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/

The Oxford Corpus of Old Japanese

Kerri L Russell
kerri.russell@orinst.ox.ac.uk
HiCor Network meeting, 15 May 2013

Outline

- The Oxford Corpus of Old Japanese (OCOJ)
- Markup of linguistic information
- Markup of historical information
- The Lexicon

The OCOJ

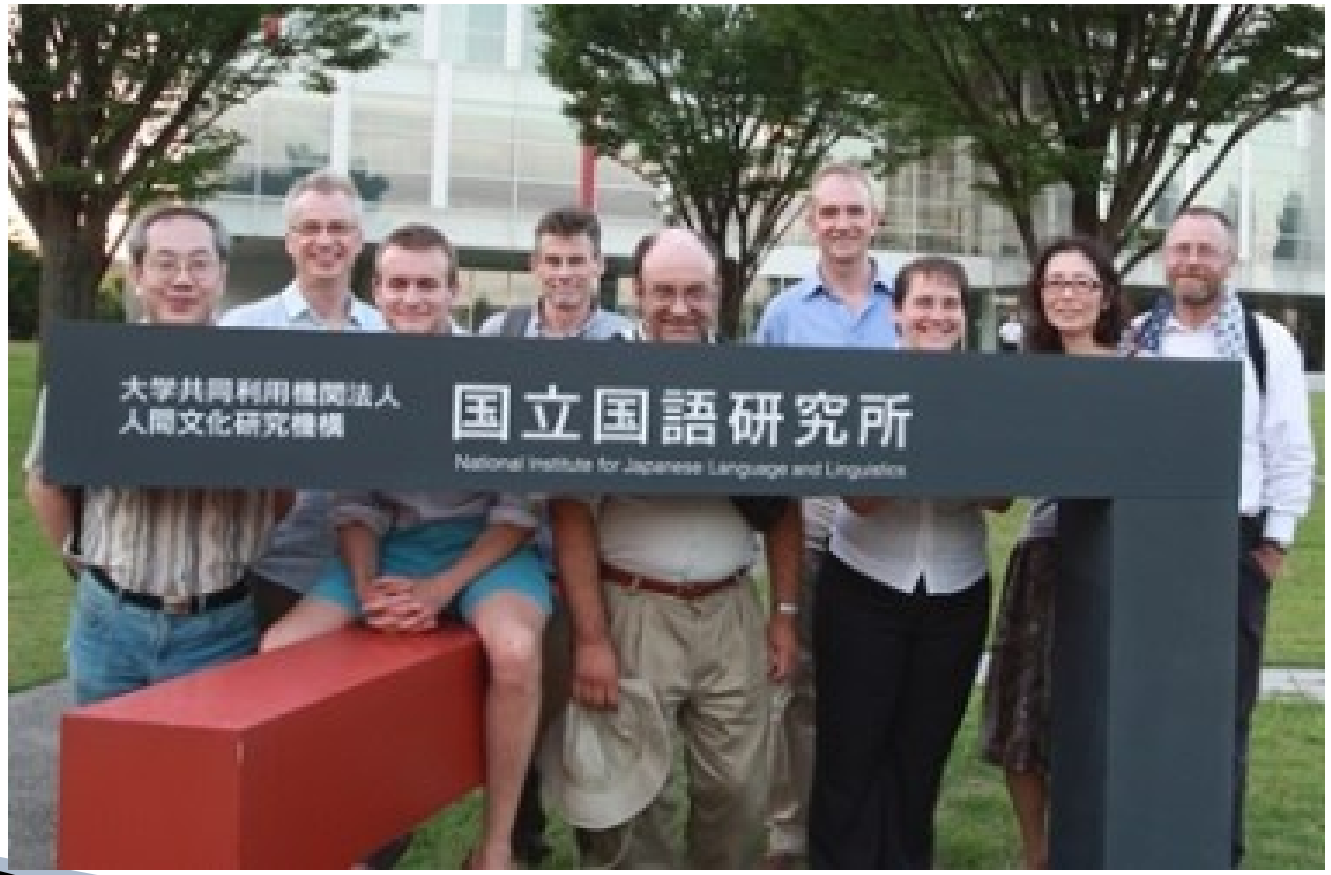
- The Oxford Corpus of Old Japanese (abbreviated OCOJ) is a long-term research project which aims to develop a comprehensive annotated digital corpus of all extant texts in Japanese from the Old Japanese (OJ) period.
- OJ is the earliest attested stage of the Japanese language, largely the Japanese language of the Asuka and Nara periods of Japanese history (7th and 8th century CE).
- This is the formative literate period upon which the development of Japanese civilization is based, and these texts are of paramount importance for the study and understanding of the origins and development of civilization of Japan, including language, writing, literature, religion, history, and culture.

The OCOJ

- Funding:

The OCOJ

- People:



The OCOJ

- More information can be found on the OCOJ webpage: http://vsarpj.orinst.ox.ac.uk/corpus/
  - A fully romanized version of all OJ texts
  - Markup and display conventions

Markup of linguistic information

- The OCOJ is marked up with XML tags following the guidelines of the Text Encoding Initiative.
- XML tags:
  - <s>…</s> for sentences
  - <cl>…</cl> for clauses
  - <phr>…</phr> for phrases
  - <w>…</w> for word-like things
  - <m>…</m> for morphemes
  - <c>…</c> for characters and orthography

Markup of linguistic information

- A poem (MYS.8.1606)

君待跡
吾戀居者
我屋戸乃
簾令動
秋之風吹

Markup of linguistic information

- A poem (MYS.8.1606)

君待跡	kimi matu *to*
吾戀居者	wa ga kwopwi-woreba
我屋戸乃	wa ga yadwo *no*
簾令動	sudare ugokasi
秋之風吹	aki no kaze puku

Markup of linguistic information

- Encoding information about how a word was written: logographically, phonographically, or not written in text.
  - <c type="logo">**kimi**</c>
  - <c type="logo">**matu**</c>
  - <c type="phon">**to**</c>

  - <c type="logo">**wa**</c>
  - <c type="noLogo">**ga**</c>
  - <c type="logo">**kwopwi**</c>
  - <c type="logo">**woreba**</c>

Markup of linguistic information

- Part of speech, lexeme, and morpheme identification
  - <w lemma="L004266">**kimi**</w>
  - <w type="verb" inflection="adnconc" lemma="L031644a">**matu** </w>
  - <w type="particle" subtype="conj" lemma="L000531a">**to**</w>

```
<cl>
   <cl>
      <phr type="arg">
         <w lemma="L004266">
            <c type="logo">kimi</c>
         </w>
      </phr>
      <w type="verb" inflection="adnconc" lemma="L031644a" lemmaRef="35830">
         <c type="logo">matu</c>
      </w>
      <w type="particle" subtype="conj" lemma="L000531a">
         <c type="phon">to</c>
      </w>
   </cl>
   <lb xml:id="MYS.8.1606-trans_1" corresp="#MYS.8.1606-orig_1"/>
   <phr type="arg">
      <w lemma="L042057" lemmaRef="41100">
         <c type="logo">wa</c>
      </w>
      <w type="particle" subtype="case" function="gen" lemma="L000503" lemmaRef="7889">
         <c type="noLogo">ga</c>
      </w>
   </phr>
   <w>
      <w type="verb" inflection="stem" lemma="L030731a" lemmaRef="52566">
         <c type="logo">kwopwi</c>
      </w>
      <w type="verb" inflection="provisional" function="progressive" lemma="L031957a" lemmaRef="5360">
         <c type="logo">woreba</c>
      </w>
   </w>
</cl>
```

Markup of historical information

```
  <sourceDesc>
    <xi:include href="texts/Manyoshu.xml" parse="xml" xmlns:xi="http://www.w3.org/2001/XInclude"/>
  </sourceDesc>
</fileDesc>
<profileDesc>
  <creation>
    <date/>
  </creation>
  <langUsage>
    <language ident="ojp">Old Japanese</language>
  </langUsage>
  <textClass>
    <catRef target="#manyoshu"/>
  </textClass>
```

```
<phr type="arg">
   <cl ana="mk1">
      <phr type="arg">
         <w>
            <w type="adjective" inflection="stem">
               <c type="phon">awo</c>
            </w>
            <w>
               <c type="phon">ni</c>
            </w>
         </w>
      </phr>
      <w>
         <w type="adjective">
            <c type="phon">yo</c>
         </w>
         <m type="adjcop" lemma="L000033" inflection="conclusive">
            <c type="phon">si</c>
         </m>
      </w>
   </cl>
   <lb xml:id="KK.58-trans_5" corresp="#KK.58-orig_5"/>
   <w lemma="pln1">
      <c type="phon">nara</c>
   </w>
   <w type="particle" subtype="case" function="acc" lemma="L000534" lemmaRef="41407">
      <c type="phon">wo</c>
   </w>
</phr>
```

The Lexicon

- The lexicon has two main functions:
  - It stores information about items in the corpus.
  - It allows us to retrieve information.

```
<superEntry xml:id="L031840-main">
    <entry xml:id="L031840a">
        <form type="stem">
            <orth stage="I">yuk-</orth>
            <gramGrp>
                <pos>verb</pos>
                <iType type="QD"/>
                <gram type="class">motion</gram>
            </gramGrp>
        </form>
        <sense n="1">
            <def>go</def>
        </sense>
        <re type="related">
            <form>
                <orth>
                    <ref target="L030145a">ik-</ref>
                </orth>
            </form>
        </re>
    </entry>
    <entry xml:id="L031840b">
        <form type="noun">
            <orth stage="I">yuki</orth>
            <gramGrp>
                <pos>noun</pos>
            </gramGrp>
```

The Lexicon

- Place names:

```
<superEntry xml:id="pln1-main">
    <entry xml:id="pln1">
        <form>
            <orth stage="I">nara</orth>
        </form>
        <sense n="1">
            <def>Nara</def>
        </sense>
    </entry>
</superEntry>
```

```
<superEntry xml:id="per1-main">
    <entry xml:id="per1a">
        <form>
            <orth stage="I">tenno kooken</orth>
        </form>
        <sense n="1">
            <def>Empress Kooken</def>
        </sense>
    </entry>
    <entry xml:id="per1b">
        <form>
            <orth stage="I">daijoutenno kooken</orth>
        </form>
        <sense n="1">
            <def>Retired Empress Kooken</def>
        </sense>
    </entry>
    <entry xml:id="per1c">
        <form>
            <orth stage="I">tenno syootoku</orth>
        </form>
        <sense n="1">
            <def>Empress Syootoku</def>
        </sense>
    </entry>
</superEntry>
```

User-friendly views

- Lexicon entry

yuk-
L031840a

Part of speech: **verb**
Conjugation class: **quadrigrade**
Verb classification: motion

1 Gloss: **go**

Related forms:
See also: ik-'go'

statistics attestations clause structure

yuki
L031840b

Part of speech: **noun**

statistics attestations

User-friendly views

- Plain text view

MYS.8.1606 gloss tree

君待跡	kimi matu *to*
吾戀居者	wa ga kwopwi-woreba
我屋戸乃	wa ga yadwo *no*
簾令動	sudare ugokasi
秋之風吹	aki no kaze puku

User-friendly views

- Glossed view

« { { { [kimi (L004266 lord)] matu (verb adnconc L031644a 35830 wait) *to* ( ) }

[wa (L042057 41100 1st person pronoun) ga (L000503 [genitive case particle])] kwopwi (verb stem L030731a 52566 love) -woreba (verb provisional progressive L031957a 5360 be sitting) }

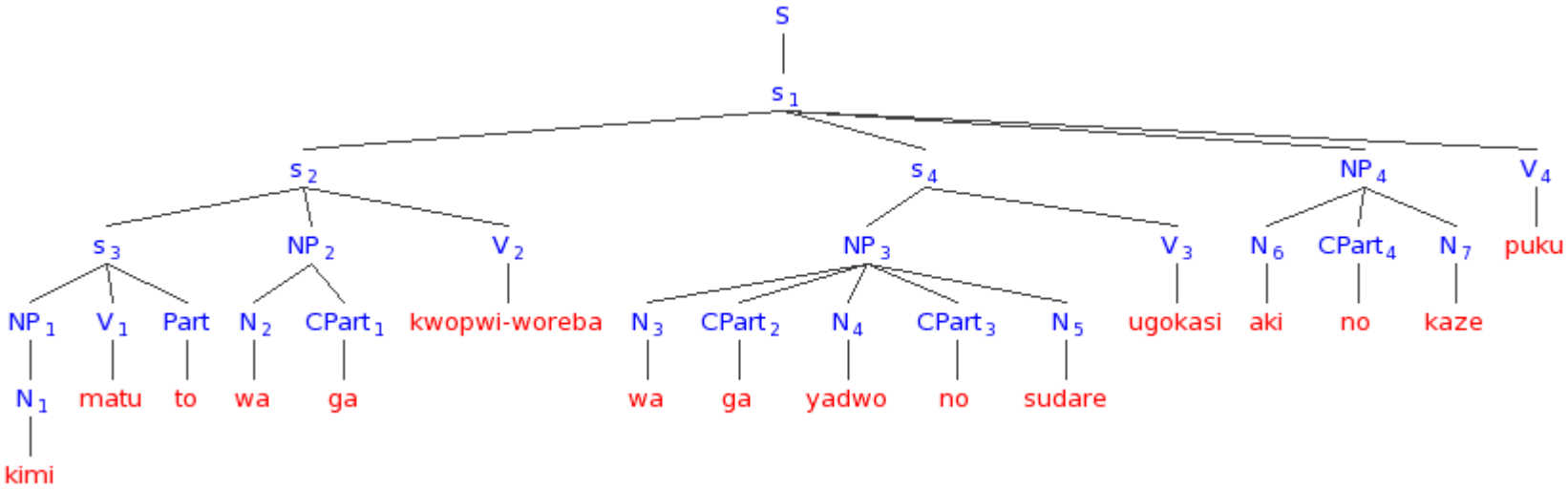
{ [wa (L042057 41100 1st person pronoun) ga (L000503 [genitive case particle]) yadwo *no* (L000520 [genitive case particle])

sudare] ugokasi (verb infinitive L030247a 3094 move t) }

[aki no (L000520 [genitive case particle]) kaze] puku (verb adnconc L031516a 32591 blow) } »

The Lexicon

- Tree view



Questions and Comments Welcome

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