『UniDic』と『分類語彙表』の見出し対応表データの構築

著者	近藤 明日子, 田中 牧郎
雑誌名	言語資源活用ワークショップ発表論文集
巻	1
ページ	79-86
発行年	2017
URL	http://doi.org/10.15084/00001460

『UniDic』と『分類語彙表』の見出し対応表データの構築

近藤 明日子（国立国語研究所コーパス開発センター）†
田中 牧郎（明治大学国際日本学部）‡

Construction of a Correspondence Table between Headwords of UniDic and Headwords of "Word List by Semantic Principles"

KONDO Asuko (National Institute for Japanese Language and Linguistics)
TANAKA Makiro (Meiji University)

要旨

日本語の大規模コーパスへの網羅的・系統的な語義情報付与を目的として、各種大規模コーパスの構築に利用されている形態素解析辞書の元データである電子化辞書 UniDic の見出し（語彙素）と、大規模な現代日本語のシソーラス『分類語彙表増補改訂版データベース』の見出しとを対応づける表形式データの構築を行った（2017 年公開予定）。対応付け作業は UniDic・分類語彙表両者の見出しの読み・表記・類に基づき人手により行い、2017 年 1 月時点で、UniDic 語彙素 50,122 と分類語彙表見出し 64,045 の多対多の関連を表す対応表が構築できている。一方で、見出しの単位設計の違いにより、UniDic 語彙素と対応付けできない分類語彙表見出しの存在も明らかになった。さらに、本対応表を用いた大規模コーパスへの網羅的な語義情報付与に向けて、今後検討すべき課題の存在も明らかになった。

１．はじめに

日本語のコーパスに対する語義情報の付与は、言語研究・自然言語処理の両分野で必要度の高い課題である。意味の面から日本語の語彙全体を分析するためには、日本語の語彙を構成する語が表しうる意味の世界を系統的に分類した語義情報が付与されることが望まれる。また、語義情報を付与するコーパスは日本語の代表性を担保する大規模コーパスとし、さらにそのコーパスを構成する語すべてに網羅的に語義情報を付与することも望まれる。

そのコーパスの形態素解析に使われる形態素解析辞書の見出しデータに語義情報を付与することができれば、その解析結果であるコーパスの各語に語義情報を付与することが可能となる。そこで本研究では、複数の大規模コーパスの形態素解析に利用されている形態素解析辞書の元となる電子化辞書 UniDic の見出しデータと、大規模な現代日本語のシソーラスである国立国語研究所（2004）『分類語彙表増補改訂版データベース』（ver.1.0）[1]（以下、「分類語彙表 DB」という）において意味項目が付与された見出しデータとを対応づけた表形式データを構築した。この対応表を介して、分類語彙表 DB の意味項目を UniDic の見出しに語義情報として付与し、ひいてはそれに対応づけられたコーパスを構成する各語にも語義情報を付与することができるようになる。

２．分類語彙表 DB

まず、対応表の一方に配する分類語彙表 DB のデータについて概説する。分類語彙表 DB は、本格的な現代日本語のシソーラスの先駆である国立国語研究所（編）（1964）『分類語

† kondo@ninjal.ac.jp
‡ makiro@meiji.ac.jp

1 http://pj.ninjal.ac.jp/corpus_center/archive.html#bunruidb

彙表』を増補改訂した国立国語研究所（編）（2004）『分類語彙表増補改訂版』のデータベース版である。分類語彙表 DB での意味分類方式は、番号（以下、「分類番号」という）を用いてそれぞれの分類項目の体系的位置づけを示したところに特徴がある（国立国語研究所（編）2004、p.3）。分類番号は「1.3131」のような 5 桁の数字として表記され、各数字あるいはその組み合わせが「類」「部門」「中項目」「分類項目」という 4 階層の意味的範疇を示す構造となっている（図 1）。

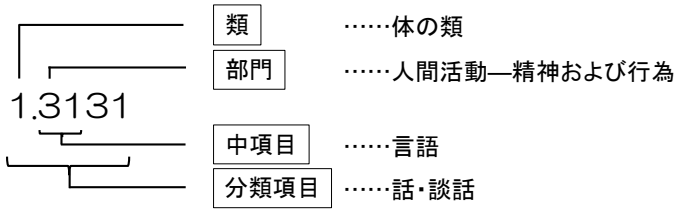

図 1　分類番号の構造

そして、この分類番号と分類項目の中をさらに分類する「段落番号」「小段落番号」、および「小段落番号」内の配列順序を表す「語番号」のもとに 98,241 の見出し[2]を配列するのが分類語彙表 DB のデータである（表 1）。

表 1　分類語彙表 DB データ例

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	見出し本体	読み
体	活動	言語	話・談話	1.3131	1	1	1	話	はなし
体	活動	言語	話・談話	1.3131	1	1	2	話	わ
体	活動	言語	話・談話	1.3131	1	1	3	トーク	とおく
体	活動	言語	話・談話	1.3131	1	2	1	談話	だんわ
体	活動	言語	話・談話	1.3131	1	2	2	談	だん
(…中略…)									
体	活動	言語	問答	1.3132	1	1	1	問答	もんどう
体	活動	言語	問答	1.3132	1	1	2	自問自答	じもんじとう
体	活動	言語	問答	1.3132	1	1	3	一問一答	いちもんいっとう
体	活動	言語	問答	1.3132	1	1	4	応酬	おうしゅう
体	活動	言語	問答	1.3132	1	2	1	禅問答	ぜんもんどう

各見出しは「分類番号」「段落番号」「小段落番号」「語番号」の 4 列により一意となる。多義語の場合は各意味の分類番号が与えられるため、1 語が複数の分類番号に配列され、それぞれ別見出しとなる。

3．UniDic

次に、対応表のもう一方に配する UniDic のデータについて概説する。UniDic とは国立国語研究所が整備している電子化辞書である。コーパスの日本語研究への応用を志向し開発された現代語に対応した辞書（伝ほか 2007）をはじめとして、「近代文語 UniDic」「中古和文 UniDic」（小木曽ほか 2013）等、各時代・文体に対応した複数の形態素解析器 MeCab[3]用

2 分類語彙表 DB 収録の全 101,070 レコードから、書籍版の分割前の見出しであることを表すレコード種別が「B」の 2,589 レコードと意味的区切り「＊」を表す 240 レコードを除いた数。
3 http://taku910.github.io/mecab/

辞書として提供されている[4]。

UniDic の特長として以下の 2 点があげられる（小椋ほか 2011、上 pp.9-10）。

(1) 見出しの単位として「短単位」を採用する。短単位とは、例えば「国立国語研究所に勤務している。」というテキストであれば、「国立｜国語｜研究｜所｜に｜勤務｜し｜て｜いる｜。」と分割する、短い語の単位である。単位の基準が分かりやすく揺れが少ないという長所がある。

(2) 表記や語形の違いにかかわらず、同じ語であれば同一の見出しを与える方針のもと、語を階層化した形で登録する。最上層に国語辞典の見出しに相当する「語彙素」、その下に語形の違いを区別する「語形」、その下に表記の違いを区別する「書字形」を設ける（表 2）。

表 2　UniDic の階層構造

語彙素	語形	書字形
ヤハリ【矢張り】	ヤハリ	矢張り
		やはり
		矢張
	ヤッパリ	やっぱり
		ヤッパリ
		やッぱり
	ヤッパシ	やっぱし
	ヤッパ	やっぱ

UniDic による形態素解析辞書で形態素解析したコーパスとして、国立国語研究所で構築された大規模コーパス『日本語話し言葉コーパス』（CSJ）[5]、『現代日本語書き言葉均衡コーパス』（BCCWJ）[6]、『日本語歴史コーパス』（CHJ）[7]がある。これらのコーパスは短単位によりテキストが区切られ、各短単位に対して UniDic のデータが形態論情報として付与されている。UniDic とコーパスの形態論情報はともに国立国語研究所の形態論情報データベース（小木曽・中村 2011）で管理され、UniDic とコーパスに出現する短単位が対応づけられている。よって、UniDic 見出しに語義情報が付与できれば、コーパスの各短単位に語義情報を付与することが可能となる。

４．分類語彙表 DB と UniDic の見出しの対応付け作業

ここから、本研究の主旨である分類語彙表 DB 見出しと UniDic 見出しとを対応付けた表の構築について述べる。

分類語彙表 DB の各見出しに対応づけるのは UniDic の語の複数の階層のうち語彙素とした。語彙素は、語源が同一であり、かつ意味の違いを生じていない複数の語形をまとめあげるもので（小椋ほか 2011、下 p.78）、語義情報を付与するのに適当な階層である。UniDic 語彙素は「語彙素」「語彙素読み」「語彙素細分類」「類」「語種」の 5 列により一意となる。

分類語彙表 DB の各見出しと UniDic の各語彙素との同語判別を行い、同語であれば対応付けを行った。同語判別に使う条件として以下の(A)～(C)を設けた。

4 http://unidic.ninjal.ac.jp/
5 http://pj.ninjal.ac.jp/corpus_center/csj/
6 http://pj.ninjal.ac.jp/corpus_center/bccwj/
7 http://pj.ninjal.ac.jp/corpus_center/chj/

(A) 分類語彙表 DB 見出しの「見出し本体」と UniDic 語彙素の「語彙素」が一致する[8]
(B) 分類語彙表 DB 見出しの「読み」と UniDic 語彙素の「語彙素読み」が一致する[9]
(C) 分類語彙表 DB 見出しの「類」と UniDic 語彙素の「類」との対応（表 3）が一致する

表 3　分類語彙表 DB と UniDic の「類」の対応

分類語彙表DB	UniDic
体	体 固有名 人名 姓 名 地名 国 数 接尾体
用	用 接尾用
相	相 接尾相
他	他

「類」とは品詞の上位概念に相当するもので、分類語彙表 DB では「体の類」「用の類」「相の類」「その他の類」の 4 種を設ける。UniDic 語彙素にも「類」が設けられており、その区分は分類語彙表 DB よりも細かい。そのため両者の「類」の定義や所属する見出しを参照し、表 3 の対応を設定した。

(A)～(C)の条件がすべて満たされれば同語とすることを原則とした（図 2）。

分類語彙表DB				UniDic		
見出し本体	読み	類		語彙素	語彙素読み	類
事	こと	体	—	事	コト	体

図 2　(A)～(C)による対応付け例

ただし、以下の①～③の例外ルールを設け、(A)～(C)の条件が満たされなくとも同語としたものがある。

① (A)の条件が満たされない場合でも、分類語彙表 DB の「分類項目」や UniDic 語彙素に対応づけられるコーパスの用例等を参照し、同語と判断される場合は同語とする（図 3）。

8 分類語彙表 DB の一部の「見出し本体」には「－周年」「…ている」のように UniDic 語彙素の「語彙素」との同定に不要な記号が含まれているため、この記号を除いたデータを作成し同定した。
9 分類語彙表 DB「読み」と UniDic「語彙素読み」では表記に違いがあるため、「読み」の表記を「語彙素読み」の表記にあわせて変換したデータを作成し同定した。

分類語彙表DB			
見出し本体	読み	類	分類項目
これ	これ	体	こそあど

UniDic		
語彙素	語彙素読み	類
此	コレ	体

図３　①による対応付け例

② (B)の条件が満たされない場合でも、分類語彙表 DB 見出しの「読み」と UniDic 語彙素に所属する「語形」とが一致する場合は同語とする（図 4）

分類語彙表DB		
見出し本体	読み	類
依存	いそん	体

UniDic			
語彙素	語彙素読み	類	語形
依存	イゾン	体	イソン

図４　②による対応付け例

③ (C)の条件が満たされない場合でも、UniDic 語彙素に所属する語形の「品詞」や語彙素に対応づけられるコーパスの用例等を参照し、同語と判断される場合は同語とする（図 5）。

分類語彙表DB		
見出し本体	読み	類
リアル	リアル	相
正式	せいしき	体

UniDic			
語彙素	語彙素読み	類	品詞
リアル	リアル	体	名詞−普通名詞−形状詞可能
正式	セイシキ	相	形状詞−一般

図５　③による対応付け例

以上のルールに則った同語判別作業は、専用の作業用ツールを用いて、人による判断を交え行った。分類語彙表 DB 見出しのうち UniDic に登録されていない語は、UniDic の設計上登録可能であれば新たに UniDic に登録し対応付けを行った。

５．対応づけ作業結果

2017 年 1 月現在、分類語彙表 DB の全見出しについて、ひととおり UniDic 語彙素との同語判別を終え、同語判別を保留している分類語彙表 DB 見出し約 700 を除き、分類語彙表 DB の 64,045 見出しと UniDic の 50,122 語彙素との多対多の関連を表す対応表が構築できている。

対応表に見られる対応付けの例として、分類語彙表 DB 見出しが多で UniDic 語彙素が一の対応の例を図 6 にあげる。

分類語彙表DB				
見出し本体	読み	類	分類番号	分類項目
出す	だす	用	2.3832	出版・放送
出す	だす	用	2.3770	授受
出す	だす	用	2.1531	出・出し
出す	だす	用	2.1521	移動・発着
ー出す	だす	用	2.1502	開始
出す	だす	用	2.1211	発生・復活
出す	だす	用	2.1210	出没

UniDic		
語彙素	語彙素読み	類
出す	ダス	用

図 6　分類語彙表 DB 見出しが多、UniDic 語彙素が一の対応例

このような分類語彙表 DB 見出しが多で UniDic 語彙素が一の対応は例が多く、分類語彙表 DB 見出しと対応づけられた UniDic 語彙素 50,122 の 21%にあたる 10,490 語彙素がそれぞれ複数の分類語彙表 DB 見出しと対応づけられた。一つの UniDic 語彙素に対して対応づけられる分類語彙表 DB 見出しの最大数は 13 にのぼる（表 4）。

表 4　分類語彙表 DB 見出しと対応する UniDic 語彙素数

対応する分類語彙表DB見出し数	UniDic語彙素数
1	39,632
2	8,321
3	1,421
4	501
5	113
6	72
7	28
8	15
9	8
10	6
11	2
12	2
13	1
計	50,122

逆に、分類語彙表 DB 見出しが一で UniDic 語彙素が多の対応の例を図 7 にあげる。

分類語彙表DB				
見出し本体	読み	類	分類番号	分類項目
小じゅうと	こじゅうと	体	1.2140	兄弟

UniDic		
語彙素	語彙素読み	類
小舅	コジュウト	体
小姑	コジュウト	体

図 7　分類語彙表 DB 見出しが一、UniDic 語彙素が多の対応例

このような分類語彙表 DB 見出しが一で UniDic 語彙素が多の対応は例が少なく、分類語彙表 DB の 21 見出しに限られ、対応づけられる UniDic 語彙素の最大数も 2 にとどまる。

ところで、分類語彙表 DB 全 98,241 見出しの 35%に相当する 34,822 見出しが UniDic 語彙

素と対応付けができなかったことになるが、これは見出しの単位設計の相違によるものである。UniDic は短単位の語を見出しとするのに対し、分類語彙表 DB は「有機物質」「図示する」「詭弁を弄する」といった短単位を複数つなげた合成語や連語・慣用句の類も見出しとして収録する。このような見出しは UniDic には設計上登録できないので対応付けできなかった。

６．今後の課題

構築した対応表を用いた大規模コーパスへの網羅的な語義情報付与を目指す上で、今後検討すべき課題について述べる。

第一に、コーパスへの網羅的語義情報付与のために必要な、UniDic 語彙素に対する網羅的分類番号付与についての課題がある。５．で述べたとおり分類語彙表 DB 見出しと対応づけられた UniDic 語彙素数は 50,122 であり、これは、UniDic に登録されている全 181,241 語彙素[10]（2017 年 1 月時点）の 28%に過ぎない。残る 131,119 語彙素は分類語彙表 DB に未収録の語のため、分類語彙表 DB 見出しと対応付けがとれず、分類番号が付与できない。これらの語彙素への分類番号の付与は今後の課題である。古典語であれば、宮島ほか（編）(2014)『日本古典対照分類語彙表』のデータと UniDic との対応表を別途作成し、UniDic 語彙素に分類番号を付与する方法が考えられる。それでも分類番号が付与されない語彙素については、人手による付与等の方法を検討する必要がある。

第二に、一つの UniDic 語彙素が複数の分類語彙表 DB 見出しと対応づけられる多義語についての課題がある。多義語がテキストの文脈内で用いられる場合、一般には複数の語義のうち一つが用いられる。よって、コーパスの各短単位に付与される分類番号は通常 1 種類ずつとなる。コーパスへの語義情報付与作業では、複数の分類番号が対応づけられる UniDic 語彙素の場合、その中の一つの分類番号を選択する工程が必要となる。人手による選択、語義の曖昧性解消の技術を用いた自動選択等の方法を今後検討する必要がある。

第三に、語義情報を付与する単位についての課題がある。本対応表によって実現するコーパスへの語義情報の付与は短単位に対するものである。しかし、コーパスの利用目的によっては、たとえば「勤務する」を「勤務」と「する」の短単位に分割してそれぞれに分類番号を付与するのではなく、「勤務する」全体に分類番号を付与することが要求される場合もあるだろう。この対処法として、UniDic の設計にあるもう一つの単位「長単位」に対して語義情報を付与することが考えられる。長単位は文節を自立語部分と付属語部分とに分割して得られる長い語の単位で（小椋ほか 2011、上 p.4)、たとえば「国立国語研究所に勤務している。」というテキストは「国立国語研究所｜に｜勤務し｜ている｜。」と分割される。CSJ・BCCWJ・CHJ には長単位による形態論情報も付与されており、これに語義情報を付与することは理論上可能である。分類語彙表 DB には長単位に相当する見出しも収録されており、これを利用して長単位による対応表を作成することも考えられるが、長単位の異なり語数は短単位より多くなるため、分類語彙表 DB 収録の見出しだけではコーパスへの網羅的語義情報付与には対処できない。今後別の方法の検討が必要である。

７．おわりに

以上、大規模コーパスへの網羅的・系統的な語義情報付与を目的とした、UniDic・分類語彙表見出し対応表データの構築について延べた。本対応表は 2017 年中に公開予定である。本対応表を用いた BCCWJ への語義情報付与作業は既に始まっており、６．にあげた多義語の語義選択や長単位への語義情報付与といった課題に対する検討も実作業を通じて行われつつある（加藤ほか 2017 予定)。今後、本対応表を用いて、網羅的・系統的に語義情報が

10 分類語彙表 DB に積極的に収録されない固有名詞・助詞・助動詞・記号類を除いた数。

付与されたコーパスの構築が進展することが期待される。

謝辞

本研究は、科研費特定領域研究「言語政策に役立つ、コーパスを用いた語彙表・漢字表等の作成と活用」（18061008）および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」、国立国語研究所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」の成果の一部である。

参考文献

小木曽智信・小町守・松本裕治（2013）「歴史的日本語資料を対象とした形態素解析」『自然言語処理』20(5), pp.727-748

小木曽智信・中村壮範（2011）『『現代日本語書き言葉均衡コーパス』形態論情報データベースの設計と実装 改訂版』国立国語研究所（特定領域研究「日本語コーパス」平成 22 年度研究成果報告書）

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕（2011）『『現代日本語書き言葉均衡コーパス』形態論情報規程集 第 4 版（上）（下）』（特定領域研究「日本語コーパス」平成 22 年度研究成果報告書）

加藤祥・浅原正幸・山崎誠（2017 予定）「『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の試行」『言語資源活用ワークショップ 2016 予稿集』

国立国語研究所（2004）『分類語彙表増補改訂版データベース』（ver.1.0）
http://pj.ninjal.ac.jp/corpus_center/archive.html#bunruidb

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

伝康晴・峯松信明・小木曽智信・内本清貴・小椋秀樹・小磯花絵・山田篤（2007）「コーパス日本語学のための言語資源―形態素解析用電子化辞書の開発とその応用」『日本語科学』22, pp.101-123

宮島達夫・石井久雄・安部清哉・鈴木泰（編）（2014）『日本古典対照分類語彙表』笠間書院