「分類語彙表番号－UniDic語彙素番号対応表」の構築

著者	近藤 明日子, 田中 牧郎
雑誌名	国立国語研究所論集
号	18
ページ	77-91
発行年	2020-01
URL	http://doi.org/10.15084/00002542

国立国語研究所論集 (NINJAL Research Papers) 18: 77–91 (2020)
ISSN: 2186-134X print/2186-1358 online

「分類語彙表番号－UniDic 語彙素番号対応表」の構築

近藤明日子[a]　　田中牧郎[b]

[a] 国立国語研究所 コーパス開発センター 非常勤研究員
[b] 明治大学／国立国語研究所 共同研究員

要旨

日本語の大規模コーパスへの網羅的・体系的な語義情報付与を目的として，語義の体系的な分類を示す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，各種大規模コーパスの構築に利用されている電子化辞書 UniDic の見出し（語彙素）との同語関係による対応を表す表形式データの構築を行った。同語判別の作業は分類語彙表・UniDic 両者の見出しの表記・読み・類の対応に基づいて人手により行い，その結果，『分類語彙表』の 64,759 見出しと UniDic の 50,795 語彙素との同語関係による多対多の対応を表す「分類語彙表番号－UniDic 語彙素番号対応表」を構築した。本対応表を活用して大規模コーパスへの網羅的な語義情報付与作業が始まっており，また，形態素解析結果に分類語彙表番号を付与する機能を実装した形態素解析ツールも開発された。一方で，本格的な大規模コーパスへの語義情報の網羅的付与に向けて，対応表の拡張や多義語の語義選択といった課題への対処も必要である *。

キーワード：分類語彙表，UniDic，対応表，大規模日本語コーパス，語義情報付与

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体系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，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）（以下，「分類語彙表 DB」と呼ぶ）において意味項目が付与された見出しリストと，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見出しの元データである電子化辞書 UniDic の見出しリストとの，同語関係による多対多の関連を表す表形式データ「分類語彙表番号－ UniDic 語彙素番号対応表」（以下，「対応表」と呼ぶ）を構築した。この

* 本研究の内容は，言語資源活用ワークショップ 2016（2017 年 3 月 7–8 日開催）での発表（近藤・田中 2017a），言語処理学会第 23 回年次大会（2017 年 3 月 13–17 日開催）での発表（近藤・田中 2017b），The 8th Conference of Japanese Association for Digital Humanities（2018 年 9 月 9–11 日開催）での発表（Kondo, Tanaka, and Asahara 2018）に基づいている。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」（プロジェクトリーダー：浅原正幸）の研究成果である。

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 UniDic 見出しに紐付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。

本研究では，対応表の構築方法を中心として，対応表における分類語彙表 DB と UniDic との同語関係の分布，対応表の活用，大規模コーパスへの語義情報の網羅的付与に向けての課題について論ずる。

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 のデータについて説明する。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」と呼ぶ）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: 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）。

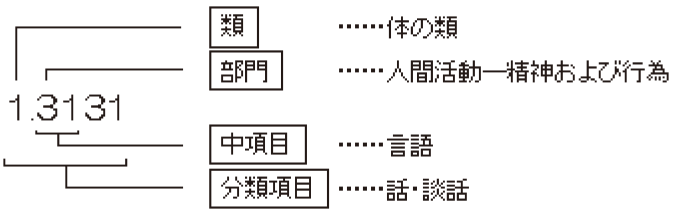

図 1　分類番号の構造

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 [1] を配列するのが分類語彙表 DB のデータである（表 1）。

[1] 分類語彙表 DB 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。

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 列により一意となり，4 列の数字を連結した「分類語彙表番号」が各見出しに ID として付与されている。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。また，「少食」「小食」のような表記の違いや「昆布（こんぶ）」「昆布（こぶ）」のような語形の違いもそれぞれ区別して別見出しとなる。見出しの単位は，「話」「談話」「トーク」のような UniDic の採用する短単位（後述）と同じものから，「禅問答」「一問一答」「ここだけの話」「口をへの字に結ぶ」のような短単位を超えるもの，あるいは「客－」（「客冬」の「客」）のような短単位未満のものまで幅広い。

3. UniDic

次に，対応表のもう一方に配する UniDic のデータについて説明する。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。「現代書き言葉 UniDic」「現代話し言葉 UniDic」「近代文語 UniDic」「中古和文 UniDic」等，各時代・文体の日本語に対応した複数の形態素解析器 MeCab 用辞書として提供されている。

UniDic の特長として以下の 2 点があげられる。

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（小椋・小磯・冨士池ほか 2011，上：9）。

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）（小椋・小磯・冨士池ほか 2011，上：10）。

表 2　UniDic の階層構造の例

語彙素	語　形	書字形
矢張り	ヤハリ	やはり 矢張り
	ヤッパリ	やっぱり 矢っ張り
	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），『現代日本語書き言葉均衡コーパス』（BCCWJ），『日本語歴史コーパス』（CHJ）等がある。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紐付けられている。よって，UniDic の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。

表 2 に示したように UniDic の語には 3 階層があるが，そのうち分類語彙表 DB の見出しに対応付けるのは語彙素とする。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・小磯・冨士池ほか 2011，下：78），語義情報を付与するのに適当な階層である。各語彙素は「語彙素」「語彙素読み」「語彙素細分類」「類」「語種」の 5 列により一意となり，「語彙素 ID」が付与されている（表 3）。

表 3　UniDic 語彙素データ例

語彙素 ID	語彙素	語彙素読み	語彙素細分類	類	語種
71	あい	アイ		他	和
72	アイ	アイ	I	体	外
73	合い	アイ		体	和
74	合い	アイ		接尾体	和
75	哀	アイ		体	漢
		(…中略…)			
153	アイスランド	アイスランド		国	固
154	アイスラー	アイスラー		人名	固
155	愛する	アイスル		用	混
156	合図	アイズ		体	混
157	哀惜	アイセキ		体	漢
158	愛惜	アイセキ		体	漢
159	愛想	アイソ		体	漢

UniDic 語彙素は国立国語研究所でのコーパス構築に連動して日々増補されており，2019 年 5 月時点で約 263,000 語彙素が登録されている。

4. 対応表の構築

本研究で構築する対応表とは，分類語彙表 DB の 98,241 見出しと UniDic の約 263,000 語彙素との間の多対多の同語関係を記述するものである。これらの見出し・語彙素のうち対応表への登録対象となるのは，UniDic 語彙素と対応付けの可能な短単位の分類語彙表 DB 見出しということになる。そのため，まず分類語彙表 DB の全 98,241 見出しに対して，「1 短単位」「複数短単位」「短単位未満」の 3 種の単位情報を付与する作業を人手で行い，分類語彙表 DB 見出しの中から短単位の見出し 64,759 を抽出した（表 4）。

表 4　分類語彙表 DB の単位別見出し数

単位	見出し数
1 短単位	64,759
複数短単位	33,477
短単位未満	5
計	98,241

この 64,759 見出しについて人手により UniDic 語彙素との同語判別を行い，同語関係が認められれば，その見出し・語彙素の対を 1 レコードとして対応表に登録した。対応表は多対多の同語関係を記述する表であるから，一つの分類語彙表 DB 見出しが複数の UniDic 語彙素と同語関係にある場合や，一つの UniDic 語彙素が複数の分類語彙表 DB 見出しと同語関係にある場合もあり，これらについてもそれぞれの見出し・語彙素の対を 1 レコードとして登録した。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 の設計に沿って新たに UniDic に登録し，分類語彙表見出しとの対応付けを行った。これにより構築した対応表は，分類語彙表 DB の 64,759 見出しと UniDic の 50,795 語彙素の間の多対多の同語関係を表す全 65,043 レコードからなるものとなった[2]。

対応表構築作業の主眼となるのは，分類語彙表 DB 見出しと UniDic 語彙素との間の同語判別である。同語判別は最終的にはすべて人手により行ったが，作業の効率化を図るため，同語判別の手がかりとして以下の条件（A）～（C）を設定し，条件（A）～（C）をすべて満たす関係にある分類語彙表 DB 見出しと UniDic 語彙素とが同語の可能性が非常に高いという見込みのもと，優先的に作業を行った。

（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[3]
（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[4]
（C）分類語彙表 DB 見出しの「類」と UniDic 語彙素の「類」との対応が一致する

[2] 以下に示すレコード数等の対応表に関するデータは，バージョン 1.0.2 の対応表に基づく。
[3] 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。
[4] 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，分類語彙表 DB「読み」の表記を UniDic「語彙素読み」の表記に合わせて変換したデータを作成し同定した。

（A）は表記の一致を見るもの，（B）は読み（語の外形）の一致を見るもの，（C）は「類」の一致を見るものである。（C）の「類」とは品詞の上位概念に相当するもので，分類語彙表 DB 見出しでは「体」「用」「相」「他」の 4 種を設けるのに対し，UniDic 語彙素では 25 種を設け，分類語彙表 DB よりも細分化されており，両者の「類」の直接の一致をとることはできない。そのため，両者の「類」の定義や各「類」に所属する見出し・語彙素の概覧により，両者の「類」の対応関係が判断できるものについては対応表を作成し，それにより両者の「類」の一致を判定した。ただし，UniDic 語彙素の「類」のうち，分類語彙表 DB 見出しの「類」との対応関係が判断できないものについては対応は未設定とした（表 5）。

表 5　分類語彙表 DB と UniDic の「類」の対応表

分類語彙表 DB 見出し「類」	UniDic 語彙素「類」
体	体 固有名 人名 姓 名 地名 国 数 接尾体 助数
用	用 接尾用
相	相 接尾相
他	他
（対応未設定）	助動 格助 副助 係助 接助 終助 準助 接頭 記号 補助

条件（A）～（C）をすべて満たし，同語と判別され，対応表にレコードとして登録した見出し・語彙素の対の例を表 6 にあげる。

表 6　条件（A）～（C）をすべて満たす対応表レコード例

分類語彙表 DB 見出し			UniDic 語彙素		
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
－日	か	体	日	カ	助数
関する	かんする	用	関する	カンスル	用
－めく	めく	用	めく	メク	接尾用
正しい	ただしい	相	正しい	タダシイ	相
－的	てき	相	的	テキ	接尾相
且つ	かつ	他	且つ	カツ	他

ただし，分類語彙表 DB の分類番号や UniDic 語彙素に紐付けられたコーパスの用例を参照し，条件（A）～（C）をすべて満たす見出し・語彙素の対であっても同語関係にないと判断されるものについては，対応表には登録しなかった。例えば，表 7 にあげる分類語彙表 DB 見出しと UniDic 語彙素の対は，分類語彙表 DB 見出しのほうは分類番号により表される類－部門－中項目－分類項目が「体－活動－心－方法」であることから，「聞く方法」の意の語と判断されるのに対し，UniDic 語彙素のほうは短単位の規定[5]や語彙素に紐付けられたコーパスの用例から，「聞く側・聞き手」の意で用いられる語と判断され，両者の間に同語関係は認められない。このような対は対応表には登録しなかった。

表 7　条件（A）～（C）をすべて満たすが同語関係にない見出し・語彙素の対の例

分類語彙表 DB 見出し				UniDic 語彙素		
見出し本体	読み	類	分類番号	語彙素	語彙素読み	類
聞き方	ききかた	体	1.3081	聞き方	キキカタ	体

一方で，条件（A）～（C）のいずれか，あるいはすべてを満たさなくとも，同語関係を認め対応表に登録した見出し・語彙素の対もある。

まず，条件（A）を満たさない場合でも，分類語彙表 DB の分類番号や UniDic 語彙素に紐付けられたコーパスの用例等を参照し，同語と判断した場合がある。そのレコード例を表 8 にあげる。

表 8　条件（A）を満たさないレコード例

	分類語彙 DB 見出し			UniDic 語彙素		
	見出し本体	読み	類	語彙素	語彙素読み	類
(1)	これ	これ	体	此	コレ	体
(2)	する	する	用	為る	スル	用
(3)	若若しい	わかわかしい	相	若々しい	ワカワカシイ	相
(4)	篤い	あつい	相	厚い	アツイ	相

[5] 短単位の規定では，「聞く方法」の意の「聞き方」は「聞き｜方」と 2 短単位に分割される。よって，1 短単位を 1 語として登録する設計の UniDic には「聞く方法」の意の「聞き方」は登録できない。

分類語彙表 DB の「見出し本体」と UniDic 語彙素の「語彙素」の不一致の多くは，表 8 の（1）～（3）のように「見出し本体」と「語彙素」の表記法の違いに由来する。その他には（4）の「篤い」と「厚い」の対のように，分類語彙表 DB では意味の違いに応じて「厚い」と「篤い」を別見出しとしているのに対し，UniDic では多義の 1 語と見なし 1 語彙素として扱っているという，分類語彙表 DB 見出しと UniDic 語彙素との間の同語判別の基準の違いに由来するものもある。

次に，条件（B）を満たさない場合でも，分類語彙表 DB 見出しの「読み」と UniDic 語彙素の下層に所属する「語形」とが一致する場合は同語とした。そのレコード例を表 9 にあげる。

表 9　条件（B）を満たさないレコード例

分類語彙表 DB 見出し			UniDic 語彙素			UniDic 語形
見出し本体	読み	類	語彙素	語彙素読み	類	語形
依存	いそん	体	依存	イゾン	体	イソン
行き詰まる	いきづまる	用	行き詰まる	ユキヅマル	用	イキヅマル
尊い	たっとい	相	尊い	トウトイ	相	タットイ

分類語彙表 DB「読み」と UniDic「語彙素読み」の不一致は，表 9 にあげた例のように，分類語彙表 DB では「読み」が異なれば別見出しとするのに対し，UniDic では語形の違いにかかわらず同じ語であれば同一の語彙素とし，異語形は語彙素に包含されるという同語判別の基準の違いに由来するものである。

次に，条件（C）を満たさない場合でも，UniDic 語彙素の下層に所属する語形の「品詞」や語彙素に紐付けられたコーパスの用例等を参照し，同語と判断した場合がある。そのレコード例を表 10 にあげる。

表 10　条件（C）を満たさないレコード例

	分類語彙表 DB 見出し			UniDic 語彙素			UniDic 語形
	見出し本体	読み	類	語彙素	語彙素読み	類	品詞
(1)	正式	せいしき	体	正式	セイシキ	相	形状詞 - 一般
(2)	リアル	りある	相	リアル	リアル	体	名詞 - 普通名詞 - 形状詞可能
(3)	今度	こんど	相	今度	コンド	体	名詞 - 普通名詞 - 副詞可能
(4)	ぴよぴよ	ぴよぴよ	他	ぴよぴよ	ピヨピヨ	相	副詞

分類語彙表 DB「類」と UniDic 語彙素「類」の不一致の多くは，表 10 の（1）～（3）の例に見られるように，分類語彙表 DB と UniDic で「体」と「相」の認定基準が異なることに由来する。分類語彙表 DB では，名詞と形容動詞語幹との間の「類」の認定に多少の動揺を許容し，また，時に関する語は「体」「相」両方に重出させる（国立国語研究所（編）2004: 12）。一方 UniDic では，名詞と形容動詞語幹（UniDic では「形状詞」と称する）の両方の用法をもつ語のために「名詞－普通名詞－形状詞可能」という品詞を設け，時を表す語のように名詞と副詞の両用法をもつ語のために「名詞－普通名詞－副詞可能」という品詞を設け，一括して名詞＝「体」の類として

扱うという違いがある。その他に（4）のような，動物の鳴き声は分類語彙表 DB では「他」，UniDic では「相」とする「類」の認定基準の違いに由来するもの等もある。

表 8・表 9・表 10 には条件（A）（B）（C）のいずれか一つを満たさないレコードの例をあげたが，それ以外に複数の条件を満たさない対でも同語と判断し，対応表のレコードとした場合がある。それらのレコード例を表 11 にあげる。

表 11　複数の条件を満たさないレコード例

	分類語彙表 DB 見出し			UniDic 語彙素		
	見出し本体	読み	類	語彙素	語彙素読み	類
条件（A）（B）を満たさない	（1） メディアン	めでぃあん	体	メジアン	メジアン	体
	（2） 信じる	しんじる	用	信ずる	シンズル	用
条件（A）（C）を満たさない	（3） 抑えめ	おさえめ	体	押さえめ	オサエメ	相
	（4） 近ごろ	ちかごろ	相	近頃	チカゴロ	体
条件（B）（C）を満たさない	（5） 幅広	はばびろ	体	幅広	ハバヒロ	相
	（6） 非力	ひりょく	相	非力	ヒリキ	体
条件（A）（B）（C）を満たさない	（7） インタラクティブ	いんたらくてぃぶ	相	インターラクティブ	インターラクティブ	体
	（8） 休め	やすめ	体	休む	ヤスム	用
	（9） いらっしゃい	いらっしゃい	他	いらっしゃる	イラッシャル	用

表 11 の（1）～（7）の例に見られるように，複数条件を満たさないレコードの多くは，これまでに述べた分類語彙表 DB と UniDic との間の表記法・同語判別基準・類認定基準の違いが複合しているものである。その他に注意すべきレコードとして，（8）（9）のように分類語彙表 DB で活用語が終止形以外の活用形で見出しとなっているものがある。このような分類語彙表 DB 見出しは終止形にした形を UniDic 語彙素と対応付け，同語関係を認めた。

以上見てきたように，対応表のレコードとして登録した，同語関係が認められる分類語彙表 DB 見出し・UniDic 語彙素の対は，条件（A）（B）（C）の適合・不適合のパターンにより分類することができる。そのパターン別レコード数を表 12 に示す。表中，○は該当条件を満たすレコード，×は該当条件を満たさないレコードであることを表す。レコード数の降順にパターンを掲出する。

表 12　条件の適合・不適合のパターン別レコード数

条件（A）	条件（B）	条件（C）	レコード数
○	○	○	52,158
×	○	○	9,283
○	○	×	2,015
×	×	○	677
○	×	○	490
×	○	×	345
×	×	×	47
○	×	×	28
計			65,043

表 12 から，条件（A）～（C）をすべて満たすレコードは 52,158 と全体の 80% を占めることが分かり，「条件（A）～（C）をすべて満たす関係にある分類語彙表 DB 見出しと UniDic 語彙素は同語の可能性が非常に高い」という対応表構築作業当初の見込みが妥当であったことが確認できる。次にレコード数の多いものは，条件（B）（C）を満たし条件（A）は満たさない 9,283 レコードで全体の 14% にあたる。次いでレコード数の多いものは条件（A）（B）を満たし条件（C）を満たさない 2,015 レコードで全体の 3% にあたる。

条件ごとに見ると，条件（A）を満たさないレコードの合計は 10,352 レコードで全体の 19%，条件（B）を満たさないレコードの合計は 1,242 レコードで全体の 2%，条件（C）を満たさないレコードは 2,435 レコードで全体の 4% にあたる。このことは，分類語彙表 DB 見出しと UniDic 語彙素とでは表記法の違いが比較的大きく，同語判別基準や類認定基準の違いは比較的小さいことを示している。

5. 分類語彙表 DB・UniDic 間の多対多の同語関係の分布

次に，対応表によって表される分類語彙表 DB 見出しと UniDic 語彙素との間の多対多の同語関係の分布について見ていく。

まず，一つの UniDic 語彙素に対して同語関係にある分類語彙表 DB 見出し数の分布を表 13 に示す。

表 13　一つの UniDic 語彙素と同語関係にある分類語彙表 DB 見出し数の分布

同語関係にある分類語彙表 DB 見出し数	UniDic 語彙素数
1	40,113
2	8,449
3	1,449
4	529
5	111
6	77
7	28
8	18
9	9
10	6
11	2
12	3
13	1
計	50,795

表 13 から分かるように，一つの分類語彙表 DB 見出しとのみ同語関係にある UniDic 語彙素が 40,113 と最も多く，全体の 79% を占める。残る 21% にあたる 10,682 語彙素が複数の分類語彙表

DB 見出しと同語関係にあり，同語関係にある見出し数は最大 13 にのぼる。複数の分類語彙表 DB 見出しと同語関係にある UniDic 語彙素の例を表 14 にあげる。

表 14　複数の分類語彙表 DB 見出しと同語関係にある UniDic 語彙素の例

分類語彙表 DB 見出し					UniDic 語彙素		
見出し	読み	類	分類番号	部門 - 中項目 - 分類項目	語彙素	語彙素読み	類
－掛ける［話し～･働き～］	かける	用	2.1110	関係 - 類 - 関係	掛ける	カケル	用
－かける	かける	用	2.1502	関係 - 作用 - 開始			
掛ける	かける	用	2.1513	関係 - 作用 - 固定・傾き・転倒など			
懸ける	かける	用	2.1513	関係 - 作用 - 固定・傾き・転倒など			
架ける	かける	用	2.1513	関係 - 作用 - 固定・傾き・転倒など			
掛ける	かける	用	2.1535	関係 - 作用 - 包み・覆いなど			
掛ける［手間暇～］	かける	用	2.1600	関係 - 時間 - 時間			
掛ける	かける	用	2.3064	活動 - 心 - 測定・計算			
掛ける	かける	用	2.3710	活動 - 経済 - 経済・収支			
賭（か）ける	かける	用	2.3710	活動 - 経済 - 経済・収支			
掛ける	かける	用	2.3730	活動 - 経済 - 価格・費用・給与など			
架ける	かける	用	2.3823	活動 - 事業 - 建築			
ちょっと	ちょっと	相	3.1600	関係 - 時間 - 時間	一寸	チョット	相
ちっと	ちっと	相	3.1910	関係 - 量 - 多少			
ちと	ちと	相	3.1910	関係 - 量 - 多少			
ちょっと	ちょっと	相	3.1910	関係 - 量 - 多少			
ちょいと	ちょいと	相	3.1910	関係 - 量 - 多少			
ちょっと	ちょっと	相	3.1920	関係 - 量 - 程度			
ちょっと	ちょっと	相	3.3000	活動 - 心 - 心			
ちょいと	ちょいと	他	4.3200	*- 呼び掛け - 呼び掛け・指図			
ちょっと	ちょっと	他	4.3200	*- 呼び掛け - 呼び掛け・指図			

表 14 のような分類語彙表 DB 見出しと UniDic 語彙素の多対一の同語関係は，表 16（後述）にあげる一対多の同語関係より多い。これは，分類語彙表 DB が多義語の意味ごとに見出しを立て日本語の表しうる意味の世界を示そうとするシソーラスであるのに対し，UniDic が微妙な意味の差やそれに対応する語の書き分けをできるだけまとめあげて 1 語彙素とし，形態素解析の精度を保持しようとする形態素解析辞書用データであるという両者の使用目的の違いによりもたらされたものである。

次に，一つの分類語彙表 DB 見出しに対して同語関係にある UniDic 語彙素数の分布を表 15 に示す。

表 15　一つの分類語彙表 DB 見出しと同語関係にある UniDic 語彙素数

同語関係にある UniDic 語彙素数	分類語彙表 DB 見出し数
1	64,489
2	256
3	14
計	64,759

表 15 から分かるように，一つの UniDic 語彙素とのみ同語関係にある分類語彙表 DB 見出しが 64,489 と最も多く，全体の 99% 以上を占める。複数の UniDic 語彙素と同語関係にある分類語彙表 DB 見出しは 270 に過ぎず，同語関係にある UniDic 語彙素の最大数も 3 にとどまる。複数の UniDic 語彙素と同語関係にある分類語彙表 DB 見出しの例を表 16 にあげる。

表 16　複数の UniDic 語彙素と同語関係にある分類語彙表 DB 見出しの例

	分類語彙表 DB 見出し					UniDic 語彙素		
	見出し	読み	類	分類番号	部門－中項目－分類項目	語彙素	語彙素読み	類
(1)	骨	こつ	体	1.5606	自然－身体－骨・歯・爪・角・甲	骨	コツ	体
						骨	コツ	接尾体
						骨	コツ	接頭
(2)	至極［～便利・迷惑～］	しごく	相	3.1920	関係－量－程度	至極	シゴク	体
						至極	シゴク	相
(3)	小じゅうと	こじゅうと	体	1.2140	主体－家族－兄弟	小姑	コジュウト	体
						小舅	コジュウト	体

複数の UniDic 語彙素と同語関係にある分類語彙表 DB 見出しのほとんどを占めるのは，表 16 の（1）のように，名詞・接頭辞・接尾辞のいずれの用法を指しているのか明示的でないものである。分類語彙表 DB 見出しの中には「非－」「－的」のように表記に「－」を用いて接頭辞・接尾辞の用法を明示的に指すものがある一方，(1)「骨」のように「－」は用いていないが名詞だけでなく接頭辞・接尾辞の用法を含むと考えられる見出しがあり，これについては UniDic 語彙素の類「体」「接尾体」「接頭」それぞれと同語関係にあると認め，対応表に登録した。その他に，(2) のように「体」「相」の類認定の違いに由来するものや，(3) のように表記法・同語判別基準の違いに由来するものも少数ある。

6. 対応表の活用によるコーパスへの語義情報の付与

対応表の構築の目的は，日本語の大規模コーパスへの網羅的・体系的な語義情報付与であった。現在，対応表を活用して BCCWJ の書籍・新聞・雑誌データに対し『分類語彙表』に基づいた語義情報付与が行われ，そのデータが公開された（加藤・浅原・山崎 2019）。また，現代語のコーパスだけでなく CHJ に対して対応表を活用して『分類語彙表』に基づいた語義情報付与を行う

作業も進められている（浅原・加藤・鈴木ほか 2018）。

また，対応表の活用により，形態素解析辞書「現代書き言葉 UniDic」「現代話し言葉 UniDic」による形態素解析結果に分類語彙表番号を付与する機能を実装した形態素解析ツール「ChaMame」が開発･公開された[6]。これにより，既存のコーパスだけでなく，今後構築されるコーパスにも語義情報を自動付与できることになり，語義情報付きコーパスの構築がさらに広がることが期待される。

7. コーパスへの網羅的な語義情報付与に向けての課題

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の拡大を目指す上で，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な分類番号付与についての課題がある。4 で述べたとおり，対応表で分類語彙表 DB 見出しと対応づけられた UniDic 語彙素数は 50,795 であり，これは UniDic に登録されている約 263,000 語彙素の 2 割程度に過ぎない。残る 8 割の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 見出しと対応がとれず，分類番号が未付与である[7]。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。加藤・浅原・山崎（2019）は，実際の BCCWJ への語義情報付与作業において，コーパスに出現した語彙素に対応する語義が対応表に登録されていない場合，適当な分類番号を作業者が選択し，分類語彙表 DB に適当な分類番号がない場合は新たな分類番号を作成して語義情報を付与するという方法をとり，網羅的な語義情報付与を行った。そして，実際のコーパスへの語義情報付与作業を通して対応表を拡充することが可能であることを示した。また，古典語であれば，分類番号を付与した古典語の語彙表（宮島・石井・安部ほか（編）2014；宮島（編）2015）のデータと UniDic との対応表を別途作成し，UniDic 語彙素に新たに分類番号を付与することが考えられる[8]。

第二に，一つの UniDic 語彙素が複数の分類語彙表 DB 見出しと同語関係にあ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，複数の分類語彙表 DB 見出しと同語関係にある 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。浅原・加藤・鈴木ほか（2018），加藤・浅原・山崎（2019）では人手による選択の方法が報告されており，対処法の一例として参考となる。他には，鈴木・古宮・浅原ほか（2019）等に示される語義の曖昧性解消の技術を用い

[6] 分類語彙表番号付与機能を実装した「ChaMame」は「現代書き言葉 UniDic」「現代話し言葉 UniDic」に同梱されている。また，ウェブページから単体でダウンロードすることも可能である。
[7] なお，加藤・浅原・山崎（2017）では，BCCWJ の新聞の 54 サンプル中の自立語全 33,725 短単位（延べ）のうち，対応表の UniDic 語彙素とマッチしたものは 28,696 短単位であったとの報告がある。ここから，実際のコーパスに出現する短単位のうち対応表によって何らかの分類番号が付与されるものの割合は 8 割以上あり，延べ語数ベースでは比較的高い割合を対応表がカバーしていることが分かる。
[8] 浅原・加藤・鈴木ほか（2018）に宮島・石井・安部ほか（編）（2014）と UniDic との対照表の整備が報告されている。

た自動選択の方法も検討する必要があろう。

第三に，語義情報を付与する単位についての課題がある。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・小磯・冨士池ほか 2011，上：4），例えば「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[9]。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。

8. おわりに

以上，大規模コーパスへの網羅的・体系的な語義情報付与を目的とした「分類語彙表番号－UniDic 語彙素番号対応表」の構築について論じた。対応表は国立国語研究所ウェブサイト内の『分類語彙表』に関するページ（https://pj.ninjal.ac.jp/corpus_center/goihyo.html）で公開中である。対応表を用いた大規模コーパスへの語義情報付与は既に始まっており，その実作業を通じてコーパスへの網羅的な語義情報付与に向けての課題への対処法も示されつつある。今後，対応表を用いて，網羅的・体系的に語義情報が付与されたコーパスの構築が拡大し，意味の面からの日本語の語彙研究が進展することが期待される。

参照文献

浅原正幸・加藤祥・鈴木泰・池上尚（2018）「『日本語歴史コーパス』4 作品に対する分類語彙表番号付与とその分析」『日本語学会 2018 年度秋季大会予稿集』167–172.

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』306–309.

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』15(2): 134–141.

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』東京：秀英出版.

国立国語研究所（編）（2004）『分類語彙表増補改訂版』東京：大日本図書.

近藤明日子・田中牧郎（2017a）「『UniDic』と『分類語彙表』の見出し対応表データの構築」『言語資源活用ワークショップ 2016 発表論文集』79–86.

近藤明日子・田中牧郎（2017b）「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

[9] 加藤・浅原・山崎（2017）に BCCWJ の長単位への分類番号の付与の実践について報告がある。

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』90–93.
Kondo, Asuko, Makiro Tanaka and Masayuki Asahara (2018) Alignment table between UniDic and 'Word list by semantic principles'. *Proceedings of the 8th Conference of Japanese Association for Digital Humanities* 125–128.
宮島達夫（編）（2015）『万葉集巻別対照分類語彙表』東京：笠間書院.
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』東京：笠間書院.
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』（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）.
鈴木類・古宮嘉那子・浅原正幸・佐々木稔・新納浩幸（2019）「概念辞書の類義語と分散表現を利用した教師なし all-words WSD」『自然言語処理』26(2): 361–379.

関連 Web サイト

分類語彙表番号－ UniDic 語彙素番号対応表　https://pj.ninjal.ac.jp/corpus_center/goihyo.html
ChaMame　https://ja.osdn.net/projects/chaki/releases/p15635
『現代日本語書き言葉均衡コーパス』　https://pj.ninjal.ac.jp/corpus_center/bccwj/
MeCab　https://taku910.github.io/mecab/
『日本語話し言葉コーパス』　https://pj.ninjal.ac.jp/corpus_center/csj/
『日本語歴史コーパス』　https://pj.ninjal.ac.jp/corpus_center/chj/
UniDic　https://unidic.ninjal.ac.jp/

Construction of an Alignment Table between 'Word List by Semantic Principles' and UniDic

KONDO Asuko[a]　TANAKA Makiro[b]

[a]Adjunct Researcher, Center for Corpus Development, NINJAL
[b]Meiji University / Project Collaborator, NINJAL

Abstract
In this study, we have constructed an alignment table between 'Word List by Semantic Principles (revised and enlarged edition)' (hereafter WLSP) and UniDic to develop large-scale Japanese corpora which is comprehensively annotated with systematic word senses. WLSP is an extensive contemporary Japanese thesaurus with systematic semantic categories. UniDic is a vast lexicon used for Japanese morphological analysis and is utilized in the development of large-scale Japanese corpora. The alignment table defines n-to-n same word relations between 64,759 WLSP entries and 50,795 UniDic lexemes. These relations were manually verified based on scripts, readings, and classes of the WLSP entries and the UniDic lexemes. The development of word-sense annotated Japanese corpora has commenced with the use of the table. A Japanese morphological analysis tool to annotate word-sense was also developed with the table. Meanwhile, for a full-scale development of word-sense annotated Japanese corpora, it is necessary that problems, such as enlargement of the table and identification of word senses in corpora are effectively dealt with.

Key words: 'Word List by Semantic Principles', UniDic, alignment table, large-scale Japanese corpus, word-sense annotation