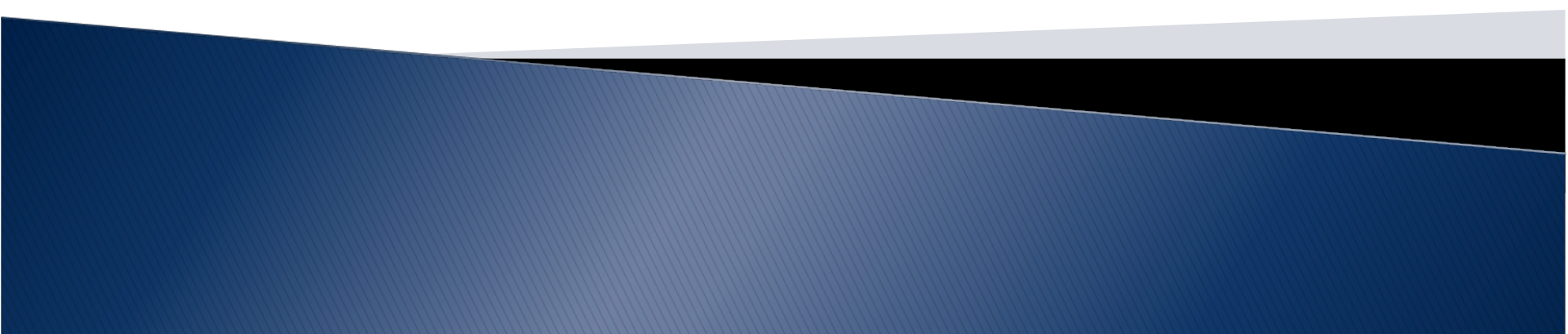


Pre-Modern Japanese Syntax: A Corpus Based Study

Kerri L Russell
West Oxford Academy (16 November 2010)



The VSARPJ Corpus

- Background
 - Funding: Arts & Humanities Research Council
 - Goals
 - Create a tool for linguistic analyses
 - Investigate argument structure for verbs both synchronically and diachronically
 - People
 - XML & Text Encoding Initiative (TEI)
 - Texts



The VSARPJ Corpus: Texts

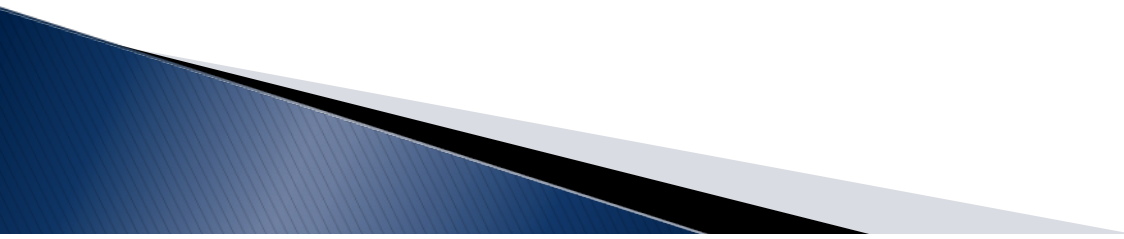
- Old Japanese (abbreviated 'OJ', approximately 700-800)
 - Kojiki kayō 古事記歌謡 712
 - Nihon shoki kayō 日本書紀歌謡 720
 - Fudoki kayō 風土記歌謡 730s
 - Bussokuseki-uta 仏足石歌 after 753
 - Man'yōshū 万葉集 after 759
 - Jōgū shōtoku hōō teisetsu 上宮聖徳法王帝説 unknown (early Heian?)
 - Shoku nihongi kayō 続日本紀歌謡 797
 - Norito 祝詞 927 & Senmyō 宣命 697-791

The VSARPJ Corpus: Texts

- Early Middle Japanese ('EMJ', 800-1200)
 - Kokin wakashū preface 古今和歌集仮名序 914
 - Ise monogatari 伊勢物語 early 10th century
 - Tosa nikki 土佐日記 935
 - Taketori monogatari 竹取物語 mid 10th century
 - Kagerō nikki 蜻蛉日記 second half of 10th century
 - Ochikubo monogatari 落窪物語 late 10th century
 - Makura no sōshi 枕草子 c. 1000
 - Genji monogatari 源氏物語 1001-1010
 - Sarashina nikki 更級日記 1059-1060
 - Konjaku monogatari-shū 今昔物語集 1120

The VSARPJ Corpus: Texts

- Late Middle Japanese ('LMJ', 1200-1600)
 - Hōjōki 方丈記 1212
 - Esopo no fabulas 1593
 - Feiqe monogatari 1593



Orthography

君待跡 / 吾戀居者 / 我屋戸乃 / 簾令動 / 秋之風吹
きみまつと / わがこひをれば / わがやどの / すだれう
ごかし / あきのかぜふく (MYS 8.1606)

君之行 / 氣長成奴 / 山多都祢 / 迎加将行 / 待尔可将待
きみがゆき / けながくなりぬ / やまたづね / むかへか
ゆかむ / まちにかまたむ (MYS 2.85)

如此許 / 戀乍不有者 / 高山之 / 磐根四卷手 / 死奈麻死
物呼

かくばかり / こひつつあらずは / たかやまの / いはねし
まきて / しなましものを (MYS 2.86)



Orthography

君待跡 / 吾戀居者 / 我屋戸乃 / 簾令動 / 秋之風吹
きみまつと / わがこひをれば / わがやどの / すだれう
ごかし / あきのかぜふく (MYS 8.1606)

君之行 / 氣長成奴 / 山多都祢 / 迎加将行 / 待尔可将待
きみがゆき / けながくなりぬ / やまたづね / むかへか
ゆかむ / まちにかまたむ (MYS 2.85)

如此許 / 戀乍不有者 / 高山之 / 磐根四卷手 / 死奈麻死
物呼
かくばかり / こひつつあらずは / たかやまの / いはねし
まきて / しなましものを (MYS 2.86)

Orthography

君待跡 / 吾戀居者 / 我屋戸乃 / 簾令動 / 秋之風吹
きみまつと / わがこひをれば / わがやどの / すだれう
ごかし / あきのかぜふく (MYS 8.1606)

君之行 / 氣長成奴 / 山多都祢 / 迎加将行 / 待尔可将待
きみがゆき / けながくなりぬ / やまたづね / むかへか
ゆかむ / まちにかまたむ (MYS 2.85)

如此許 / 戀乍不有者 / 高山之 / 磐根四卷手 / 死奈麻死
物呼

かくばかり / こひつつあらずは / たかやまの / いはねし
まきて / しなましものを (MYS 2.86)

Orthography

君待跡 / 吾戀居者 / 我屋戸乃 / 簾^レ動 / 秋之風吹
きみ^マつと / わがこひをれば / わがやどの / すだれ^ウ
ごかし / あきのかぜふく (MYS 8.1606)

君之行 / 氣長成奴 / 山多都祢 / 迎加^サ行 / 待^マ爾可^サ待^マ
きみがゆき / けながくなりぬ / やまたづね / むかへか
ゆかむ / ^マちにか^サまた^マむ (MYS 2.85)

如此許 / 戀乍不有者 / 高山之 / 磐根四卷手 / 死奈麻死
物呼

かくばかり / こひつつあらずは / たかやまの / いはねし
まきて / しなましものを (MYS 2.86)

<div xml:id="MYS.8.1606">

<ab type="original">君待跡<lb/>

吾戀居者<lb/>

我屋戸乃<lb/>

簾令動<lb/>

秋之風吹</ab>

<ab type="transliterated">kimi matu to<lb/>

wa ga kwopwi woreba<lb/>

wa ga yadwo no<lb/>

sudare ugwokasi<lb/>

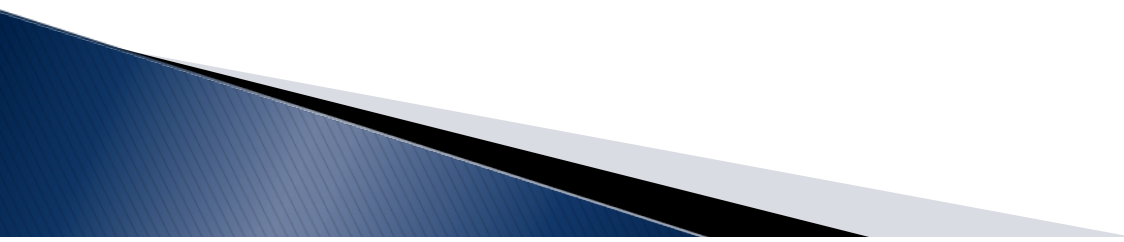
aki no kaze puku</ab>

</div>



Orthography

- Preserve information about how a text was written
 - Logographic
 - Phonographic
 - Not recorded in text
 - Large or small phonograms
- Character tags: `<c>word</c>`
- Character attributes: `@type="some value"`



<c type="logo">kimi</c>

<c type="logo">matu</c>

<c type="phon">to</c>

<lb/>

<c type="logo">wa</c>

<c type="noLogo">ga</c>

<c type="logo">kwopwi</c>

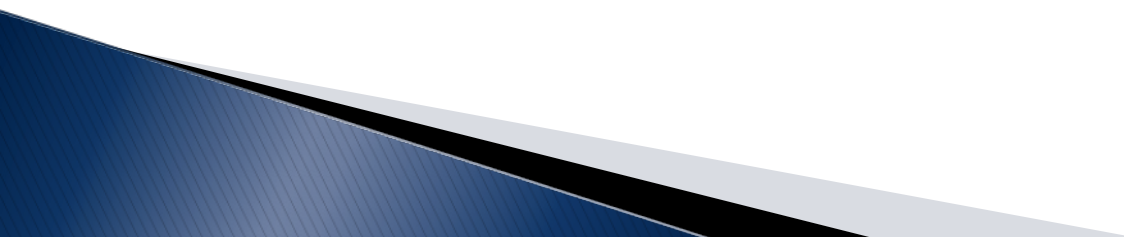
<c type="logo">woreba</c>

<lb/>



Morphology

- Information about words and morphemes
 - Part of speech
 - Inflection
 - Function



Unique Identification Numbers

- Search words through time
- Search inflected forms
- All words and morphemes are assigned IDs
- This number links to an entry in our dictionary which contains:
 - Unique identification number
 - Pronunciation changes
 - Part of speech
 - Usage
 - Start/end
 - Dialect (EOJ)



<w type="noun" ana="#L042066">

<c type="logo">kimi</c>

</w>

<w type="particle" subtype="case" function="gen" ana="#L000503">

<c type="logo">ga</c>

</w>

<w type="verb" inflection="infinitive" ana="#L031840">

<c type="logo">yuki</c>

</w>



<w type="noun" ana="#L050033">

<c type="phon">ke</c>

</w>

<w>

<w type="adjective" ana="#L007004">

<c type="logo">naga</c>

</w>

<m type="adjcop" inflection="infinitive" ana="#L000033">

<c type="logo">ku</c>

</m>

</w>

<w>

<w type="verb" inflection="stem" ana="#L031317">

<c type="logo">nari</c>

</w>

<m type="auxiliary" function="perf" inflection="conclusive" ana="#L000018">

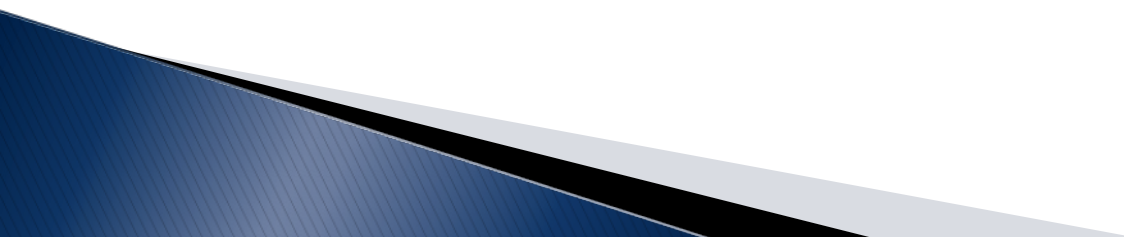
<c type="phon">nu</c>

</m>

</w>

Syntax

- Syntactic Markup
 - phrases
 - clauses
 - sentences
- Phrases and clauses can be arguments or adjuncts
- Phrases are assigned semantic roles



So...there's a corpus. Now what?

- Statistics

- poems: 4976
- words: 89,356
- morphemes: 15,795
- verbs: 22,855
- adjectives: 3,601
- phrases: 26,927
- argument phrases: 11,265
- argument phrases with no case particle: 6012

- Searches

Questions and Comments Welcome

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

