

ヒトによる多義的形容詞に対する類似性の評価データベース構築 — 「長い」と「短い」の事例から —

西内沙恵†
筑波大学[院]

浅原正幸
国立国語研究所

加藤祥
国立国語研究所

1. 本研究の概要

多義的に用いられる現代日本語形容詞の用例を『現代日本語書き言葉均衡コーパス』(以下, BCCWJ) などから収集し, 用例間の類似性を調査した。調査により, (1) のような, ある形容詞の多義的使用 (<空間>か<時間>か) がどのくらい似ていると判定されるかがわかる。

- (1) a. 人がやっとすれ違える程度の、細くて暗く【長い】トンネル。<空間> (PN2d_00004,12570)
b. 【長い】打ち合わせだった。気がつくと2時間以上経過していた。<時間> (PM25_00084,51940)

クラウドソーシングで大規模調査した類似性評価のデータベースを構築した。調査結果は, クラウドソーシングの生データ・集計データ・デンドログラムとして公開している。「長い」と「短い」を事例に公開内容を報告する。

2. 多義的形容詞の類似性調査

調査は, (2) の手順でデータセットを作成し, (3) の方法で実施した。調査協力者は, 20 歳以上の Yahoo! Japan ID 所有者である。

- (2) a. BCCWJ および『国語研日本語ウェブコーパス』から用例を抽出する。
b. 日本国語大辞典・分類語彙表などから語源と語義を付与する。
c. 述定用法・装定用法・副詞的用法の用法別に分類する。
d. 語義と用法別の用例分布を整理し, データセットを作成する。
(3) a. クラウドソーシングで数千名規模の調査協力者を募集し, 調査を実施する。
b. データセットの用例について, ある1文(指標文)から見てほかの文(判定文)が似ているかどうかを6段階でチェックしてもらう(図1)。
c. 回答を語義・用法別に集計する。

[] 内の語の意味が例文とどのくらい似ているか判定してください。

(例文) : 3: 人がやっとすれ違える程度の、細くて暗く
指標文 【長い】トンネル。 <空間>

判定文
14: フランスからの独立をめざして【長く】戦う
この島の人々はフランス人というよりコルシカ人
としての意識が強いのだという。

0: まったく違う 1
2 3
4 5: まったく同じ

類似度: 0%

35: 仕事中は、動きやすく足が【長く】見える
パンツが大活躍!

0: まったく違う 1
2 3
4 5: まったく同じ

類似度: 0%

図1 「長い」の調査画面の一部

(2b) では, 本研究に先駆する山崎・柏野(2017)と加藤ほか(2019)を参照している。山崎・柏野(2017)は, 分類語彙表の多義語に代表義を付与した。加藤ほか(2019)は, BCCWJ の新聞・書籍・雑誌データの自立語に人手で文脈的に妥当な分類語彙表番号に基づく意味情報を付与した。調査では, これらの代表義情報付与済み多義語リストと意味情報付与済み均衡コーパスのデータも援用する。

クラウドソーシングとは, 不特定多数の作業員への簡単な仕事の外注である。Yahoo!クラウドソーシングやランサーズなどのクラウドソーシングサイトで, 非常に小さい単位のタスクを数千人規模で依頼できるサービスが提供されている。大量の語彙・用例に対する大人数のアンケート調査を安価に実施できる(浅原2019)。

なお, タスクの性質上, 指示文や用例を読まずに回答できる。このような不適切な回答を排除するために, 調査協力の同意確認を兼ね, 「同意する」・「同意しない」をランダムに配置したチェック設問を設けた。「同意しない」を選択した回答者は「落選」となり, 回答が回収されない¹⁾。

† snishiuchi[at]ninjal.ac.jp

¹⁾ 「落選」した作業員は, 「長い」で5人(回答者数723人), 「短い」で3人(回答者数733人)であった。

3. 理論的視座とデータ活用

本研究は、ヒトの判定に基づく心理実験的な手法をとる。田中（1987）によれば、文脈に依存しない意味を得るためには、文脈が満たされたデータから帰納的に推測しなければならない。通常、ヒトは、単語を文脈から孤立した形で用いないためである。言語分析の一般的なプロセスは、経験世界の範例である *e* (exemplar) をもとに可能世界の理論値 *E* (Exemplar) を推測し、*X* (カテゴリー) を探る方法である。この方法は、言語学的なテストによる論証に相当すると考えられる。対して、本研究がとる心理実験的手法では、田中（1987）と同様、私たち一人一人が経験世界で学習した *Se* をもとに *X* と *E* を追究する（図2）。*Se* もまた、「単語の意味に対する言語的直感を実証的に研究する際のデータ・ベースとなり、また、私たちの *X* 理解のベースになっているものである。」（田中 1987²: 127）

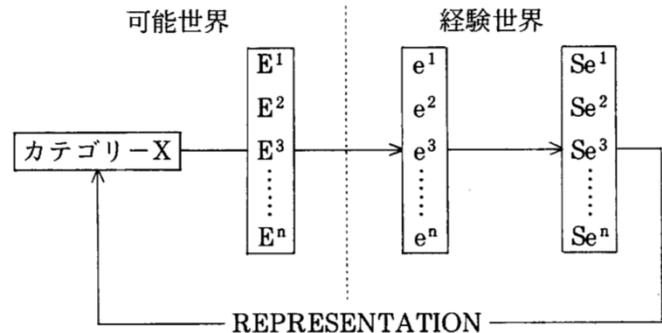


図2 語彙意味論のカテゴリー／
イグゼンプラー・モデル（田中 1987:127）

多義語分析の主要な目的は、多義の認定・プロトタイプの意味の認定・派生関係の解明である。いずれの手法としても、文法・語彙・論理の観点から言語学的テストが編まれてきた。ただし、早瀬（2018）に指摘されるように、テストには欠点がある。適用範囲が限定されたり機能が剛柔であったりと、ただ一つのテストでは妥当性が懸念されることが多い。現状では、多様な手法による並列的な検証が必要である。しかし、言語学的テストは、容認性判断などの違いから研究者に応じて恣意性の高い結論が導かれる。そこで、多義間の類似性評価が役立てられると考えられる。

類似性については、中本ほか（2004）や李ほか（2007）で、多義の認定への活用可能性が論じられている。また、西内ほか（2020）で用例の双方向評価がプロトタイプの意味認定の基準になりうると主張した。さらに、類似の認められやすさが多義間の派生関係の分析に活用できると考えられる。以上により、類似性評価の調査が多義語の分析に役立てられる。

本研究で提案する評価データベースは、言語学的なテストの課題を（4）のように方法論の観点からも解決できる。まず、研究者の容認性判断や調査協力者群の特性などの違いに関係なく、同じ環境での調査が可能である（4a）。次に、質的な論証で一般性が担保されない場合に、類似性評価の量的データを補完できる（4b）。さらに、共時的な多義性は通時的な意味拡張の一時点に過ぎない。再調査可能な手法を確立することで、時間の経過に応じて多義性が異なっている可能性を検討できる（4c）。最後に、これまでの多義研究では、一つの研究で扱える対象が数語であった。語の特性に応じて分析手法が臨機応変に選択されていたことが一つの要因だと考えられる。類似性評価が量的調査として統一的に機能する（4d）。大規模調査の計量から、高い再現性での実証が期待される。

- (4) a. サービスを利用した不特定多数の作業員への外注であり、再調査が可能である。
- b. 言語学的なテストによる質的な論証を量的に補完しうるデータである。
- c. 通時的な多義と共時的な多義が相違する可能性を検討できる。
- d. 複数の語を統一的な手法で扱える。

4. 評価データベースの構築と公開

調査結果は、クラウドソーシングの生データ・集計データ・デンドログラムを Google Sites (<http://bit.do/tagimsubi>) において公開する。本研究では、「長い」と「短い」を事例に公開内容を報告する³。「長い」、「短い」各 36 例の全順序付き 2 つ組 1260 対を、対毎に 50 人（延べ 2100 人：異なり「長い」723 人、「短い」733 人）実施した結果である。

² 田中（1987）は、多義間の接点を探るコア理論について論じている。コアは、西内ほか（2020）などで研究するプロトタイプの意味と厳密には異なる概念であるが、同様の心理実験的アプローチが有効と考えられる。

³ 2019 年 2 月 5 日 8 時に開始し、同日 15 時 25 分に終了した結果である。

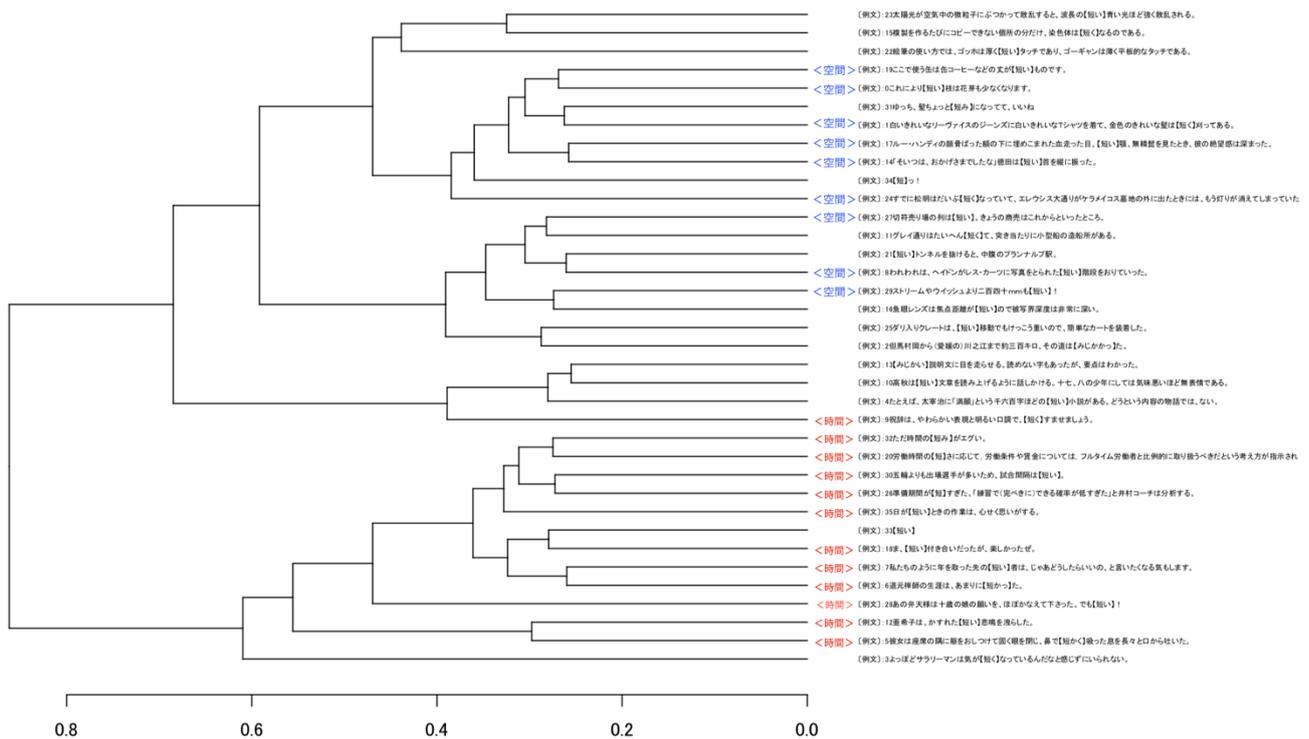


図3 「短い」のデンドログラム

5. おわりに

本研究は、多義の言語直感を実証研究に活用することを目的に、多義間の類似性を調査し、その結果を公開した。Google Sites (<http://bit.do/tagimusubi>) に、クラウドソーシングの生データ・集計データ・デンドログラムを掲載している。多義間の類似性評価が、多義語分析の狙いである多義認定・プロトタイプの意味認定・派生関係解明に役立てられる。本研究で提案する評価データベースは、言語学的なテストの問題を補完・解決できる。現在「長い」と「短い」を含む33語を調査済みである。順次調査結果を公開していく。

謝辞

本研究は、国立国語研究所コーパス開発センター共同研究プロジェクトおよび時間生成学（言語による時間生成）、国立国語研究所所長裁量経費 2018, JSPS 科研費 18H05521, 19K00591 によるものです。

文献

- 浅原正幸 (2019) 「クラウドソーシング結果の可視化手法と統計処理」 『日本言語学会第 158 回大会予稿集』 379-384.
- 加藤祥・浅原正幸・山崎誠 (2019) 「分類語彙表番号を付与した『現代日本語書き言葉均衡コーパス』の書籍・新聞・雑誌データ」 『日本語の研究』 15(2): 134-141.
- 田中茂範 (1987) 「多義語の分析—コアとプロトタイプ—」 『茨城大学教養部紀要』 19: 123-158.
- 中本敬子・野澤元・黒田航 (2004) 「動詞「襲う」の多義性—カード分類と意味素性評定に基づく検討—」 『日本認知心理学学会第 2 回大会発表論文集』 38.
- 西内沙恵・加藤祥・浅原正幸 (2020) 「語義間類似度の双方向評定に基づくプロトタイプの意味の解明—クラウドソーシングを用いた量的調査による多義的形容詞分析—」 『日本認知言語学会論文集』 20: To Appear.
- 早瀬尚子 (2018) 「言語表現の意味とその指示対象」 早瀬尚子 (編) 『言語の認知とコミュニケーション—意味論・語用論、認知言語学、社会言語学—』 29-44. 開拓社.
- 山崎誠・柏野和佳子 (2017) 「『分類語彙表』の多義語に対する代表義情報のアノテーション」 『言語処理学会第 23 回年次大会発表論文集』 302-305.
- 李在鎬・鈴木幸平・永田由香 (2007) 「動詞「流れる」の語形と意味の問題をめぐって」 『計量国語学』 26(2): 64-74.

関連 URL

- コーパス検索アプリケーション『中納言』 <https://chunagon.ninjal.ac.jp/>
『分類語彙表』 https://pj.ninjal.ac.jp/corpus_center/goihyo.html
Yahoo! Japan クラウドソーシング <https://crowdsourcing.yahoo.co.jp>