Alignment Table between 'Word List by Semantic Principles' and 'Annotated Corpus of Iwanami Japanese Dictionary Fifth Edition 2004'

『分類語彙表』と『岩波国語辞典第五版タグ付きコーパス 2004』の対応表

呉 佩珣 (筑波大学)
近藤 森音 (東京大学)
森山 奈々美 (津田塾大学・国立国語研究所)
荻原 亜彩美 (津田塾大学・国立国語研究所)
加藤 祥 (国立国語研究所)
浅原 正幸 (国立国語研究所)

Alignment Table between 'Word List by Semantic Principles' and 'Annotated Corpus of Iwanami Japanese Dictionary Fifth Edition 2004'

Wu Peihsun (University of Tsukuba)
Morine Kondo (University of Tokyo)
Nanami Moriyama (Tsuda University)
Asami Ogiwara (Tsuda University)
Sachi Kato (National Institute for Japanese Language and Linguistics)
Masayuki Asahara (National Institute for Japanese Language and Linguistics)

要旨

『分類語彙表』の見出し語と『岩波国語辞典第五版タグ付きコーパス 2004』に含まれる国語辞典見出し語との対応表を作成した。分類語彙表は統語・意味に基づいて見出し語を分類したシソーラスであるが、その語義を規定する語釈文を含んでいない。そこで、岩波国語辞典の見出し語と対照させることで対応表を構築し、統語・意味分類と語釈文を結びつける作業を行った。作業は、見出し語表記による 2 部グラフを構成し、対応する見出し語対を抽出することによる。本作業は 5 人の作業者により平行して進めた。本作業結果により、『現代日本語書き言葉均衡コーパス』に付与された 2 種類の語義情報（分類語彙表番号・岩波語義タグ）との対照比較ができるようになった。本発表では、情報付与作業の方法と基礎情報を報告する。

1. はじめに

『分類語彙表』(国立国語研究所 1964) は、語義を体系的に整理した現代日本語の大規模シソーラスである。2004 年に増補改訂版 (国立国語研究所 2004) が公刊されたほか、電子化データ『分類語彙表増補改訂版データベース』(ver. 1.0) （以下「分類語彙表 DB」）[(1)]が公開され

[(1)] https://pj.ninjal.ac.jp/corpus_center/goihyo.html

ている。同データと形態素解析用電子化辞書 UniDic [(2)]の対応表 [(3)] (近藤・田中 2020) が整備され、GUI ツール ChaMame [(4)]を用いることにより、形態素解析と同時に可能な分類語彙表番号を付与できるようになった。

分類語彙表は統語・意味分類により整理されているが、語釈文は付与されていない。そこで『岩波国語辞典第五版』(西尾ほか 1994) の電子化データ『岩波国語辞典第五版タグ付きコーパス 2004』(以下「岩波データ」)[(5)] の見出し語と分類語彙表の見出し語を対照し、岩波データの語釈文と分類語彙表の統語・意味分類とを見出し語レベルで結びづける作業を行った。

本稿では、情報付与作業の方法を示すとともに、構築された対応表の基礎統計を示す。

2. 各データの構造と作業の概要

2.1 『分類語彙表増補改訂版データベース』

分類語彙表 DB は『分類語彙表増補改訂版』(国立国語研究所 2004) の電子化データである。学術研究用は無償で利用できる。商用利用については、商用契約することにより利用できる。

表 1 分類番号の構造「前（まえ)」(分類番号： 1.1670)

類	部門	中項目	分類項目
体 (1)	関係 (.1)	時間 (.16)	時間的前後 (.1670)

分類語彙表は表 1 のような構造を持つ分類番号により、見出し語の語義を統語・意味分類に基づいて整理する。統語分類である「類」は、分類番号の小数点より左にある数字によって定義され、体 (1)・用 (2)・相 (3)・他 (4) の 4 つに分類される。意味分類は、分類番号の小数点より右にある 4 つの数字により定義され、上位 1 ケタによる「部門」、上位 2 ケタによる「中項目」、上位 4 桁による「分類項目」の 3 階層で定義される。表 1 の例「前（まえ)」には、分類番号 1.1670 が付与されている。このうち (1) が「体」の類、(.1) が「関係」の部門、(.16) が「時間」の中項目、(.1670) が「時間的前後」の分類項目を意味する。

表 2 『分類語彙表増補改訂版データベース』

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	見出し	見出し本体	読み	逆読み
体	関係	時間	時間的前後	1.1670	01	02	01	前（まえ）	前	まえ	えま
体	関係	空間	左右･前後･たてよこ	1.1740	05	02	01	前（まえ）	前	まえ	えま
体	関係	量	助数接辞	1.1962	06	02	02	－前（まえ）	－前	まえ	えま

「前（まえ)」は多義であり、他に空間を表す分類番号 1.1740 のものや、助数接辞である分類番号 1.1962 のものがある。表 2 に例を示す。分類番号と分類項目以下、さらに「段落番号」「小段落番号」に分類され、小段落番号内の語を一意に決める「語番号」が付与されている。各

(2) https://unidic.ninjal.ac.jp/
(3) https://github.com/masayu-a/wlsp2unidic
(4) https://ja.osdn.net/projects/chaki/releases/70372
(5) https://www.gsk.or.jp/catalog/gsk2010-a/

レコードには「見出し」「見出し本体」「読み」「逆読み」が付与されている。他に、表には掲載していないが、「レコード ID 番号」・「見出し番号」・「レコード種別」の情報が付与されている。

2.2 『岩波国語辞典第五版タグ付きコーパス 2004』

表 3 『岩波国語辞典第五版タグ付きコーパス 2004』

かな	表層形	ID	ID2	語釈文
まえ	前	48488.0.1	1-1	視線・顔が向いている方。⇔うしろ・しりえ。空間的に、後ろ・横でない方向や場所。「家の前」「駅前広場」「銅像の前に立つ」「前（＝陰部）を隠す」「目方」の意。貴人のおん前の意で貴人をさしたことから、身分の高い女性の名に添えた語。「玉藻の前」
まえ	前	48488.0.2	1-2	順序が先の方。⇔のち・あと。本体より先。「前置き」「前ぶれ」時間的に早い方。「疑う前によく捜せ」現在に先立つ（ある）時。「三年前の事」「前に聞いた話」
まえ	前	48488.0.3	1-3	多くは名詞の下に付けてそれ相当のもの。「そんなことは覚悟の前さ」「一人前の男」「三人前の料理」「男前がいい」「腕前」

『岩波国語辞典第五版タグ付きコーパス 2004』は、岩波国語辞典第五版の約 56,000 項目のデータに形態素・構文・照応・共参照情報を付与したデータである。本作業では、各項目の語義の ID と語釈文を用いる。表 3 に項目の例を示す。「かな」「表層形」のほかに、「語彙項目の ID」(「大分類」「中分類」「小分類」をピリオドで連結した 「ID」) 、語釈文の細分類 「ID2」、「語釈文」からなるように整形したものを用いる。

2.3 作業の概要

作業は対応可能な候補を全展開して、実際に対応するものを抽出する作業を行った。

図 1 1 次作業シート

1 次作業では、分類語彙表 DB の「読み」と岩波データの「かな」が一致する全組合せ 349498 件について、5 人の作業者により並行して行った。図 1 に 1 次作業シートの例を示す。最左列に分類語彙表 DB の項目を、3 列目以降に岩波データの項目を並べ、対応する場合に 2 列目に印をつける作業を行った。

2 次作業では、カタカナ語の長音記号・分かち書き記号「／」や接尾辞の「－」などの表記ゆれでマッチしなかったものに対して、対応条件を緩めたうえで、2 人の作業者により作業を行った。図 2 に 2 次作業シートの例を示す。2 次作業では、左に岩波データの項目、右に分類語彙表 DB の項目を配置した。1 次作業と反対の配置にし、見直し作業を行った。

3 次作業では、作業者によって判断に揺れがある場合の統制を行う。基本的に、1 次・2 次作業では並行作業で抽出する作業を進めていたが、3 次作業では分類語彙表から見て、岩波の語義が多く付与されている順に、誤って付与された例を削除する作業を行う。2019 年 9 月 27 日現在 3 次作業中である。

アイス		00088-1.0.1	1-1	氷。ｉｃｅ	051708,49694,A,体,生産物,食料,菓子,1.4340,14,02,02,アイス,アイス,あいす,すいあ
アイス		00088-1.0.1	1-1	氷。ｉｃｅ	○ 059296,57063,A,体,自然,物質,水・乾湿,1.5130,15,01,02,アイス,アイス,あいす,すいあ
アイス		00088-1.0.2	1-2	俗語利貸。ｉｃｅの訳「氷」と雪が通ずるので、明治時代にもじって使った。	051708,49694,A,体,生産物,食料,菓子,1.4340,14,02,02,アイス,アイス,あいす,すいあ
アイス		00088-1.0.2	1-2	俗語利貸。ｉｃｅの訳「氷」と雪が通ずるので、明治時代にもじって使った。	059296,57063,A,体,自然,物質,水・乾湿,1.5130,15,01,02,アイス,アイス,あいす,すいあ
アイス		00088-1.0.3	1-3	「アイス　キャンデー」・「アイス　クリーム」の略。	○ 051708,49694,A,体,生産物,食料,菓子,1.4340,14,02,02,アイス,アイス,あいす,すいあ
アイス		00088-1.0.3	1-3	「アイス　キャンデー」・「アイス　クリーム」の略。	059296,57063,A,体,自然,物質,水・乾湿,1.5130,15,01,02,アイス,アイス,あいす,すいあ
アイス／キャンデー		00088-2.0		果汁などを冷凍した、一種の氷菓子。日本でｉｃｅとｃａｎｄｙとを合わせて作った語。	051713,49699,A,体,生産物,食料,菓子,1.4340,14,03,01,アイスキャンデー,アイスキャンデー,あい
アイス／クリーム		00088-3.0		牛乳、卵の黄身に砂糖・香料を加え、まぜ合わせて凍らせた菓子。氷菓子。	051707,49693,A,体,生産物,食料,菓子,1.4340,14,02,01,アイスクリーム,アイスクリーム,あいすく
アイス／ボックス		00088-4.0		氷を使って食品などを冷やす、（携帯用の）冷蔵庫。	053528,51463,A,体,生産物,住居,家具,1.4470,07,03,01,アイスボックス,アイスボックス,あいすぼ
アイス／ホッケー		00088-5.0		氷上でスケートをはいてするホッケー。	038536,36881,A,体,活動,生活,スポーツ,1.3374,14,02,05,アイスホッケー,アイスホッケー,あいす

図 2　2 次作業シート

3.　データの概要

表 4　1 つの岩波データの項目に付与される分類語彙表 DB の語義数

付与される分類語彙表 DB の語義数	左の語義数をもつ岩波データ項目数と割合	
1	44002	53.34%
2	4758	5.76%
3	661	0.80%
4	133	0.16%
5	21	0.03%
6	27	0.03%
7	4	0.00%
8	4	0.00%
9	7	0.00%
10	1	0.00%
12	1	0.00%
14	1	0.00%
合計	49620	60.15%

以下では 2019 年 9 月 27 日現在の基礎統計を示す。

表 4 に岩波データの項目側から見た、分類語彙表 DB の語義数を示す。岩波データの語釈文の総数 82491 項目のうち、60.15% の 49620 項目に分類語彙表番号が付与された。岩波データ 44002 項目については、1 つの分類語彙表 DB の項目のみ割り当てられているが、5618 項目については、2 つ以上の分類語彙表の項目が割り当てられており、最大 14 項目と対応するものもあった。分類語彙表のほうが細かい語義が設定されている事例として「いたむ」の例を表 5 に示す。岩波データでは、「痛む」「傷む」「悼む」の異表記を同語として扱うほか、肉体的な痛みと精神的な痛みを区別していないために、分類語彙表 DB の 14 項目に対応する。

表 6 に分類語彙表 DB の項目側から見た、岩波データの語義数を示す。分類語彙表 DB 101070 項目のうち、48.47% の 48995 項目に対して語釈文がついたことになる。 分類語彙表

表 5　岩波データ・分類語彙表 DB 1:14 対応 （例）「いたむ」

岩波データの項目
いたむ 痛む・傷む・悼む 02216.1.1 1-1 痛くなる。肉体的な痛さを感ずる。痛「傷が痛む」精神的な苦しみ・打撃を受ける。それによって悩み苦しむ。痛・傷「心が痛む」「ふところが痛む」（予想外の出費などで、つらい）
対応する分類語彙表 DB の項目
075158,72011,B, 用, 活動, 心, 感覚,2.3001,06,01,01, いたむ（痛・傷）, いたむ, いたむ, むたい
075159,72011,1, 用, 活動, 心, 感覚,2.3001,06,01,02, 痛む, 痛む, いたむ, むたい
075160,72011,2, 用, 活動, 心, 感覚,2.3001,06,01,03, 傷む, 傷む, いたむ, むたい
075926,72707,B, 用, 活動, 心, 苦悩・悲哀,2.3014,08,01,03, いたむ（痛・傷・悼）, いたむいたむ, むたい
075927,72707,1, 用, 活動, 心, 苦悩・悲哀,2.3014,08,01,04, 痛む, 痛む, いたむ, むたい
075928,72707,2, 用, 活動, 心, 苦悩・悲哀,2.3014,08,01,05, 傷む, 傷む, いたむ, むたい
075929,72707,3, 用, 活動, 心, 苦悩・悲哀,2.3014,08,01,06, 悼む, 悼む, いたむ, むたい
076220,72978,A, 用, 活動, 心, 敬意・感謝・信頼など,2.3021,08,02,01, 悼む, 悼む, いたむ, たい
090567,86078,B, 用, 自然, 生命, 障害・けが,2.5720,01,01,01, いたむ（痛・傷）, いたむ, たむ, むたい
090568,86078,1, 用, 自然, 生命, 障害・けが,2.5720,01,01,02, 痛む, 痛む, いたむ, むたい
090569,86078,2, 用, 自然, 生命, 障害・けが,2.5720,01,01,03, 傷む, 傷む, いたむ, むたい
090641,86148,B, 用, 自然, 生命, 病気・体調,2.5721,04,01,01, いたむ（痛・傷）, いたむ, いたむ, むたい
090642,86148,1, 用, 自然, 生命, 病気・体調,2.5721,04,01,02, 痛む, 痛む, いたむ, むたい
090643,86148,2, 用, 自然, 生命, 病気・体調,2.5721,04,01,03, 傷む, 傷む, いたむ, むたい

表 6　1 つの分類語彙表 DB の項目に付与される岩波データの語釈数

付与される岩波データ語釈割り当て数	左の語義数をもつ分類語彙表 DB 項目数と割合	
1	42862	42.40%
2	5330	5.27%
3	699	0.69%
4	102	0.11%
5	2	0.00%
計	48995	48.47%

表 7　岩波データ・分類語彙表 DB 5:1 対応 （例）「手踊り」

分類語彙表 DB 項目				
037900,36251,A, 体, 活動, 生活, 遊楽,1.3370,13,05,02, 手踊り, 手踊り, ておどり, りどおて				
対応する岩波データの項目				
ておどり	手踊（り）	34919.0.1	1-1	すわったまま手だけ動かしてする簡単な踊り。
ておどり	手踊（り）	34919.0.2	1-2	三味線に合わせてする踊り。
ておどり	手踊（り）	34919.0.3	1-3	所作事のうち、手に何も持たないでする踊り。
ておどり	手踊（り）	34919.0.4	1-4	盆踊りなどの、大勢そろって同じ手振りでする踊り。
ておどり	手踊（り）	34919.0.5	1-5	表情に重きを置かない、ちょっとした踊り。

DB の 42865 項目について、1 つの岩波データの語釈文が対応していた。また、分類語彙表 DB の 6133 項目については、2 つ以上の語釈文が対応していた。岩波データに複数の語義が設定されている事例として「手踊り」の例を表 7 に示す。本項目は、岩波データにおいて語釈文を細かく分割しているため 1:5 対応となっている。

4. おわりに

本稿では、分類語彙表と岩波国語辞典の語釈文の対応表について紹介した。分類語彙表の項目に岩波国語辞典の語釈文が付与され、定義-被定義間情報を用いた語彙調査が可能になる (正津ほか 2001, 野呂・徳田 2007, 岡久ほか 2019)。また同データを用いて、単語心象性の調査（語義ごとの親密度調査）もすすめたい。

なお、同データは、完成後 `https://github.com/masayu-a/wlsp2iwanami` にて配布する。『分類語彙表』および『岩波国語辞典第五版タグ付きコーパス 2004』の双方のライセンスに基づいて利用されたい。

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 17H00917, 18H05521, 18K18519, 19K00591, 19K00655 によるものです。

文　献

国立国語研究所 (編) (1964).『分類語彙表』　秀英出版.

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

近藤明日子・田中牧郎 (2020).「「分類語彙表番号-UniDic 語彙素番号対応表」の構築」　国立国語研究所論集, pp. To appear.

西尾実・岩淵悦太郎・水谷静夫 (編) (1994).『岩波国語辞典第五版』　岩波書店.

正津康弘・白井清昭・徳永健伸・田中穂積 (2001).「国語辞典の語釈文の解析と語義のシソーラスへのマッピング」　第 15 回人工知能学会全国大会, pp. 2B2–01, 1–4.

野呂智哉・徳田雄洋 (2007).「語釈文記述のための日本語定義語彙の構築に関する一考察」　言語処理学会第 13 回年次大会発表論文集, pp. 626–629.

岡久太郎・久保圭・水谷勇介・河原大輔・黒橋禎夫 (2019).「クラウドソーシングにより収集した語釈文を基にした単語の基本度推定」　言語処理学会第 25 回年次大会発表論文集, pp. 1499–1502.