Research Centre for Japanese Language and Linguistics
University of Oxford

オックスフォード大学 日本語研究センター
www.orinst.ox.ac.uk/research/jap-ling/

通時コーパスで使用する辞書の構築について

Kerri L Russell
ケリー・ラッセル

第45回　NINJAL（国語研）サロン
2011年5月10日

目次

- オックスフォード大学で行われる日本語史の研究プロジェクト (VSARPJ) の紹介
- lexicon の紹介
- UniDic の紹介
- 通時辞書を構築する問題

VSARPJ の紹介

- 「日本語史における動詞の意味構造と項の具現化：日本語史におけるシンタクスの基礎に関する包括的研究」 Verb semantics and argument realization in pre-modern Japanese: A comprehensive study of the basic syntax of pre-modern Japanese
- 基金: Arts and Humanities Research Council, UK （英国芸術・人文リサーチカウンシル）
- 期間: ２００９年１月から２０１３年１０月まで
- 人事：
  - プロジェクトリーダー Frellesvig (Principal Investigator), Sells (Co-Investigator)
  - フルタイム PD 研究員二人 Horn, Russell
  - 大学院生１人
  - その他、十人以上の日本・アメリカ・イギリスの研究者のメンバー
- url: http://vsarpj.orinst.ox.ac.uk

VSARPJ の紹介

- 日本語史における動詞の項の具現化を詳細に調査し、８世紀初頭から１７世紀初頭の史的段階における文法の基本的側面に対し包括的な説明を与えようとするものである。
- プロジェクトの初期段階は記述的・分析的研究の基礎をなす各時代の代表的テキストの、文法的アノテーションを含む電子コーパスの作成である。

VSARPJ の紹介

- 文法的アノテーション
  - 原文の表記法
    - 表意文字
    - 表音文字
  - 品詞
  - 形態素解析
    - 接頭語、接尾語
    - 活用
  - 語彙素と形態素の ID 番号
  - シンタクス
    - 文、節、項
    - 項の意味役割

VSARPJ の紹介

- 現在はオックスフォード大学と国語研は研究協力することになった。
- オックスフォード大学は上代語コーパスを作成中。
  - 項の意味役割のタグ付け作業中。
  - http://vsarpj.orinst.ox.ac.uk/corpus/
- 国語研の古典コーパスを使って、シンタクスのアノテーションを含む。
- 1 つの問題はオックスフォード大学の使っている ID 番号は国語研の ID 番号と一致していないことである。
  - ID 番号は lexicon.xml にある。

lexicon の紹介

- 今のところは見出し語が 2514 しかない。
  - 助詞、助動詞、動詞、代名詞、数詞は全部 lexicon にある。
  - 形容詞、副詞、名詞は少ししかない。
- 見出しの情報
  - ID 番号
  - 各時代の発音表記
  - 品詞
  - 意味
  - 使用
    - いつから、いつまで
    - 方言
- 研究のために、辞書を構築することになった。
  - 研究成果も含む予定。

```
<superEntry xml:id="L000006">  ID番号
  <entry>
    <form>
      <orth stage="I" alt="yes">-n-</orth>
      <orth stage="I" alt="yes">-zu</orth>
      <orth stage="EMJ" alt="yes">-n</orth>
      <gramGrp>
        <pos>auxiliary</pos>  品詞
        <iType type="irr" affixType="aStem"/>
                                              活用形
      </gramGrp>
    </form>
    <sense n="1">
      <def>[negative]</def>  意味
    </sense>
  </entry>
  <entry>
    <form>
      <usg type="geo">EOJ</usg>  東語
      <orth stage="OJ" alt="yes">-nana</orth>
      <orth stage="OJ" alt="yes">-nani</orth>
      <gramGrp>
```

```
<superEntry xml:id="L000011">
  <entry>
    <form>
      <usg only="OJ"/>
      <orth stage="I">-ye-</orth>
      <gramGrp>
        <pos>auxiliary</pos>
        <iType type="LB" affixType="aStem"/>
      </gramGrp>
    </form>
    <sense n="1">
      <def>[passive]</def>
    </sense>
  </entry>
</superEntry>
```

UniDic の紹介

- UniDic は国語研の形態素解析用の辞書
  - ID 番号
  - 品詞
  - 活用形
  - 用例

UniDic Explorer

ミル | 検索 | 検索履歴 | DDWin | 要注意語 | 削除語形

モード 標準
出典

Ver.2010-08-02

語彙素ID 36920

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

- 素 36920見る《用》
  - 形 1181441ミル（動詞-非自立可能）
    - 書 302448897みる
    - 書 302448898見る
    - 書 302448899看る
    - 書 302448900視る
    - 書 302448901観る
    - 書 302448902診る
    - 書 302448903覧る
    - 書 302448904瞥る
    - 書 302448905試る
    - 書 302448906?る
    - 書 302448907看護る
    - 音 302448897ミル
  - 形 1181442ミレル（動詞-非自立可能）
    - 書 302449153見れる
    - 書 302449154みれる
    - 書 302449155診れる
    - 書 302449156観れる
    - 音 302449153ミレル
  - 形 1181443ミル（動詞-非自立可能）
    - 書 302449409見る
    - 書 302449410みる
    - 書 302449411看る
    - 書 302449412睹る
    - 書 302449413賭る
    - 書 302449414覩る
    - 書 302449415観る
    - 書 302449416觀る
    - 書 302449417視る
    - 書 302449418覧る
    - 書 302449419瞻る

ツリーを折り畳む | ツリーを展開

語彙素

語彙素ID 36920 | 用例 c2702:r1014:w641:b720:n | 類 用 | 最小単位 ミル/
語彙素読み ミル | コメント
語彙素 見る | 語義 | 語種 和
分類語彙表番号 | 出典 [Macpr | 状態 | 評価 | 選択

語形

語形ID 1181442 | 用例 c15:r2:y28:D14:N2:B1 | 品詞 動詞-非自立可能
語形 ミレル | Check | 活用型 下一段-ラ行 | 一般
語頭変化型: | 語頭変化結合型:
語末変化型: | 語末変化結合型: | コメント
代表性 ら | 出典 M_ | 状態 | 評価 ○ | 選択

書字形

書字形ID 302449156 | 用例 y2:B4:Y126:(133) | 特殊活用形 | 仮名形 ミレル
書字形 観れる | 語彙素 | コメント
代表性 | 出典 y | 状態 | 評価 | 選択

レコード: 1 / 4 | フィルタなし | 検索

発音形

発音形ID 302449153 | 1 | アクセント型 2
発音形 ミレル | アクセント結合型 C1
コメント:
代表性 | 出典 M_ | 状態 | 選択

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

レコード: 1 / 5 | 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
語彙素 36920
見る
ミル

クリア | 削除 | へ | クリア | コピー | 移動

データソース SELECT * FROM 短単位語彙素 WHERE 語彙素読み like N'ミル' ORDER BY 短単位語彙素.語彙素ID

現在の総レコード

UniDic の紹介

- 現代語コーパスのために構築したので、現代語の観点から設計している。
  - lexicon.xml は上代語の観点から設計している。
  - ID 番号の問題がある。

通時辞書を構築する問題

- 通時辞書の目標
  - 共時の使用を調べられること。
  - 通時の使用を調べられること。
- 問題
  - UniDic にない語
    - 多くは奈良時代だけの言葉。
  - 時代により品詞などが変わった語
  - 単純語や複合語
  - lexicon での見出し語一語が UniDic では複数の別語として区別されている。
  - UniDic での見出し語一語が lexicon では複数の別語として区別されている。

通時辞書を構築する問題

- 時代により品詞などが変わった語の例
  - 如し
    - UniDic によると、「如し」は助動詞。
    - 上代語では形容詞であった。
      - 「ごと」、「ごとし」、「ごとく」などの形容詞の活用で発生した。

通時辞書を構築する問題

- 時代により品詞などが変わった語の例
  - 如し

例１：(万葉集　11.2352)

玉如	tama no goto
所照公乎	teras-eru kimi wo

例２：(万葉集　6.961)

湯原尓	yu no para ni
鳴蘆多頭者	naku asitadu pa
如吾	wa ga goto-ku
妹尓戀哉	imo ni kwopure ya

通時辞書を構築する問題

- 時代により品詞などが変わった語の例
  - ながら
    - UniDic によると、接続助詞である。
    - 上代語では接尾語だろう。
      - 名詞
        - kamu-nagara (神ながら)
        - yama-nagara (山ながら)
        - umi-nagara (海ながら)
      - 動詞
        - tuki-nagara (附きながら)
        - tutuke-nagara (続けながら)
        - uresibwi-nagara (嬉しびながら)

通時辞書を構築する問題

- 時代により品詞などが変わった語の問題
  - lexicon は usage という情報を付与。１つの語形に対し、どの時代にはどのように使われたかを入力。
  - UniDic は品詞が違ってしまえば、別語として登録せざるをえない。
  - 解決案としては UniDic にも歴史的変化に関する情報を追加することはできないか。

通時辞書を構築する問題

- 単純語や複合語の問題
  - 「顧みる」、「奉る」
  - 現代語の観点から見れば単純語である。
  - 上代語の観点からは複合語とみるべき可能性もある。
  - 更なる研究が必要。

```
<superEntry xml:id="L030517">
  <entry>
    <form type="stem">
      <orth stage="I">kapyerimi-</orth>
      <orth stage="II">kaperimi-</orth>
      <orth stage="V">kawerimi-</orth>
      <orth stage="VII">kaerimi-</orth>
      <gramGrp>
        <pos>verb</pos>
        <iType type="UM"/>
      </gramGrp>
    </form>
    <re>
      <form type="full">
        <lbl>compound of</lbl>
        <orth>
          <ref target="#L030518">kapyer-</ref>
        </orth>
        <orth>
          <ref target="#L031712">mi-</ref>
        </orth>
      </form>
    </re>
    <sense n="1">
      <def>look back</def>
```

```
<cl>
  <w ana="#L030517a" lemma="5975">
    <w type="verb" inflection="stem" ana="#L030518a" lemma="5979">
      <c type="logo">kapyeri</c>
    </w>
    <w type="verb" inflection="stem" ana="#L031712a" lemma="36920">
      <c type="logo">mi</c>
    </w>
    <m type="auxiliary" function="neg" ana="#L000006" inflection="infconc" lemma="19587">
      <c type="logo">zu</c>
    </m>
  </w>
</cl>
```

通時辞書を構築する問題

- 単純語や複合語の問題の解決案としては
  - UniDic もリンクを追加することはできないか。

通時辞書を構築する問題

- lexicon での見出し語一語が UniDic では複数の別語として区別されている。
- その１
  - 漢字や意味で区別した例

lexicon	UniDic
tomaru	泊まる、止まる
naku	鳴く、泣く
twopu	問う、訪う

  - 本来の意味が不明。
  - 解決案としては UniDic に上代語の見出しを作って、現代語の見出し語のリンクを追加することはできないか。

通時辞書を構築する問題

- lexicon での見出し語一語が UniDic では複数の別語として区別されている。
- その２
  - 語頭の母音の有無による差異。

lexicon	UniDic
ideru, deru	いでる、でる
imasu, masu	います、ます
idasu, dasu	いだす、だす

  - が、 omopu, mopu (思う) は UniDic では１つの見出し語。
  - 上代語は表意文字で書かれた場合は発音が分からない。
  - 解決案としては UniDic にリンクを追加して、複数の見出し語間のつながりがたどれるようにできないか。

通時辞書を構築する問題

- lexicon での見出し語一語が UniDic では複数の別語として区別されている。
- その３
  - 動詞と動詞の名詞形。
  - 区別が容易にできる。
  - 解決案としては
    - UniDic にリンクを追加する。
    - lexicon の ID 番号を変更する。
      - 動詞： ID 番号 +a
      - 動詞の名詞形： ID 番号 +b

```
<superEntry xml:id="L030465-main">
  <entry xml:id="L030465a" corresp="6509">
    <form type="stem">
      <orth stage="I">kasum-</orth>
      <gramGrp>
        <pos>verb</pos>
        <iType type="QD"/>
      </gramGrp>
    </form>
    <sense n="1">
      <def>be misty</def>
    </sense>
  </entry>
  <entry xml:id="L030465b" corresp="6506">
    <form type="noun">
      <orth stage="I">kasumi</orth>
      <gramGrp>
        <pos>noun</pos>
      </gramGrp>
    </form>
  </entry>
</superEntry>
```

通時辞書を構築する問題

- UniDic での見出し語一語が lexicon では複数の別語として区別されている。
- その１
  - 多くの例は上代語は自動詞と他動詞があって、現代語は自他の区別しない。

自動詞	他動詞
分く（四段）	分く（下二段）
染む（四段）	染む（下二段） *
惑ふ（四段）	惑ふ（四段）
垂る（四段）	垂る（下二段）

* 現代語で「染みる」もある。

通時辞書を構築する問題

- UniDic での見出し語一語が lexicon では複数の別語として区別されている。
- その２
  - 活用形で区別できる例
    - 与ふ（四段）　　与ふ（下二段）
    - 悲しむ（四段）　　悲しぶ（上二段）
    - 忘る（四段）　　忘る（下二段）
    - 頼む（四段）　　頼む（下二段）

通時辞書を構築する問題

- UniDic での見出し語一語が lexicon では複数の別語として区別されている。
- その３
  - **発音で区別できる例。**
    - 「問ふ」と「妻問ふ」
    - 表意文字で書かれた時は「問」を使った例はある
    - 表音文字の場合は「問ふ」は twopu （甲類の「と」）で、「妻問ふ」 tumadopu（乙類の「と」）であった。だから、別語である。

通時辞書を構築する問題

- UniDic での見出し語一語が lexicon では複数の別語として区別されている。
- その３
  - 発音で区別できる例。
    - 「ゆく」と「いく」
    - ゆく：用例＝ 668 ；表意文字＝ 469 ；表音文字：199
    - いく：用例＝ 9 ；表意文字＝ 2 ；表音文字： 7
      - 複合語の例は１つだけ。
      - kwoye-ika-mu (MYS 15.3722)

通時辞書を構築する問題

- UniDic での見出し語一語が lexicon では複数の別語として区別されている。
- その４
  - 意味で区別できる例。
    - かれる：　「枯れる」
      「涸れる」
    - しく：「（波、花など）敷く／頻く」（自動詞）
      「（髪、着物）敷く」（他動詞）
      「敷く」（治めるという意味）

通時辞書を構築する問題

- UniDic での見出し語一語が lexicon では複数の別語として区別されている。
- つまり
  - lexicon で複数の見出し語として区別されているものが Unidic では 1 つの見出し語になっている場合。
  - 上代語では活用形の違いがあり、それが自他の区別など、意味の違いに対応していることが多い。現代語ではそのうちの一方のみしか使われていない場合がある。

通時辞書を構築する問題

- UniDic での見出し語一語が lexicon では複数の別語として区別されている。
- UniDic での登録状態
  - 現代語から出発した UniDic はこれらの活用形、自他の差異について、登録にばらつきがある。

上代語		現代語
piraku （四段）	→	開く
piraku （下二段）	→	開ける
tanomu （四段）	→	頼む／頼める
tanomu （下二段）	→	

UniDic Explorer

ヒラク 検索 検索履歴 DDWin 要注意語 削除語形 モード 標準 出典 語彙素ID 31814 Ver.2010-

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

- 素 31814開く《用》
  - 形 1018049ヒラク（動詞-一般）
    - 書 260620545ひらく
    - 書 260620546拓く
    - 書 260620547開く
    - 書 260620548啓く
    - 書 260620549披く
    - 書 260620550開らく
    - 書 260620551排く
    - 書 260620552展く
    - 書 260620553ヒラく
    - 書 260620554墾く
    - 音 260620545ヒラク
  - 形 1018050ヒラケル（動詞-一般）
    - 書 260620801開る
    - 書 260620802ひらける
    - 書 260620803開ける
    - 書 260620804拓ける
    - 書 260620805開開る
    - 音 260620801ヒラケル
  - 形 1018051ヒラク（動詞-一般）
    - 書 260621057開く
    - 書 260621058拓く
    - 書 260621059ひらく
    - 書 260621060啓く
    - 書 260621061闡く
    - 書 260621062開く
    - 書 260621063披く
    - 書 260621064展く
    - 書 260621065ヒラク
    - 音 260621057ヒラク
  - 形 1018052ヒイラク（動詞-一般）

ツリーを折り畳む ツリーを展開

語彙素

語彙素ID 31814 用例 c56:r437:w9:b81:n176:y15 類 用 最小単位 ヒラク/

語彙素読み ヒラク コメント

語彙素 開く 語義 語種 和

分類語彙表番号 出典 Iacr 状態 評価 選択

語形

語形ID 1018049 用例 c56:r437:w9:b81:n175 品詞 動詞-一般

語形 ヒラク Check 活用型 五段-カ行 一般

語頭変化型: 語頭変化結合型: 語末変化型: 語末変化結合型: コメント

代表性 代 出典 Iacr 状態 評価 選択

書字形

書字形ID 260620554 用例 B1:*(1) 特殊活用形 仮名形 ヒラク

書字形 墾く 語彙素 コメント

代表性 出典 b 状態 評価 選択

レコード: 1 / 10 フィルタなし 検索

発音形

発音形ID 260620545 1 アクセント型 2

発音形 ヒラク アクセント結合型 C1

コメント:

代表性 出典 Iacr 状態 選択

レコード: 1 / 1 フィルタなし 検索

レコード: 1 / 4 フィルタなし 検索

レコード: 1 / 2 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 クリア 削除 へ クリア コピー 移動

データソース SELECT * FROM 短単位語彙素 WHERE 語彙素読み like N'ヒラク' ORDER BY 短単位語彙素.語彙素ID 現在の総レコード

UniDic Explorer

ヒラケル 検索 検索履歴 DDWin 要注意語 削除語形 モード 標準 出典 語彙素ID 31815 Ver.2010-

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

- 素 31815開ける《用》
  - 形 1018081ヒラケル（動詞-一般）
    - 書 260628737ひらける
    - 書 260628738開ける
    - 書 260628739開る
    - 書 260628740展ける
    - 書 260628741開らける
    - 書 260628742拓ける
    - 音 260628737ヒラケル
  - 形 1018082ヒラク（動詞-一般）
    - 書 260628993ひらく
    - 書 260628994開く
    - 書 260628995拓く
    - 書 260628996啓く
    - 書 260628997展く
    - 書 260628998啓らく
    - 音 260628993ヒラク

ツリーを折り畳む ツリーを展開

語彙素

語彙素ID 31815 用例 c4:r8:b5:n3:y1:z5:j1:D2:B2 類 用 最小単位 ヒラケル/

語彙素読み ヒラケル コメント

語彙素 開ける 語義 語種 和

分類語彙表番号 出典 IMUcr 状態 評価 選択

語形

語形ID 1018082 用例 V23:近36:古13:(391) 品詞 動詞-一般

語形 ヒラク Check 活用型 文語下二段-カ行

語頭変化型: 語頭変化結合型: 語末変化型: 語末変化結合型: コメント

代表性 出典 活 状態 評価 選択

書字形

書字形ID 260628997 用例 V6:(7) 特殊活用形 仮名形 ヒラク

書字形 展く 語彙素 コメント

代表性 出典 近 状態 k 評価 選択

レコード: 1 / 6 フィルタなし 検索

発音形

発音形ID 260628993 1 アクセント型 2

発音形 ヒラク アクセント結合型 C1

コメント:

代表性 出典 活 状態 選択

レコード: 1 / 1 フィルタなし 検索

レコード: 1 / 2 フィルタなし 検索

レコード: 1 / 1 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 クリア 削除 へ クリア コピー 移動

データソース SELECT * FROM 短単位語彙素 WHERE 語彙素読み like N'ヒラケル' ORDER BY 短単位語彙素.語彙素ID 現在の総レコード

UniDic Explorer

タノム 検索 検索履歴 DDWin 要注意語 削除語形 モード 標準 出典 語彙素ID 22580

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

UniDicツリー

- 素 22580頼む《用》
  - 形 722561タノム（動詞-一般）
    - 書 184975617たのむ
    - 書 184975618恃む
    - 書 184975619頼む
    - 書 184975620嗔む
    - 音 184975617タノム
  - 形 722562タノメル（動詞-一般）
    - 書 184975873たのめる
    - 書 184975874頼める
    - 書 184975875恃める
    - 音 184975873タノメル
  - 形 722563タノム（動詞-一般）
    - 書 184976129頼む
    - 書 184976130たのむ
    - 書 184976131恃む
    - 書 184976132頼む
    - 音 184976129タノム
  - 形 722564タノム（動詞-一般）
    - 書 184976385頼む
    - 書 184976386たのむ
    - 音 184976385タノム

語彙素

語彙素ID 22580 用例 c61:r26:b46:n9:y11:z13:j2: 類 用 最小単位 タノム/

語彙素読み タノム コメント

語彙素 頼む 語義 語種 和

分類語彙表番号 出典 IUcpr 状態 評価

語形

語形ID 722564 用例 古55:(58) 品詞 動詞-一般

語形 タノム Check 活用型 文語下二段-マ行

語頭変化型: 語頭変化結合型: コメント

語末変化型: 語末変化結合型:

代表性 出典 古 状態 o 評価

書字形

書字形ID 184976385 用例 古46:(47) 特殊活用形 仮名形 タノム

書字形 頼む 語彙素 コメント

代表性 代 出典 古 状態 o 評価

レコード: 1 / 2 フィルタなし 検索

発音形

発音形ID 184976385 1 アクセント型 2

発音形 タノム アクセント結合型 C1

コメント:

代表性 出典 古 状態 o

レコード: 1 / 1 フィルタなし 検索

レコード: 4 / 4 フィルタなし 検索

レコード: 1 / 1 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

語形 722564

タノム

クリア へ クリア コピー

通時辞書を構築する問題

- UniDic での見出し語一語が lexicon では複数の別語として区別されている。
- lexicon の ID を UniDic の語形（見出し語の下位）と対応させるのか
  - できれば見出し語（語彙素）と対応させたい。
  - 理由
    - 現代でも自他の区別などがついているものについては lexicon の見出しと UniDic の見出し（語彙素）が対応している。
    - UniDic の語形は通時的な研究にはやや狭すぎる。

通時辞書を構築する問題

- UniDic での見出し語一語が lexicon では複数の別語として区別されている。
- 解決案として、現代語では使わない形でも、語彙素レベルで見出しを立てて 統一できないか。

まとめ

- 通時的な研究者にとっては、言葉の通時的な使用を調査が容易にできると研究も容易に進む。
- これからはオックスフォード大学と国語研が研究協力して、研究者のために通時的な辞書を構築すると役に立つだろう。

Research Centre for Japanese Language and Linguistics
University of Oxford

オックスフォード大学 日本語研究センター
www.orinst.ox.ac.uk/research/jap-ling/

ご清聴ありがとうございました

Kerri L Russell
kerri.russell@orinst.ox.ac.uk