『分類語彙表』に対する反対語情報付与

荻原 亜彩美　森山 奈々美　浅原 正幸 *　加藤 祥　山崎 誠
津田塾大学　国立国語研究所　国立国語研究所　国立国語研究所　国立国語研究所

1. はじめに

国立国語研究所では『分類語彙表増補改訂版』(国立国語研究所 2004)（以下、分類語彙表）を中心とした言語資源整備を進めている。分類語彙表は「語を意味によって分類・整理したシソーラス (類義語集)」で、意味の区切り行を入れて 101,070 件からなる。語句に対して、5 桁の数字からなる分類番号、2 桁以下の数字からなる段落番号、2 桁以下の数字からなる小段落番号、2 桁以下の数字からなる語番号が付与されている（表 1)。語句はレコード ID 番号によって一意に指定できるだけでなく、分類番号・段落番号・小段落番号・語番号によっても階層的に一意に指定できる。

本研究では、分類語彙表中の反対語を枚挙し、反対語の程度情報を整備したので報告する。

分類語彙表の分類では、類義語だけでなく、反対語・対義語なども同じカテゴリに分類される。表 1 の例では、「有性」⇔「無性（むせい)」や「同性」⇔「異性」などが、反対語に相当する。しかしながら、「同性」「異性」に関しては、どちらも「両性」というカテゴリーに属していることもあり、反対語か否かの判定は作業者により揺れるだろう。何を反対語としてとらえるかは、人によって異なる。例えば二律関係だけでなく、三律関係を反対語としてとらえる人もいる。また反対語の中にも、文脈によって置き換え不可なものもある。このような情報を評価するために、分類語彙表から反対語対候補を収集し、「反対語としての認識」「文脈による置き換えの可否」についてクラウドソーシングを用いた調査による分布情報を付与する。

表 1　分類語彙表の構造

類	部門	中項目	分類項目	分類番号	段落番号	小段落番号	語番号	レコード ID 番号	見出し
体	自然	生物	生物	1.5300	1	2	1	61426	有性
体	自然	生物	生物	1.5300	1	2	2	61427	無性（むせい）
体	自然	生物	生物	1.5300	1	3	1	61428	両性
体	自然	生物	生物	1.5300	1	3	2	61429	同性
体	自然	生物	生物	1.5300	1	3	3	61430	異性

関連するデータとして、日本語単語類似度データベース (Sakaizawa and Komachi 2018)、日本語類似度・関連度データセット (猪原・内海 2018) など、類似度のデータがあるが、分類語彙表に基づく反対語のデータベースを新たに構成することで、多角的な語義の評価が可能になると考える。

2. 反対語データベース構築作業

2.1 1 次作業：分類語彙表からの反対語対候補の抽出

反対語データベース構築作業は、分類語彙表データベースからの反対語対候補の抽出（1 次作業）とクラウドソーシングによる反対語らしさの評価（2 次作業）の 2 段階による。

分類語彙表から反対語対候補を抽出する作業は、4 人の作業者により分担して行った。まず、表 1 の同じ小段落番号の語句群から反対語対候補を抽出した。次に段落番号が同じ語句まで拡大して、再度反対語対候補を追加した。この際、小段落番号で抽出漏れがあった場合には抽出した。

本作業は 2017 年 6 月に開始し、2018 年 11 月に終了した。作業にあたり『三省堂 反対語対立語辞

* masayu-a@ninjal.ac.jp

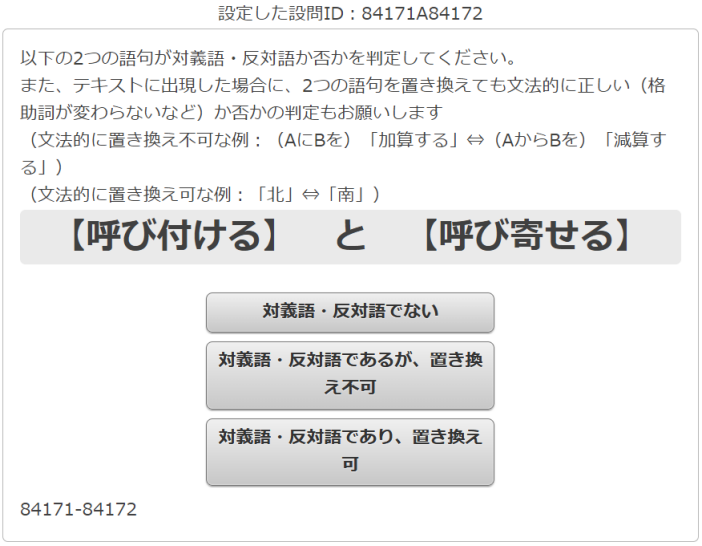

図 1　クラウドソーシング画面例

典』(三省堂編修所 2017) を参照した。3 節の基礎統計で詳細に示すが、1 次作業で 7658 件の反対語対を抽出し、このうち小段落一致は 3405 件、段落番号一致（小段落番号不一致）は 4253 件であった。

2.2　2 次作業：抽出した反対語対の評価

1 次作業で抽出した反対語対 7658 件（小段落番号一致 3405 件、段落番号 4253 件）に、抽出されなかった小段落番号一致のランダムな単語対 4342 件を追加し、 あわせて 12000 件について、Yahoo! クラウドソーシングを用いて反対語対の評価を行った。1 語対あたり 20 人の作業者に判断を依頼した。作業者は図 1 に示す画面で、「対義語・反対語でない」・「対義語・反対語であるが、置き換え不可」・「対義語・反対語であり、置き換え可」の 3 択のいずれかを選択する。

「対義語・反対語であるが、置き換え不可」の例として、格が変化する（AにBを）「加算する」⇔（AからBを）「減算する」を呈示した。「対義語・反対語であり、置き換え可」の例として、「北」⇔「南」、「暑い」⇔「寒い」を呈示した。

本作業は 2018 年 12 月 17 日 08:03 に開始し、同日 21:25 に終了した（13 時間 21 分）。1597 人（異なり数）の作業者が従事した。

3.　基礎統計

本節では、構築したデータの基礎統計を表 2 に示す。1 語あたり 20 人分の回答を得て、回答数の比率を割合で示す。1 次作業で反対語とみなした語については約 35% が「反対語でない」と判断され、フィラーとして追加した反対語でない語については、約 84% が「反対語でない」と判断された。反対語か否かの判定において、1 次作業で「小段落番号一致」で抽出したものと「段落番号一致」で抽出したものとの差は 1.07% と小さかった。一方、「反対語であり置き換え可」の回答数は、「小段落番号一致」が 43.47% であるのに対して、「段落番号一致」が 37.89% であり、差が 5.58% と大きかった。

次に語ごとの回答の割合について検討する。図 2 に 1 次作業で反対語対として認定した語ごとの「反対語であり置き換え可」の割合の頻度を示す。100% （20 人中 20 人）置き換え可と回答された例は、「同姓」⇔「異姓」・「既婚」⇔「未婚」・「暖流」⇔「寒流」の 3 例であった。図 3 に「反対語であるが置き換え不可」の割合とその頻度を示す。図 4 に「反対語ではない」の割合とその頻度を示す。グラフの

表 2　反対語データベースの基礎統計

	反対語でない	置き換え不可	置き換え可	回答数	語数
1 次作業で反対語（小段落番号一致）	34.32%	22.19%	43.47%	68100	3405
1 次作業で反対語（段落番号一致）	35.39%	26.71%	37.89%	85060	4253
1 次作業で反対語でない	83.76%	8.37%	7.86%	86840	4342
調査対象すべて	52.59%	18.79%	28.61%	240000	12000

ように連続的に変化しているため、「反対語であるか否か」「文脈において置き換え可か不可か」をきれいに分離することが困難であることがうかがえる。

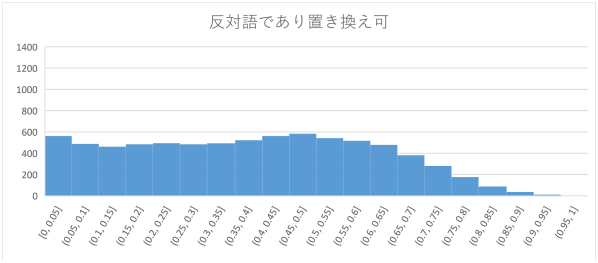

図 2　「反対語であり置き換え可」の割合と頻度

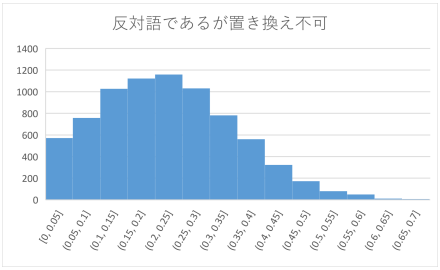

図 3　「反対語であるが置き換え不可」の割合と頻度

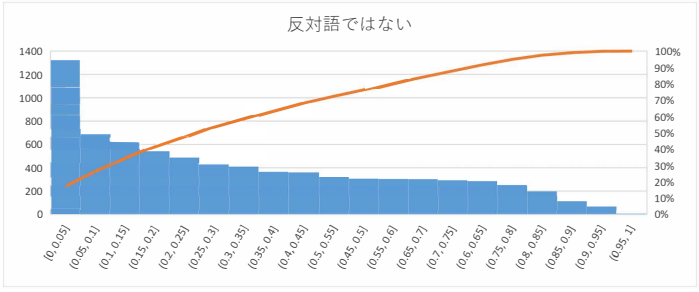

図 4　「反対語ではない」の割合と頻度

また表 3 に分類番号ごとの集計結果を示す。1 次作業において抽出した反対語対が 40 件以上の分類番号について、反対語でない割合の昇順で示す。「体-関係-空間-内外」「体-関係-空間-方向・方角」など 1 次元的な軸が仮定できるものは、反対語であり、かつ、置き換え可である割合が高い。「用-関係-作用-増減・補充」「用-関係-作用-作用・変化」などは、反対語と認定される割合が高い一方、格標示が変化する可能性があり、置き換え不可であるものが増えている。一方、「用-活動-心-表情・態度」「用-活動-待遇-脅迫・中傷・愚弄など」は、反対語と認定されない傾向が高い。これらは、ある語に対する反対の概念が多様であり、特定の反対語を認定することが難しいのではないかと考えられる。

4.　おわりに

本稿では、分類語彙表に対する反対語情報付与について解説した。1 次作業により反対語対候補を抽出したのち、候補に対してクラウドソーシングに基づく反対語か否かおよび置き換え可能か否かについての判断を 20 人分ずつ収集した。同データは `https://github.com/masayu-a/WLSP-antonym` からダウンロードできる。ライセンスは分類語彙表にならい Creative Commons BY-NC-SA 3.0 (表示-非

表 3　分類番号ごとの集計（反対語対が 40 件以上のもの）

分類番号	分類項目	反対語でない	置き換え不可	置き換え可	語対の数
1.1770	体-関係-空間-内外	14 %	22 %	64 %	47
2.1580	用-関係-作用-増減・補充	16 %	35 %	49 %	60
2.1730	体-関係-空間-方向・方角	26 %	20 %	53 %	55
1.2340	体-主体-人物-人物	27 %	21 %	51 %	48
2.1500	用-関係-作用-作用・変化	28 %	35 %	37 %	49
2.3000	用-活動-心-心	29 %	34 %	36 %	45
3.1340	用-関係-様相-調和・混乱	32 %	37 %	30 %	45
1.2450	体-主体-成員-その他の仕手	33 %	21 %	47 %	46
1.3100	用-活動-言語-言語活動	33 %	35 %	32 %	73
3.1520	用-関係-作用-進行・過程・経由	35 %	31 %	34 %	49
1.2120	体-主体-家族-親・先祖	37 %	20 %	43 %	44
1.3790	用-活動-経済-貧富	38 %	30 %	32 %	52
2.1527	用-関係-作用-往復	38 %	29 %	32 %	58
2.3620	用-活動-待遇-人事	41 %	31 %	28 %	59
3.3390	用-活動-生活-立ち居	43 %	30 %	28 %	47
1.3113	体-活動-言語-文字	48 %	17 %	36 %	42
2.3040	用-活動-心-信念・努力・忍耐	48 %	30 %	22 %	53
2.3151	用-活動-言語-書き	59 %	23 %	18 %	54
3.3030	用-活動-心-表情・態度	66 %	17 %	17 %	53
1.3683	用-活動-待遇-脅迫・中傷・愚弄など	67 %	22 %	11 %	159

営利-継承) とする。商用利用については、分類語彙表の商用ライセンスを取得されたい。

基礎統計の傾向から、反対語か否かの二律背反的な定義は困難であることが伺える。本データのように、大規模な評定調査を行うことで、反対語か否かを判定の分布により表現することができる。より精度を高めるためには、クラウドソーシングの作業者毎の揺れをランダム要因として取り入れた回帰が必要だと考える。ランダム要因により、1597 人の作業者のバイアスを吸収し、適切な閾値を設定することにより用途に応じたデータの補整を行うことができる。今後この作業をすすめる。

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科研費 JP17H00917, JP18H05521, JP18K18519 によるものです。

文　献

国立国語研究所 (編) (2004).『分類語彙表－増補改訂版－』　大日本図書.

Yuya Sakaizawa, and Mamoru Komachi (2018). “Construction of a Japanese Word Similarity Dataset.” *1th edition of the Language Resources and Evaluation Conference (LREC 2018)*, pp. 948–951.

猪原敬介・内海彰 (2018).「日本語類似度・関連度データセットの作成」　言語処理学会第 24 回年次大会, pp. 1011–1015.

三省堂編修所 (編) (2017).『反対語対立語辞典』　三省堂.