



『現代日本語書き言葉均衡コーパス』への 意味情報(分類語彙表番号と助動詞用法)付与



加藤 祥(目白大学・元国立国語研究所コーパス開発センター)

概要

どのジャンルで
どの意味が現れる？

集計結果を
意味で分析したい！

特定の意味用法の
用例を集めたい！

多義語の
どの意味が高頻度？

類語を
まとめて検索したい！

同じ意味の語は
どのくらい使われる？

『現代日本語書き言葉均衡コーパス』に意味情報を付与することにしました

作業

UniDic語彙素ID(小木曾・中村, 2014)に
分類語彙表番号を手対で対応させたデータ(近藤・田中, 2020)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可能

※多義語の場合、文脈的にどの意味なのか判断が必要
※分類語彙表には機能語の掲載がほとんどない:助動詞に対応が必要

『分類語彙表増補改訂版』(2004)の
分類語彙表番号

『現代語の助詞・助動詞』(1951)の
助動詞用法

文脈上適切な語義(分類語彙表番号)を手対で選択・入力！
※適切な語義(分類語彙表番号)がなければ新たな分類番号を追加

『現代日本語書き言葉均衡コーパス』(Maekawa et al, 2014)の
書籍・新聞・雑誌データ(約35万語)

文脈的な意味分類によってコーパスを調査することが可能になりました

<https://github.com/masayu-a/BCCWJ-WLSP/>

集計例

ジャンルごとの意味分類(大分類)分布の例

媒体	1:関係	2:主体	3:活動	4:生産物	5:自然	対象外	総計
書籍 (54,474語)	46.25%	12.07%	28.97%	4.32%	5.22%	3.17%	100.00%
雑誌 (60,786語)	47.68%	10.99%	28.41%	4.94%	5.18%	2.80%	100.00%
新聞 (66,906語)	45.61%	16.45%	29.22%	3.08%	2.82%	2.81%	100.00%
総計 (182,166語)	46.49%	13.32%	28.88%	4.07%	4.33%	2.91%	100.00%

【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「コーパスアプリケーションの拡張・統合・自動化に関する基礎研究」によるものです。本研究の一部はJSPS科研費 17H00917、18H05521、19K00591、19K00655の助成を受けました。

【参考文献・資料】
小木曾智信・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.
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』vol. 15, No. 2, pp. 134-141.
加藤 祥・浅原 正幸・山崎 誠. (2019)『現代日本語書き言葉均衡コーパス』新聞・書籍・雑誌データの助動詞に対する用法情報付与』、『日本語学会2019年度春季大会』, pp. 169-174.
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』, (18), 77-91.
F. Bond, T. Baldwin, 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.
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://jnl.nijl.ac.jp/corpus_center/archive/jnlBunruidb
『現代日本語書き言葉均衡コーパス』短単位語彙表 ver.1.1. 品詞構成表 ver.1.1

助動詞「う・よう」の用法別前接語の意味(分類語彙表番号中項目)・用法(数値は用例数)

用法・中項目	使役	過去	完了	断定	断定(丁寧)	丁寧	受身	11 関係類	12 関係存在	13 関係様相	15 関係作用	30 活動心	31 活動言語	32 活動芸術	33 活動生活	34 活動行為	35 活動交わり	36 活動待遇	37 活動経済	38 活動事業	57 自然生命	その他	計
① 意志	8					35	1	2	10	1	69	34	12	1	9	76	8	7	17	4	1		295
② 勧誘						5					5	6				4	1			1			22
③ 推量・想像		2	2	191	106	2	1		29	1	6					2			1				347
④ 疑問・質問・反語		2		70	43				5				1		1	4							126
計	8	4	2	261	149	42	2	2	44	2	80	40	13	1	10	86	9	7	18	5	1	4	790