『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

加藤祥[1] 浅原正幸[1] 山崎誠[12]
1国立国語研究所 コーパス開発センター 2国立国語研究所 研究系言語変化研究領域

1．はじめに

類語の調査はもちろん、比喩をはじめとする表層的な表現と意味の差を研究する際など、意味的な情報の付与されたコーパスが有用なリソースとなる。また自然言語処理の分野では、語義曖昧性解消のタスクの学習・評価データとして様々な語義タグつきデータが整備されてきた。日本語では、古くは新聞記事を対象とした EDR コーパスや RWCP コーパスなどが国語辞典に基づく語義タグを付与していた。また、代表性を持つコーパスとして、『現代日本語書き言葉均衡コーパス』(Maekawa et. al. 2014)（以下 BCCWJ）の一部に対しても、岩波国語辞典の語義が付与され、SemEval-2010 Japanese WSD Task (Okumura et. al. 2011) では、日本語の語義曖昧性解消の基礎データとして用いられてきた。また、シソーラスに基づくデータとして、日本語ワードネットに基づく語義タグ付きコーパス(Bond et. al. 2012) が整備されている。このコーパスは、英語データを翻訳したものであり、代表性をもつ自然な日本語コーパスに対する、シソーラスに基づく語義タグつきデータは管見の限りない。

国立国語研究所では、BCCWJ コアデータに『分類語彙表増補改訂版』（2004）の分類語彙表番号を悉皆付与する作業に着手した。現在進めているアノテーション基準と作業状況を報告する。

2．アノテーション

2.1 概要

アノテーション作業対象として、コアデータに含まれる新聞サンプル 54 ファイル（部分集合 A：PN(A)）から、アノテーション優先順位に基づき順次作業に着手した。分類語彙表番号を人手で UniDic 語彙素番号（小木曽・中村, 2014）に対応させたデータ（近藤・田中, 2017）により、BCCWJ の言語単位（短単位・長単位）に対応可能性のある分類語彙表番号を列挙したうえで、人手で正しい語義を選択する作業を進めている。本付与作業にあたっては、分類語彙表の 5 桁目までの番号を付与する。（例は表 1。図 1 における「3.1010 こそあど」の分類部分が該当する。）

表 1 分類番号の構造（例：この（分類番号：3.1010））

類	部門	中項目	分類項目
用（3）	関係（.1）	真偽（.10）	こそあど（.1010）

アノテーション作業は、短単位と長単位のそれぞれについて行う。列挙された分類語彙表番号の選択肢から、該当する意味分類が選択可能であれば選択し、選択できない場合や、語彙素に対応する分類語彙表番号がない場合には、新たに適切な番号を付与する。以下では、それぞれの単位のアノテーション作業基準について示す。

3 相 の 類
3.1 抽象的関係
3.10 真偽
3.1010 こそあど・他
01 この こんな こういう こうした
かかる こう かく かよう
こうこう かくかく
しかじか このまま
かくのごとく／のごとき このとおり
02 その そんな そういう そうした さる しかる
そう さ さよう
連ほどさようにしかく

図 1 分類語彙表（部分例）

2.2 短単位に対する分類語彙表番号アノテーション

2.2.1 概要

機能語を除く短単位に分類語彙表番号を付与する。分類語彙表番号の付与される短単位の機能語は助動詞と助詞の一部に限られるためである。分類語彙表番号が付与される機能語の内訳を表 2 に示す。頻度は BCCWJ PN A サンプルのものである。

表 2 分類語彙表番号が付与される機能語

頻度	語彙素番号	語彙素	品詞
261	40741	れる	助動詞
214	27905	など	助詞-副助詞
87	35891	まで	助詞-副助詞
62	39787	られる	助動詞
49	20355	せる	助動詞
47	21652	たい	助動詞
41	23122	だけ	助詞-副助詞
26	22727	たり	助詞-副助詞
8	10403	くらい	助詞-副助詞
7	34770	ほど	助詞-副助詞
6	30577	ばかり	助詞-副助詞
4	19641	ずつ	助詞-副助詞
2	29213	のみ	助詞-副助詞
2	24320	つ	助詞-副助詞
1	14185	させる	助動詞

語彙素に対応して列挙された番号（曖昧性：1～11 種類）がある場合、作業者は該当番号を選択する（図 2）。

短単位	品詞	記入欄	分類語彙表番号（選択肢）			
研究	名詞-普通名詞-サ変可能		1.3065:体-活動-心	研究・試験・調査・検査など		
費	接尾辞-名詞的-一般	1.3721				
を	助詞-格助詞					
受け入れる	動詞-一般		2.3430:用-活動-行	2.3532:用-活動-交	2.3770:用-活	

図 2　番号付与作業例

新聞サンプル 54 ファイル（部分集合 A：PN(A)）における短単位数は表 3 の通りである。すなわち、短単位 56,922 のうち、アノテーション対象となり得る自立語は 33,725 語（59.2%）あり、そのうち UniDic-分類語彙表データと語彙素番号がマッチした 28,696 語（50.4%）について、選択可能な番号が列挙されている。

表 3　BCCWJ PN(A)集合の短単位内訳

PN(A)短単位	のべ	56922
	機能語	23197
	自立語	33725
UniDic-分類語彙表データにマッチしたもの	全て	29513
	機能語	817
	自立語	28696

また、UniDic-分類語彙表データにマッチした自立語の、選択肢数（分類番号の曖昧性）を表 4 に示す。複数選択肢の列挙された短単位（曖昧性 2 以上）は 12,857（44.8%）ある。なお、曖昧性が 8 となる短単位数の頻出はサ変動詞「する」の頻度の影響による。

表 4　分類番号の曖昧性（BCCWJ PN(A)サンプル）

曖昧性	短単位数	曖昧性	短単位数
1	16656	6	237
2	7479	7	134
3	2621	8	1253
4	826	9	49
5	237	10	21

短単位の番号付与にあたっては、最小限の文脈に依拠した意味とし、比喩的・慣用的な表現などは語源的な意味とする。内容に即した意味は、長単位で対応する。

「名詞-普通名詞-形状詞可能」「名詞-普通名詞-副詞可能」のような品詞の語については、体（1.で始まる分類語彙表番号）・相（3.で始まる分類語彙表番号）のどちらとも読み取れるが、BCCWJ コアに付与された人手による「名詞」「形状詞」などの用法情報に従う。

2.2.1　UniDic-分類語彙表対応のない場合

列挙された選択肢に、文脈上適切な番号がないと判断される場合は、新たな番号を付与する。新たに番号を付与する場合は、『分類語彙表増補改訂版』を参照し、適切な意味分類を検討する。UniDic の語彙素に対応する番号がなく、そもそも選択する番号のない場合も、分類語彙表の意味分類を確認し、適切な番号を付与する。

UniDic の語彙素に対応する番号がない例としては、未知語、固有名詞、略語などがある。未知語には「ロック」「カム」「トゥゲザー」のような外来語も多く含まれるが、それぞれ外来語の意味に相当する意味分類を選択し、分類語彙表番号として付与する。

なお、用法によっては、分類語彙表に既存の分類番号がない場合がある。その場合、分類語彙表に存在しない番号を新設して付与することもあり得る。

固有名詞

人名についてはアノテーション対象外とするが、地名や普通名詞を含む「名古屋タワープラザホール」「岡山ホテル」「阪急グランドビル」のような固有名詞については、それぞれ短単位ごとの意味分類が可能と考え、「名古屋タワープラザホール」であれば、「名古屋」「タワー」「プラザ」「ホール」のそれぞれに分類語彙表番号を付与する。

略語・掛詞等

略語についても、元の語形を考慮し、該当する分類語彙表番号を付与する。但し、「厚労」「自民」のように複数語義の組み合わせが一短単位となっている場合もある。このような場合は、「厚労」は「厚生」「労働」、「自民」は「自由」「民主」のそれぞれの短単位に相当する複数の分類語彙表番号を付与する。掛詞やダジャレなど、一短単位について複数の意味が読み取れる場合にも、複数の意味について分類語彙表番号を付与する。

その他

「一個」「一口」のように短単位での登録がある語は、その単位での分類語彙表番号候補がある場合、文脈上「一」「口」や「一」「個」が別の短単位と読むことが適切にも関わらず、「一個」「一口」を一短単位として選択肢（番号の候補）が挙がる。このような場合は、「一」「口」を一短単位と判断し、各々に分類語彙表番号を付与する。また、副詞用法の語であるが分類語彙表番号に体の類しかない場合には、対応する相の番号を新たに付与する。

2.3 長単位アノテーション

短単位と同様に、長単位についても分類語彙表番号を付与する。短単位作業時に、対応した長単位があればマークを表示し、長単位作業のあることを示している。

長単位に対応して列挙された選択可能な番号（1～11種類）がある場合は、該当する番号を選択する。文脈上、適切な番号がないと判断される場合や、語彙素番号に対応する番号がない場合は、同様に適切な意味分類を行い、新たに番号を付与する。

「ていく」「てくる」をはじめ、「にとって」など、助動詞扱いとなるが短単位と異なる意味分類となる場合などは、機能語であっても分類語彙表番号を付与する。また、長単位より大きな単位（慣用句など）として分類語彙表番号がある場合には、メモとして番号を付与する。

3． 作業状況

3.1 現在までの作業概要

作業者と担当サンプルにより、作業ペースに差が生じるが、1時間あたり100語～300語程度のアノテーションが可能である。以下の表5にこれまでの作業における番号付与作業量を示す。

表5　番号付与作業量

付与対象	作業内容	作業量
短単位	番号選択（自立語）	全短単位の47%
	新規番号追加（自立語）	全短単位の5%
長単位（短単位の15%程度）	番号選択	長単位の14%
	新規番号追加	長単位の76%

短単位へのアノテーション作業の内訳は、全短単位の47%において番号選択、5%で新規番号追加である。両作業をあわせ、52%の短単位に番号付与を行っている。長単位は、短単位の31%が番号付与作業対象となっている。

なお、長単位は全短単位の15%程度に付与作業を行うこととなる。長単位におけるアノテーション作業の内訳は、14%で番号選択、76%で新規番号追加となり、新規番号追加作業の割合が高い。

3.2 作業における問題点等

作業者から質問のある点については、作業者間に揺れが生じることや、作業結果に影響のあることが予想される。現在までの作業では、作業者とQAを共有しており、作業者の記入した質問に、発表者が回答している。ここでは、作業者とのQAに見られる傾向から、作業において問題となる可能性のある点について報告する。

複数の読みが可能な場合

文脈上、複数の読みが可能な場合、付与する1つの番号をいずれと定めるのかを作業者個人の判断にゆだねるため、作業者によって同様の文脈でも揺れの生じる可能性が考えられる。

長単位の文法的分類

助動詞扱いになっている場合をはじめ、どの部分が主となる複合語であるのかの判断にあたり、長単位を、体・用・相のいずれに分類するのかが問題となりがちである。意味の番号は等しい場合でも、作業者によって文法的な分類が異なってくる可能性が考えられる。

UniDic-分類語彙表対応データの部分的な不備や不足などの影響による作業者の迷いや揺れも散見されるが、これらは作業が進むことで、付与作業済みデータを用いたUniDic-分類語彙表対応データの補填や拡充が可能となることが期待され、今後解消され得ると考えられる。

4．進捗

これまで（2016年10月～12月）に付与作業の完了したデータは以下である。

表6　番号付与済みデータ（2016年10月～12月）

集合	短単位	付与数	選択	追加等
PN	50837	25524	24005	3074
その他	13656	6823	6505	318
総計	64493	32347	30510	3392

作業者によって、追加等作業数に差があるが、今後作業者間の揺れなどの確認を進め、整理と統一を行う予定である。PNに付与された分類語彙表番号の類は、体の類（1）が71%、用の類（2）が20%、相の類（3）が8%となっており（図3）、PNの他の集合でもこの割合は概ね等しい。

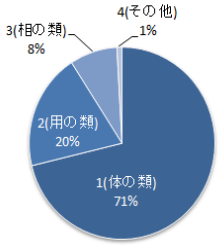

図3　これまでに付与した分類語彙表番号の類（PN）

次に、PNに付与された上位頻度（1.0%以上）の番号

を以下の表 7 に示す。アノテーションしたサンプルが新聞であることから、数記号が 1 割に及ぶ。また、地名や人間や団体の活動や存在に関する分類番号が多いという結果になっている。

表 7　上位頻度付与番号（PN, 1.0%以上）

番号	頻度	例
1.1960	10.4%	数記号
2.3430	4.3%	行為・活動（する）
1.2590	3.3%	固有地名
2.1200	2.4%	存在（ある）
1.2000	1.7%	人間
1.1962	1.4%	助数接辞
1.1000	1.4%	事柄
1.2760	1.0%	同盟・団体
3.1010	1.0%	こそあど・他

５．おわりに

本稿では、現代日本語書き言葉均衡コーパス（BCCWJ）に分類語彙表番号を付与する作業について、アノテーション基準と現在までの作業状況を報告した。現在まで、月に 2 万単位（短単位）ほどのアノテーションが進行中である。

これらのデータ整備によって、BCCWJ が意味的な情報によって検索可能となり、従来用例の収集が困難であった意味上のグループに応じた分類を要する研究における新たな可能性が期待される。たとえば、比喩研究では、隠喩のように明示された比喩指標のない用例を収集するために、結合する要素のずれを判定する必要がある。文脈と意味分類を対照することで、このような隠喩の用例は格段に収集しやすくなるはずである。

作業によって、未知語をはじめとする分類語彙表にない語への番号付与が進んでいるほか、UniDic-分類語彙表対応データの補完となり得る番号付与はもちろん、分類語彙表にない番号の新設が要される場合もあり、既存のデータの拡充も可能となる。

今後、本データを利用した語義の曖昧性解消を自然言語処理研究者により進められることを望む。そのために国語研は本アノテーションデータとともに以下の言語資源を提供する予定である：

- 分類語彙表代表義データ（山崎・柏野 2017）
- UniDic 語彙素番号-分類語彙表番号対応表 （近藤・田中 2017）
- 『国語研日本語ウェブコーパス』に基づく word2vec モデル（浅原・岡 2017）
- 『日本語歴史コーパス』に対する分類語彙表番号アノテーション

この他、『日本語歴史コーパス』平安時代編の相の類についての分類語彙表番号アノテーション （池上 2017）や、L1 学習者作文コーパスに対する分類語彙表番号アノテーションが進められている。これらのデータに基づく、通時適応モデルの開発や作文支援システムの構築が期待される。

謝辞

本研究の一部は国語研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」によるものです。

参考文献等

F. Bond, T. Boldwin, R. Fothergill, and K. Uchimoto. 2012. “Japanese SemCor: A Sense-tagged Corpus of Japanese” in The 6th International Conference of the Global WordNet Association (GWC-2012)

K. Maekawa, M. Yamazaki, T. Ogiso, T. Maruyama, H. Ogura, W. Kashino, H. Koiso, M. Yamaguchi, M. Tanaka and Y. Den, 2014. “Balanced corpus of contemporary written Japanese”, Language Resources and Evaluation, 48:2, 345-371.

M. Okumura, K. Shirai, K. Komiya and H. Yokono. 2011. “On SemEval-2010 Japanese WSD Task”,『自然言語処理』18(3), 293-307.

浅原正幸・岡照晃. 2017.「NWJC2vec:『国語研日本語ウェブコーパス』に基づく単語の分散表現データ」, 言語処理学会第 23 回年次大会発表論文集.

池上尚. 2017.「『日本語歴史コーパス 平安時代編』出現形容詞に対する古典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.

小木曽智信・中村壮範. 2014.「『現代日本語書き言葉均衡コーパス』形態論情報アノテーション支援システムの設計・実装・運用」,『自然言語処理』21(2), 301-332.

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.

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.

国立国語研究所（編）. 2004.『分類語彙表増補改訂版データベース』http://pj.ninjal.ac.jp/corpus_center/archive.html#bunruidb