『複数短単位対応版「分類語彙表番号-UniDic」対応表』の構築と活用

国立国語研究所論集 (NINJAL Research Papers) 28: 179–195 (2025)
ISSN 2186-1358

『複数短単位対応版「分類語彙表番号 -UniDic」対応表』の構築と活用

菊池そのみ[a]　片山久留美[b]　髙橋雄太[c]　小木曽智信[d]

[a] 筑波大学／国立国語研究所 共同研究員
[b] 国立国語研究所 共同研究員
[c] 国立国語研究所 言語資源開発センター
[d] 国立国語研究所 研究系

要旨

本稿は『分類語彙表増補改訂版データベース』における見出し（分類語彙表番号）と形態素解析用辞書 UniDic の見出し（語彙素）とを対応づける試みの一つとして『複数短単位対応版「分類語彙表番号 -UniDic」対応表』（Ver.1.0）の構築とこれを活用した分析の事例とについて報告するものである。まず，『分類語彙表増補改訂版データベース』（Ver.1.0.1）における見出しのうち，複数の短単位からなるものを対象に分類語彙表番号と語彙素とを対応づけ，既に公開されている『分類語彙表番号 -UniDic 語彙素番号対応表』と合わせて『複数短単位対応版「分類語彙表番号 -UniDic」対応表』（Ver.1.0）として公開した。次にこの『複数短単位対応版「分類語彙表番号 -UniDic」対応表』を用いて意味範疇・見出しの長さ・品詞・語種の観点から『分類語彙表増補改訂版データベース』の見出しを計量的に分析した。特に長い見出しにおける語種・品詞の構成についての詳細を報告すると共に，見出しの長さと見出しの数との関係について意味範疇ごとに考察した *。

キーワード：分類語彙表，UniDic，複数短単位，対応表，語義情報付与

1. はじめに

日本語研究においては国立国語研究所『日本語話し言葉コーパス』や『現代日本語書き言葉均衡コーパス』に端を発し，断続的にコーパスの構築・拡充が進められており，利用できるコーパスの量・種類の増加が著しい状況にある。これと並行して近年，言語研究・自然言語処理の分野においては近藤・田中（2020）の述べるように日本語の大規模なコーパスに語義情報を付与する取り組みも進められている。実際に加藤・浅原・山崎（2019）や浅原・池上・鈴木・市村・近藤・加藤・山崎（2023）によって国立国語研究所『現代日本語書き言葉均衡コーパス』や同『日本語歴史コーパス』に語義情報を付与する試みが進められている。

このように大規模なコーパスに語義情報を付与するためには，語彙の一覧（シソーラス）とコーパス構築に用いられている「形態論情報」（見出し語・品詞・語種等の情報）とを対応づける必要があり，その成果の一つが近藤・田中（2020）によって構築された国立国語研究所『分類語彙表番号 -UniDic 語彙素番号対応表』である。これは国立国語研究所（編）（2004）『国立国語研究

* 本稿は国立国語研究所の共同研究プロジェクト「多様な語彙資源を統合した研究活用基盤の共創」（プロジェクトリーダー：小木曽智信）における研究成果の一部である。また，本稿は言語処理学会第 30 回年次大会（2024 年 3 月 13 日，神戸国際会議場，片山・髙橋・菊池・小木曽 2024）と日本語学会 2024 年年度春季大会（2024 年 6 月 2 日，東京外国語大学，菊池・片山・髙橋・小木曽 2024）とにおいて発表した内容に基づくものである。

所資料集 14 分類語彙表―増補改訂版』のデータベースである『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 と呼ぶ）における見出しの「分類語彙表番号」と形態素解析用辞書 UniDic の短単位とを対応づけて整理したものである。この『分類語彙表番号-UniDic 語彙素番号対応表』（以下，1 短単位版対応表と呼ぶ）は分類語彙表 DB における見出しのうち，UniDic の 1 短単位と一致するもの（例：「猫」「鳴く」）のみを対象とした対応表であり，UniDic の 1 短単位と一致せず，複数の短単位からなる見出し（例：「選挙管理委員会」「頭の中が真っ白になる」）については扱われていない。近藤・田中（2020）において扱われているのは分類語彙表 DB 全 98,241 見出しのうち [1]，64,759 見出しであり [2]，残りの 33,482（約 34%）の見出しは対応づけがなされていないという現状にある。

そこで本稿は分類語彙表 DB において複数の短単位からなる見出しを対象として UniDic との対応表（以下，複数単位版対応表と呼ぶ）を構築し，1 短単位版対応表と併せて国立国語研究所『複数短単位対応版「分類語彙表番号 -UniDic」対応表』（Ver.1.0）として整備し，公開する。更に複数単位版対応表と 1 短単位版対応表とを活用して分類語彙表 DB の見出しについて全体を概観し，体・用・相の各類における品詞・語種の構成や見出しの長さと「意味範疇」の分類との関係について分析する。

2. 対応表の概要

ここでは分類語彙表 DB における見出し語の「分類語彙表番号」と形態素解析用辞書 UniDic の短単位とを対応づけ，表として整備する方法について概要を述べる。

まず，分類語彙表 DB においては各見出しに「1.5501」のような「分類番号」が付与されており，これは整数位に「類」が，小数点以下に「部門」「中項目」「分類項目」が順に示されることによって階層構造をなしている（国立国語研究所（編）2004）。分類番号の構造の例を表 1 に示す。

表 1　分類語彙表 DB における分類番号の構造 [3]

見出し	分類番号	類	部門	中項目	分類項目
猫	1.5501	体（1）	自然物および自然現象（.5）	動物（.55）	哺乳類（.5501）
鳴く	2.3031	用（2）	精神および行為（.3）	心（.30）	声（.3031）
めでたい	3.3310	相（3）	精神および行為（.3）	生活（.33）	人生・禍福（.3310）
選挙管理委員会	1.2730	体（1）	人間活動の主体（.2）	機関（.27）	議会（.2730）
頭の中が真っ白になる	2.3001	用（2）	精神および行為（.3）	心（.30）	感覚（.3001）
やむにやまれず	3.1230	相（3）	抽象的関係（.1）	存在（.12）	必然性（.1230）

[1] 近藤・田中（2020: 78，注 1）によれば，「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」が 98,241 であるとのことである。
[2] 後述するように分類語彙表 DB の複数の見出しが UniDic の一つの語彙素と対応する場合（多対一の対応関係）と分類語彙表 DB の 1 見出しが UniDic の複数の語彙素と対応する場合（一対多の対応関係）とがあることから，それらを合わせた結果，1 短単位版対応表には 65,043 の見出しが採録されている。
[3] 表 1 は加藤・浅原・山崎（2019: 135）における表 1 を参考に作成したものである。

また，近藤・田中（2020）によれば，分類語彙表 DB においては分類項目内の通し番号である「段落番号」「小段落番号」と「小段落」内の通し番号である「語番号」とが各見出しに付与されており,「分類番号」「段落番号」「小段落番号」「語番号」の数字を連結させた「分類語彙表番号」によって見出しが一意に指定されるとのことである。

次に形態素解析用辞書 UniDic は「短単位」という言語単位に基づいて語を認定するものである（小椋・小磯・冨士池・宮内・小西・原 2011）。この UniDic について小椋・小磯・冨士池・宮内・小西・原（2011: 10）によれば，「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という方針を取り，語を階層化した形で登録している」とのことであり，具体的には最上層として国語辞典の見出しに相当する「語彙素」, その下の階層に語形を区別する「語形」，さらにその下の階層に表記を区別する「書字形」が設けられている。例えば，語彙素「攪拌（カクハン）」には「カクハン」と「コウハン」との 2 つの語形があり，更に各語形について実際の表記を示す「書字形」と発音を示す「発音形」とがある。このように語形や書字形の別を含めて同一の語として整理したのが語彙素であり，各語彙素には「語彙素 ID」という固有の番号が付与されている [4]。

この分類語彙表 DB（分類語彙表番号）と形態素解析用辞書 UniDic の見出し（語彙素）とを対応づける試みとして近藤・田中（2020）によって 1 短単位版対応表が構築された。これは分類語彙表 DB の見出しと UniDic の見出しとについて同語関係にあると考えられるものを対象として「分類語彙表番号」と「語彙素 ID」とを対応づけたものである。その構築に際しては分類語彙表 DB の全 98,241 見出しのうち，1 短単位からなると見られる見出しを人手で抽出している。これは例えば，前掲表 1 における「猫」（1.5501）「鳴く」（2.3031）「めでたい」（3.3310）のように分類語彙表 DB の見出しが UniDic の短単位の切れ目と一致するものである。なお，1 短単位版対応表においては短単位の認定規則によって 1 短単位と認定されない見出し [5]（計 33,477）と短単位未満の単位からなる見出し（計 5）とが対象から除外されている。前者は前掲表 1 における「選挙管理委員会」（1.2730）「頭の中が真っ白になる」（2.3001）「やむにやまれず」（3.1230）等であり，後者は「－す」（2.1110）「－ぐむ」（2.1302）「－らむ」（2.1302）「－まる」（2.1500）「－す」（2.3670）の 5 見出しである。

3. 複数短単位版対応表の構築

前述した対応表の概要を踏まえ，ここでは分類語彙表 DB における見出しのうち複数短単位からなるものを対象に分類語彙表番号と語彙素 ID とを紐付けた対応表を構築する手順について見出しの形態素解析，活用形・異語形の処理，データの公開形式の順に述べる。

[4] UniDic については伝・小木曽・小椋・山田・峯松・内元・小磯（2007）に詳しい。
[5] 後述するようにこの 33,477 の見出しには，1 短単位版対応表の構築において複数短単位からなる見出しであると判定されたものの，本稿において検討した結果，1 短単位からなる見出しであると判定された 104 の見出しが含まれている。

3.1 見出しの形態素解析

まず，1 短単位対応表において対象にされていない 33,482 見出しについて XML 化し，タグの属性値に分類語彙表番号を付与した上で「現代書き言葉 UniDic」を用いて形態素解析を実施した。次に形態素解析によって認定された形態論情報について国立国語研究所「形態論情報データベース」（小木曽・中村 2014）を用いて人手で誤りを修正した。

3.2 活用形・異語形の処理

続いて形態論情報の修正に当たって問題となる活用形・異語形の処理について述べる。複数短単位からなる見出しにおいては，主格の標示形式によって見出しの末に位置する活用語に終止形以外の活用形が想定される場合や，UniDic における同語彙素・異語形の関係にある語が別の見出しとして立項される場合がある。

前者について例えば，「威勢がよい」という見出しにおける形容詞「良い」は終止用法が想定されるが[6]，「頭のよい」という見出しにおける形容詞「良い」は連体用法のみが想定され，終止形とは認定し難い。UniDic によって付与された形態論情報を（1）に示す。複数短単位からなる見出しについてはこのような主格の標示形式による述語の活用形の差異を考慮する必要があるという点で 1 短単位からなる見出しと異なっている。

（1）　a.　威勢　　語彙素「威勢」名詞 - 普通名詞 - 一般
　　が　　語彙素「が」助詞 - 格助詞
　　よい　　語彙素「良い」形容詞 - 非自立可能　終止形 - 一般
　　b.　頭　　語彙素「頭」名詞 - 普通名詞 - 一般
　　の　　語彙素「の」助詞 - 格助詞
　　よい　　語彙素「良い」形容詞 - 非自立可能　連体形 - 一般

後者については例えば，「攪拌（カクハン）する」「攪拌（コウハン）する」における「攪拌」のように UniDic においては同一の語彙素「攪拌」の異なる語形であると認定されるものが分類語彙表 DB においてはそれぞれ別の見出しとして掲出されている場合がある。UniDic によって付与された形態論情報を（2）に示す。

（2）　a.　攪拌　　語彙素「攪拌（カクハン）」 語形「カクハン」 発音形「カクハン」
　　する　　語彙素「為る」
　　b.　攪拌　　語彙素「攪拌（カクハン）」 語形「コウハン」 発音形「コーハン」
　　する　　語彙素「為る」

本稿においてはこのような活用形や異語形についても形態論情報として反映させ，後述する語彙表 ID（lid）を用いた詳細な分析が可能となるようなデータを整備した。

[6] 当然，「威勢がよい人」のような連体用法も想定し得るが，「威勢がよい」という見出しにおける「よい」を敢えて連体形であると捉える必要は乏しいものと言える。

3.3 データの公開形式

続いてデータの公開形式について取り上げる。本稿においては新たに構築した複数短単位版対応表と既に公開されている 1 短単位版対応表とを合わせて『複数短単位対応版「分類語彙表番号-UniDic」対応表』（Ver.1.0）として整備し，目的に応じた利用を可能にするために 3 つの公開形式を用意した。このデータは 2024 年 3 月に国立国語研究所学術情報リポジトリにおいて公開済みであり，無償でダウンロードすることが可能である。

まず，（3）のように 1 短単位版対応表と同様に分類語彙表番号と UniDic の語彙素 ID とを併記したテキストデータの形式を用意した。これは語彙素 ID を UniDic の辞書アーカイブ（「現代語用 UniDic」のフルパッケージ）の語彙素一覧（lex.csv）に含まれる語彙素 ID 列に対応づけることで，語彙素・語彙素読みや語種，品詞の情報を取り出すことが可能なものである。（3）に示す情報のうち，短単位数とは分類語彙表 DB における当該の見出しがいくつの短単位からなるかを示す数値であり，（3）の「威勢がよい」の例では 3 つの短単位からなることが判る。1 短単位版対応表と同様に語彙素に基づいて語を同定することを目的とする場合には，このデータが有用である。

（3） 分類語彙表見出し，分類語彙表番号，短単位数，語彙素 ID-1，語彙素 ID-2，語彙素 ID-3
例：威勢がよい，3.3430-02-02-02，3，1949，7889，38988

次に分類語彙表番号と UniDic の語彙表 ID とを併記したテキストデータの対応表を用意した。前述の通り，複数短単位版対応表においては活用形や語形を区別して形態論情報を付与している。しかし，語彙素 ID はあくまで語彙素を一意に識別するものであり，活用形や語形の情報は得られない。そこで，分類語彙表番号と UniDic の語彙表 ID（lid）とを対応させ，併記したテキストデータも公開することとした。この語彙表 ID（lid）は UniDic 用語集によれば[7]，「UniDic DB 中の各エントリ（短単位）を一意識別するための ID（主キー）」であり，語形・書字形・発音形・活用形等の違いも区別することができるという点で語彙素 ID と異なっている。例えば，「威勢がよい」「頭のよい」について語彙素 ID を用いた対応表においては「よい」にいずれも「38988」という同じ ID が付与されるが，語彙表 ID（lid）を用いた対応表においては活用形を表す末尾部分の数字が異なった ID がそれぞれ付与されている。このように語彙表 ID（lid）を用いた対応表を活用することによって語の詳細な情報を得ることが可能となる。

これらのテキストデータに加え，XML 形式のデータも併せて用意した。この XML 形式のデータにおいては分類語彙表 DB の各見出しに基づき，それを構成する短単位を子要素とし，短単位の形態論情報を属性値として記述してある。具体的には各見出しを構成する短単位について語彙素 ID，語彙素，語形，品詞，発音形，語種，語彙表 ID を他のデータと関連づけることなく参照することが可能である。

[7] 国立国語研究所「UniDic」の Web ページにおける「用語集」（https://clrd.ninjal.ac.jp/unidic/glossary.html）による。なお，語彙素 ID・語彙表 ID の生成については小木曽・中村（2011，2014）に詳しい。

4. 複数短単位版対応表の概要

ここでは構築した複数短単位版対応表について語彙統計，分類語彙表 DB における見出しと UniDic の語彙素との対応関係，複数短単位版対応表の利点の 3 点を報告することを通してその概要を説明する。

4.1 複数短単位版対応表の語彙統計

前述の手順によって構築した複数短単位版対応表における 33,482 の見出しについて構成する短単位数別の見出し数を表 2 に示す。なお，1 短単位からなる見出しについては「鬼がわら」「貫き通す」等の近藤･田中（2020）において複数短単位と認定されて対象から除外されたもの（104 見出し）と「－ぐむ」のように短単位未満のもの（5 見出し）とを合算してある。

表 2 から 2 短単位からなる見出しが 25,743 であって最も多く，全体の約 77% を占めること，短単位数は 9 短単位が最大であることが判る。この 9 短単位からなる見出しは「一つかまの飯を食った仲」(1.2200)「居ても立っても居られない」(3.3013)「目の中に入れても痛くない」(3.3020)「横の物を縦にもしない」(3.3040）の 4 見出しであった。また，2 短単位からなる 25,743 見出しのうち, 9,603 見出しは「勉強する」「アウトプットする」「いらいらする」等のように語彙素「為る」を後項に取る動詞であった。

表 2　構成短単位数別の見出し数

短単位数	見出し数	割合
1	109	0.33%
2	25743	76.89%
3	6162	18.40%
4	989	2.95%
5	364	1.09%
6	78	0.23%
7	27	0.08%
8	6	0.02%
9	4	0.01%
総計	33482	100.00%

更に複数短単位版対応表における 33,482 の見出しについて「類」別の見出し数を表 3 に示す[8]。表 3 から体の類が全体の約 46%，用の類が約 43%，相の類が約 10% をそれぞれ占めていることが読み取れる。

[8] 国立国語研究所（編）(2004: 4）によれば，「体の類は，名称を表す語で，名詞の類。用の類は，存在・活動を叙述する語で動詞の類。相の類は，状態を叙述する語で，形容詞・形容動詞・副詞・連体詞の類。その他の類は，一部の副詞，接続詞，感動詞である」とのことであり，「慣用句･連語については，個々の慣用句･連語の文法的な機能を考慮して，それぞれ適切な「類」に配置した」とのことである。

表 3　類別の見出し数

類	見出し数	割合
体（1.）	15310	45.73%
用（2.）	14280	42.65%
相（3.）	3466	10.35%
その他（4.）	426	1.27%
総計	33482	100.00%

4.2 分類語彙表 DB の見出しと UniDic の語彙素との対応関係

続いて分類語彙表 DB における見出しと UniDic の語彙素との対応関係について述べる。分類語彙表 DB においては別の見出しであり，異なる分類番号が与えられている場合でも UniDic においては同一の語彙素（語彙素 ID）の組み合わせからなると判断されるものがある。例えば,「平行する」という見出しは（4）のように分類語彙表 DB において 5 つの見出しとしてそれぞれに異なる分類番号が付与されているが，UniDic の語彙素としてはいずれも語彙素「平行」と語彙素「為る」との組み合わせとして処理されるのである。つまり，これは分類語彙表 DB における見出しが複数であり，UniDic の語彙素が単数であることから多対一の対応関係にあるということになる。複数短単位版対応表においては 3,727 パターンの語彙素 ID の組み合わせにおいてこのような多対一の対応関係が認められた。

（4）　a.　分類語彙表 DB：　平行する　2.1120　用 - 関係 - 類 - 相対
　　平行する　2.1525　用 - 関係 - 作用 - 連れ・導き・追い・逃げなど
　　平行する　2.1570　用 - 関係 - 作用 - 成形・変形
　　平行する　2.1573　用 - 関係 - 作用 - 配列・排列
　　平行する　2.1730　用 - 関係 - 空間 - 方向・方角
　　b.　UniDic：　「平行」（語彙素 ID: 33855）＋「為る」（語彙素 ID: 19537）

これに対して分類語彙表 DB における見出しが単数であり，それに対応する UniDic の語彙素が複数であるという一対多の対応関係にある場合は認められなかった。近藤・田中（2020: 88）においては 1 短単位版対応表の構築に当たって一対多の対応関係にある場合が報告されており，その多くが「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」であると説明している。その例として，分類語彙表 DB における見出し「骨（こつ，体，1.5606，自然 - 身体 - 骨・歯・爪・角・甲）」が UniDic の語彙素では「骨（コツ，体）」「骨（コツ，接尾体）」「骨（コツ，接頭）」に対応するものであることが挙げられている。これに加えて「類認定の違いに由来するもの」や「表記法・同語判別基準に由来するもの」も見られること（近藤・田中 2020: 88）が報告されている。このことを踏まえると，複数短単位版対応表において一対多の対応関係にある場合が認められなかったのは，複数の用法や類が想定される語（1 短単位）であっても複数短単位からなる見出しの構成要素である場合には前後の語との関係によってそのうちの一つが指定されている（つまり，複数の用法や類にまたがる曖昧性は回避されている）ためであ

ると言える[9]。このように複数短単位版対応表においては一対多の対応関係に該当する見出しが見られないという点で 1 短単位版対応表と異なっていると言える[10]。

4.3 複数短単位版対応表の利点

本稿において構築した複数短単位版対応表を活用すると UniDic によって解析された文章について分類語彙表 DB における意味範疇の分類（語義情報）を使用した分析が可能になるという点が大きな利点である。例えば，「頭のよい大学生が一堂に会して議論する」（作例）という文について短単位，語彙素 ID，1 短単位版対応表における分類番号，複数短単位版対応表における分類番号を図 1 に示す。この文は 12 の短単位からなる文であり，1 短単位版対応表によってそのうちの「頭」「よい」「大学」「会する」「議論」「する」の 6 つの短単位と分類語彙表 DB との対応づけが可能である。一方で「一堂」については対応しておらず，分類番号を利用することができない。この点について複数短単位版対応表においては「一堂に会する」という見出しが採録されているため，この分類番号「2.3510」を活用することが可能になるのである。

短単位	頭	の	よい	大学	生	が	一堂	に	会し	て	議論	する
語彙素 ID	741	28989	38988	22965	19690	7889	2096	28178	5775	24874	9967	19537
1 短単位版対応表	1.3421		3.1332	1.2630					2.3510		1.3133	2.3320
複数短単位版対応表	3.3421			1.2419			2.3510				2.3133	

図 1　短単位と分類番号との対応関係の例

また,「大学 | 生」の「大学」は 1 短単位版対応表においては「1.2630」（体 - 主体 - 社会 - 社寺・学校）と認定されるが，複数短単位版対応表においては「1.2419」（体 - 主体 - 成員 - 学徒）と認定されることから，複数短単位版対応表を用いることで文脈に照らして適切な語義分類を付与することが可能になると言える。

これらに加えて「議論する」のように後項に「する」を取る動詞について見ると，1 短単位版対応表を用いる場合には「議論」と「する」との 2 語として扱わざるを得ないことに加え，動詞「する」に 8 つの見出しが立てられている（分類番号が 8 つある多義語である）ことから，前接語や文脈を踏まえて個別に語義を検討する必要があるという課題があった。しかし，複数短単位版対応表においては「議論する」が 1 見出しとして立てられ，一つの分類番号が付与されていることから，「する」の語義情報を個別に検討する必要がないという点も有用であると言える。

[9] 例えば，複数短単位版対応表に見られる「頭蓋 | 骨」の「骨」は UniDic の認定において接尾辞であることが一意に定まっているということである。

[10] なお，近藤・田中（2020: 87）は 1 短単位版対応表において多対一の対応関係の方が一対多の対応関係よりも多く見られることを指摘し，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」であることに起因すると説明している。

5. 複数短単位版対応表を用いた分類語彙表 DB の分析

本稿において構築した複数短単位版対応表を活用し，分類語彙表 DB の見出しに関する新たな分析を試みる。ここでは 1 短単位版対応表と複数短単位版対応表とを合わせて分析する。『分類語彙表―増補改訂版』における見出しについては山崎（2004），柏野（2006），林（2009）等によって報告されているが，特に見出しの長さや見出しを構成する語の観点からの分析は進められていない。これは UniDic と対応づけることによって語を統一的に認定することが可能になったことで新たに可能になったものであり，本稿ではこの点に関する調査の事例として報告する。

なお，近藤・田中（2020）において複数短単位になるとして対象外とされた計 33,477 の見出しのうち，本稿における再検討によって 104 の見出しが 1 短単位と認定された（類の内訳は体 54，用 14，相 30，その他 6）。この 104 の見出しと短単位未満の単位からなる 5 の見出し（類の内訳は用 5）とを合わせた計 109 の見出しについては，以下の分析において 1 短単位として扱うこととする。

5.1 全体の概観

まず，分類語彙表 DB における見出し全体を概観する。分類語彙表 DB と UniDic との対応づけがなされている全 98,525 の見出しの内訳を表 4 に示す。なお，既に述べたように分類語彙表 DB の見出しと UniDic の語彙素とは多対一／一対多の対応関係にある場合があることから，分類語彙表 DB の見出しの合計 98,241 よりも対応表における見出し数の方が多くなっている。本稿における分析は対応表の見出し全てを対象としたものである。

表 4 から全体としては 1 短単位からなる見出しが 65,152（約 66%），2 短単位からなる見出しが 25,744（約 26%）であり，短単位数が増えるに連れてそれに該当する見出し数が減ることが看て取れる [11]。また，体・用・相・その他 [12] の各類について見ると，体の類が 65,893（約 67%），用の類が 22,619（約 23%），相の類が 9,126（約 9%），その他が 887（約 1%）である。更に体の類を見ると，1 短単位からなる見出し数が約 77%，2 短単位が約 20% を占め，3 〜 9 短単位が僅少であると言える。このように 1 短単位からなる見出しが最も多く，短単位数が増えていくに連れて見出し数が減る傾向は相の類・その他においても同様である。これに対して用の類は 1 短単位からなる見出し数が約 37%，2 短単位が約 45%，3 短単位が約 16% を占め，2 短単位のものが 1 短単位のものよりも多く，体の類・相の類・その他に比べて複数短単位で構成される見出しが多いことが読み取れる。

[11] これは水谷（1983: 87）が「語の長さの分布」として掲げるような語の長さと語彙頻度との関係に関する傾向と一致するものと言える。
[12] 分類語彙表 DB における「その他」に該当する見出しには〈接続〉「しかしながら」「にもかかわらず」，〈挨拶〉「それではまた」「お疲れ様」，〈動物の鳴き声〉「ぴーちくぱーちく」等がある。

表 4　分類語彙表 DB における見出し全体の内訳

	体の類		用の類		相の類		その他		計	
	見出し数	割合	見出し数	割合	見出し数	割合	見出し数	割合	見出し数	割合
1 短単位	50637	76.85%	8358	36.95%	5690	62.35%	467	52.65%	65152	66.13%
2 短単位	13353	20.26%	10240	45.27%	1944	21.30%	207	23.34%	25744	26.13%
3 短単位	1650	2.50%	3557	15.73%	841	9.22%	114	12.85%	6162	6.25%
4 短単位	196	0.30%	280	1.24%	448	4.91%	64	7.22%	988	1.00%
5 短単位	47	0.07%	156	0.69%	137	1.50%	24	2.71%	364	0.37%
6 短単位	5	0.01%	15	0.07%	48	0.53%	10	1.13%	78	0.08%
7 短単位	3	0.00%	13	0.06%	10	0.11%	1	0.11%	27	0.03%
8 短単位	1	0.00%	0	0.00%	5	0.05%	0	0.00%	6	0.01%
9 短単位	1	0.00%	0	0.00%	3	0.03%	0	0.00%	4	0.00%
計	65893	100.00%	22619	100.00%	9126	100.00%	887	100.00%	98525	100.00%

次に体・用・相の各類の見出しを構成する短単位の品詞（大分類）を 1 短単位からなる見出しと複数短単位からなる見出しとに分けて表 5 に示す。表 5 においては，複数短単位からなる見出しについてその見出しを構成する短単位ごとに品詞を計上してある。例えば，「説明 | する」は名詞 1，動詞 1 と数え，「耳 | が | 遠い」は名詞 1，助詞 1，形容詞 1 と数える。

表 5　体・用・相の類の見出しを構成する品詞（網掛けは 10% 以上の品詞）

	体の類				用の類				相の類			
	1 短単位		複数短単位		1 短単位		複数短単位		1 短単位		複数短単位	
名詞	49636	98.02%	24961	76.23%	0	0.00%	13757	41.40%	1683	29.58%	2824	30.32%
代名詞	136	0.27%	79	0.24%	0	0.00%	7	0.02%	1	0.02%	139	1.49%
形状詞	230	0.45%	225	0.69%	0	0.00%	58	0.17%	1302	22.88%	199	2.14%
連体詞	0	0.00%	77	0.24%	0	0.00%	8	0.02%	42	0.74%	23	0.25%
副詞	45	0.09%	48	0.15%	0	0.00%	333	1.00%	1655	29.09%	366	3.93%
接続詞	0	0.00%	1	0.00%	0	0.00%	0	0.00%	1	0.02%	1	0.01%
感動詞	1	0.00%	3	0.01%	0	0.00%	2	0.01%	5	0.09%	4	0.04%
動詞	1	0.00%	481	1.47%	8332	99.69%	14630	44.02%	8	0.14%	1232	13.23%
形容詞	0	0.00%	196	0.60%	1	0.01%	100	0.30%	873	15.34%	726	7.80%
助動詞	0	0.00%	89	0.27%	13	0.16%	267	0.80%	2	0.04%	953	10.23%
助詞	3	0.01%	648	1.98%	0	0.00%	3757	11.31%	15	0.26%	1985	21.31%
接頭辞	53	0.10%	1241	3.79%	0	0.00%	130	0.39%	72	1.27%	216	2.32%
接尾辞	515	1.02%	4616	14.10%	12	0.14%	179	0.54%	31	0.54%	644	6.92%
記号	17	0.03%	63	0.19%	0	0.00%	4	0.01%	0	0.00%	0	0.00%
補助記号	0	0.00%	15	0.05%	0	0.00%	0	0.00%	0	0.00%	1	0.01%
計	50637	100.00%	32743	100.00%	8358	100.00%	33232	100.00%	5690	100.00%	9313	100.00%

表 5 から体の類において 1 短単位からなる見出しは名詞が約 98% を占めており，複数短単位の場合は名詞の次に接尾辞の割合が高いことから，名詞同士や名詞と接尾辞との結合によって構成される見出しが多いことが示唆される。これに対し，用の類において 1 短単位からなる見出しはほぼ動詞であり，複数短単位の場合は動詞が約 44%，名詞が約 41% を占め，続いて助詞が約

11% を占めている。相の類において 1 短単位からなる見出しは名詞・副詞が約 30%，形状詞が約 23%，形容詞が約 15% を占めており，複数短単位の場合は名詞が約 30%，助詞が約 21%，動詞・助動詞が約 10% を占めている。各類における複数短単位の語構成の詳細は次節以降に述べる。

更に体・用・相の各類の見出しを構成する短単位の語種を 1 短単位からなる見出しと複数短単位からなる見出しとに分けて表 6 に示す。表 6 は複数短単位からなる見出しについてその見出しを構成する短単位ごとに語種を計上してある。例えば，「カバー | する」は外来語 1，和語 1 と数え，「ローマ | 字 | 書き」は固有名 1，漢語 1，和語 1 と数える。なお，UniDic における語種の認定の仕方については小椋・小磯・冨士池・宮内・小西・原（2011）に詳しい。

表 6　体・用・相の類の見出しを構成する語種

	体の類				用の類				相の類			
	1 短単位	複数短単位	計	割合	1 短単位	複数短単位	計	割合	1 短単位	複数短単位	計	割合
和語	14112	7192	21304	25.55%	7852	22106	29958	72.03%	3174	6895	10069	67.11%
漢語	28764	21039	49803	59.73%	0	10314	10314	24.80%	2043	2266	4309	28.72%
外来語	4930	3257	8187	9.82%	0	472	472	1.13%	261	54	315	2.10%
混種語	2140	255	2395	2.87%	506	331	837	2.01%	212	93	305	2.03%
固有名	573	906	1479	1.77%	0	3	3	0.01%	0	4	4	0.03%
記号	117	94	211	0.25%	0	6	6	0.01%	0	1	1	0.01%
不明	1	0	1	0.00%	0	0	0	0.00%	0	0	0	0.00%
計	50637	32743	83380	100.00%	8358	33232	41590	100.00%	5690	9313	15003	100.00%

表 6 から体の類は和語が約 26%，漢語が約 60% であるのに対し，用の類は和語が約 72%，漢語が約 25% であり，相の類は和語が約 67%，漢語が約 29% であり，類によって語種の構成も異なることが判る。また，特に用の類は 1 短単位の見出しが和語･混種語のみであるという点が体･相の類と異なっていると言える。

5.2 体の類

ここでは体の類について見る。まず，表 5 を踏まえて複数短単位からなる体の類の見出しの語構成について確認する。2 短単位においては「前代 | 未聞」のような普通名詞 + 普通名詞が最も多く，3 短単位においては「世 | の | 常」のような普通名詞 + 格助詞 + 普通名詞が最も多い。4 短単位においては「一 | つ | 二 | つ」や「七 | 転び | 八 | 起き」のような数詞と接尾辞や普通名詞とを組み合わせた見出しが多くを占める。5 短単位以上は「風 | の | 前 | の | ちり」「蚊 | の | 鳴く | よう | な | 声」「はし | が | 転ん | で | も | おかしい | 年ごろ」「飛ん | で | 火 | に | 入る | 夏 | の | 虫」「一 | つ | かま | の | 飯 | を | 食っ | た | 仲」のように名詞・動詞・形容詞・助詞・助動詞等によって末尾の名詞を修飾する型の見出しが見られる。

次に体の類における見出しの長さ（見出しを構成する短単位数）と「意味的範疇」の分類との関係について検討する。見出しを構成する短単位数について中項目ごとの分布（総計における上位 6 項目）を表 7 に示す。表 7 から体の類においては〈心（.30)〉が最も多く，1 短単位のみならず，2 ～ 7 短単位においても割合が最も高いことが読み取れる。それに対して〈作用（.15)〉

は総計としては 3 位であり，1 短単位の割合は高いものの，2 短単位を見ると総計 4 ～ 6 位の〈生活（.33)〉〈量（.19)〉〈時間（.16)〉よりも割合が低いことが判る。つまり，〈心（.30)〉のように見出しの長さ（短単位数）に関係なく見出し数が多い場合と〈作用（.15)〉のように特定の長さの見出しに偏って分布している場合とがあると言える。

表 7　体の類における中項目ごとの短単位数の分布（総計上位 6 項目）

	心（.30）		言語（.31）		作用（.15）		生活（.33）		量（.19）		時間（.16）			短単位数総計	
1 短単位	4684	9.25%	3225	6.37%	3287	6.49%	2679	5.29%	2108	4.16%	1818	3.59%	…	50637	100.00%
2 短単位	1229	9.20%	985	7.38%	284	2.13%	687	5.14%	503	3.77%	479	3.59%	…	13353	100.00%
3 短単位	247	14.97%	100	6.06%	27	1.64%	75	4.55%	52	3.15%	109	6.61%	…	1650	100.00%
4 短単位	25	12.76%	17	8.67%	2	1.02%	10	5.10%	19	9.69%	15	7.65%	…	196	100.00%
5 短単位	11	23.40%	3	6.38%	2	4.26%			4	8.51%	1	2.13%	…	47	100.00%
6 短単位	3	60.00%											…	5	100.00%
7 短単位	2	66.67%									1	33.33%	…	3	100.00%
8 短単位													…	1	100.00%
9 短単位													…	1	100.00%
総計	6201	9.41%	4330	6.57%	3602	5.47%	3451	5.24%	2686	4.08%	2423	3.68%	…	65893	100.00%

（網掛けは短単位数ごとの上位 6 項目に該当するもの）

また，「分類番号」の部門や中項目を用いて全体を概観すると，短単位数が多くなる（＝見出しが長くなる）に連れて見出し数が減る傾向にあるが，その下位分類である分類項目で見ると，この関係が逆転している場合がある。具体的には（5）の通りであり，これらの分類項目については全体の傾向と異なり，より長い見出しが属していることが窺える。

(5)　a. 2 短単位からなる見出し数＞1 短単位からなる見出し数：
〈学徒（.2419)〉「上級 | 生」「児童」
〈機関（.2700)〉「事務 | 局」「機構」
〈政府機関（.2710)〉「外務 | 省」「役場」
〈地帯（.5280)〉「南極 | 圏」「赤道」

b. 3 短単位からなる見出し数＞2 短単位からなる見出し数：
〈勢い（.1403)〉「破竹 | の | 勢い」（2 短単位からなる見出し 0）
〈節・節目（.1634)〉「ひな | の | 節句」「バレンタイン | デー」
〈国際機構（.2750)〉「世界 | 保健 | 機構」「国際 | 連合」

c. 4 短単位からなる見出し数＞3 短単位からなる見出し数：
〈数記号（.1960)〉「二 | 分 | の | 一」「お | ひと | 方」
〈議会（.2730)〉「安全 | 保障 | 理事 | 会」「教育 | 委員 | 会」

5.3 用の類

続いて用の類について見る。用の類は 2 短単位において「関連 | する」のような普通名詞（サ変可能）＋動詞（非自立可能）が全体の約 8 割を占めており，3 短単位においては「腹 | が | 立つ」

のような普通名詞＋格助詞＋動詞（一般）が全体の半数を占めている。また，4 短単位においては「胸 | を | 躍ら | せる」のような普通名詞＋格助詞＋動詞（一般）＋助動詞，5 短単位においては「目 | から | 火 | が | 出る」のような普通名詞＋格助詞＋普通名詞＋格助詞＋動詞（一般）が最も多い。なお，分類語彙表 DB には「…て | いる」「…て | しまう」「－化 | する」「せ | られる」等の見出しがあり，接続助詞・接尾辞・助動詞といった付属語が見出しの先頭に現れる場合もある。

続いて体の類と同様に見出しの長さと「意味的範疇」の分類との関係について検討する。ここでは見出しを構成する短単位数と部門との関係を表 8 に示す[13]。表 4 において見たように用の類は全体として 2 短単位からなる見出しが最も多いが，表 8 を見ると〈抽象的関係（.1)〉と〈自然物および自然現象（.5)〉とにおいては 1 短単位からなる見出しの方が多いことが読み取れる。つまり，2 短単位からなる見出しが最も多いという用の類全体の様相は，3 つの部門のうち全体の約 6 割を占める〈精神および行為（.3)〉における様相を反映したものであり，部門ごとの内実は異なっていると言える。

表 8　用の類における見出しの長さと部門との関係

	抽象的関係（.1)		精神および行為（.3)		自然物および自然現象（.5)		計	
1 短単位	3642	43.58%	4022	48.12%	694	8.30%	8358	100.00%
2 短単位	3062	29.90%	6510	63.57%	668	6.52%	10240	100.00%
3 短単位	818	23.00%	2527	71.04%	212	5.96%	3557	100.00%
4 短単位	52	18.57%	216	77.14%	12	4.29%	280	100.00%
5 短単位	26	16.67%	118	75.64%	12	7.69%	156	100.00%
6 短単位	4	26.67%	10	66.67%	1	6.67%	15	100.00%
7 短単位	0	0.00%	12	92.31%	1	7.69%	13	100.00%
計	7604	33.62%	13415	59.31%	1600	7.07%	22619	100.00%

また，分類項目について体の類は 1 短単位からなる見出しが全ての項目に属しているのに対し，用の類は（6）に示す項目において複数短単位からなる見出しのみが属している（＝ 1 短単位からなる見出しがない）場合がある。これらの項目に該当する用の類の見出しは，1 短単位では作りにくいものである可能性も示唆される[14]。

(6)　〈こそあど（.1010)〉「こう | する」「どう | か | する」
〈真偽・是非（.1030)〉「是非 | を | 分かつ」
〈出入り（.1530)〉「出入り | する」
〈見聞き（.3090)〉「視聴 | する」
〈符合（.3114)〉「マーク | する」「句読点 | を | 打つ」

[13] 分類語彙表 DB において用の類・相の類には〈人間活動の主体（.2)〉〈生産物および用具（.4)〉がないことから表 8（と後掲する表 9）から除いてある。
[14] 当然，分類語彙表における語義分類の特徴や収録されている語の偏り等によって偶然に生じたものである可能性もある。

〈旅・行楽（.3371)〉「ハイキング | を | する」
〈身上（.3410)〉「お | 里 | が | 知れ | る」「すね | に | 傷 | を | 持つ」
〈税（.3720)〉「納税 | する」「税 | が | 掛かる」
〈エネルギー（.5001)〉「放電 | する」「ぴかぴか | する」

5.4 相の類

最後に相の類について見る。相の類は 2 短単位において「間違っ | た」のような動詞＋助動詞や「本格 | 的」のような普通名詞＋接尾辞（形状詞的）が多くを占めており，3 短単位においては「耳 | が | 遠い」のような普通名詞＋格助詞＋形容詞が多くを占めている。また，4 短単位においては「訳 | の | 分から | ない」のような普通名詞＋格助詞＋動詞＋助動詞，5 短単位においては「味 | も | 素っ気 | も | ない」のような普通名詞＋係助詞＋普通名詞＋係助詞＋形容詞が多くを占めている。

続いて見出しの長さ（見出しを構成する短単位数）と「意味的範疇」の分類との関係について検討する。用の類と同様に見出しを構成する短単位数と部門との関係を表 9 に示す。

表 9　相の類における見出しの長さと部門との関係

	抽象的関係（.1）		精神および行為（.3）		自然物および自然現象（.5）		計	
1 短単位	2503	43.99%	2295	40.33%	892	15.68%	5690	100.00%
2 短単位	888	45.68%	878	45.16%	178	9.16%	1944	100.00%
3 短単位	347	41.26%	427	50.77%	67	7.97%	841	100.00%
4 短単位	225	50.22%	204	45.54%	19	4.24%	448	100.00%
5 短単位	69	50.36%	65	47.45%	3	2.19%	137	100.00%
6 短単位	27	56.25%	20	41.67%	1	2.08%	48	100.00%
7 短単位	4	40.00%	6	60.00%	0	0.00%	10	100.00%
8 短単位	2	40.00%	3	60.00%	0	0.00%	5	100.00%
9 短単位	0	0.00%	3	100.00%	0	0.00%	3	100.00%
計	4065	44.54%	3901	42.75%	1160	12.71%	9126	100.00%

表 4 において見たように相の類は全体として 1 短単位からなる見出しが最も多いが，表 9 を見ると用の類に比して〈抽象的関係（.1)〉と〈精神および行為（.3)〉との間の数値の隔たりが小さいことが判る。この点について中項目を見ると，前者の中では特に「規則 | 的」「ちょっと | し | た」等の〈様相（.13)〉と「一 | 晩 | 中」「知ら | ぬ | 間 | に」等の〈時間（.16)〉とが多くを占め，後者の中では特に「親し | げ」「気 | が | 進ま | ない」等の〈心（.30)〉が多くを占めている。

また，分類項目で見ると〈信念・努力・忍耐（.3040)〉には 7 短単位以外の全ての長さの見出しがあり，〈良不良・適不適（.1332)〉等には 7 短単位・9 短単位以外の全ての長さの見出しがある。具体例は（7）の通りである。

(7)　〈3.3040 信念・努力・忍耐〉「手 | を | 替え | 品 | を | 替え」「雨 | が | 降ろう | が | 槍 | が | 降ろう | が」「横 | の | 物 | を | 縦 | に | も | し | ない」

〈3.1332 良不良・適不適〉「中途 | 半端」「まんざら | で | も | ない」「帯 | に | 短し | たすき | に | 長し」「毒 | に | も | 薬 | に | も | なら | ない」

5.5 分析のまとめ

分類語彙表 DB の見出し全体について 1 短単位版対応表と複数短単位版対応表とを活用して分析を試みた。このような分析によって分類語彙表 DB における意味範疇を UniDic によって認定される短単位の数（＝見出しの長さ）・品詞・語種の観点から検討することが可能になったと言える。例えば，全体について見ると分類語彙表 DB は 1 短単位からなる見出しが約 67% を占め，見出しが長くなるに連れて見出しの数は少なくなる傾向にあるが，用の類（特に「2.3 精神および行為」）においては 2 短単位からなる見出しが最も多いということが捉えられた。また，1 短単位からなる見出しにおける品詞と複数短単位からなる見出しにおける品詞とのバリエーションや多寡についても体・用・相の各類によって異なっており，これは日本語における語構成・造語法に関する議論と関わるものであると考えられる。

6. おわりに

本稿は『複数短単位対応版「分類語彙表番号 -UniDic」対応表』の構築と複数単位版対応表の概要とについて述べた上で，これを活用する事例として分類語彙表 DB における見出しの分析を実施した。本稿において示した『複数短単位対応版「分類語彙表番号 -UniDic」対応表』の構築や活用は短単位という言語単位に基づいて日本語の語彙を捉えるための方法論の開拓やその際に生ずる問題点・解決策を模索するものとして位置づけられる。実際に分類語彙表 DB における見出しの長さは全てが短単位・長単位と一致する訳ではなく，いわゆる国語辞典等の見出し（見出し語）と一致する訳でもないという点に留意する必要がある。これを踏まえると，分類語彙表 DB の各見出しを一つの言語単位として扱い，文章を解析して語義情報を参照するような研究に応用し得る可能性を含むものであると言える。

また，この対応表は水谷（1983）が「語彙に関する分布問題」として提起するような語の長さと出現頻度・語彙頻度との関係や多義語と使用率との関係を捉えようとする計量語彙論的な分析に資するデータセットであり，今後も更なる活用が期待される。そして，複数短単位版対応表の構築・公開によって日本語のコーパスに分類語彙表 DB の語義情報を付与するための準備が整ったものと考えられ，加藤・浅原・山崎（2019）に類するコーパスデータの分析や語義情報を活用した文章の分析も進展することが見込まれる。

参照文献

浅原正幸・池上尚・鈴木泰・市村太郎・近藤明日子・加藤祥・山崎誠（2023）「『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とその利用」『日本語の研究』19（3）：88–96. 日本語学会.

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』（国立国語研究所内部報告書 LR-CCG-10-06）東京：国立国語研究所.

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. 言語処理学会.

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）』（国立国語研究所内部報告書 LR-CCG-10-05-01，02）東京：国立国語研究所.
柏野和佳子（2006）「『分類語彙表』の特徴と位置付け」『日本語科学』19：143–160. 国立国語研究所.
片山久留美・髙橋雄太・菊池そのみ・小木曽智信（2024）「複数短単位版「分類語彙表 -UniDic」対応表の整備と公開」『言語処理学会第 30 回年次大会発表論文集』165–170. 言語処理学会.
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. 日本語学会.
菊池そのみ・片山久留美・髙橋雄太・小木曽智信（2024）「複数短単位版「分類語彙表番号 -UniDic」対応表を用いた『分類語彙表』の分析」『日本語学会 2024 年度春季大会予稿集』137–142. 日本語学会.
国立国語研究所（編）（2004）『分類語彙表―増補改訂版』（国立国語研究所資料集 14）東京：大日本図書.
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. 国立国語研究所.
伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵（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』22：101–123. 国立国語研究所.
水谷静夫（1983）「語彙に関する分布問題」『朝倉日本語新講座 2 語彙』86–120. 東京：朝倉書店.
山崎誠（2004）「『分類語彙表―増補改訂版―』の分類の特徴について」『日本語文學』20：73–86. 日本語文學會.
林彦伶（2009）「『分類語彙表』における各意味分野の語数」『語彙研究』7：40–47. 語彙研究会.

関連 Web サイト

国立国語研究所『分類語彙表増補改訂版データベース』（ver.1.0.1）https://github.com/masayu-a/WLSP（2024 年 5 月 14 日確認）
国立国語研究所『分類語彙表番号 -UniDic 語彙素番号対応表』https://github.com/masayu-a/wlsp2unidic（2024 年 5 月 14 日確認）
国立国語研究所『複数短単位対応版「分類語彙表番号 -UniDic」対応表』（Ver.1.0）https://doi.org/10.15084/0002000248（2024 年 5 月 14 日確認）
国立国語研究所「用語集」https://clrd.ninjal.ac.jp/unidic/glossary.html（2024 年 6 月 8 日確認）
国立国語研究所「現代書き言葉 UniDic」https://clrd.ninjal.ac.jp/unidic/download.html#unidic（2024 年 6 月 9 日確認）

Construction and Application of an Alignment Table between ‘Word List by Semantic Principles’ and Single and Multiple UniDic IDs

KIKUCHI Sonomi[a] KATAYAMA Kurumi[b] TAKAHASHI Yuta[c] OGISO Toshinobu[d]

[a]Tsukuba University / Project Collaborator, NINJAL
[b]Project Collaborator, NINJAL
[c]Center for Language Resource Development, NINJAL
[d]Research Department, NINJAL

Abstract

This study reports the construction of a correspondence table between the entries in ‘Word List by Semantic Principles (WLSP: revised and enlarged edition)’ and UniDic, and provides examples. First, the WLSP headings were aligned to multiple UniDic IDs, and a correspondence table, “Alignment Table between WLSP Number and Single and Multiple UniDic IDs,” was constructed and made publicly available. Next, the correspondence table was employed to analyze the headings of the WLSP quantitatively. Therefore, we obtained insights into the composition of word types and parts of speech with longer headings.

Keywords: ‘Word List by Semantic Principles’, UniDic, multiple UniDic IDs, alignment table, word-sense annotation