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

Bjarke Frellesvig, Stephen Wright Horn, and Kerri L Russell
vsarpj@orinst.ox.ac.uk
HiCorp Workshop and 1 March 2013

OCOJ initially developed as part of the AHRC funded research project

Verb semantics and argument realization in pre-modern Japanese

1 January 2009 – 30 June 2014
appr. £1 million

url: http://vsarpj.orinst.ox.ac.uk

OCOJ

recognized as a

British Academy Research Project

from April 2012

OCOJ people and collaborating institutions

People

- Bjarke Frellesvig (Oxford)
- Peter Sells (University of York)
- Stephen Wright Horn (Oxford)
- Kerri L Russell (Oxford)

Collaborating institutions

- National Institute for Japanese Language and Linguistics, Tokyo
- University of York

Man'yōshū (万葉集) 3.341

賢跡
物言従者
酒飲而
酔哭為師
益有良之

Man'yōshū (万葉集) 3.341

賢跡	sakasi-mi*to*
物言従者	mono-ipu ywori pa
酒飲而	sake nomite
酔哭為師	wepi-naki suru *si*
益有良之	masari-taru *rasi*

"It should be better to drink saké and weep drunkenly than talking in a clever fashion"

Oxford Corpus of Old Japanese

Currently comprises all poetic texts from the Old Japanese period

approximately 90,000 words

addition of ‘prose’ texts from the period is under way

url: http://vsarpj.orinst.ox.ac.uk/corpus/

Texts in the OCOJ

Poetic texts, currently in the OCOJ

Kojiki kayō (古事記歌謡; 712)	(112 poems; 2527 words)
Nihon shoki kayō (日本書紀歌謡; 720)	(133 poems; 2444 words)
Fudoki kayō (風土記歌謡; 730s)	(20 poems; 271 words)
Bussukoseki-ka (仏足石歌; after 753)	(21 poems; 337 words)
Man'yōshū (万葉集; after 759)	(4685 poems; 83706 words)
Shoku nihongi kayō (続日本紀歌謡; 797)	(8 poems; 134 words)
Jōgū shōtoku hōō teisetsu (上宮聖徳法王帝説)	(4 poems; 60 words)

‘Prose texts’, under way

Shoku nihongi Senmyō (続日本紀宣命; 797)	(approx. 14,000 words)
Engishiki Norito (延喜式祝詞)	(approx. 6,500 words)
Kojiki (古事記; 712)	???

OCOJ

Original text

Phonemic transcription

Online digital text

website: http://vsarpj.orinst.ox.ac.uk/corpus/

Annotation

Home

Online corpus

Pre-modern Japanese syntax research project

Research Centre for Japanese Language and Linguistics

Contact

The Oxford Corpus of Old Japanese

This is a brief introduction, under continuing development, of the Oxford Corpus of Old Japanese. Old Japanese is mainly the language of the 8th century. The Oxford Corpus of Old Japanese is being constructed as part of the Verb semantics and argument realization in pre-modern Japanese project.

The corpus consists of Old Japanese texts which are supplied with a range of information in the form of xml tags, following the conventions of the Text Encoding Initiative. The texts are romanized in a phonemic transcription and supplied with information about the original orthography; original script is also retained. Lexemes and morphemes are given unique identifiers and words are part-of-speech tagged. Inflecting words are supplied with information about morphology. Finally, information about syntactic constituency is encoded.

We are posting a very simple online version of the full Old Japanese poetic corpus which

古之	inisipye no
七賢	nana no sakasi-ki
人等毛	pito-domo *mo*
欲為物者	pori se-si mono pa
酒西有良師	sake *ni si* aru *rasi*

MYS.3.341

賢跡	sakasi-mi*to*
物言従者	mono-ipu ywori pa
酒飲而	sake nomite
酔哭為師	wepi-naki suru *si*
益有良之	masari-taru *rasi*

MYS.3.342

将言為便	ipa-mu subye
将為便不知	se-mu subye sira-zu
極	kipamarite
貴物者	taputwo-ki mono pa
酒西有良之	sake *ni si* aru *rasi*

OCOJ Annotation

XML mark-up following the internationally recognized standards of the **T**ext **E**ncoding **I**nitiative (TEI)

Manual mark-up

xml tag set in the OCOJ

<c> … </c>	character
<m> … </m>	morpheme
<w> … </w>	word
<phr> … </phr>	phrase
<cl> … </cl>	clause
<s> … </s>	sentence

OCOJ Annotation

Writing 原文の表記法

xml tags: <c>

Part-of-speech 品詞

Lexeme and morpheme UID 語彙素と形態素のID番号

Morphology, inflection 活用

xml tags: <w> <m>

Syntax シンタクス (要素•構造)

xml tags: <s> <cl> <phr>

語彙素と形態素の情報
Word and Morpheme Information

xml tags: <w>, <m>

Part-of-speech 品詞

Lexeme and morpheme UID [Unique Identifier]
語彙素と形態素のID番号

Morphology, inflection 活用

Syntax シンタクス (要素・構造)

xml tags:

<s> Sentence 文

<cl> Clause 節

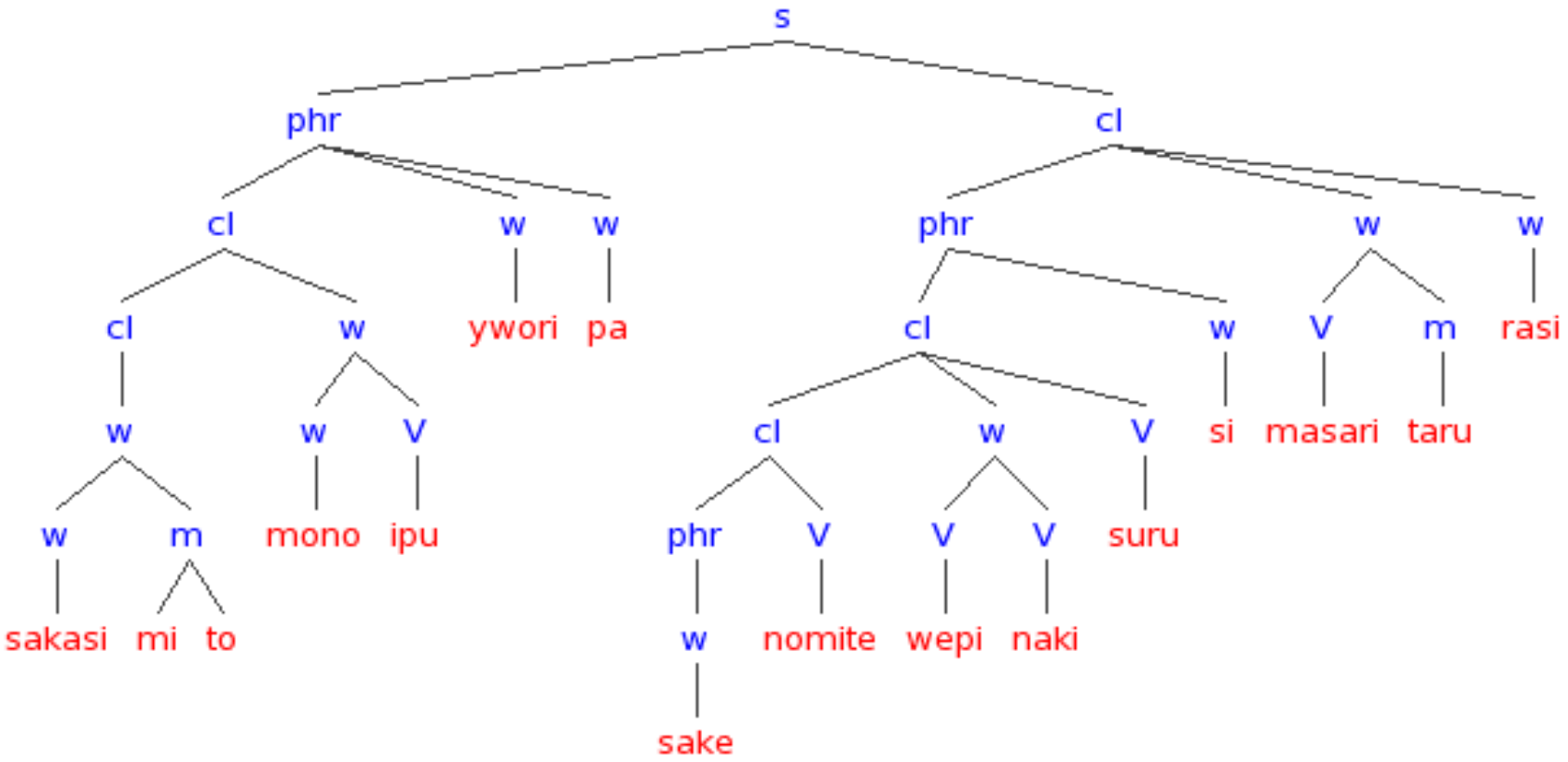
<phr> Noun phrase 名詞句
argument (項)

```
<s>
  <phr>
    <cl>
      <cl>
        <w>
          <w type="adjective">
            <c type="logo">sakasi</c>
          </w>
          <m inflection="gerund" lemma="L000034" lemmaRef="36226" type="adjcop">
            <c type="logo">mi</c>
            <c type="phon">to</c>
          </m>
        </w>
      </cl>
      <lb corresp="#MYS.3.341-orig_1" xml:id="MYS.3.341-trans_1"/>
      <w lemma="L031771a">
        <w lemma="L050042">
          <c type="logo">mono</c>
        </w>
        <w inflection="adnconc" lemma="L030199a" lemmaRef="1571" type="verb">
          <c type="logo">ipu</c>
        </w>
      </w>
    </cl>
    <w function="abl" lemma="L000540" subtype="case" type="particle">
      <c type="logo">ywori</c>
    </w>
    <w lemma="L000522" lemmaRef="29321" subtype="top" type="particle">
      <c type="logo">pa</c>
    </w>
  </phr>
  <lb corresp="#MYS.3.341-orig_2" xml:id="MYS.3.341-trans_2"/>
```

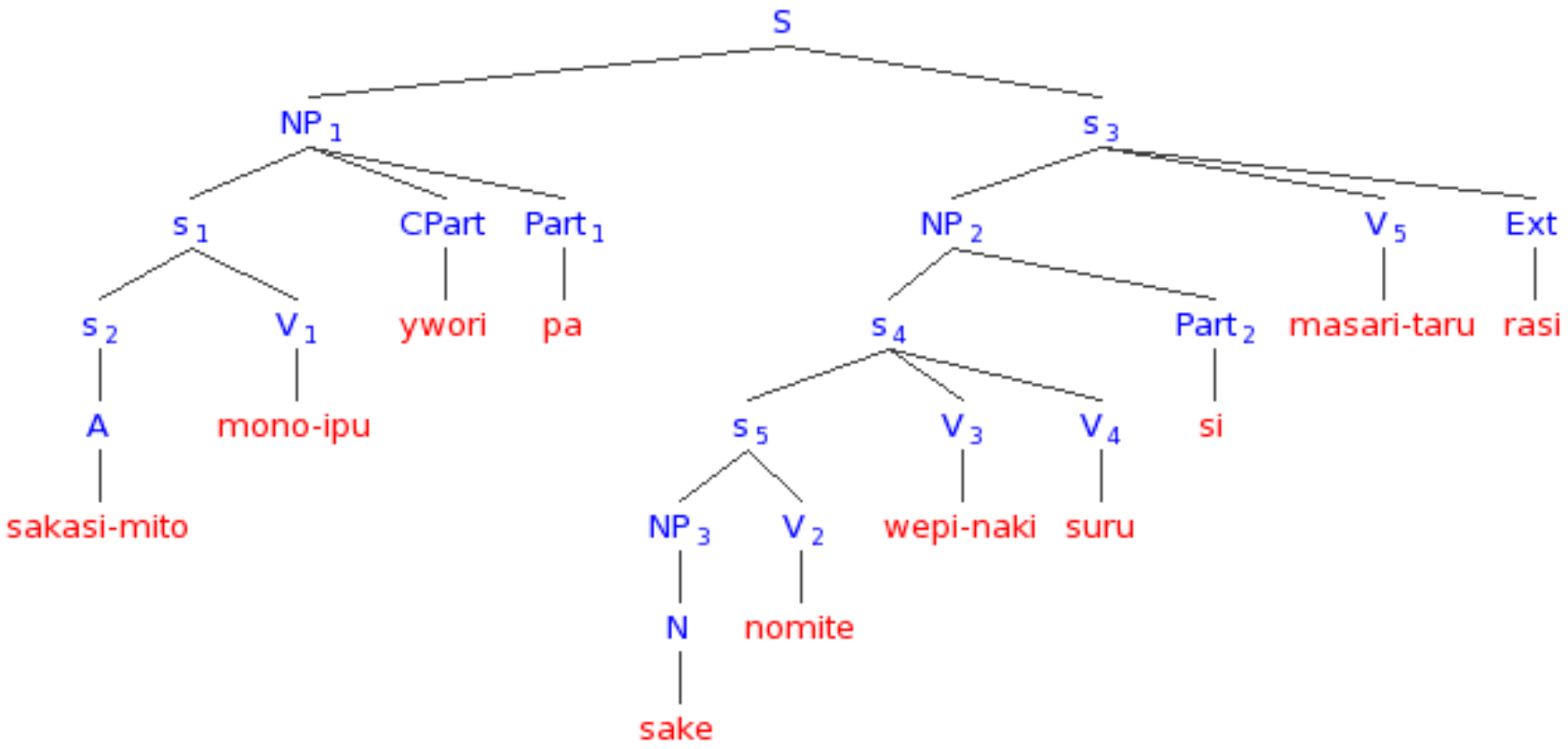
Man'yōshū (万葉集) 3.341

sakasi$_{(adjective)}$-mi*to*$_{(adjcop, gerund)}$
mono$_{(noun)}$-ipu$_{(verb, adnconc)}$ ywori$_{(abl\ case\ particle\ )}$ pa$_{(top\ particle\ )}$
sake$_{(noun)}$ nomite$_{(verb, gerund)}$
wepi$_{(verb, stem)}$-naki$_{(verb, infinitive)}$ suru$_{(verb, adnominal)}$ *si*$_{(res\ particle\ )}$
masari$_{(verb, stem)}$-taru$_{(auxiliary, stative, adnominal)}$ *rasi*$_{(extension, conclusive)}$

Man'yōshū (万葉集) 3.341



Man'yōshū (万葉集) 3.341



検索
Searching

検索　Searching

Lemmatization allows us to search the corpus for words regardless of their inflected form by doing a simple KWIC (KeyWord In Context) search:

nomu	drink.conclusive	‘I drink’
nomite	drink.gerund	‘I am drinking and’
noma-nu	drink-negative.adnominal	‘I who am not drinking’

Lexical Item: lemma search L031386a

Orthography: phonographically attested / logographically attested / all[default]

Source: cOJ / EOJ / all[default]

Constituent boundary: sentence / clause / phrase

search

Lexical Item Information

Search lemma	L031386a
Lexical class	verb
Stem form	*nom-*
Inflection type	QD
Definition	drink

Search Result Information

Source	all
Orthography	all
Number of attestations	22
Constituent boundaries	sentence
	clause
	phrase

italic phonologically attested
plain logographically attested
bold verb
♣ link to tree images
«,» sentence boundary
{,} clause boundary
more

[Updated: 2012-Dec-0

No.	Source	Before	Search lemma	After
1	MYS.3.338♣	i] na-ki} mono *wo*] **omopa**-zupa} [{[pito-tuki] *no*} {**nigor**-eru} sake *wo*]	nomu	be-ku **aru** *rasi*}»
2	MYS.3.341♣	«[{{sakasi-mi*to*} mono-**ipu**} ywori pa] {{{{sake]	nomite	}**wepi-naki suru**} *si*] **masari**-taru *rasi*}»
3	MYS.3.344♣	«{ [*ana*] miniku}»«{{{{{sakasi-*ra wo*] **su** *to*} [sake]	noma	-nu} pito *wo*] {yo-ku} **mireba**} } [saru *ni ka mo*] {**ni**-mu}»
4	MYS.3.346♣	«{{{{{{yworu] **pikaru**} tama] Ø *to*} **ipu** *tomo*} {[sake]	nomite	} [{{kokoro *wo*] **yaru**} *ni*] [ani] ***sika***-*me ya mo*}»
5	MYS.3.350♣	«[{{{moda] **worite**} [sakasi-*ra*] **suru**} pa] {{{{sake]	nomite	}**wepi-naki suru**} *ni*] [napo] **sika**-zu-*kyeri*}»
6	MYS.4.555♣	«{{{{kimi ga tame] **kami**-*si*} mati-zake] {*yasu* no nwo *ni*] [pito-ri ya]	noma	-mu} {{{tomo] na-si} *nisite*}»
7	MYS.5.821♣	«[{{{{{{*awo-yanagwi*] {*ume to*]]*no pana wo*] ***wori*** ***kazasi***}	nomite	}*no noti pa*] {{***tiri***-*nu tomo*} *yo-si*}»
8	MYS.5.833♣	*ru no*] ***ki***-*taraba*} {*kaku si koso*] {{*ume wo*] ***kazasite***} {*tanwosi-ku*}	noma	-*me*}»
9	MYS.6.973♣	**ipu**}»«{{uti-**nade** *so*} ***negwi*-tamapu**}»«{{{{**kapyeri-ko**-mu} pi] a	noma	-mu} ki] Ø *so*} [ko no toyo-mi-ki pa]»
10	MYS.6.995♣	«{{{kaku] **situtu**} **aswobi**	nomi	**-koso**}»«{ [kusa-kwi sura] {{paru pa] {**opwi*tutu***}}} [aki pa] {**tiri-yuk**
11	MYS.7.1128♣	[{*asibi*] *nasu*} **sakaye**-*si*} kimi ga] **pori**-*si*} wi no ipa-*wi* no midu pa] {{	nomedo	}**aka**-nu *kamo*}»
12	MYS.7.1142♣	} saki-*ku*} yo-kye-mu to} [{**ipabasiru**} tarumi no midu *wo*] **musubite**}	nomi	-*tu*}»
13	MYS.7.1295♣	ı} *mikasa no* yama *ni*] {{tukwi] no} pune] **idu**}»«{{{miyabwi-wo no]	nomu	} sakaduki *ni*] {{kage] *ni*} **mi**-yetutu}»
14	MYS.8.1656♣	«[{{{saka-duki *ni*] [ume no pana] **ukabe**} [{**omopu**} doti]	nomite	}no noti pa] {{**tiri**-*nu tomo*} *yo-si*}»
15	MYS.8.1657♣	«[tukasa *ni mo*] {**yurusi-tamap**-yeri}»«{{{[ko-yopi nomwi]	noma	-mu} sake] Ø *kamo*}»«{**tiri-*kosu*** *na* [*yume*] }»
16	MYS.12.2925♣	«{{{{[midori-kwo no tame koso] } [*omo pa*] {**motomu** to} } **ipe**} [ti]	nome	ya}»«{ [kimi ga] [*omo*] **motomu** *ramu*}»
17	MYS.13.3260♣	ı-na-ku *so*} } [pito pa] {**kumu** to} } **ipu**}»«{{{{toki*zi-ku so*} } [pito pa] {	nomu	to} } **ipu**}»«{{{{{**kumu**} pito] no} ma-na-ki ga} goto} {{{{{**nomu**} pito

検索　Searching

Advanced searches
(combining search criteria)

例：all clauses (節)

whose predicate is in the conclusive form (終止形)

which contain a noun phrase (NP) that contains

a WH-word (疑問詞　有) , but has

no focus particle (係助詞　無)

必要な情報
Necessary information

Writing 表記法

Part-of-speech 品詞

Morphology, inflection 活用

Clause　節

Noun phrase 名詞句

UID (Lemmatization) and Lexicon

Another important function of the UIDs is to provide a link between the texts and the Lexicon associated with the corpus.

The Lexicon contains basic lexical information about each lexeme and morpheme, including POS, meaning or function, etc.

The Lexicon also contains information about the occurrence in the corpus of each lexeme and morpheme

Lexicon

nom-
L031386a

Part of speech: **verb**
Conjugation class: **quadrigrade**

1 Gloss: **drink**

statistics attestations clause structure

FURTHER INFORMATION to be added to the OCOJ,

in particular as part of the British Academy Research Project

INFORMATION to be added to the OCOJ

- The annotation will be enriched to include amongst other information such as the following, both ***in*** and ***about*** the texts
  - literary
  - historical
  - geographical
  - and other information.

- The texts will be supplied with translations into English.

DICTIONARY
辞書

The dictionary associated with the OCOJ will be expanded to constitute a real bilingual

Old Japanese -- *English dictionary*

The dictionary will be hyperlinked to the texts, making cross-reference in both directions possible.

THANK YOU