞書を構成する連想距離は人間の記憶における単語間関係を表しており，言語データにおける係り受け関係から単語と単語の近さを表しているモデル[4]と比べて，より人間の記憶に基づいた連想という要素から成り立っている．このような視点から連想概念辞書は，既存のコーパスやシソーラスとは異なる面を持っている．

ここで人間が日常で使用する文脈の中で動作や変化を表す品詞である動詞は意味理解の重要な役割を担っていることから，動詞を中心とした連想概念辞書，即ち動詞連想概念辞書の必要性が考えられる．そこで本研究では動詞を刺激語とした連想実験を行い，得られたデータを用いて刺激語と連想語における単語間距離である連想距離を定量化することで動詞連想概念辞書の構築を試みた．

本稿では動詞連想概念辞書の構築について述べた後に約 5 億文の Web テキストから自動的に構築された大規模格フレーム[3]と比較し，動詞連想概念辞書の特徴について言及していく．

2. 連想実験

小学校の国語の教科書で扱われる 1123 語の動詞[1]を基本動詞と見なした上で，サ変動詞の除外や他動詞の優先などで減らし 798 語とした．これらの基本動詞の一部に対して，「位置移動動詞」「所有移動動詞」「感情動詞」「知覚動詞」「構築動詞」「破壊動詞」という意味的な分類を行った．この分類は連想語に及ぼす動詞の種類の影響の有無を明確にし，動詞連想概念辞書の今後の拡張に生かすためである．そして後述の連想課題の組み合わせを考慮し，計 54 語の代表的な動詞を刺激語と設定した．また連想課題はフレーム意味論における深層格の一部を参考にしている．深層格とは，構文の表層的な情報を表す表層格とは異なり，構文中で動詞と特定の意味関係を持つ格であり，これに基づいた連想実験データを利用することで動詞まわりの知識を体系化できる可能性が考えられる．また，幾つかの連想課題については複数の深層格を統合させて被験者の理解の簡易化を図り，表 1 のような 10 種類の連想課題を設定して連想実験を行った．尚，この連想実験には慶應義塾大学湘南藤沢キャンパス（SFC）の学内ネットワークを利用したオンライン上で稼動する実験システムを用いており，1 刺激語 1 連想課題につき被験者を 40 人としている．被験者は SFC に所属する学部生と大学院生を対象とした．

表 1 連想課題の内容

連想課題	意味内容
動作主	動作を行う主体
対象	動作の対象
始点	動作の始点・起点
終点	動作の終点・目標
時点	動作が行われる時刻・時間
場所	動作が行われる場所・空間
手段	動作を行うための道具・材料
様相	動作の様態・様子・程度など
理由	動作の理由・原因
目的	動作の目的

3. 連想距離の定量化

連想実験の被験者が 20 人の規模で構築した動詞連想概念辞書[7]の時と同様な手法で連想距離の定量化を図るため，連想距離 D を連想頻度の逆数 F，連想順位 S，連想時間を対数で表した値 T の線形結合で表わす．

$$D = \alpha \times F + \beta \times S + \gamma \times T \quad (1)$$

F は連想された頻度の逆数を表し，連想した人数に補正値を加えた値で被験者数を割った値であり，補正値を分母に加えることで正規化を行っている．これにより被験者数を大幅に増加させた時に，連想した人数が少ない場合でも F の値が極端に大きくなるのを防ぎ，連想距離 D の極端な変動を抑えることが可能となっている．また連想順位 S は各被験者が連想した語の順位の相加平均である．ここで，連想実験システムで刺激語と連想課題が提示された時から被験者が連想語を入力し始めるまでの時間を連想時間と見なし，その相加平均を対数で表した値 T を用いることで被験者ごとに生ずる大幅な個人差を

解消している．このようにして正規化を行い，式(1)の係数 α，β，γ を求めるために目的関数（式(2)）と制約式（式(3)～(5)）を設定し，線形計画法を用いて目的関数の値 Z を最小にする時の係数の最適解を求める．

$$Z = c_1 \times \alpha + c_2 \times \beta + c_3 \times \gamma \quad (2)$$

$$\begin{cases} a_{11} \times \alpha + a_{12} \times \beta + a_{13} \times \gamma = D_1 & (3) \\ a_{21} \times \alpha + a_{22} \times \beta + a_{23} \times \gamma = D_2 & (4) \\ \alpha, \beta, \gamma \geq 0 & (5) \end{cases}$$

制約式を(a_{11}, a_{12}, a_{13}, D_1) = (20/21, 1, 1, 1)，(a_{21}, a_{22}, a_{23}, D_2) = (10, 9, 5, 10)とした時、目的関数において(c_1, c_2, c_3)=(10, 8, 3)としてシンプレックス法によって最適解を求めた．その結果(α, β, γ) = (7/10, 1/3, 0)の解を得た（式(6)）．T を計算する際に被験者ごとに生じる大幅な個人差などの影響を小さくしたが，連想頻度の逆数 F と連想順位 S の値と比べると，ばらつきが大きいため目的関数の c_3 を低く設定したためである．ゆえに最適解を求める上で信頼性の低い γ が 0 となったと考えられる．
このようにして刺激語と連想語における連想距離を定量化し，動詞連想概念辞書を構築した．

$$D = \frac{7}{10} \times F + \frac{1}{3} \times S \quad (6)$$

4. 動詞連想概念辞書

4.1 連想語数と異なり語数

構築した動詞連想概念辞書は，刺激語 54 語に対して連想語数が 22992 語，異なり語数が 9441 語となっている．連想語数とは連想された語の全ての合計数であり，異なり語数とは刺激語と連想課題の組み合わせが異なっていたとしても同じ語が連想されている場合は同じ単語として数えた合計数である．図 1 より明らかに「動作主」「時点」「場所」「様相」の連想課題に関して連想語数と異なり語数の差が顕著であり，異なり語数が連想語と比べてかなり少なくなっている．これは色々な刺激語から多くの同一の語が連想されていることを表している．反対に「始点」「終点」「理由」「目的」の連想課題は連想語数と異なり語数の差が小さい関係であることから，各刺激語に対して連想課題特有の語が連想されている．以上のことから，連想課題によって連想語が受ける刺激語の影響の度合いが異なっていることが分かる．

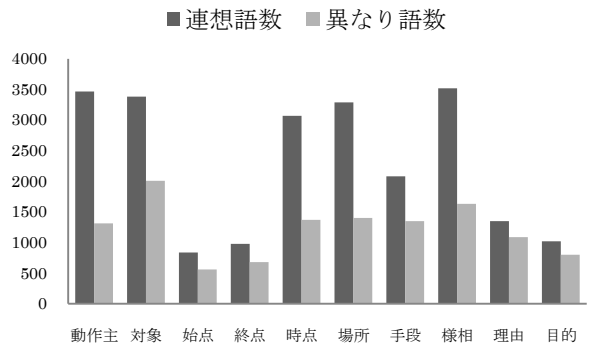

図 1 連想課題ごとの連想語数と異なり語数

4.2 連想語と連想距離の特徴

図 2 は反意語関係（双対関係）にある「借りる」と「貸す」について動詞連想概念辞書から「動作主」「対象」「様相」「場所」における連想語のデータを抽出して連想語と連想距離の関係を図に表したものであり，連想距離の値が小さい程，刺激語動詞と連想語が近い関係になっている．

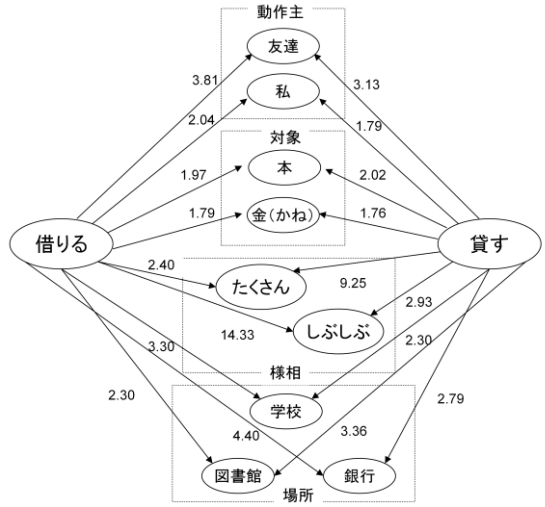

図 2「借りる」と「貸す」の関係

図 2 から分かるように反意語の関係にある刺激語にも関らず，「借りる」と「貸す」の「動作主」「対象」「場所」に対して同じ語が連想されており，それらの連想距離の大小の関係も似ている．つまりこれらの動作を行う「動作主」や動作が行われる「場所」，そして動作の「対象」について被験者が同じようなことを共通して連想したことが示唆されている．また動作の様態や程度などを表す「様相」に注目すると“たくさん”と“しぶしぶ”の連想距離の大小関係が入れ替わっていることが分かる．このように人間の心理的な面が連想語と連想距離の関係によって表わされていると考えられる．

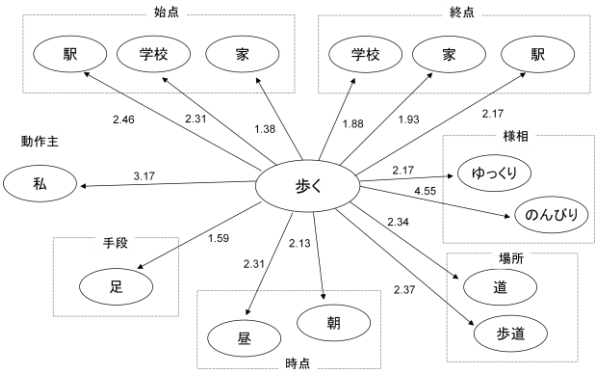

図 3「歩く」と連想語の関係

また，図 3 は「歩く」に対して「動作主」「始点」「終点」「時点」「場所」「手段」「様相」の各連想課題についてそれぞれ連想距離が短い連想語と連想距離の関係を図に表したものである．この図より「“私”は

“家”から“学校”まで“ゆっくり”歩く」や「“朝”“駅”まで“歩道”を“のんびり”歩く」などの意味合いが取れた場面を連想距離の短い連想語で構成することができる．更に図 2 でも「“私”は“図書館”で“本”を“たくさん”借りる」や「“友達”は“学校”で“金（かね）”を“しぶしぶ”貸す」など図 3 と同様に意味合いが取れた場面を作成できる．これは即ち動詞連想概念辞書の連想語が動詞と意味的な関係を持つ深層格の情報を表していることが言え，このことから動詞まわりの背景にある知識を連想語と連想距離が反映していると考えられる．

5. 動詞連想概念辞書と大規模格フレームの比較

5.1 大規模格フレーム

大規模格フレームとは Web 上の約 5 億文の日本語テキストから自動的に構築され，約 5 万用言から成っている．用言と関係する名詞を整理した格フレームを各用言について検索することが可能であり，現在は β バージョン[1]が公開されている．尚，用言とその直前の格要素を格フレーム収集の単位とすることで，同じ表記の用言で複数の意味や用法がある場合において異なる格フレームを別々に得ることを可能にしている[2]．このように大規模格フレームは用言（ここでは動詞）を中心とした格情報をまとめたものであることから，本研究で構築した動詞連想概念辞書と比較する対象とした．

5.2 順位相関係数による比較

大規模格フレームは主に「ガ格」「ヲ格」「ニ格」の他にも「ノ格」や「カラ格」など数々の格フレームがあるが，必ずしも同じ格フレーム内にある格要素がそのまま同じ深層格であるとは限らない．そのため，深層格を基にして連想課題を設定している動詞連概念辞書とそのまま比較する上で意味関係がほとんど一致する格フレームを選ぶ必要がある．ゆえに連想課題「動作主」に対して格フレーム「ガ格」を，連想課題「対象」に対して格フレーム「ヲ格」を比較対象とし，動詞連想概念辞書の刺激語動詞 54 語について大規模格フレームの「ガ格」と「ヲ格」のデータをそれぞれ抽出した．尚，ここではノンパラメトリックな手法として，スピアマンの順位相関(Spearman’s rank correlation)を計算し，両者の関係を比較検討していく．

5.2.1 連想課題「動作主」と格フレーム「ガ格」

動詞連想概念辞書の「動作主」に関する連想語と大規模格フレームの「ガ格」の要素にある名詞の内，共通する語を連想距離の小さい順とスコア（格フレームを含む文の数）の大きい順にそれぞれ上位 10 語でソートした時の順位相関係数(r_s)を求めた．動詞 54 語中 10 語以上共通語を得られたのは 32 語であるが，その理由は 5.3 節の考察で述べる．因みにかなり正の相関関係がある動詞が 11 語($r_s \geqq 0.4$)，やや正の相関関係がある動詞が 11 語($0.4 > r_s \geqq 0.2$)，ほとんど相関関係がない動詞が 9 語($0.2 > r_s \geqq -0.2$)，かなり負の相関関係がある動詞が 1 語($-0.4 > r_s \geqq -0.7$)であり，表 2 はそれらの一部を表したものである．

表 2 「動作主」と「ガ格」の順位相関係数(r_s)

動詞	r_s
貸す	0.687879
建てる	0.636364
走る	0.090909
買う	0.090909
滑る	-0.457580

表 2 のように「動作主」と「ガ格」において「貸す」や「建てる」が正の相関関係が強い理由としては動詞に対する「動作主」が特有であるためである．一般的に「動作主」或いは「ガ格」に関しては「私」や「人」などが多いのだが，「貸す」や「建てる」では，他の動詞に関する「動作主」ではあまり見られない語がある．例えば「貸す」では「銀行」が，「建てる」では「大工」が挙げられ，このような語がある場合は他の動詞と比べて相関関係が高い．一方で「走る」「買う」は相関関係がほとんどないが，これは動詞連想概念辞書の「私」や「親」の連想距離が小さく上位 10 語に入らなかったためである．逆に大規模格フレームでは「車」などの無生物や「馬」や「客」などの順位が低い．さらに「滑る」においても大規模格フレームでは「足」や「車」，「ギャグ」などの順位が小さく，「私」や「自分」は比較的順位が高いのに対して動詞連想概念辞書は「私」や「人」の順位が小さく，「足」や「車」の順位が大きくなっているため「滑る」は負の相関関係がかなり生じている．これらの原因としては，動詞連想概念辞書と大規模格フレームのデータ元の相違が挙げられる．前者は，連想実験であるため被験者はどうしても自分自身の場合をあてはめて連想語を答えてしまう傾向にあり，自分を表す「私」や身近な存在である「親」などの連想距離が小さい．それに対して後者は Web 上の日本語テキストであるため，ブログなどから「私」などのデータが得られる場合があるが，一部なため前者ほどの大きな特徴にはならないと考えられる．

5.2.2 連想課題「対象」と格フレーム「ヲ格」

5.2.1 項と同様にして動詞連想概念辞書の「対象」に関する連想語と大規模格フレームの「ヲ格」の要素にある名詞について順位相関係数(r_s)を求めた．10 語以上の共通語があったのは動詞 54 語中 46 語の動詞であり，強い正の相関関係がある動詞が 5 語($r_s \geqq 0.7$)，かなり正の相関関係がある動詞が 19 語($0.7 > r_s \geqq 0.4$)，やや正の相関関係がある動詞が 8 語($0.4 > r_s \geqq 0.2$)，ほとんど相関関係がない動詞が 10 語($0.2 > r_s \geqq -0.2$)，やや負の相関関係がある動詞が 4 語($-0.2 > r_s \geqq -0.4$)である．表 3 は一部の動詞の順位相関係数を表したものである．動詞連想概念辞書の「対象」と大規模格フレームの「ヲ格」は，「動作主」と「ガ格」の場合よりも相関関係が強い動詞が多い．これは先の 5.2.1 項で述べたような「動作主」に関して連想実験が起因となる偏りがあまり生じないためである．また表 3 より，正の相関が非常に強い「傷める」に注目すると「心」「足」「腰」「膝」「手」など，共通の上位語 10 語は全

[1] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/caseframe.html

て身体の一部に関する語であることから，「傷める」という動詞によって「身体の一部」という「対象」の属性が決まることが考えられる．他にも「耕す」や「建てる」に関しても前者は「畑」や「田んぼ」などの「土地」，後者は「家」や「ビル」などの「建物」のように，動詞に対して「対象」となる語はある程度まとまった属性を持っていると考えられる．そのためこのような動詞に関しては，動詞連想概念辞書と大規模格フレームのデータ内容は共通する語が多くなり，且つそれらの順位に関して相関が強くなる．これらの動詞とは逆に「借りる」と「貸す」は相関関係がほとんどない．「借りる」と「貸す」のそれぞれにおいて順位が上位となっている語を確認すると，「借りる」では「鉛筆」や「ノート」が動詞連想概念辞書では順位が 3 位と 4 位で低いのに対して大規模格フレームでは 10 位と 9 位となっていて順位が高い．また「貸す」では「ペン」と「ノート」が動詞連想概念辞書では 3 位と 5 位だが大規模格フレームではそれぞれ 10 位と 6 位になっている．この違いの理由としては，連想実験の被験者が大学生と大学院で学生であるため「借りる」と「貸す」における「対象」として鉛筆やノートなどの文房具を連想する被験者が多かったと考えられる．

表 3 「対象」と「ヲ格」の順位相関係数(r_S)

動詞	r_S
傷める	0.936364
耕す	0.818182
建てる	0.700000
借りる	0.169697
貸す	-0.00303

5.3 考察

動詞連想概念辞書と大規模格フレームを 5.2 節で求めた「動作主」と「ガ格」，「対象」と「ヲ格」の順位相関係数だけでは動詞連想概念辞書の特徴を一概には割り出せないが，少なくとも上記の範囲において相関がある程度強い動詞に関しては動詞連想概念辞書の連想距離と大規模格フレームのスコアが順位を決める上で同じ役割を果たしていると考えられる.しかし，これは同じ動詞において動詞連想概念辞書と大規模格フレームの両方に十分なデータが無ければ当てはまらず，実際は該当しない動詞が幾つも存在したことをないがしろにはできない．そこで 5.2.1 項にて動詞 54 語中 10 語以上共通語を得られたのは 32 語であり，残りの 22 語については得られなかった理由について考察を述べたい．まず始めに動詞連想概念辞書と共通する単語が少なく 10 語に満たない動詞が 16 語あったことが挙げられる．これは「砕く」「傷める」などの動詞に対して「ヲ格」を取る文は多いが「ガ格」を取り得る文が少ないためである．次に「ちぎる」「聞く」などの動詞に対して「ガ格」が全く無かったことが挙げられる．これは自動構築する際に元になった文にこれらの用言が含まれていなかったと考えられる．また 5.2.2 項についても同様の理由が当てはまるであろう．

ここで順位相関係数では比較しきれない点について考えると，動詞と意味関係を持つ計 10 種類の連想課題について連想語と連想距離から成り立っている動詞連想概念辞書は，格フレームとその要素から成り立っている大規模格フレームと比べて動詞まわりの意味情報をより明確にまとめていることが言える．なぜならば，後者は構文中の格フレームごとに名詞を分けることにより「デ格」や「ニ格」の内容を前者の「場所」や「手段」にまで区別することが不可能であると考えられるためである．また，後者は構文中の名詞だけを格要素にしているが前者は連想課題の「様相」で形容詞や形容動詞なども連想語としているため，これらの理由からも前者は動詞まわりの意味情報を詳しく反映させることを可能にしている．

以上から動詞連想概念辞書の特徴としては，連想実験から得られたデータと大規模格フレームの約 5 億文のデータを比べるとかなり小さい規模だが，それでも連想距離で表わされる動詞と連想語の関係は大量のテキストデータから計算して求められたスコアを表す名詞と用言の関係に似ていることが挙げられる．また深層格に基づいて設定された連想課題により，動詞まわりの意味情報を体系化していると言えよう．

6. おわりに

現段階の動詞連想概念辞書は 1 刺激語が 40 人あたりのデータであるが，刺激語に関しては 54 語のため応用を視野に入れた上でデータの拡充を考えなければならない．そのため今後は更に多くの基本動詞について連想実験を行いたい．また概念辞書として連想語の表記を修正する必要があると考えており，例えば「私」などの人称代名詞に関しては「人間」などの対応する概念を設けて表記を変換するつもりである．最後に，今回の大規模格フレームとの比較では共通する語の上位 10 語を比較対象に設定したが，上位 20 語位まで範囲を広げて順位が下位の語との関係についても比較を検討し，今後の課題として全体的な評価をしていきたいと考えている．

参考文献

[1] 甲斐睦朗, 松川利広, “語彙指導の方法 - 語彙表編 - ”, 光村図書 (2001).

[2] 河原大輔, 黒橋禎夫, “格フレーム辞書の漸次的自動構築”, 自然言語処理, Vol.12, No.2 (2005).

[3] 河原大輔, 黒橋禎夫, “高性能計算環境を用いた Web からの大規模格フレーム構築”, 情報処理学会 自然言語処理研究会 171-12(2006).

[4] 工藤拓, 松本裕治, “相対的な係りやすさを考慮した日本語係り受け解析モデル”, 情報処理学会論文誌, Vol.46, No.4(2005).

[5] 岡本潤, 石崎俊, “概念間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(2001).

[6] 岡本潤, 石崎俊, “連想概念辞書の距離情報を用いた重要文の抽出”, 自然言語処理, Vol.10, No. 5(2003).

[7] 寺岡丈博, 岡本潤, 石崎俊, “動詞連想概念辞書の構築とその応用”, 第 7 回情報科学技術フォーラム 一般講演論文集(2008).