文責：岡照晃

電子化辞書『UniDic』を中心に見たリレーショナル・データベースによる統合的言語資源管理環境

国立国語研究所　コーパス開発センター

○　岡　照晃

中村 壮範

2018/03/16 (Fri)　言語処理学会第24回年次大会ワークショップ
形態素解析の今とこれから @ 岡山コンベンションセンター (1Fホール右)

注意：　文責は岡のみにあります。

勢いに任せて作ったスライドなので、
途中から非常に感情的になっていきます。
また、お堅い発表は好きくないので、
内容は大分カジュアルです。

あらかじめ、ご了承ください。

Q: 『UniDic』、ご存知ですか？

About UniDic

これはUniDicの１つの側面でしかありません。

UniDicは、電子化された便利な卓上辞書です。

文責：岡照晃

UniDicは、テキストに目盛をふるためのものさしです。

UniDicとは？

UniDicとは、国語研の規定した斉一な言語単位（短単位）と 階層的見出し構造に基づく電子化辞書の

■ 設計方針

およびその実装としてのリレーショナルデータベース

■ UniDicデータベース

と、そのデータベースからエクスポートされた短単位をエントリとする形態素解析器MeCab用の解析用辞書

■ 解析用UniDic

の総称です。

UniDic
Electronic Dictionary with Uniformity and Identity.

①設計方針

【階層的な見出し構造】

異表記・異形態に対して同一の見出しを与える構造

辞書見出し*

語彙素読み：
　オオキイ
語彙素：
　大きい
類：相

語形（基本形）	書字形・発音形（基本形）	
オオキイ	書	大きい
	書	おおきい
	音	オーキー
オッキイ	書	おっきい
	音	オッキー

* 語彙素読み，語彙素，類を合わせてはじめて当該の階層構造が一意に定まる。

【斉一な言語単位】

単位認定が，ある場合は長く，ある場合は短いという不揃いをなくす

斉一でない言語単位で分割した例：

「梅雨 / 空」が２単位に分割されるのに対し，「青空」は１単位として扱われる。

梅雨 / 空 / の / 間 / の / 青空 / だ / 。

斉一な言語単位で分割した例：

梅雨空 / の / 間 / の / 青空 / だ / 。

実装

②UniDicデータベース

エクスポート

③解析用UniDic

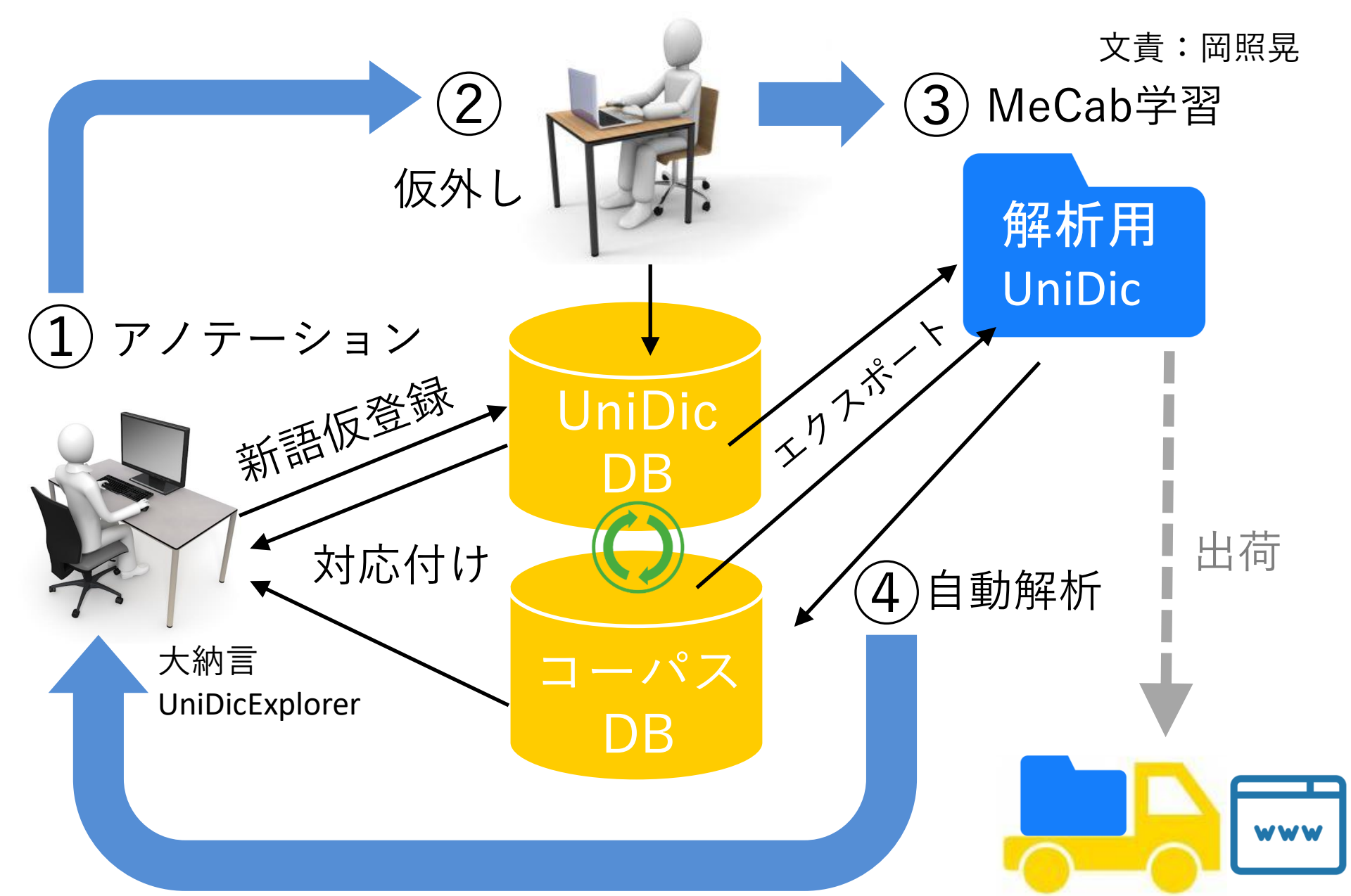
書字形（基本形）	発音形（基本形）	語形（基本形）	語彙素読み	語彙素	…
大きい	オーキー	オオキイ	オオキイ	大きい	…
おおきい	オーキー	オオキイ	オオキイ	大きい	…
おっきい	オッキイ	オッキイ	オオキイ	大きい	…

『UniDic』を使ったコーパス構築の流れ

- CSJ
- BCCWJ
- CHJ
- I-JAS
- CEJC
- 名大会話コーパス
- 女性のことば・男性のことば(職場編)

Workflow and my works

文責：岡照晃


②
仮外し
③ MeCab学習
解析用
UniDic
① アノテーション
新語仮登録
UniDic
DB
エクスポート
対応付け
④ 自動解析
出荷
大納言
UniDicExplorer
コーパス
DB
www

文責：岡照晃

② 仮外し

③ MeCab学習

解析用 UniDic

① アノテーション

各コーパス構築プロジェクトのお仕事

新語登録

UniDic DB

エクスポート

CCDのおしごと！

対応付け

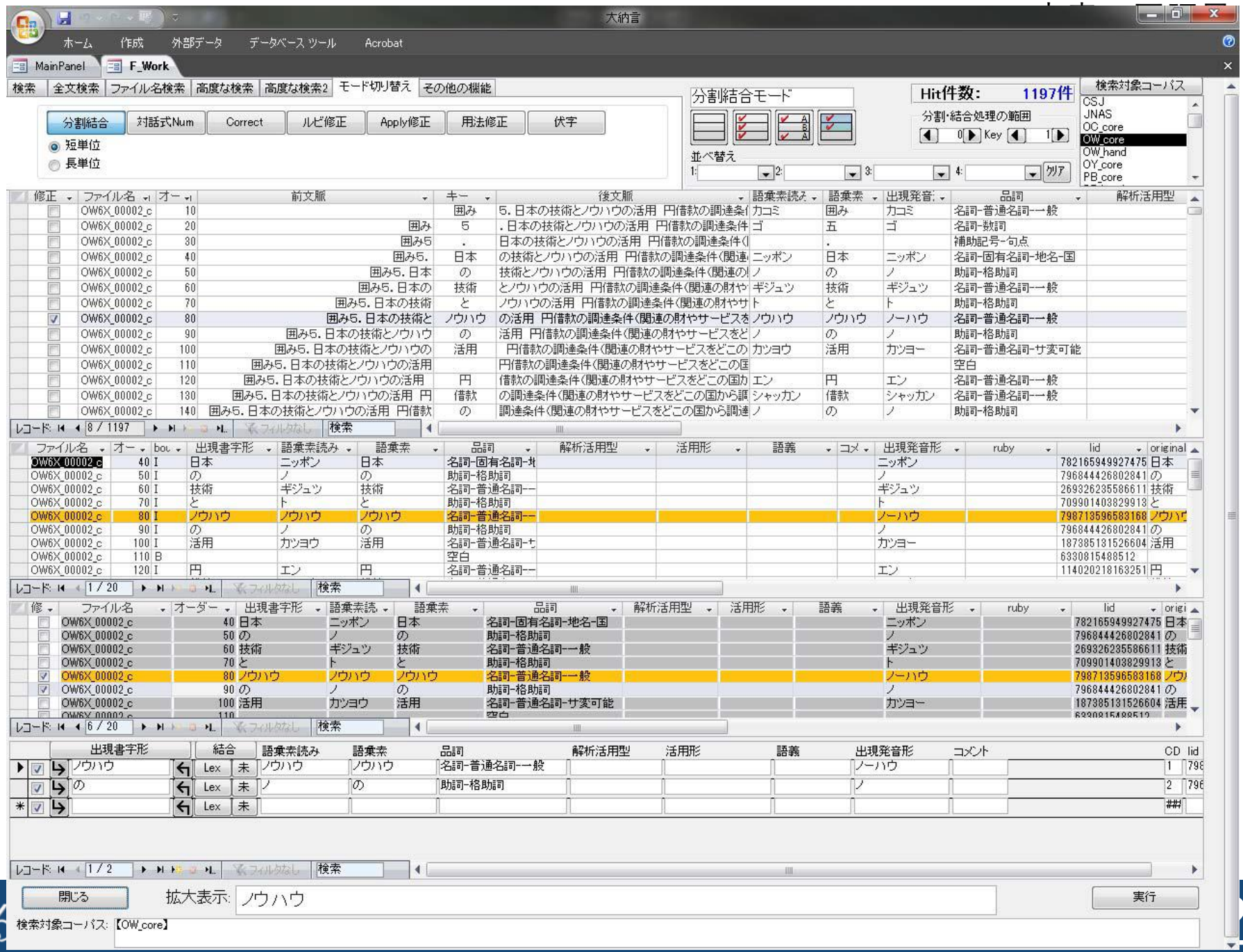
④ 自動解析

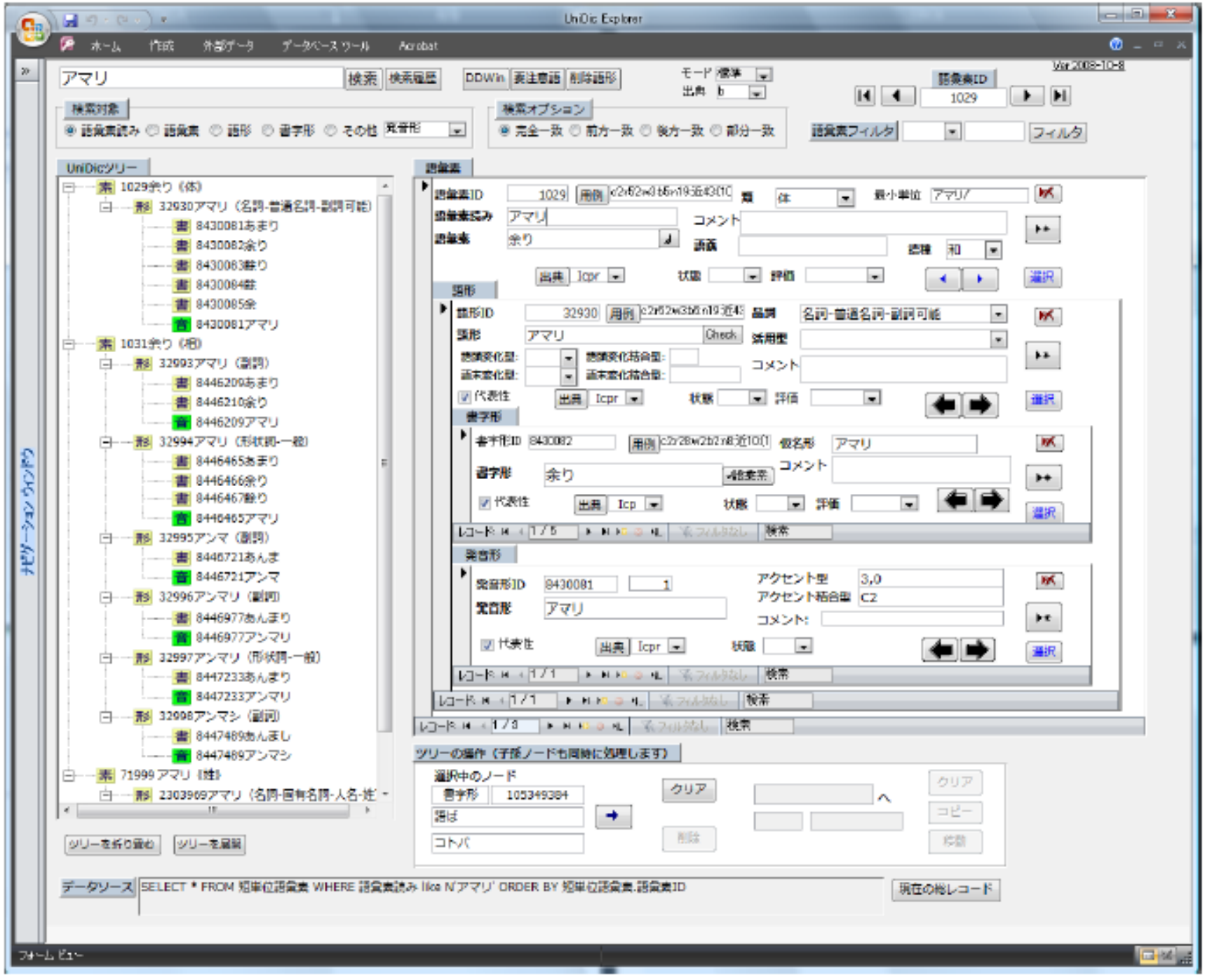
出荷

大納言
UniDicExplorer

コーパス DB

www

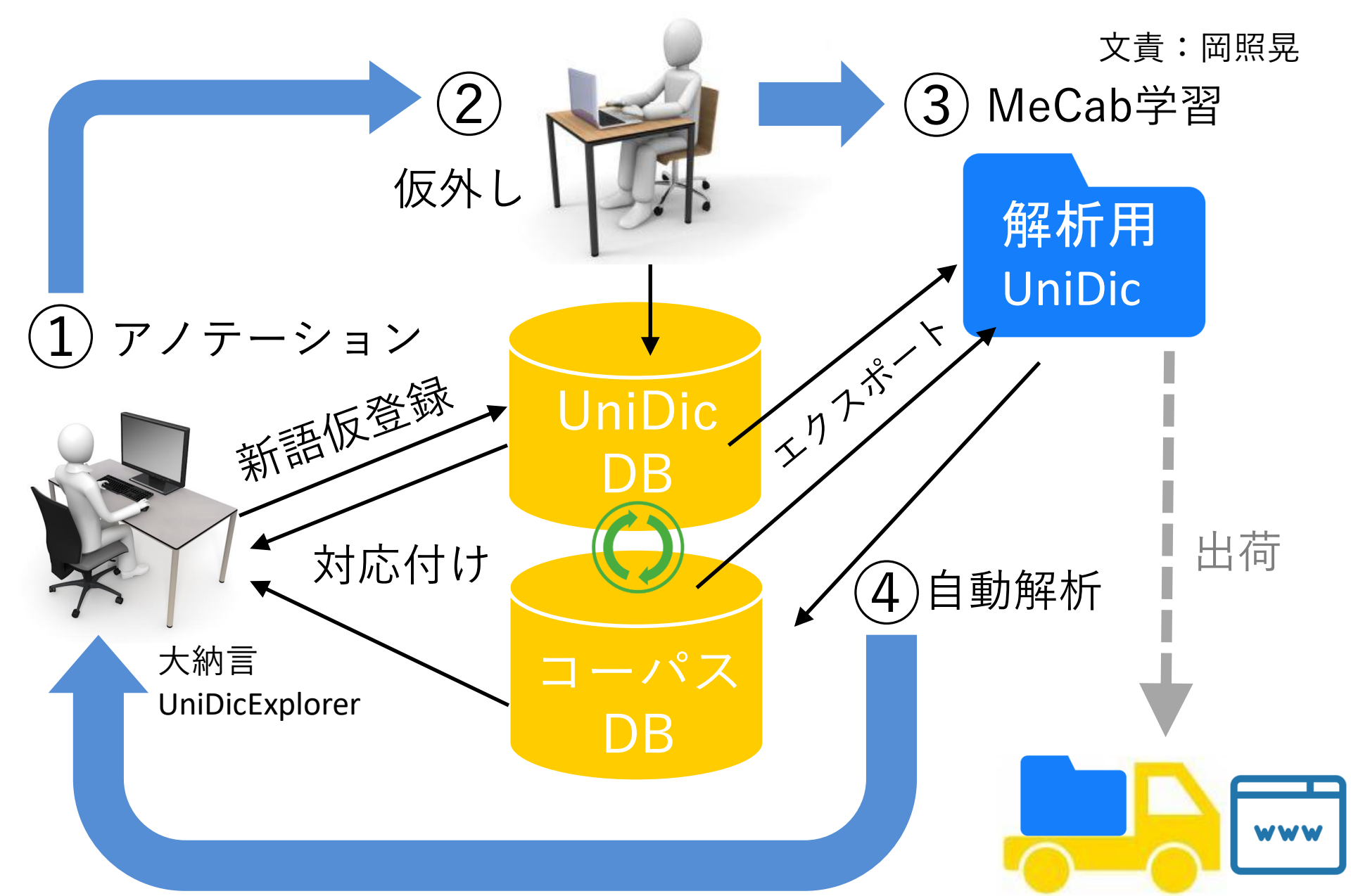

大納言
ホーム
作成
外部データ
データベース ツール
Acrobat
MainPanel
F_Work
検索
全文検索
ファイル名検索
高度な検索
高度な検索2
モード切り替え
その他の機能
分割結合
対話式Num
Correct
ルビ修正
Apply修正
用法修正
伏字
短単位
長単位
分割結合モード
Hit件数: 1197件
分割・結合処理の範囲
Key
並べ替え
クリア
検索対象コーパス
CSJ
JNAS
OC_core
OW_core
OW_hand
OY_core
PB_core
修正
ファイル名
オー
前文脈
キー
後文脈
語彙素読み
語彙素
出現発音形
品詞
解析活用型
OW6X_00002_c 10 囲み 5. 日本の技術とノウハウの活用 円借款の調達条(カコミ 囲み カコミ 名詞-普通名詞-一般
OW6X_00002_c 20 囲み 5 . 日本の技術とノウハウの活用 円借款の調達条件 ゴ 五 ゴ 名詞-数詞
OW6X_00002_c 30 囲み5 . 日本の技術とノウハウの活用 円借款の調達条件(. 補助記号-句点
OW6X_00002_c 40 囲み5. 日本 の技術とノウハウの活用 円借款の調達条件(関連 ニッポン 日本 ニッポン 名詞-固有名詞-地名-国
OW6X_00002_c 50 囲み5. 日本 の 技術とノウハウの活用 円借款の調達条件(関連の ノ の ノ 助詞-格助詞
OW6X_00002_c 60 囲み5. 日本の 技術 とノウハウの活用 円借款の調達条件(関連の財や ギジュツ 技術 ギジュツ 名詞-普通名詞-一般
OW6X_00002_c 70 囲み5. 日本の技術 と ノウハウの活用 円借款の調達条件(関連の財やサ ト と ト 助詞-格助詞
OW6X_00002_c 80 囲み5. 日本の技術と ノウハウ の活用 円借款の調達条件(関連の財やサービスを ノウハウ ノウハウ ノーハウ 名詞-普通名詞-一般
OW6X_00002_c 90 囲み5. 日本の技術とノウハウ の 活用 円借款の調達条件(関連の財やサービスをど ノ の ノ 助詞-格助詞
OW6X_00002_c 100 囲み5. 日本の技術とノウハウの 活用 円借款の調達条件(関連の財やサービスをどこの カツヨウ 活用 カツヨー 名詞-普通名詞-サ変可能
OW6X_00002_c 110 囲み5. 日本の技術とノウハウの活用 円借款の調達条件(関連の財やサービスをどこの国 空白
OW6X_00002_c 120 囲み5. 日本の技術とノウハウの活用 円 借款の調達条件(関連の財やサービスをどこの国か エン 円 エン 名詞-普通名詞-一般
OW6X_00002_c 130 囲み5. 日本の技術とノウハウの活用 円借款 の調達条件(関連の財やサービスをどこの国から調 シャッカン 借款 シャッカン 名詞-普通名詞-一般
OW6X_00002_c 140 囲み5. 日本の技術とノウハウの活用 円借款 の 調達条件(関連の財やサービスをどこの国から調達 ノ の ノ 助詞-格助詞
レコード: 8 / 1197
フィルタなし
検索
ファイル名 オー bou 出現書字形 語彙素読み 語彙素 品詞 解析活用型 活用形 語義 コメ 出現発音形 ruby lid original
OW6X_00002_c 40 I 日本 ニッポン 日本 名詞-固有名詞-地 ニッポン 782165949927475 日本
OW6X_00002_c 50 I の ノ の 助詞-格助詞 ノ 796844426802841 の
OW6X_00002_c 60 I 技術 ギジュツ 技術 名詞-普通名詞-一 ギジュツ 269326235586611 技術
OW6X_00002_c 70 I と ト と 助詞-格助詞 ト 709901403829913 と
OW6X_00002_c 80 I ノウハウ ノウハウ ノウハウ 名詞-普通名詞-一 ノーハウ 798713596583168 ノウハウ
OW6X_00002_c 90 I の ノ の 助詞-格助詞 ノ 796844426802841 の
OW6X_00002_c 100 I 活用 カツヨウ 活用 名詞-普通名詞-サ カツヨー 187385131526604 活用
OW6X_00002_c 110 B 空白 6330815488512
OW6X_00002_c 120 I 円 エン 円 名詞-普通名詞-一 エン 114020218163251 円
レコード: 1 / 20
フィルタなし
検索
修 ファイル名 オーダー 出現書字形 語彙素読 語彙素 品詞 解析活用型 活用形 語義 出現発音形 ruby lid origi
OW6X_00002_c 40 日本 ニッポン 日本 名詞-固有名詞-地名-国 ニッポン 782165949927475 日本
OW6X_00002_c 50 の ノ の 助詞-格助詞 ノ 796844426802841 の
OW6X_00002_c 60 技術 ギジュツ 技術 名詞-普通名詞-一般 ギジュツ 269326235586611 技術
OW6X_00002_c 70 と ト と 助詞-格助詞 ト 709901403829913 と
OW6X_00002_c 80 ノウハウ ノウハウ ノウハウ 名詞-普通名詞-一般 ノーハウ 798713596583168 ノウハ
OW6X_00002_c 90 の ノ の 助詞-格助詞 ノ 796844426802841 の
OW6X_00002_c 100 活用 カツヨウ 活用 名詞-普通名詞-サ変可能 カツヨー 187385131526604 活用
OW6X_00002_c 110
レコード: 6 / 20
フィルタなし
検索
出現書字形 結合 語彙素読み 語彙素 品詞 解析活用型 活用形 語義 出現発音形 コメント CD lid
ノウハウ Lex 未 ノウハウ ノウハウ 名詞-普通名詞-一般 ノーハウ 1 798
の Lex 未 ノ の 助詞-格助詞 ノ 2 796
Lex 未 ##
レコード: 1 / 2
フィルタなし
検索
閉じる
拡大表示: ノウハウ
実行
検索対象コーパス: 【OW_core】


UniDic Explorer
ホーム
作成
外部データ
データベース ツール
Acrobat
アマリ
検索
検索履歴
DDWin
表注意語
削除語形
モード
出典
語彙素ID
1029
検索対象
語彙素読み
語彙素
語形
書字形
その他
発音形
検索オプション
完全一致
前方一致
後方一致
部分一致
語彙素フィルタ
フィルタ
UniDicツリー
語彙素
語彙素ID
1029
用例
品詞
名詞-普通名詞-副詞可能
最小単位
アマリ/
語彙素読み
アマリ
コメント
語彙素
余り
語種
和
出典
Jcpr
状態
評価
選択
語形
語形ID
32930
語形
アマリ
Check
活用型
語彙素化型:
語彙素化結合型:
コメント
書字形化型:
書字形化結合型:
代表性
書字形
書字形ID
8430082
仮名形
アマリ
書字形
余り
語彙素
コメント
代表性
出典
Icp
状態
評価
レコード
1 / 5
フィルタなし
検索
発音形
発音形ID
8430081
1
アクセント型
3,0
アクセント結合型
C2
発音形
アマリ
コメント:
代表性
出典
Jcpr
状態
レコード
1 / 1
フィルタなし
検索
レコード
1 / 1
フィルタなし
検索
レコード
1 / 3
フィルタなし
検索
ツリーの操作（子孫ノードも同時に処理します）
選択中のノード
書字形
105349384
語彙素
コトバ
クリア
削除
へ
クリア
コピー
挿入
ツリーを折り畳む
ツリーを展開
データソース
SELECT * FROM 短単位語彙素 WHERE 語彙素読み like N'アマリ' ORDER BY 短単位語彙素.語彙素ID
現在の総レコード
フォーム ビュー
ナビゲーション ウィンドウ
1029余り（体）
32930アマリ（名詞-普通名詞-副詞可能）
8430081あまり
8430082余り
8430083餘り
8430084餘
8430085余
8430081アマリ
1031余り（形）
32993アマリ（副詞）
8446209あまり
8446210余り
8446209アマリ
32994アマリ（形状詞-一般）
8446465あまり
8446466余り
8446467餘り
8446465アマリ
32995アンマ（副詞）
8446721あんま
8446721アンマ
32996アンマリ（副詞）
8446977あんまり
8446977アンマリ
32997アンマリ（形状詞-一般）
8447233あんまり
8447233アンマリ
32998アンマシ（副詞）
8447489あんまし
8447489アンマシ
71999アマリ（姓）
2303969アマリ（名詞-固有名詞-人名-姓

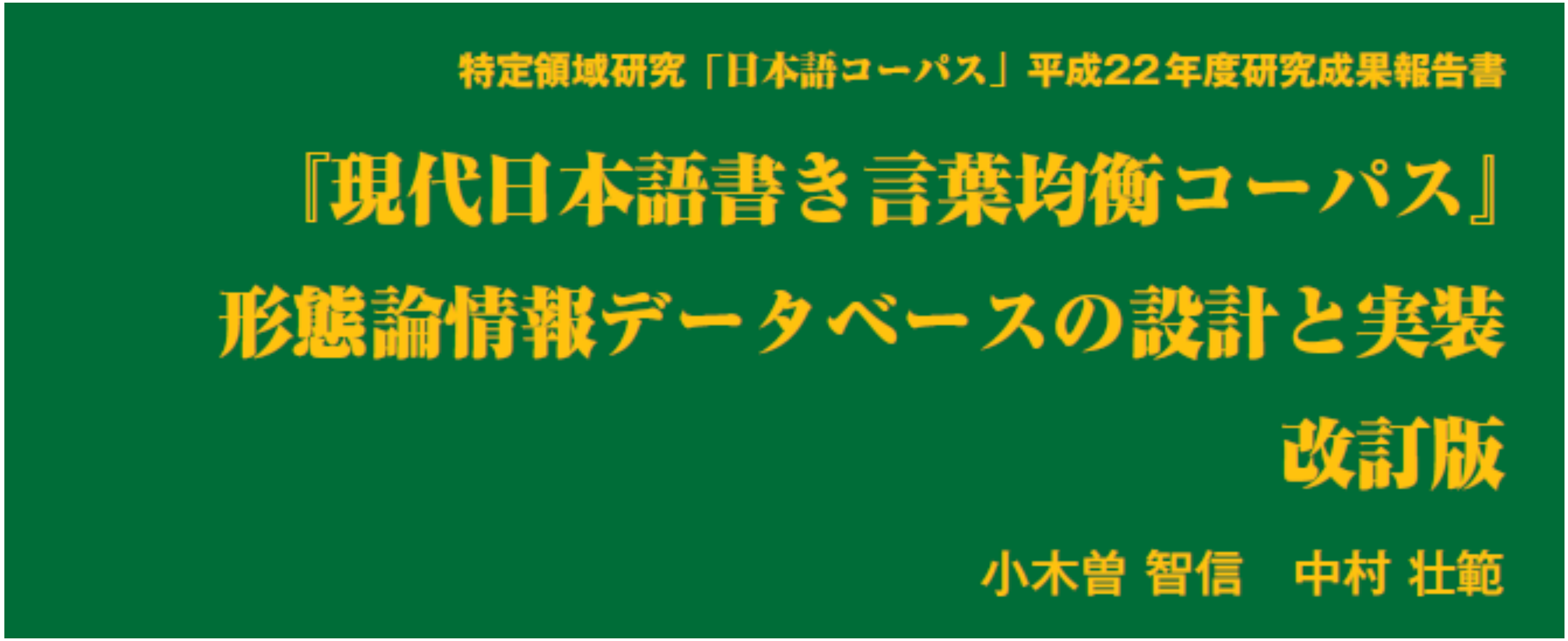
晃

文責：岡照晃

UniDicは、電子化された便利な卓上辞書です。


②
仮外し
③ MeCab学習
解析用
UniDic
① アノテーション
新語仮登録
UniDic
DB
エクスポート
対応付け
④ 自動解析
出荷
大納言
UniDicExplorer
コーパス
DB
www

文責：岡照晃



DBの詳しい実装や、大納言・ UniDicExplorerについては
この文献を参照

文責：岡照晃

② 仮外し

③ MeCab学習

解析用
UniDic

① アノテーション

各コーパス構築プロジェクトのお仕事

新語登録

対応付け

大納言
UniDicExplorer

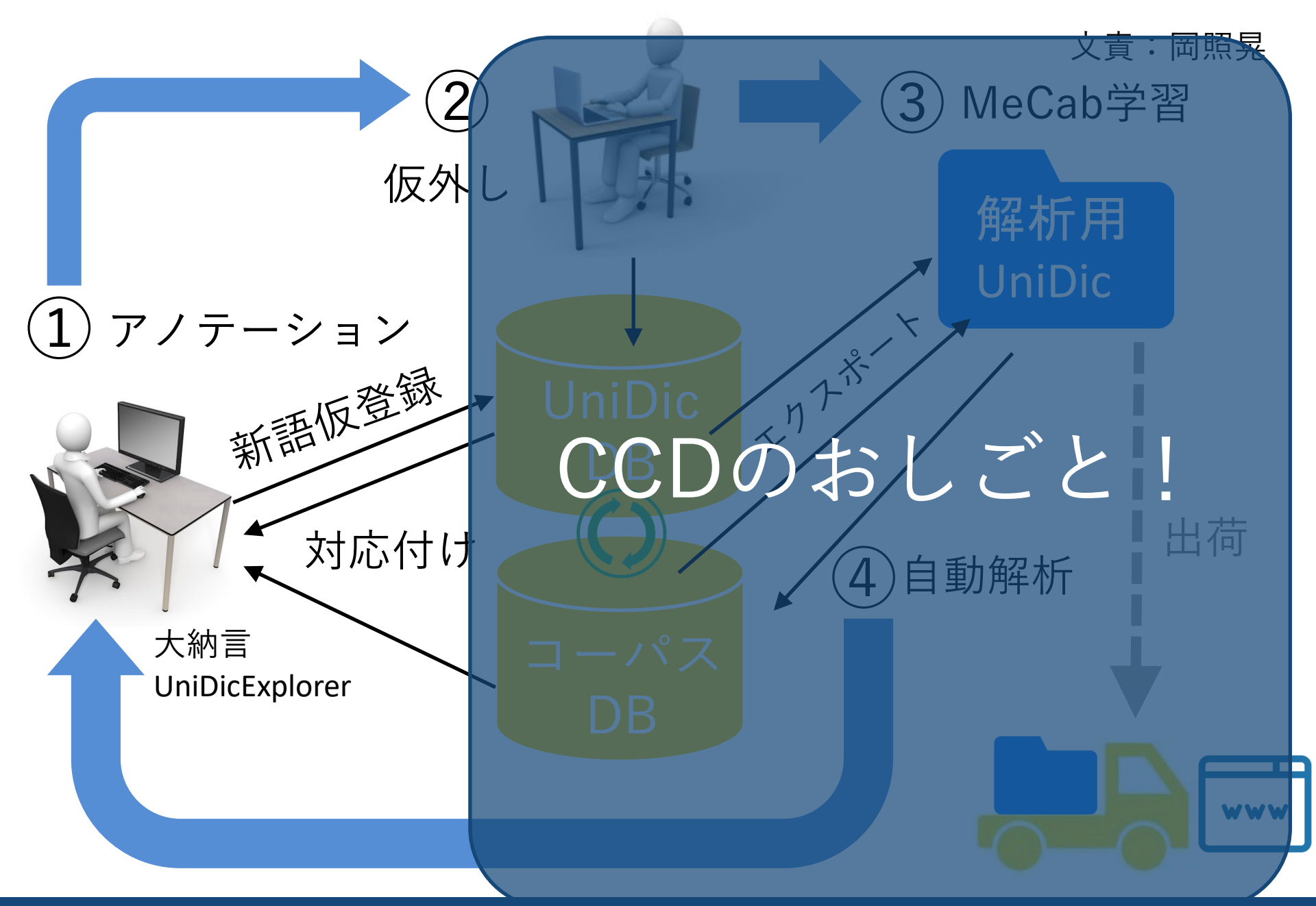
UniDic
DB

エクスポート

④ 自動解析

コーパス
DB

出荷


文責：岡照晃
② 仮外し
③ MeCab学習
解析用 UniDic
① アノテーション
新語仮登録
UniDic DB
エクスポート
CCDのおしごと！
対応付け
出荷
④ 自動解析
大納言
UniDicExplorer
コーパス DB
www

ポイント１：
UniDicはCorpus Drivenで拡張されていく辞書

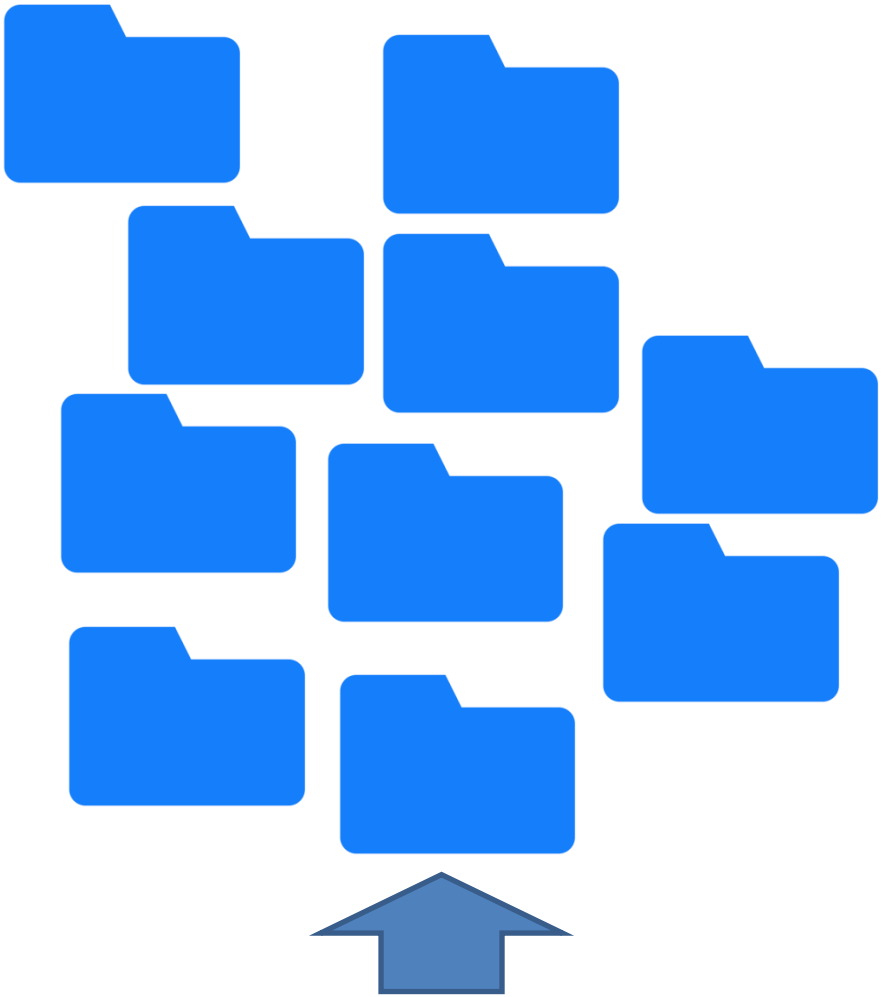
ポイント２：
UniDicの目的は、国語研で構築している
コーパスアノテーションを支援すること

理由：非コアデータの作成のため

- 例：BCCWJ
  - コーパスの規模は1億語（短単位）
  - ただし、人手で形態論アノテーションしたコアデータは、全体の約100分の1
  - 残りは、非コアデータ＝自動解析結果を提供
    - 解析用UniDicを使用

- 解析用UniDicを使えば、誰でも非コアデータ相当のものが作れますよ～　という気分。

実は、公開されていたのはごく一部



文責：岡照晃

出荷先も１つじゃなかった

Web茶まめが使用している、時代別UniDic辞書を公開しています。

ダウンロード

現代語UniDic	ver.2016.3	同意してダウンロード
現代語話し言葉UniDic	ver.2016.3	同意してダウンロード
旧仮名口語UniDic	ver.2016.3	同意してダウンロード
近代文語UniDic	ver.2016.3	同意してダウンロード
近世口語（洒落本）UniDic	ver.2016.3	同意してダウンロード
中世口語（狂言）UniDic	ver.2016.3	同意してダウンロード
中世文語（説話・随筆）UniDic	ver.2016.3	同意してダウンロード
中古和文UniDic	ver.2016.3	同意してダウンロード
上代（万葉集）UniDic	ver.2016.3	同意してダウンロード

UniDic

概要 ▾ ダウンロード ソースコード ▾ チケット ▾ コミュニケー

プロジェクトの説明 レビューする

画像一覧

UniDicとは

- UniDicは日本語テキストを単語に分割し，形態論情報を付与するための電子化辞書です。
- unidic-mecabは形態素解析器MeCabの辞書として利用できます。
- UniDicは国立国語研究所の現代日本語書き言葉均衡コーパスにも利用されています。

UniDicの特長

- 国立国語研究所で規定した「短単位」という揺れがない斉一な単位で設計されてい
- 語彙素・語形・書字形・発音形の階層構造を持ち，表記の揺れや語形の変異にかかわらず同一の見出しを
- 語種をはじめとする言語研究に有用な情報を付与することができます。

ライセンス

- 形態素解析辞書としてのUniDicは、Ver.2.0.1以降、完全なフリーソフトウェアになりました。
- GPL/LGPL/BSD Licenseのトリプルライセンスです。

PUKIWIKI

UniDic/近代文語UniDic

Top / UniDic / 近代文語UniDic

所外での野生化も進む。

現在はここで一元管理中

探せるだけ探し出して、権利関係クリアしたものはすべてここに置きました。

http://unidic.ninjal.ac.jp/

ここにないものは、サポート対象外。

ワークフロー内の岡の仕事

- 解析用UniDicの作成
  - 随時、改良も加えてます。
    - 学習用コーパスの選定
    - MeCab設定ファイルの改善
      - feature.def
      - rewrite.def
      - char.def
      - dicrc
- 解析用UniDicを使った自動解析結果の提供
- 解析用UniDicの出荷
  - 含、公開用サイト作成、更新、用語集・FAQ整備

CCDのおしごと！

② 仮外し

③ MeCab学習

解析用 UniDic

① アノテーション

新語仮登録

対応付け

UniDic DB

エクスポート

④ 自動解析

出荷

コーパス DB

大納言
UniDicExplorer

② 仮外し

③ MeCab学習

解析用
UniDic

① アノテーション

新語仮登録

対応付け

UniDic
DB

コーパス
DB

エクスポート

おかの
おしごと！

④ 自動解析

出荷

大納言
UniDic explorler

文責：岡照晃

UniDicって、よくディスられますよねー（守ってあげたくなる系辞書）

Disrespect for UniDic

UniDicよくあるdis

- 短く切りすぎ。短く切って分割性能F1値99って、ズルい。
- 係り受け解析に使えない。
- 斉一性を高めることが目的になってる。
  - 可能性に基づく品詞体系多用しすぎ。
  - 本当に解いて欲しい問題を巧みに回避している。
- 斉一とか言いつつ、「苺狩り」は１単位なのに「葡萄|狩り」が２単位、とかある。

UniDicよくあるdis

- 短く切りすぎ。短く切って分割性能F1値99って、ズルい。
- 係り受け解析に使えない。
- 斉一性を高めることが目的になってる。
  - 可能性に基づく品詞体系多用しすぎ。
  - 本当に解いて欲しい問題を巧みに回避している。
- 斉一とか言いつつ、「苺狩り」は１単位なのに「葡萄|狩り」が２単位、とかある。

UniDicDBを使ったコーパス作成では「操作主義的立場」をとる

「これこれこういうものを『～単位』とする」という規定をするだけで、 その「～単位」が言語学的にどういうものなのか、 単語なのか、単語でないとするなら、どこが単語と違うのか といった問題にはまったく触れない 。

- 短単位：　意味を持つ最小の単位（≒形態素）である最小単位の２個までの結合で認定
- 長単位：　文節を自立部と付属部に分けることで認定

文責：岡照晃

短単位と長単位

<table>
<tr><td rowspan="2">文節</td><td colspan="8">国立国語研究所の</td></tr>
<tr><td colspan="8">↓ 文節を自立部と付属部に分けることで認定（トップダウン）</td></tr>
<tr><td rowspan="2">長単位</td><td colspan="7">国立国語研究所</td><td>の</td></tr>
<tr><td colspan="7">← 長単位を越える短単位は認定しない →</td><td></td></tr>
<tr><td rowspan="2">短単位</td><td colspan="2">国立</td><td colspan="2">国語</td><td colspan="2">研究</td><td>所</td><td>の</td></tr>
<tr><td colspan="8">↑ 最小単位の結合により認定（ボトムアップ）</td></tr>
<tr><td>最小単位</td><td>国</td><td>立</td><td>国</td><td>語</td><td>研</td><td>究</td><td>所</td><td>の</td></tr>
</table>

UniDicは短単位の辞書なので、短く切って当然。

UniDicよくあるdis

- 短く切りすぎ。短く切って分割性能F1値99って、ズルい。
- 係り受け解析に使えない。
- 斉一性を高めることが目的になってる。
  - 可能性に基づく品詞体系多用しすぎ。
  - 本当に解いて欲しい問題を巧みに回避している。
- 斉一とか言いつつ、「苺狩り」は１単位なのに「葡萄|狩り」が２単位、とかある。

形態素解析の目的には２つの立場がある [吉田, 84]

① 形態素解析を後段の構文解析・意味解析へ進むための準備段階と捉え、機械翻訳・質問応答・情報検索など、より下流の解析処理を目指していく立場。

② 形態素解析の段階での結果を最終的なものとして使用する立場。

短単位のアノテーション補助を目的に作られているUniDicにとっては、②の立場が当てはまる。

UniDicよくあるdis

- 短く切りすぎ。短く切って分割性能F1値99って、ズルい。
- 係り受け解析に使えない。
- 斉一性を高めることが目的になってる。
  - 可能性に基づく品詞体系多用しすぎ。
  - 本当に解いて欲しい問題を巧みに回避している。
- 斉一とか言いつつ、「苺狩り」は１単位なのに「葡萄|狩り」が２単位、とかある。

はい。斉一性を高めることが目的です。

斉一性が担保されてない場合

- コーパス中で、もし、
  - ＼国立国語研究所＼
  - ＼国立＼国語研究所＼
  - ＼国立＼国語研究＼所＼
  - ＼国立＼国語＼研究＼所＼

  という分割が混在していて、
- 「このコーパスのサイズは123,456語です！」

  と言われたら、「は？（怒）」ってなりますよね？
- 「数出すなら、切り方、統一しろよ（怒）」

UniDicは、テキストに目盛をふるためのものさしです。

短”単位”は、目盛の幅。
メートルとか、kgとおなじ。

だけれど、

- アノテータの中には、「私の考える『単語』」がどうしてもある。
- そして、短単位は「コーパスから切り出すことで認定する」（Corpus Driven）
  - 文脈付きの文字列から切り出すため、文脈によって、切り方が揺れる場合がある。
- UniDicをものさしにするには、コーパスアノテーション時に、「私の考える『単語』」の混入を防いで、１目盛をアノテータ間で共有してないといけない。

そのため、アノテータの統制が必要

- UniDic DBとコーパスデータベースの対応付けによる、コーパスアノテーション
- 短単位の規定集を作り、新しい短単位をUniDic DBに追加する際に参照
- 新しい短単位は仮登録とし、「仮外し」の作業を専門家が行う

可能性に基づく品詞体系の多用

- 名詞-普通名詞-サ変可能（e.g., 運動、アクセス）
- 名詞-普通名詞-形状詞可能（e.g., 安全、健康、アクティブ）
- 名詞-普通名詞-サ変形状詞可能（e.g., 安心、おしゃれ、オーバー）
- 名詞-普通名詞-副詞可能（e.g., 今日、毎日、以上、今度）
- 名詞-普通名詞-助数詞可能（e.g., 円、ドル、メートル、グラム、時間、箇月、条）
- 動詞-非自立可能（e.g., する、くる、いく）
- 形容詞-非自立可能（e.g., ない、欲しい、よい）
- 接尾辞-名詞的-サ変可能（e.g., 化、ナイズ、分）
- 接尾辞-名詞的-形状詞可能（e.g., 三昧、深）
- 接尾辞-名詞的-サ変形状詞可能（e.g., unidic-mecab-2.1.2_srcに0例）
- 接尾辞-名詞的-副詞可能（e.g., 当り、中、後）

これらは文脈に応じて、品詞が変わる。

文脈見たらわかるでしょ？
いいえ、わからない例もあります。

- 「出勤するときはいつもコンビニに寄っていく」の「いく」は自立？　非自立？

- そこであえて一方に倒してしまうのではなく、最終的な判断はその分野の研究者に任せ、その判断自体を研究の対象としてもらうことをUniDicでは目指している。

脱文脈化

UniDicよくあるdis

- 短く切りすぎ。短く切って分割性能F1値99って、ズルい。
- 係り受け解析に使えない。
- 斉一性を高めることが目的になってる。
  - 可能性に基づく品詞体系多用しすぎ。
  - 本当に解いて欲しい問題を巧みに回避している。
- 斉一とか言いつつ、「苺狩り」は１単位なのに「葡萄/狩り」が２単位、とかある。

文責：岡照晃

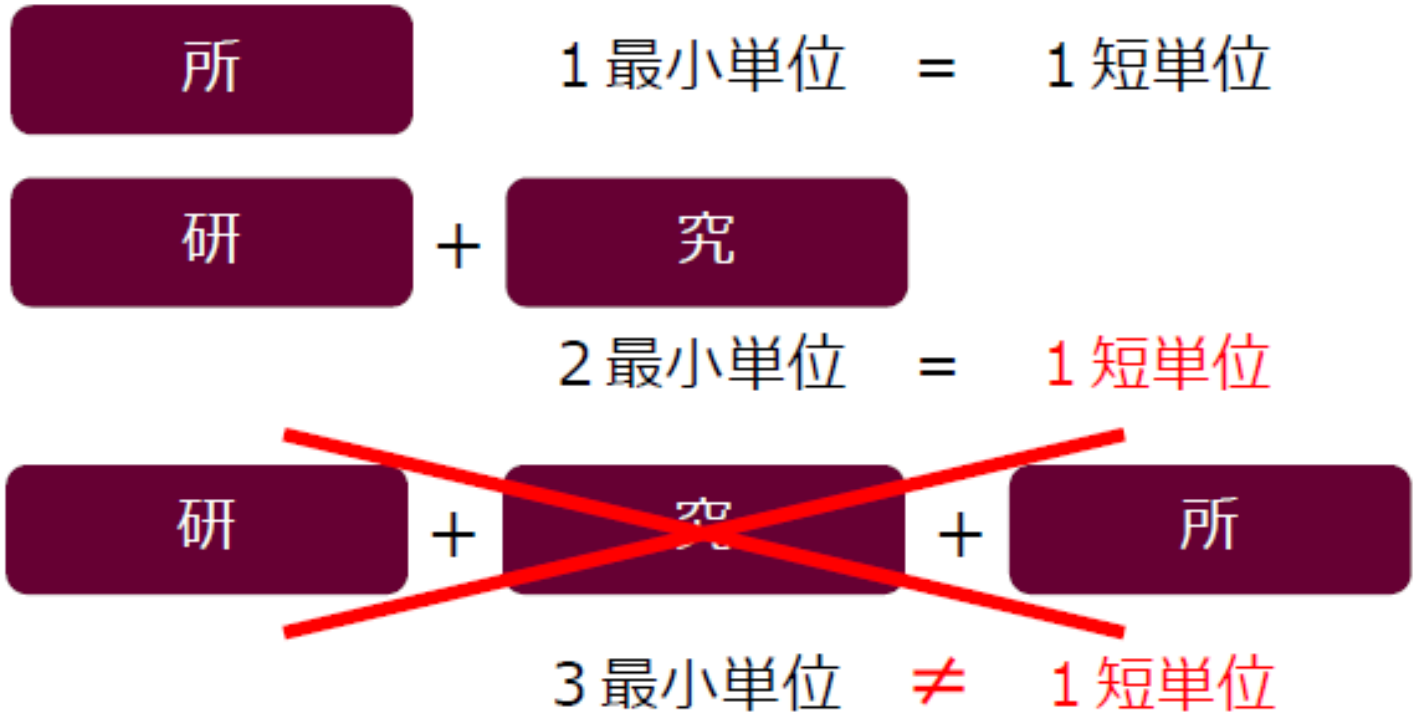
短単位を認定するための最小単位結合規定①

① 和語＋和語の結合は２つまで

文責：岡照晃

短単位を認定するための最小単位結合規定②

② 漢語＋漢語の結合は２つまで



この規定に則れば、操作的にこうなる

- 苺（最小単位：和語）

＋狩り（最小単位：和語）

＝苺狩り（短単位：和語）

- 葡（最小単位：漢語）

＋ 萄（最小単位：漢語）

＝ 葡萄（短単位：漢語）

/ 狩り（最小単位：和語→短単位：和語）

最小単位は意味を持つ最小の単位。
漢字は基本１文字１最小単位。

和語とか、漢語とかの判定とあわせ、
岩波の辞書や日本国語大辞典
を参考に認定しています。

反論

- やっぱり、気持ち悪い。
- 「狩り」は動詞由来の名詞なんだから、動詞由来の名詞に関しては、和語和語でも結合しないとかいうルール作れよ。

それに反論

- 「息継ぎ」とか「息＼継ぎ」に切るのかよ！
- あと、ルール複雑にすると、今度はどのルールを適用するかで、揺れが入るんだよ！

今月出荷する新しい解析用UniDic
・現代書き言葉UniDic 2.3.0
・現代話し言葉UniDic 2.3.0

UniDic 2.3.0

UniDic 2.3.0

- feature.defにアクセント素性の追加
- feature.defの余分な素性の削除
- 日常会話コーパスを学習に使った「現代話し言葉UniDic」の性能向上・書き言葉コーパスとの差分化
- 規定集の一部改変（BCCWJの規定からの分離）
- 規定にそぐわない短単位の削除
- nwjcから獲得した新語の追加
- NFKC対応
- 短単位-分類語彙表番号の対応表公開
  - https://github.com/masayu-a/wlsp2unidic
- Windows用GUIツールChaMame同梱
  - https://ja.osdn.net/projects/chaki/releases/69159

まとめ

- 解析用UniDicの特徴：
  - 斉一性を重視した短単位をエントリとするMeCab用の辞書
- 解析用UniDicの開発方針：
  - 新規の短単位の追加は（基本）Corpus Driven
  - エントリが「短単位」であることを重視
    - コーパスDBとUniDicDBを使ったアノテーション、規定集の整備、仮外し作業
- 解析用UniDicの想定ユーザ：
  - コーパスを使いたい日本語の研究者（文系寄り）
- 解析用UniDicと他のシステムとの違い：
  - 解析結果がゴール

謝辞

- 本研究は国立国語研究所コーパス開発センターの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」(2016-2021 年度) の成果である。