『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

山崎誠　柏野和佳子
国立国語研究所
{yamazaki, waka}@ninjal.ac.jp

1 はじめに

本発表では,『分類語彙表増補改定版データベース』（国立国語研究所 2004，以下,『分類語彙表』）の日本語研究における利便性を高めるため，多義語に対して代表的な意味を付与するアノテーション作業について報告する.

2 日本語研究と『分類語彙表』

『分類語彙表』は，現代日本語のシソーラスであり，日本語研究，特に語彙研究によく用いられている言語資源である．1964 年に初版[1]，2004 年に増補改訂版が刊行されており, 2015 年には増補改定版に基づく電子データ[2]が公開されている.

3 『分類語彙表』の問題点

『分類語彙表』を日本語研究で利用する場合，いくつか不便に感じる点がある．例えば，(1)語の表記が代表的なもの 1 つしか示されていない[3]ため，実際にデータに表れた表記とマッチングを取ろうとすると一致しないことがある，(2)見出し語に ID を与えていないので，同じ語か異なる語かの判断を解析者自身が行わなければならない，(3)言語単位が接辞，単語，複合語，慣用句と混在しており，それらを区別する情報がついていない，(4)例文がないため，多義語の意味を特定しにくい場合がある, (5)2004 年以降の新語が収録されていない，などである．本発表では，そのような問題点のうち，(4)の多義語に関する問題を取り上げる[4]．多義語の問題は，自然言語処理の立場からは，語義曖昧性解消の問題としてとらえることができるが，本発表では辞書学的な立場から検討した結果を示す.

4 『分類語彙表』における多義

4.1 多義の認定

上述のように,『分類語彙表』には語に ID が付与されていないため，どのようにして多義を認定するかを前もって決めておかなければならない．まず，語の認定については，見出し語の表記と読みとが一致しているものを同じ語とした．多くの語では表記と読みとで語の認定が可能だが，それだけでは不十分な場合がある．例えば,「バス(bath)」「バス(bus)」「バス(bass)」が同じ語と見なされたりする．逆に,「依存（いそん）」と「依存（いぞん）」のような，意味的な違いではない語形のゆれが別語になったりする不都合があるが，本発表の調査に関しては大勢に影響はないと思われる.

本発表では，語を上記のように認定した上で,『分類語彙表』の各分類項目[5]を語義のまとまり考え,「異なる分類項目にわたって出現する」場合を多義と見なし，そのような語を「多義語」とした.

たとえば,「セット（set）」は次の 3 つの分類項目に出現している.

1.1951　群・組・対
1.1962　助数接辞
1.3334　保健・衛生

1.1951 の「セット」は，一揃いの意味，1.1962 の「セット」も意味は同じだが, 助数詞としての用法, 1.1951 は,「髪をセットする（整える）」の意味である[6].

[1] 初版の電子データは 1994 年に市販されたが, 品切れになったため 2015 年に以下のサイトで公開した.
[2] http://pj.ninjal.ac.jp/corpus_center/archive.html
[3] 特に，和語で複数の表記をまとめた場合に平仮名表記であることが一致を低下させている.
[4] 語の ID については，UniDic との対応をとることである程度解決するものと思われる．具体的には，近藤・田中（2017）を参照されたい.
[5] 分類項目とは,『分類語彙表』を構成する分類階層の 1 つで，5 桁の数字で表されるものである．全部で 895 の分類項目がある.
[6] スーパー大辞林 3.0』（Web 更新版）では，これらのほかに, 映画のセット, テニスなどの試合の単位, 用意・準備することの意味があるが，これらに対応する「セット」はいずれも『分類語彙表』には収録されていない.

多義と言っても，上の 1.1951 と 1.1962 は用法の違いであり，意味の違いではない．また，分類番号が 1 で始まる「体の類（名詞）」と 3 で始まる「相の類（形容詞，形容動詞）」も，ほとんどの場合，多くの国語辞書で「(名・形動)」と記されることからも分かるように，これらは品詞の違いであって，意味の違いではないと考えられる．

4.2 多義語の数と分布

表 1 に，多義の分布を示す．『分類語彙表』には，101,070 行のデータがあるが，そのうち 240 行は区切りを示すものであり，また，2,589 行は複数の見出しを分割した際の元の見出しなので分析には不要である．これらを除いた 98,241 行（98,241 語）が延べ語数に該当する．それに対して異なり語数は表 1 の合計欄に示すように 81,250 語である．そのうち 83.2%が語義の数が 1 である．すなわち，多義の語は 17.8%（13,648 語）となる[7]．

表 1：『分類語彙表』における多義の分布

語義の数	語数	割合
1	67,602	83.20
2	11,227	13.81
3	1,839	2.26
4	439	0.54
5	82	0.10
6	36	0.04
7	18	0.02
8	5	0.01
10	1	0.00
11	1	0.00
合計	81,250	99.98

次に，多義語における語義の組み合わせについて観察する．ここでは，2 つの分類項目の組で集計する．前出の「セット」の例で見ると，「1.1951　群・組・対」「1.1962　助数接辞」「1.3334　保健・衛生」の 3 つの分類項目に出現していたが，これらから 2 つの部類項目の組を抜き出した場合のすべての組み合わせをとり，以下の 3 つの組を得る．

（1.1951 群・組・対，1.1962 助数接辞）
（1.1951 群・組・対，1.3334 保健・衛生）
（1.1962 助数接辞，1.3334 保健・衛生）

[7] 語義数がもっとも多かったのは「手（て）」，次に多かったのは「立てる」であった．

このようにしてすべての多義語における語義の組み合わせを集計すると，表 2 のようになる．表 2 から語義の組み合わせが 1 つだけのものが半数以上（64.9%）であることが分かる．さらに 7,465 組のうち 5,291 組は 1 つの語とのみ対応している．例えば，「2.1131 連絡・所属」と「2.1571 切断」の組み合わせは「断ち切る」にのみ見られた．

表 2: 多義語を構成する語義の組み合わせの数

組の数	語数	割合
1	7,465	64.86
2	2,039	17.72
3	841	7.31
4	430	3.74
5	235	2.04
6	160	1.39
7	111	0.96
8	71	0.62
9	44	0.38
10	28	0.24
11 以上	86	0.75
合計	11,510	100.00

表 3 は，分類番号の最初の 1 桁，すなわち「類」で見た場合の多義の組み合わせ数の分布である．1 と 1（名詞と名詞）の組み合わせが最も多く，2 と 2（動詞と動詞）がそれに次ぐ．1 と 3（名詞と形容詞・形容動詞）が多いのは品詞の認定の仕方のためであり，形容動詞を「～だ」という終止形で見出しにすれば，この組み合わせはなくなる

表 3: 類で見た語義の組み合わせの数

類の組み合せ	組み合せ数	割合
1 と 1	10,186	47.70
1 と 2	19	0.09
1 と 3	1,982	9.28
1 と 4	142	0.66
2 と 2	6,790	31.79
2 と 3	23	0.11
2 と 4	1	0.00
3 と 3	1,959	9.17
3 と 4	158	0.74
4 と 4	96	0.45
合計	21,356	100.00

多義の観察の最後に，単義語の場合との比較を行う．多義語を構成する語義を分類項目で集計した場合，いちばん多い語義は，「1.1962 助数接辞」で 195

回出現している．以下，「2.3392 手足の動作」「1.2340 人物」「2.1570 成形・変形」「2.1532 入り・入れ」と続く．同様に単義語の場合の語義は，「1.2340 人物」がいちばん多く（867 回），以下「1.5721 病気・体調」「1.3374 スポーツ」「1.2410 専門的・技術的職業」「1.5402 草本」と続く．単義語は，専門的な意味を持つ分類項目に多いことが分かる．

5 多義語の選択方法

『分類語彙表』では多義語をそれぞれの語義の意味に対応した分類項目に配置している．これはシソーラスの性質上当然のことであるが，そのため，分析対象となるデータを形態素解析した結果に対して意味情報を付与しようとしたとき，どの分類項目を選んだらよいか，迷う場合が出てくる．これを解決するには次の 2 つの方法が考えられる．どちらの方法を採っているかは分析結果にとって重要なことであるが,『分類語彙表』を利用した研究ではあまり言及されることがない．

(1)当該の文脈に合った，1 つの分類項目を選ぶ[8]．

(2)当該の文脈にかかわらず，特定の 1 つの分類項目を選ぶ．

例えば，動詞「落ちる」は，「2.1251 除去」「2.1540 上がり・下がり」「2.1584 限定・優劣」「2.1931 過不足」「2.3321 学事・兵事」「2.5701 生」の 6 つの分類項目に出現するが，原文が「汚れが落ちる」なら 2.1251，「品質が落ちる」なら，2.1584 というようにそれぞれに合った分類項目を付与する方法である．(2)は，どの語義であっても一つの分類項目（例えば，2.1540）を付与する方法である．

(1)は，文脈上の意味を重視しており，きめの細かい分析ができるが，一つ一つ語義を特定しなければならないため，かなり手間がかかる．一方，(2)は，多義語がどの語義で使われていても 1 つの語義を選ぶため，ごく粗い分析にとどまるが，手間はかからず効率的である．

(1)の方法には，もうひとつ問題点がある．それは，文脈を見ても分類項目が決めがたいものがあることである．例えば，「冷房する」は，「2.3850 技術・設備・修理」「2.5170 熱」の 2 つの分類項目に出現するが，どちらが妥当かは文脈でも分からないのではないだろうか．このような観点の違いによる多義は(1)の方法では解決できない．

[8] この方法を採っているデータについては加藤他（2017）を参照されたい．

そこで，本発表では，現時点での現実的な方法として(2)の方法を採用することにし，語に対して特定の分類項目を 1 つ決める際に，「代表義」というものを定めることにした．この代表義は，認知意味論で言うプロトタイプ的意味と言ってもよいだろう．代表義の性質として，当該の語についての典型的な語義であること，派生的な意味よりも原義に近いこと，使用頻度が高いことなどが考えられる．辞書記述で言えば，最初に出てくる①に相当する語義である．そのような考えのもとに『分類語彙表』で複数の分類項目に出現している見出し語について，代表義に相当する 1 つの分類項目をアノテーションしていく作業を行った．

なお，代表義という見方で考えると，『分類語彙表』で 1 つの分類項目にしか現れていない語についても，それが正しく代表義の項目に出現しているかの確認も必要であるが，それは行っていない．

6 代表義の決定方法

6.1 作業基準

代表義の決定が恣意的にならないように，以下の作業基準を作成した．これらの間には一定の優先順位を設け，代表義がなるべく客観的に決まるようにした．

①高頻度

使用頻度が高いものを基本義とする．直観的に使用頻度に差がありそうなものは内省で判断したが，コーパスでの頻度も参考にし．例えば，「切り出す」は，「2.1571 切断」「2.3100 言語活動」「2.3810 農業・林業」の 3 つの分類項目に現れるが,『現代日本語書き言葉均衡コーパス』における語義の分布は以下のようになっているため，「2.3100 言語活動」を選択する．

2.1571	切断	132 例	15.9%
2.3100	言語活動	655 例	79.0%
2.3810	農業・林業	42 例	5.1%

②具体性

抽象的なものより具体性があるものを優先する．例えば，「傾く」は，「2.1513 固定・傾き・転倒など」と「2.1583 進歩・衰退」の 2 つの分類項目に現れるが，具体的な意味を持つ 2.1513 のほうを優先する．

③分類項目名

分類項目名と一致しているものがあれば，その分類項目を選ぶ．例えば，「上がる」は，以下の 8 つの分類項目に出現している．「2.1540 上がり・下がり」「2.1503 終了・中止・停止」「2.1580 増減・補充」「2.1651

終始」「2.3000 心」「2.3331 食生活」「2.3520 応接・送迎」「2.3700 取得」．これらの中で分類項目に「上がり」がある 2.1540 を選ぶ．

④類義語数

当該の語が所属する段落に類義語があるものを選ぶ．『分類語彙表』は，分類項目に対応する意味領域の語を配列しているが，その意味領域に対して中核的な語と周辺的な語とがある．周辺的な語はどちらかというと，その回りに類義語が少なく，グループを形成していない．

⑤慣用句

『分類語彙表』には，慣用句も相当数収録されている．慣用句と字義通りの意味の両方がある場合は慣用句としての意味を優先する．

⑥辞書の第一義

作業用に用いた『岩波国語辞典第 5 版』(以下,『岩国』と略す）の第一義と一致するものを選択する[9]．

⑦複合語後項

日本語の複合語は意味的な中心が後項にあるため，原則として後項要素の意味を選ぶ．

⑧段落内出現順

段落内でより先頭に近いものを選ぶ．『分類語彙表』は分類項目がいくつかの段落から構成されている．その段落内での語の配列は，「なるべく意味・用法の広いほうから狭いほうへ配列しているが，必ずしも厳密ではない．」(『分類語彙表増補改定版』P.4）とされている．そこで，段落内で前のほうに出現している語を基本義と考える．

以上の作業基準は，原則として以下の優先順位で適用している．

高頻度＞分類項目名＞慣用句＞類義語数＞辞書の第一義＞段落内出現順

6.2 作業結果

作業結果として，14,102 語（異なり）に対して代表義を与えることができた．内訳は，体（名詞）6,570 語，用（動詞）6,192 語，相（形容詞・形容動詞）1,283 語，他（副詞・接続詞・感動詞等）57 語である．代表義が与えられた語について 6.1 の作業基準が適用された数は，以下のとおりである．以下に示す数字には作業基準が重複して該当した場合も含むため，合計は 14,102 語と一致しない．

表 4：作業基準が適用された語数

作業基準	適用された語数
高頻度	3,998
具体性	863
分類項目名	708
類義語数	3,931
慣用句	274
辞書の第一義	1,121
複合語後項	665
段落内出現順	969
その他	3,080

7 おわりに

形態素解析の精度が向上し，多義語に対して文脈に即した意味を自動で付与できるようになれば，このような代表義を決めておく必要はないかもしれない．しかし，文脈を持たず，語だけが並んでいる語彙リスト（例えば古辞書など）のようなデータを扱う場合には，意味を 1 つに特定したほうが便利であるし，時間の経過による意味の変化を考えた場合，代表義が変化したと捉えて記述できるメリットがある．また，将来，辞書における語義の配列に当たる情報を付与する際の基礎となる情報でもある．

謝辞

本研究は科学研究費・基盤研究（C）「語彙分類の理論的整備に基づくシソーラスの改良に関する研究」（課題番号:24520520, 研究期間:平成 24～26 年度，研究代表者：山崎誠）によるものである．本アノテーション作業には，田嶋明日香，立花幸子，平本智弥の 3 名の協力を得ている．また，本研究の一部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「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」・言語変化研究領域共同研究プロジェクト「通時コーパスの構築と日本語史研究の新展開」によるものでもある．

参考文献

加藤祥・浅原正幸・山崎誠（2017）「『現代日本語書き言葉均衡コーパス』に対する分類語彙表番号アノテーション」言語処理学会第 23 回年次大会発表論文集

近藤明日子・田中牧郎（2017）「分類語彙表・UniDic 見出し対応表の構築―コーパスへの網羅的・系統的な語義情報付与を目指して―」言語処理学会第 23 回年次大会発表論文集

[9] 『岩波国語辞典第 5 版』は，電子化データ（岩波国語辞典第 5 版タグ付きコーパス 2004）が研究用に公開されている(http://www.gsk.or.jp/catalog/).