UniDic非コアデータ：解析用UniDicのID情報にひも付く追加情報の公開について

UniDic 非コアデータ
-解析用 UniDic の ID 情報にひも付く追加情報の公開について-

岡 照晃 (国立国語研究所コーパス開発センター) *

UniDic Non-core Data: Release of additional information corresponding to ID information of UniDic

Teruaki Oka (National Institute for Japanese Language and Linguistics)*

要旨

本発表では，形態素解析器『MeCab』用の電子化辞書である短単位自動解析用辞書『UniDic』（『解析用 UniDic』）のアペンドデータの公開について紹介を行う。『UniDic』は『MeCab』用の辞書の配布という外部公開形式をとっているが，v2.2 からその解析結果中に各短単位の ID 情報を出力するようになった。この情報を使えば，所外の研究者が自ら拡張した新たなカラムの情報を『UniDic』短単位にひも付く形で配布することができ，研究者間での共有も可能になる。本発表では，短単位の ID 情報について詳説し，それにひも付け，公開を行なっているアペンドデータ『UniDic 非コアデータ』を紹介する。

1. はじめに

国立国語研究所（以下，国語研）では，『現代日本語書き言葉均衡コーパス（BCCWJ）』をはじめ，さまざまなコーパス構築が行われている [Maekawa et al., 2014] [Maekawa et al., 2000] [近藤, 2012] [Asahara et al., 2014]。国語研で整備されているコーパスの多くは，形態論情報付きコーパスであり，**短単位** [近藤, 2015] と呼ばれる独自に設定した斉一な言語単位に分かち書きされ，各短単位に品詞や活用，発音やアクセントといった情報が付与されている。コーパスの構築時に重要な点の一つとして，アノテーションの斉一性の確保がある。分かち書きの粒度をそろえるだけでなく，各短単位に付与されている形態論情報もコーパス全体を通して一貫していない（**例** コーパス中の異なる位置に出現した同一の短単位に対して，活用など，一部の情報を異なって付与されている）と，ユーザが実際に検索用途でコーパスを利用する際など，検索漏れが生じ得る。そのため国語研では短単位の一覧をデータベース上で一元管理する仕組みを用いてアノテーションの統制を行なっている。このデータベースが『電子化辞書 UniDic（**UniDicDB**）』である [伝ら, 2007][岡, 2019]。UniDicDB は所内のコーパスデータベースと参照関係にあり，コーパスデータベース中の短単位は，UniDicDB に登録されており，UniDicDB 中の一意のエントリを参照する（リンク付けられている）状態になっている（図 1）。こうしたコーパスと辞書を統合したシステム運営の利点として，先に挙げた斉一性の確保だけでなく，

*teruaki-oka {at} ninjal.ac.jp

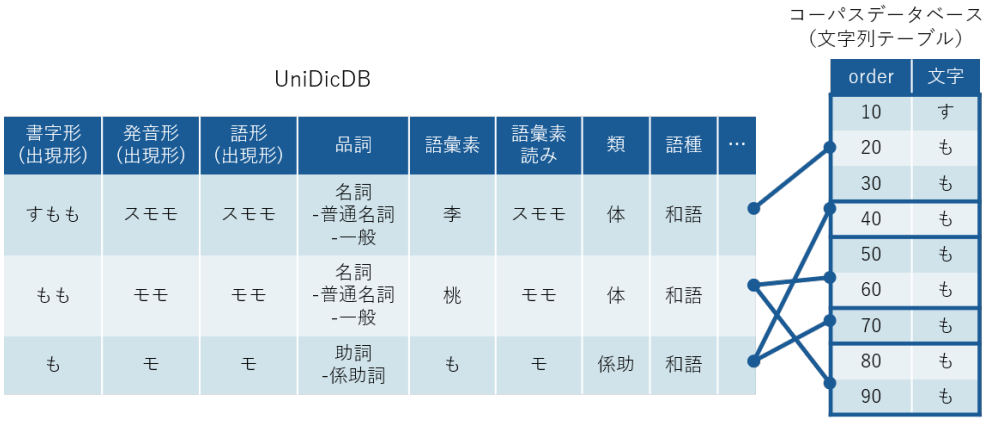

図 1　UniDicDB とコーパスデータベースのリンク関係。コーパス中の同一短単位は UniDicDB 中の一意のエントリを参照する状態となっている。

現時点の UniDicDB に存在しない情報（項目）が，新たに UniDicDB へ追加されると，その情報がデータベース間のリンクでコーパス全体へ瞬時に反映（新項目の追加）できる。

例えば [Kono et al., 2015][鴻野知暁ら, 2014] では，日本語歴史コーパス [近藤, 2012] 構築の際に各短単位の使用されていた（いる）期間の視点と終点を表す情報として「自至情報」を新たな項目として追加している。この情報を使えば，指定期間にのみ使用された語彙を UiDicDB から抽出できるだけでなく，各時代別にコーパス中の短単位はそれぞれどのくらいの期間使用され続けていたか？といった調査を行うこともできる。

[Kondo et al., 2018] は UniDicDB 中の短単位と分類語彙表（**WLSP**）[国立国語研究所, 2004] の番号との対応表（**wlsp2unidic**）を公開している(1)。UniDic の所外公開方式は，形態素解析器 MeCab[Kudo et al., 2004] 用の解析用辞書（解析用 UniDic（UniDic for Morphological Analyzer: **UniDicMA**））(2)であり，ライセンスはフリーライセンス（GPL/LGPL/BSD）となっている。これに対し，分類語彙表はクリエイティブコモンズライセンス（CC BY-NC-SA 3.0）で公開されている(3)。そのため wlsp2unidic は分類語彙表の番号付きの UniDicMA を公開するのではなく，各短単位の UniDicDB 上での識別 ID(4)と分類語彙表番号との対応表を別途クリエイティブコモンズライセンスで公開することを採用している。これはつまり，国語研所外の研究者であっても，[Kondo et al., 2018] と同様，UniDicMA を拡張するためのデータを自由に作り，各自の方法で公開できることを示唆している。

そこで本稿では，その先導的試みとして公開した UniDicMA の拡張データ『**UniDic 非コアデータ**』を紹介する（図 2）。「非コア」データというフレーズには、UniDic の短単位形態論情報が，専門家によって人手で厳密に付与されているのに対し，「専門家の手作業を介さず、非専門家の手によって作成されたものであり、公式のデータではない」という意図が込めてあ

(1) https://github.com/masayu-a/wlsp2unidic
(2) https://unidic.ninjal.ac.jp/
(3) https://github.com/masayu-a/wlsp
(4) ここでは語彙素 ID（lemma_id）。

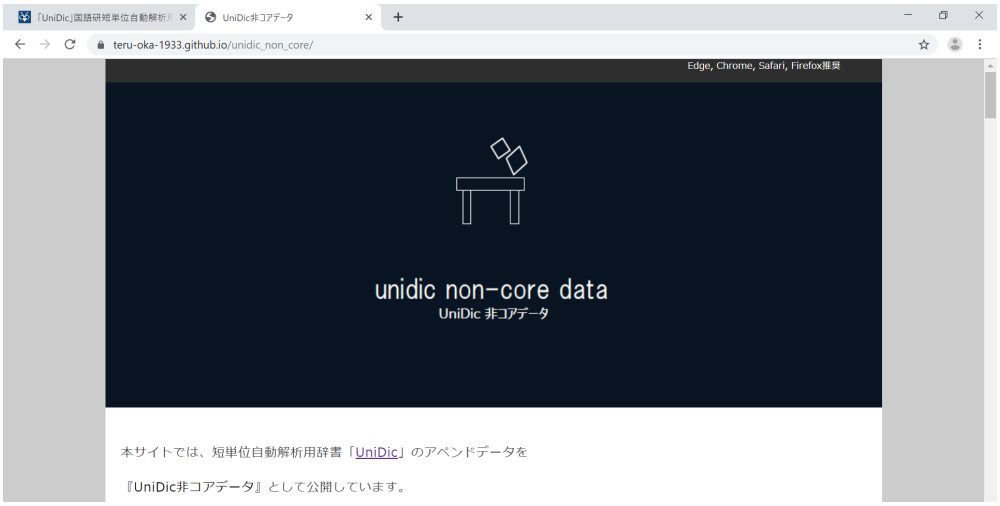

図 2　UniDic 非コアデータ公開ページ（https://teru-oka-1933.github.io/unidic_non_core/）。GitHub 上にて，フリーライセンス（Apache 2.0）で公開を行なっている。

る。意味合いは少々異なるが，BCCWJ の「非コアデータ」のような言葉の使い方だと思ってほしい。実際，今回紹介するデータはいずれもクラウドソーシングを利用して作成したデータである。クラウドソーシングは，ウェブ上の不特定多数の作業者たちに，なにかしら作業を分散・並行して依頼できる仕組みであり，大規模なアンケート調査やアノテーション作業も数百〜千人の規模で，数日のうちに完了することができる。反面，ワーカーの大多数は言語の専門家でなく，辞書整備に携わったこともない。中には作業意図を無視する悪質な作業者も存在する。そのため辞書項目の質的保証の観点から，そのデータを直接 UniDicDB に追加することはできない。しかし，クラウドソーシングを使って作成したデータであることを明記した上で，[Kondo et al., 2018] のような対応表を別途公開する分には問題なく，それを利用するか否かは UniDicMA の各ユーザの判断にゆだねられる。公開しているデータは，2019 年 8 月現在，全 3 種であり，次節で UniDicDB 上での短単位一意識別 ID について解説した後，それぞれのデータについて紹介していく。

2.　語彙表 ID と語彙素 ID

UniDicMA に ver.2.2.0 から，新たな素性[(5)]として，**語彙表 ID**（**lid**）と**語彙素 ID**（**lemma_id**）が追加された。これらは UniDicDB の中で各短単位を一意識別している ID で，

- 語彙表 ID： 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）
- 語彙素 ID： UniDic DB 中の各語彙素（語彙素-語彙素細分類，語彙素読み，語彙素類）を一意識別するための ID

となっており，基本的には UniDicMA のバージョンが上がっても変化しない情報である。

これら 2 種類の ID はもちろん，適当につけられたものでなく，UniDic の階層的見出し構造 [岡, 2019] を反映したものとなっている。例えば，

(5) 出力カラムのこと。MeCab 用語。

10進数	31678
2進数	111101110111110

図 3　「百」の語彙素 ID（10 進表記と 2 進表記）

ヒャク	8707590993355264
ビャク	8707590993363456
ヒャッ	8707590993355776
ビャッ	8707590993372168

図 4　書字形出現形「百」の発音形出現形 「ヒャク」「ビャク」「ヒャッ」「ビャッ」の語彙表 ID（10 進表記）

ヒャク	1111011101111100000100000010000000010001000100000000
ビャク	1111011101111100000100000010000000010010000100000000
ヒャッ	1111011101111100000100000010000000010001001000000000
ビャッ	1111011101111100000100000010000000010010001000000000
語彙素ID	111101110111110

図 5　書字形出現形「百」の発音形出現形 「ヒャク」「ビャク」「ヒャッ」「ビャッ」それぞれの語彙表 ID（2 進表記）

- 語彙素-語彙素細分類: 百
- 語彙素読み：ヒャク
- 語彙素類：数

という “語彙素” の語彙素 ID は「31678」である。この 10 進数を 2 進数（0 と 1 の bit 列）に変換すると，図 3 のようになる。

次にこの語彙素にぶら下がる書字形出現形「百」の発音形出現形「ヒャク」「ビャク」「ヒャッ」「ビャッ」の語彙表 ID を見ていく。10 進表記のままでは図 4 のようになるが，これらも語彙素 ID と同じく 2 進数に変換してみると，図 5 のようになる。

2 進表記を上下で見比べてみると，頭の 15 桁（15bit）はどれも同じで， しかもこれらの語彙素 ID と同じ 0 と 1 の列となっている。つまり，語彙表 ID を 2 進数に変換した bit 列中，先頭のこの部分は，語彙素 ID を表している。

実は語彙素 ID も語彙表 ID も UniDicMA の出力の上では 10 進数で表示されているものの，その本質は 2 進数に変換した長い 0 と 1 の bit 列のほうにある。bit 列の各部は図 6 のような意味を持っている。

	語彙素ID n bit (可変長最大25bit)	語形 SubID 5 bit	書字形 SubID 8 bit	発音形 SubID 8 bit	語頭変化形ID 4 bit	語末変化形ID 4 bit	活用形ID 9 bit
ヒャク	111101110111110	00001	00000010	00000001	0001	0001	000000000
ビャク	111101110111110	00001	00000010	00000001	0010	0001	000000000
ヒャッ	111101110111110	00001	00000010	00000001	0001	0010	000000000
ビャッ	111101110111110	00001	00000010	00000001	0010	0010	000000000

図 6　百の語彙表 ID（2 進表記）詳細

これをみると，4 つの発音形出現形で異なっているのは 語頭変化形 ID と語末変化形 ID の部分だけである。語頭が濁音化している「ビャク」と「ビャッ」では， 語頭変化形 ID が「0010」だが， そうでない清音の「ヒャク」と「ヒャッ」は「0001」となっている。また語末が促音になっている「ヒャッ」と「ビャッ」の 語末変化形 ID は「0010」だが， そうでない「ヒャク」と「ビャク」は「0010」である。「百」は名詞のため，語末変化の扱いが語末変化形で扱われているが，動詞や形容詞など活用のある短単位の場合，最後の 9bit の活用形 ID で語末変化の区別が行われる。

少し専門的な話になるが，これはすなわち次のことを意味している。

- 当該短単位の語彙表 ID を左に 33bit シフトすることで，その語形基本形 ID が得られる（当該短単位の語形基本形への変換・まとめ上げが可能）
- 当該短単位の語彙表 ID を左に 38bit シフトすることで，その語彙素 ID が得られる（当該短単位の語彙素への変換・まとめ上げが可能）

語彙表 ID の設定に関しては， 実際には上で説明した以上に細かな仕様がある。詳しく知りたい方は文献 [小木曽ら, 2014][小木曽ら, 2011] で確認してほしい。

3.　UniDic 非コアデータ（2019 年 8 月現在）

本節では，前述した短単位の ID 情報にひも付ける形で UniDicMA を拡張するためのデータ『UniDic 非コアデータ』の紹介を行う。2019 年 8 月現在公開しているデータは，

- 内部に原言語からの省略形を含むカタカナ語彙素のリスト_2017_02
- 複数（2 つ）の短単位に分割可能な複合語のリスト_2019_03
- 関連語リスト_2019_03

の 3 種であり，それぞれ以下のような内容となっている。

3.1　内部に原言語からの省略形を含むカタカナ語彙素のリスト_2017_02

パトカー＝パト（patrol）カー（car）のように，原言語からの省略形を内部に含むカタカナ短単位語彙素のリストである（図 7）。リスト中の記法を使うと，パトカーの「パト」は「patorol」の省略形であるため「abbr」，対して，パトカーの「カー」は「car」からの省略は起きていないので，「!abbr」マークが付けられている。公開している各ファイルの内容は以下の通り。

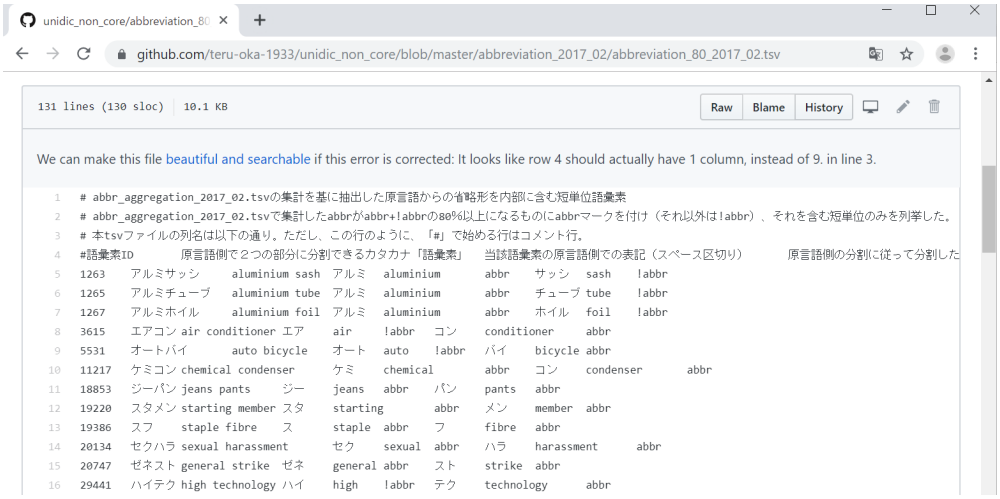

図 7　abbreviation_80_2017_02.tsv の先頭部分。

1) abbr_src_2017_02.zip（展開 → abbr_src_2017_02.tsv）
クラウドソーシングを使い，原言語からの省略が起きているかどうかアンケート調査した結果の raw データ（非識別加工済み）。1 行が 1 ユーザの一問一答となっており，クラウドソーシングで使用した質問テンプレートも含むため，少々サイズの大きいファイルとなっている。そのため，zip 圧縮形式で公開している。
2) abbr_inter_2017_02.tsv の先頭部分。
1) のデータから不要な部分を除き，閲覧しやすく加工したもの（図 8）。
3) abbr_aggregaton_2017_02.tsv
1)2) のデータは 1 行が 1 ユーザの一問一答であるので，それを集計した結果（図 9）。例えば，abbr:80，!abbr:20 の場合，その語彙素の当該箇所が原言語からの省略形であるとみなしたユーザが 80 人，そうではないとみなしたユーザが 20 人いたことを表している。
4）abbreviation_○○_2017_02.tsv
3) の集計結果から，abbr の値が abbr+!abbr 中の○○%以上になった箇所で原言語からの省略が起きているとみなし，改めて値なしの abbr マークだけを付与しなおし（○○%未満の場合!abbr マークだけ），付与しなおした中で，abbr マークを含む行だけを列挙したファイル（図 7）。

3.2　複数（2 つ）の短単位に分割可能な複合語のリスト_2019_03

短単位は 1 最小単位，もしくは複数の最小単位の結合からなる単位である（最小単位と短単位の関係については [岡, 2019] を参照）。そのため 1 短単位であっても複数の最小単位の結合からなっている場合，その最小単位それぞれが別個の短単位として UniDicDB に登録されていることがある。

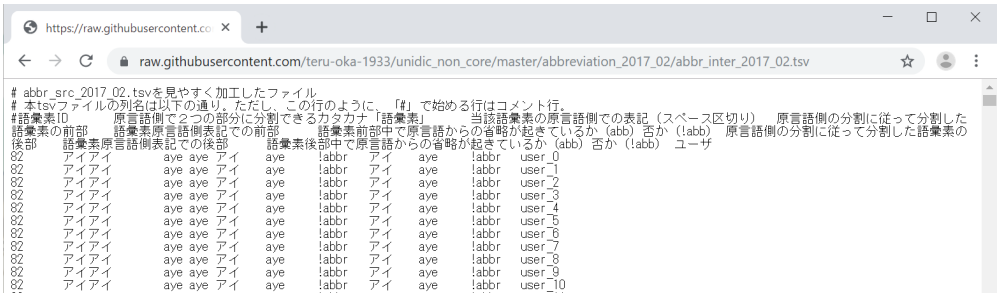

図 8　abbr_inter_2017_02.tsv の先頭部分。

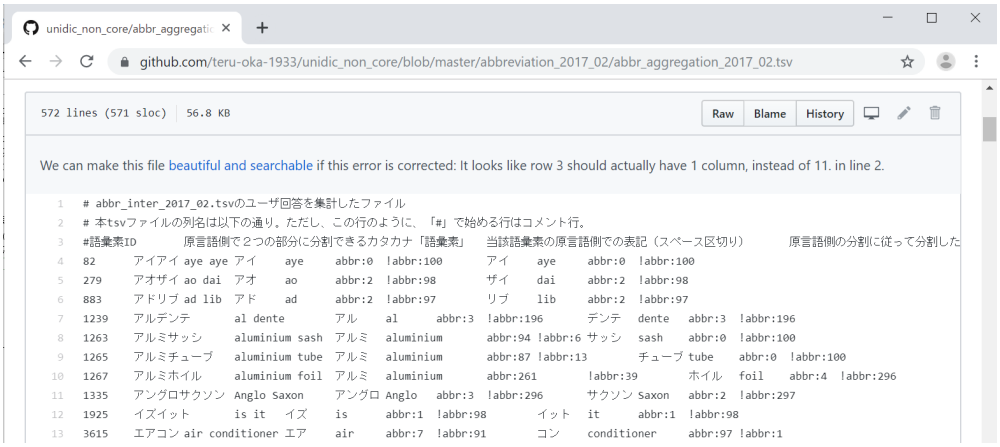

図 9　abbr_aggregation_2017_02.tsv の先頭部分。

例：（）は語彙素 ID
走り過ぎる（90675）　→　走る（29712）　過ぎる（19108）
夢見る（38909）　→　夢（38906）　見る（36920）

本データは，上のような複数短単位に分割可能な短単位語彙素のうち，“2 短単位” に分割可能なものを自動形態素解析によって候補に挙げ，クラウドソーシングを使ってその分割が正しいか否かを選別したリストである（図 10）。このデータを使うことで，現状のコーパスでは単純に「走る」でコーパスを短単位検索しても「走り過ぎる」が検索から漏れるといった問題が解決できる。公開している各ファイルの内容は以下の通り。

1) cons_src_2019_03.zip（展開 → cons_src_2019_03.tsv）
   自動形態素解析によって 2 短単位に分割した短単位を，分割の確信度に基づき並び替え，上位のもの（確信度が高い分割）を抜き出し，その分割が正しいか否かクラウドソーシングを使ってアンケート調査した結果の raw データ（非識別加工済み）。1 行が 1 ユーザの一問一答となっており，クラウドソーシングで使用した質問テンプレートも含むため，少々サイズの大きいファイルとなっている。そのため，zip 圧縮形式で公開している。

図 10　cons_all_2019_03.tsv の先頭部分。

2) cons_all_2019_03.tsv

   1) のデータに対し，クラウドソーシングの品質評価手法（Multi-class 版の GLAD : [Whitehill et al., 2009] の多クラス拡張）を適用し，各分割に行われた人手の判定に確率で確信度（0.0〜1.0）を与えたもの（図 10）。

3) constituent_〇. 〇_2019_03.tsv の先頭部分。

   2) の集計結果から，確信度が〇. 〇以上の分割の行だけを列挙したファイル。

3.3　関連語リスト_2019_03

同義や，類義，包含など，意味的に近しい関係を持つ（と思われる）字面の近しい短単位語彙素ペアのリストである（図 11）。

例：（）は語彙素 ID

アレンジメント-arrangement（1285）　アレンジ-arrange（1284）

読み聞かせる（174056）　読み聞かす（191472）

笑い声（41331）　笑う（41336）

公開している各ファイルの内容は以下の通り。

1) ref_src_2019_03.zip（展開 → ref_src_2019_03.tsv）

   次のような短単位のペアをランダムに取り出し，取り出したペアが “似たような意味を持つか否か？” クラウドソーシングを使ってアンケート調査した結果の raw データ（非識別加工済み）。

   – 片方の短単位がもう片方の短単位と全前方完全一致する
   – 2 つの短単位の先頭文字列が一致する
   – 2 つの短単位間の編集距離が近い

   1 行が 1 ユーザの一問一答となっており，クラウドソーシングで使用した質問テンプレートも含むため，少々サイズの大きいファイルとなっている。そのため，zip 圧縮形

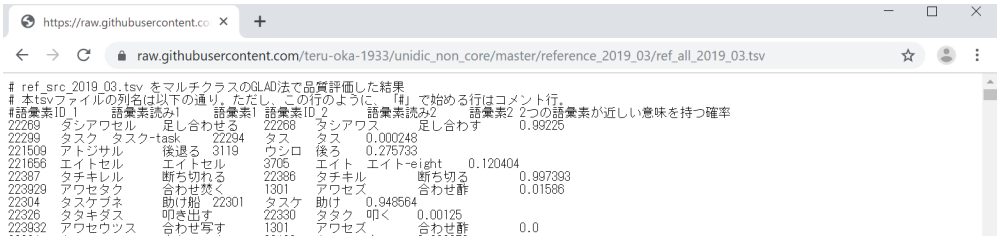

図 11　ref_all_2019_03.tsv の先頭部分。

式で公開している。

2) ref_all_2019_03.tsv

1) のデータに対し，クラウドソーシングの品質評価手法（Multi-class 版の GLAD : [Whitehill et al., 2009] の多クラス拡張）を適用し，各ペアに行われた人手の判定に確率で確信度（0.0〜1.0）を与えたもの（図 11）。

3) reference_〇. 〇_2019_03.tsv

2) の集計結果から，確信度が〇. 〇以上のペアの行だけを列挙したファイル。

4. おわりに

本稿では，国語研が所外に向けて公開している電子化辞書 UniDicMA の拡張データについて述べた。UniDic 非コアデータと名付けたこのデータは，UniDicDB 上で各短単位が保持している一意識別のための ID を利用し，それにひもづける形で UniDicMA へ新規項目の追加を実現するものである。この公開形式であれば，所外の研究者であっても UniDic へ新規情報を追加し，その情報を自由に研究者間で共有することができる。その先導的試みとして，UniDic 非コアデータは GitHub 上でフリーライセンス（Apache 2.0）での公開とした。本発表により，これまでそれぞれの研究者が独自に拡張を行なっていた UniDic データが盛んに共有されていくこと，もしくはこれから新たに UniDic データ作ろうする動きが活発になることを期待する。

謝　辞

本研究は国立国語研究所の所長裁量経費の助成を受けたものです。

文　献

[Asahara et al., 2014] Asahara, M., Maekawa, K., Imada, M., Kato, S., and Konishi, H. (2014). Archiving and Analysing Techniques of the Ultra-large-scale Web-based Corpus Project of NINJAL, Japan *Alexandria*, 25(1-2), pp.129–148.

[Kondo et al., 2018] Kondo, A., Tanaka, M., and Asahara, M. (2018). Alignment Table between UniDic and 'Word List by Semantic Principles' In *Proceedings of the 8th Conference of Japanese Association for Digital Humanities* (*JADH 2018*), pp.125–128.

[Kono et al., 2015] Kono, T. and Ogiso, T. (2015). Improving an Electronic Dictionary for Morphological Analysis of Japanese: Use of historical period information In *Proceedings of the*

18th International Conference on Machine Learning (*ICML-2001*), pp.282–289.

[Kudo et al., 2004] Kudo, T., Yamamoto, K. and Matsumoto, Y. (2004). Applying Conditional Random Fields to Japanese Morphological Analysis In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (*EMNLP-2004*), pp.230–237.

[Maekawa et al., 2000] Maekawa, K., Koiso, H., Furui, S., and Isahara, H. (2000). Spontaneous speech corpus of Japanese. In *Proceedings of the Second International Conference on Language Resources and Evaluation* (*LREC'00*).

[Maekawa et al., 2014] Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., and Den, Y. (2014). Balanced corpus of contemporary written Japanese *Language Resources and Evaluation*, 48, pp.345–371.

[Whitehill et al., 2009] Whitehill, J., Ruvolo, P., Wu, T., Bergsma, J., and Movellan, J. (2009) Whose vote should count more: Optimal integration of labels fromlabelers of unknown expertise In *Advances in Neural Information Processing Systems 22* (*NIPS'09*), pp.2035–2043.

[岡, 2019] 岡 照晃 (2019) 「言語研究のための電子化辞書」コーパスと辞書，講座 日本語コーパス 7, 朝倉書店，pp.1–28.

[小木曽ら, 2011] 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 (JC-U-10-01), https://pj.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-U-10-01.pdf

[小木曽ら, 2014] 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システムの設計・実装・運用」自然言語処理，21(2), pp. 301–332.

[鴻野知暁ら, 2014] 鴻野知暁・小木曽智信 (2014). 「見出し語の時代情報を付与した電子化辞書の構築」言語処理学会 第 20 回 年次大会 発表論文集，pp. 209–212.

[国立国語研究所, 2004] 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, 大日本図書.

[近藤, 2012] 近藤泰弘 (2012). 「日本語通時コーパスの設計」NINJAL「通時コーパス」プロジェクト・Oxford VSARPS プロジェクト合同シンポジウム「通時コーパスと日本語史研究」予稿集，pp.1–10.

[近藤, 2015] 近藤泰弘 (2015). 「『日本語歴史コーパス』と日本語史研究」コーパスと日本語史研究，ひつじ研究叢書＜言語編＞，第 127 巻， ひつじ書房，pp.1–16.

[伝ら, 2007] 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 号，pp.101–123.