Chung-Hwa Buddhist Journal (2012, 25:129-148)
Taipei: Chung-Hwa Institute of Buddhist Studies
中華佛學學報第二十五期　頁 129-148 (民國一百零一年)，臺北：中華佛學研究所
ISSN:1017-7132

Verb Semantics and Argument Realization in Pre-modern Japanese: A Corpus Based Study

Kerri L Russell and Stephen Wright Horn
University of Oxford

Abstract

We are developing a corpus in order to investigate argument realization in detail for pre-modern Japanese, giving a comprehensive account of the basic grammar of each major stage of the language and allowing for both synchronic and diachronic analyses. When completed, the corpus will contain texts from the 8th century until the beginning of the 16th century. The results of the project will impact the description and understanding of pre-modern Japanese and its changes through time, furthering our understanding and interpretation of earlier texts. The project is also expected to have implications for general linguistic theory, both with regard to frameworks for understanding verb semantics and clause structure, and with regard to the application of syntactic theory to 'dead' languages. This paper focuses on the initial stages of corpus building, including methods for encoding orthography, morphology, and syntax.

Keywords:

Old Japanese, Argument Realization, Lexicon, TEI markup, Corpus-based Linguistics

前現代日語的動詞語意與論元體現:
以語料庫為基礎的研究

Kerri L Russell and Stephen Wright Horn
牛津大學

摘要
藉由對語言的每一主要階段的基本文法之詳盡說明，並考慮共時與歷時分析，我們發展一語料庫以詳細研究前現代日語之論元體現。當此完成後，此語料庫將包含從八世紀到十六世紀初的的文本。此計畫的成果將影響前現代日語的描寫與了解，及其隨著時間所造成的改變，增進我們對於早期文本的理解與解釋。此計畫也期望對一般語言學理論有所影響，包含在了解動詞語意與句法結構的架構上，以及對於不再通行的語言之語法理論的應用兩層面。此篇著重在語料庫建立的初始階段，包含編碼拼字的方法，型態與語法。

關鍵詞：古日語、論元體現、詞典、TEI標記、語料庫語言學

Introduction

This paper presents the tagging conventions used in the development of a corpus for a pre-modern Japanese syntax project at the University of Oxford. The project is entitled Verb Semantics and Argument Realization in Pre-modern Japanese: A Comprehensive Study of the Basic Syntax of Pre-modern Japanese (abbreviated as 'VSARPJ') and is funded by a grant of almost £1 million from the Arts and Humanities Research Council in the UK. An important first phase of the project is the construction of an annotated and encoded corpus of texts. While the corpus is initially constructed specifically for the purpose of serving the VSARPJ project, we believe it will eventually become useful for the investigation of many other aspects of the syntax of pre-modern Japanese.

The primary and immediate goal of the VSARPJ project is to investigate *argument realization* in detail for pre-modern Japanese. Argument realization is a fundamental aspect of the syntax of a language which concerns the ways in which verb meaning determines the number of arguments (e.g., subjects, objects, goals, etc.) in a clause and their morpho-syntactic and semantic properties. In essence, the project will contribute to a comprehensive account of the basic syntax of each of the stages of the pre-modern Japanese language, from the beginning of its recorded history in the 8th century until the end of the 16th century, and of the changes in basic syntax that have taken place over these stages.[1]

The VSARPJ project has two parts: Synchronic and Diachronic. In the synchronic part, we investigate for the main stages of pre-modern Japanese the argument realization patterns of individual verbs and of verb classes. For each verb attested in the pre-modern Japanese texts we are using for this investigation, we establish both the syntactic frames in which it can occur and also its basic argument realization pattern. An important part of this will be the determination of what counts as an argument, and to what extent a more finely graded range of categories between argument and adjunct is needed. We will also look at other grammatical phenomena than argument realization which may be explained by verb semantics, for example, aspect, auxiliary selection, ellipsis, and case drop.

The diachronic part of the project will build on the results of the synchronic part. In addition to charting changes affecting individual verbs, we will be able to establish an inventory of changes through the history of Japanese in argument realization both for individual verbs and for classes of verbs and thereby be able to investigate general patterns of change, including possible development pathways for verb meanings and argument realization.

[1] More detail about the VSARPJ project, including the framework we use for analysis is presented on our website: *http://vsarpj.orinst.ox.ac.uk/project.html*.

Apart from the intrinsic value the results of the project will have to the description and understanding of Japanese grammar and its history, the project may also be expected to yield results of more general interest, as this will be the first detailed application of the type of framework employed here to a language such as Japanese, which frequently drops case markers, has extensive argument ellipsis (pro-drop), and has fairly free word order. It will also be the first large-scale investigation of this kind to a 'dead' language, which poses particular challenges to research into syntax.

The initial stage of the VSARPJ corpus consists of building a digital corpus of texts, encoded with information about various linguistic properties. Once this stage is completed, the next stage will involve using the corpus to conduct various types of linguistic analysis. As we are currently in the initial stage of corpus construction, this paper will focus on the encoding of the corpus, and in particular, on the oldest stage of texts in the corpus, Old Japanese (OJ). In this paper we describe the contents of the VSARPJ corpus (section 2), the initial stage of marking up texts (section 3), and XML mark-up conventions (section 4).

The VSARPJ Corpus

The corpus will in the initial stage comprise a selection of texts from the three main periods of pre-modern Japanese (Old, Early Middle, and Late Middle Japanese):

Old Japanese ('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
Shoku nihongi kayō	続日本紀歌謡	797
Shoku nihongi senmyō	続日本紀宣命	697-791
Engishiki norito	延喜式祝詞	(compiled) 927

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

Late Middle Japanese (‘LMJ’, 1200-1600)

Esopo no fabulas		1593
Feiqe monogatari		1593

The corpus includes all main extant texts from the OJ period. For EMJ, the corpus focuses on texts from the period 900-1100 which are thought to a large extent to reflect the (spoken) language of the time. For large texts from this period, e.g., *Genji monogatari*, only extensive selections and not the entire texts will be included in the initial phase of the corpus. From the LMJ period, where most of the textual material is written in ‘classical Japanese’ rather than in the contemporary language and is characterized by a high degree of fossilization, we use two texts produced by the Jesuit missionaries at the end of the 16th century, the *Esopo no fabulas* and the *Feiqe monogatari*, which both reflect the contemporary language at the very end of the period, and also have the additional advantage of being written in alphabetic writing. For all periods, we follow the readings in the critical edition of *Nihon koten bungaku taikei* (NBKT), published by Iwanami Shoten.[2]

Initial Stage of Markup

The first stage of markup was completed in MS Word. This process involved romanization of texts and the use of symbols to indicate prefixes, suffixes, compounds, etc.

Romanization of Texts

First, each text was romanized to present a phonemic transcription in accordance with the phonology of the time the text is thought to have been written, and reflecting the sound changes which had been completed by that time. For example, the word which is often written by 恋, which in Modern Japanese (NJ) has the shape *koi* and which may be glossed very roughly as ‘love’. In the historical *kana* spelling (歴史的仮名遣い) this word is writtenこひ, regardless of the time from which the text dates. In a phonemic transcription, however, this word has the shape /kwopwi/ (こ$_{甲}$ひ$_{乙}$) in OJ.[3] As a result of

[2] At this stage, construction of the OJ corpus is complete. The corpus consists of nearly 5,000 poems of around 90,000 words, 20,000 of which are verbs. We have not yet decided on how much to include from other periods, so we are not yet certain of the size of the corpora we will develop for later stages of pre-modern Japanese.

[3] We use the Frellesvig & Whitman (2008) transcription system for OJ.

sound changes which took place since OJ, the shape of this word has changed as shown in (1) below with approximate dating, and the corpus uses those shapes in accordance with the dates of the texts. Thus, in the *Tosa nikki* (from 935), this word is transcribed *kopi*, but in the *Genji monogatari* (from just after 1000) it will be written *kowi*. This is a very basic point, but one which is often ignored in the presentation of pre-modern Japanese texts.

(1)		800	950	1000	1100
OJ *kwopwi* >	EMJ	*kwopi*>	*kopi* >	*kowi* >	*koi*

Further, in the process of romanizing texts, we preserved a three-way distinction found in the texts: phonographic, logographic, and "not in text" for items which are not orthographically represented in the original text. This distinction is shown in (2) from the *Man'yōshū* (MYS 1:1) with phonographically written material in italicized text, logographically written material in plain text, and items not orthographically represented in the original text ("not in text") written in underlined text.

(2)

篭	毛	與	美篭	母乳	布久思	毛	與
kwo	*mo*	*yo*	*mi*-kwo	*moti*	*pukusi*	*mo*	*yo*
basket	ETOP	EMPH	HON-basket	hold.INF	shovel	ETOP	EMPH

美夫君志	**持**	此	岳	尓	菜
mi-bukusi	**moti**	ko no	woka	*ni*	na
HON-shovel	**hold.INF**	this GEN	hill	DAT	greens

採須	兒	家	吉閑名	告紗根
tuma-*su*	kwo	ipye	*kikana*	nora-*sane*
pick-RESP.ADN	child	home	ask.OPT	tell-RESP.OPT

'Girl with your basket, with your pretty basket, with your shovel, with your pretty shovel, picking greens on this hillside, I want to ask your home. Please tell me!'

The interpretation of logographic writing relies on reading tradition and is in many respects uncertain. This is sometimes reflected in the existence of significantly different reading traditions of some texts. If a text or crucial parts of it are written logographically, we can not, strictly speaking, be certain of which words, or inflected forms, are reflected in the text. For example, in (2) above, we can not be certain that the verb written by 持 (in bold face) really is *mot-* 'to hold', nor that its inflected form really is the infinitive *moti*, as it is read according to the reading tradition, and not the adnominal *motu.* Thus, logographically written text is far less reliable than phonographically written text and can be used as linguistic evidence only with great caution.

The items which are not orthographically represented in the text are also based solely on reading tradition. The word 此 'this' in (2) is interpreted as as "ko <u>no</u>" but the genitive particle *no* is not represented by a character in the text. This issue will become particularly important when investigating argument structure in contexts where a case particle marking an argument is understood only from the reading tradition and not from the written text itself. As there is no way to prove the existence of the case particle, such examples are less reliable as evidence of case marking than those where a particle is written phonographically or even logographically.

In the initial stage of markup we indicated this three way distinction by rendering phonographically written material in lower case, logographically written material in upper case, and items not recorded in the original script written in upper case with a comment saying "not in text".[4]

Symbols Used in Markup

While romanizing the texts in this stage of marking up texts in MS Word, we added information about certain types of words with the symbols =, -, +, and ~.

The symbol "=" was used to indicate a particle. For example, "ko <u>no</u>" in (2) above was written as KO=NO to indicate that "no" is a genitive particle. Following the discussion above, a comment was also attached to "no" to mark it as not having been represented orthographically in the original text.

Next, "-" was used to indicate

1) inflecting forms following verbs and adjectives and

2) compound verbs.

The last word in (2), *norasane*, consists of the stem of the verb *nor-* 'to tell' and the optative inflection of the respect auxiliary *-(a)s-*. This was marked in our word files as NORA-sane at this stage, thus simultaneously indicating orthography and morphology.

The "+" symbol was used to indicate

1) nouns in compounds, including noun+noun and noun+verb combinations and

2) nominal and adjectival prefixes.

For example, *mikwo* in (2) above consists of the honorific prefix *mi* followed by *kwo* 'child'. This was marked as *mi*+KWO.

Last, the symbol ~ was used for verbal prefixes and circumfixes. There are no examples of this in (2), but take, for example, *sanuru* (MYS 14.3504) which consists of

[4] In hindsight, from the point of view of converting Word files to XML format, it would have been to our advantage had we indicated these three types of orthography using distinct styles in MS Word.

the prefix *sa-*[5] and the adnominal form of the verb *ne-* ‘to sleep’. This word was marked up as sa~nuru.

XML Markup Conventions

The next stage in corpus building involves XML markup according to the guidelines of the Text Encoding Initiative (TEI). The inventory of TEI coding is a small set of *tags* which are used to enclose portions of text; text enclosed by tags can further be characterized by various attributes, such as *type*, *subtype*, *function*, *inflection*, etc. The inventory of coding elements and conventions of the TEI are under constant development and improvement; they may be viewed at *http://www.tei-c.org/*. A major consideration for adopting TEI technology and guidelines for the corpus was that such standards ensure that the corpus we design will be long lasting, non-idiosyncratic, and updateable along with future changes in technology. We attempted to follow the TEI guidelines as closely as possible, however, we had to add some attributes for items we felt important for markup and which were not available in TEI. For example, we felt it important to indicate the inflection for all forms which can inflect (e.g., verbs, adjectives, copulas, auxiliaries) and created the ‘inflection’ attribute to allow us to do this. By indicating the inflection, we can easily compare all forms in any given inflection. The inflected form of the predicate also indicates clause types, so we can investigate main clauses or subordinate clauses based on the inflection of the predicate.

Most of the OJ texts were marked up using MS Word, as described above. These were then converted into XML format.[6] Our mark-up policies consist of ways to link the original and romanized version of a text (section 1), to preserve orthographic conventions (section 2), to encode information about words, morphemes, and parts of speech (section 3), to identify lexemes and morphemes (section 4), and to encode syntactic features (section 5). As an example, we also present a fully marked up poem (section 6).

1. Original and Transliterated Text

In order to reflect the crucial distinction between logographic and phonographic writing and to represent information about how words and/or morphemes were written in the original script,[7] we have adopted the following policies. First, for OJ texts we preserve the original script together with the phonemically transcribed text. Thus, reference can be made to the original script. This is done by having the original script in an <ab> (“anonymous block”) tag and assigned the @type attribute with the value ‘original’. We

[5] The function of this prefix is unclear.

[6] The scripts for converting our word files into XML were written by James Cummings.

[7] By ‘original script’ we mean the script employed in the critical edition upon which a text is based.

use “ojp” as the value for @xml:lang for texts written in Old Japanese and “ojp-Latn” for the transliterated version of the OJ texts. The romanized version of the script follows in its own <ab> tag with the @type attribute value ‘transliteration’. Line breaks (<lb>) are also linked in the original and transliterated version using @xml:id and @corresp attributes in order to make it easy to see how a line of text was rendered in the original or how a line of text should be read. The @xml:id value contains the poem and line number; “MYS.1.1” means that the poem is from the *Man’yōshū* (MYS), Book 1, poem number 1, and orig_1 defines this as the first line break in the poem. This is illustrated in (3) using an excerpt from the poem presented in (2) above.

```
(3)<ab type="original" xml:lang="ojp">
                 篭毛與
              <lb xml:id="MYS.1.1-orig_1" corresp="#MYS.1.1-trans_1"/>
                   美篭母乳
   <!-- … -->
   </ab>
              <ab type="transliteration" xml:lang="ojp-Latn">
                   kwo mo yo
              <lb xml:id="MYS.1.1-trans_1" corresp="#MYS.1.1-orig_1"/>
                 mikwo moti
   <!-- … -->
   </ab>
```

2. Encoding Orthography

To preserve the three-way writing distinction discussed above, we use the character tag <c> with the @type attribute. The possible values for @type are “phon” for items written phonographically, “logo” for those written logographically, and “noLogo” for items not orthographically represented in the original text. This is shown in (4) below with (a) presenting the original text, the phonemic transcription, and glosses, and (b) showing the markup.

(4)	a.	我		屋戸	乃
		wa	ga	yadwo *no*	
		I	GEN	hut	GEN
		‘of my hut’		(MYS 8.1606)	

b. <c type="logo"> **wa** </c>
<c type="noLogo"> **ga** </c>
<c type="logo"> **yadwo** </c>
<c type="phon"> **no** </c>

3. Words, Morphemes, and Part of Speech

Words are enclosed in ‘word(-like)’ tags, <w>, and information about part of speech is supplied by the @type attribute. The main word classes represented in this way are *noun*, *pronoun*, *adverb*, *verb*, *adjective*, *copula*, *adjectival noun*, *verbal noun* and *particle*.

Complex words can consist of more than one word, forming a compound word. And they can consist of one or more words followed or preceded by one or more morphemes. The morpheme tag <m> is used for bound forms, and is then categorized by @type attributes with the possible values of *auxiliary*,[8] *prefix*, *suffix*, *numeral*, *counter*, and *adjectival copula*.[9] The grammatical system and terminology reflected in the coding is that of Frellesvig (2010).

Several of the parts of speech are further subcategorized, notably particles and auxiliaries, which are given subtypes and functions. For example, *ga* is a word (<w>) of the @type value “particle”, @subtype value "case" with the @function value “genitive”; and *-(i)ki* is a morpheme (<m>) of the @type value “auxiliary” and with the @function value “simple past”. A full, current list of the parts of speech, including subcategories, which are distinguished throughout the corpus is available at the corpus website (*http://vsarpj.orinst.ox.ac.uk/corpus/*).

Inflecting parts of speech, such as verbs, auxiliaries, extensions, copulas, and adjectival copulas are supplied with information about their inflectional forms with the @inflection attribute. For inflectional forms which are identical in shape, we do not specify which inflecting form is shown even when the syntax allows us to chose one or the other. For example, both the adnominal and conclusive form of the verb *yuk-* ‘to go’ is *yuku*; it is impossible to tell which inflection this is just by the shape of the word. The verb in this case is assigned the @inflection value “adnconc” and not “adnominal” or “conclusive”.

Similarly, for conjugation classes which do not have a distinction between conclusive and infinitive, we mark those categories with the @inflection value “infconc”, see (5). The reason for marking only morphologically distinct categories also at the level of individual conjugation classes is that it seems likely that there is a correlation between the inflected form of a clause predicate and the marking of its arguments, and that it therefore is important to distinguish between forms which are positively identifiable by their shape and on the other hand forms which on the basis of their shape may be assigned to either of two syncretic categories.

(5)	adnconc	*yuku*
	infconc	*ari*

[8] Auxiliaries are inflecting suffixes, corresponding largely to the *jodōshi* (助動詞) of traditional Japanese grammar, e.g., negative *-(a)zu* or perfective *-(i)te-* and *-(i)n-*.

[9] The adjectival copula is the inflectional morpheme which usually follows adjective stems, with forms like conclusive *-si*, adnominal *-ki*, and infinitive *-ku*.

In (6) we give an example of markup of part of speech and inflection.

(6) a.

君	之	行	氣	長	成奴
kimi	ga	yuki	*ke*	naga-ku	nari-*nu*
my.lord	GEN	go	day	long	become-PERF

‘My lord, it has been a long time since you left’(MYS 2.85)

b.

```
<w type="noun"> kimi </w>
<w type="particle" subtype="case" function="gen"> ga </w>
<w type="verb" inflection="infinitive"> yuki </w>
<w type="noun"> ke </w>
<w>
        <w type="adjective"> naga </w>
        <m type="adjcop" inflection="infinitive"> ku </m>
</w>
<w>
        <w type="verb" inflection="stem"> nari </w>
         <m type="auxiliary" inflection="conclusive" function="perf"> nu </m>
</w>
```

4. Lexeme and Morpheme Identification

Each distinct item (word or morpheme) in the corpus is assigned a unique ID number. This has a number of advantages, in particular in making it possible to divorce searches in the corpus from actual strings of text.

- Searches for inflecting words or morphemes in the texts will not be limited to the actual inflected forms of an item. Thus, a search for the verb *sin-* ‘die’ will return all the inflected forms of that verb. However, searches can also be modified to give only a subset of forms, for example defined by specific inflected forms or combination with specific auxiliaries.
- Searches across time for items which have changed shape as a result of sound change will be straightforward. For example, as a result of sound change the verb OJ *kwopwi-* has a number of different shapes through time, as outlined above (1), and appears in texts from different periods in significantly different shapes (*kwopi-*, *kopi-*, *kowi-*, *koi-*). With unique ID numbering, it is not necessary to search for all of these shapes, but it is possible to search for all, or a specific set of, occurrences of this verb through the corpus, regardless of the actual shape of the verb at any particular stage.
- Searches are not contaminated by text strings which are identical to the intended target of a search. For example, the verb ‘request, ask’ OJ *kop-* has a number of forms which are segmentally identical with forms of ‘love’ from somewhere in the first half of the EMJ period (for example infinitive *kopi*, *kowi*, *koi*). With unique ID numbering, forms of one verb will not show up in searches for the other verb.

In our current practice, unique ID numbers consist of the letter ‘L’ and a six-digit number. They are assigned to a word (<w>) or morpheme (<m>) as an @ana attribute. For example, the form *nari-nu* (cf. (6) above) is marked as shown in (7).

```
(7)  <w>
          <w ana="#L031317"> nari </w>
          <m ana="#L000018"> nu </m>
     </w>
```

The unique IDs are stored in a separate lexicon file, which is linked to the corpus and which contains basic information about each word or morpheme, including variant shapes of a form over time, its part of speech, conjugation class (where relevant), and a simple gloss. The information currently contained within a simple lexicon entry is as shown in (8).

(8) Shapes:From the 8th century: **kwopwi-** > From 800: **kwopi-** >
From before 950: **kopi-** > From c. 950-1000: **kowi-** >
From c. 1100: **koi-**
Part of speech: **verb**
Conjugation class: **upper bigrade** (上二段)
Gloss: **love**

This information in (8) was extracted from an entry presented below in (9).

```
(9)  <superEntry xml:id="L030731">
     <entry>
               <form type="stem">
                    <orth stage="I">kwopwi-</orth>
                      <orth stage="II">kwopi-</orth>
                      <orth stage="III">kopi-</orth>
                      <orth stage="V">kowi-</orth>
                      <orth stage="VII">koi-</orth>
                      <gramGrp>
                           <pos>verb</pos>
                           <iType type="UB"/>
                      </gramGrp>
               </form>
               <def>love</def>
          </entry>
          <entry>
               <form type="noun">
                    <orth stage="I">kwopwi</orth>
                      <orth stage="II">kwopi</orth>
                      <orth stage="III">kopi</orth>
                      <orth stage="V">kowi</orth>
                      <orth stage="VII">koi</orth>
```

```
                <gramGrp>
                    <pos>noun</pos>
                </gramGrp>
            </form>
    </entry>
</superEntry>
```

Here, the <superEntry> element defines the @xml:id for the lexical item. The <entry/> element is used to indicate one or more related lexical entries. The <form> element can be further specified with the @type attribute, which we currently only use for verbs to indicate their "stem" and the derived "noun" form of a verb. Next, <orth> (orthography) presents the shape of the form (e.g., *kwopwi-*) and also has the @stage attribute corresponding to stages of phonological development in the pre-modern period. Grammatical information is presented in <gramGrp>. This includes part of speech <pos> and conjugation class <iType>. The example in (9) above is defined as @type="UB" which stands for "upper bigrade". Finally, the meaning is presented in the <def> tag; where more than one meaning is possible, the element <sense> is also used.

As the research of the VSARPJ project progresses, additional grammatical information will also be entered into the lexicon. This will include information about the possible argument realization patterns of a verb. In this way, the lexicon will also be an important tool for organizing the results of our research as they appear.

Finally, although outside the scope of the VSARPJ project, it should be mentioned that a lexicon linked to a text corpus by means of unique ID numbering has enormous potential for enriching the field of Japanese lexicography.

5. Syntax

Syntactic information is encoded by means of a minimal inventory of constituents, namely those of clause, <cl>, and phrase, <phr>. The @type attribute can be used to identify the clause or phrase as being an *argument* (predicate selected) or *adjunct* (e.g., free adverbials).

Clauses can be embedded within other clauses as subordinate clauses. Adnominal, or relative, clauses are embedded within phrases. Nominalized clauses are first wrapped as clauses to show the clausal structure and then wrapped as phrases to put them on the same level as noun phrases. Predicate-selected clauses (including but not limited to complement clauses) are categorized by the @type attribute as arguments ("arg").

Phrases can be headed by adverbs and nominalized clauses, in addition to nouns. Phrases are categorized by the @type attribute as arguments if they are clearly predicate selected, and as adjuncts ("djunct") if they are clearly free adverbials or sentence adjuncts. At this stage of markup, a large proportion of phrases are marked neither as arguments nor as adjuncts, because their status is not entirely clear. Resolving the status of such phrases, and other important issues such as the determination of whether categories may be needed which are intermediary between the poles of argument and adjunct, or whether argumenthood is a scalar property, are parts of the substantive research of the VSARPJ project. The corpus will eventually reflect the results of this research.

The structure of both clauses and phrases is generally flat.[10] The words which can form predicates of clauses are verbs, adjectives, or copulas. Within a clause, the word or words which form its predicate are identifiable by not being enclosed in phrase tags. Topics and right dislocated elements are located outside of the clauses they relate to. (10) exemplifies syntactic markup: (10a) shows a complex clause from the poem in (6a); (10b) shows the topic *pito pa*; (10c) shows the relative clause *a ga kwopuru* modifying *kimi*; and (10d) shows the right dislocated topic *ware pa*.

(10) a. Complex clause

```
<cl>
        <cl>
        <phr type="arg"> kimi ga </phr>
        yuki
        </cl>
        <cl type="arg">
        <phr type="arg"> ke </phr>
        nagaku
        </cl>
        narinu
</cl>
```

b. Topic

人	者	待跡	不来家留
pito	**pa**	mate*do*	ko-zu-*kyeru*
person	TOP	wait.CONC	come-NEG.INF-MPAST.ADN

Even though I wait for you, you do not come'(MYS 4.589)

```
<phr> pito pa </phr>
<cl>
  <cl> matedo </cl>
  kozukyeru
</cl>
```

c. Relative clause

吾	戀	流君
a ga	**kwopu*ru***	kimi
I GEN	love.ADN	lord

[10] Within phrases constituency is usually predictable from the sequence of constituents, but if not, constituency can be marked as necessary.

‘my lord, whom I love’(MYS 4.485)

```
<phr type="arg">
    <cl>
        <phr type="arg"> a ga </phr>
        kwopuru
    </cl>
    kimi
</phr>
```

d. Right dislocated topic

野嶋	我	左吉	爾	伊保里	須	和礼	波
Nwosima	*ga*	*saki*	*ni*	*ipori*	***su***	***ware***	***pa***
[place name]	GEN	cape	DAT	hut	do.CONCL	I	TOP

‘me, I make a hut on the cape of Noshima’(MYS 15.3606)

```
<cl>
      <phr> nwosima ga saki ni </phr>
      ipori su
  </cl>
  <phr>
      ware pa
</phr>
```

6. An Example of Full Markup

Finally in this section, we provide as an example the full markup of the text in (6a) above.

```
(11)  <ab type="original" xml:lang="ojp">
      君之行
      <lb xml:id="MYS.2.85-orig_1" corresp="#MYS.2.85-trans_1"/>
      氣長成奴
      <lb xml:id="MYS.2.85-orig_2" corresp="#MYS.2.85-trans_2"/>
      山多都祢
      <lb xml:id="MYS.2.85-orig_3" corresp="#MYS.2.85-trans_3"/>
      迎加将行
      <lb xml:id="MYS.2.85-orig_4" corresp="#MYS.2.85-trans_4"/>
      待尓可将待
      </ab>

            <ab type="transliteration" xml:lang="ojp-Latn">
                    <s>
                        <cl>
                            <cl>
                                <phr type="arg">
                                    <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>
              </phr>
              <w type="verb" inflection="infinitive"
ana="#L031840">
                <c type="logo">yuki</c>
              </w>
            </cl>
            <lb/>
            <cl type="arg">
              <phr type="arg">
                <w type="noun" ana="#L050033">
                  <c type="phon">ke</c>
                </w>
              </phr>
              <w>
                <w type="adjective" ana="#L007007">
                  <c type="logo">naga</c>
                </w>
                <m type="adjcop"
inflection="infinitive" ana="#L000033">
                  <c type="logo">ku</c>
                </m>
              </w>
            </cl>
            <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>
          </cl>
        </s>
        <lb/>
        <s>
          <cl>
            <phr>
              <cl>
                <cl type="djunct">
                  <phr type="arg">
                    <w type="noun"
```

```
ana="#L050034">
                                            <c
type="logo">yama</c>
                                        </w>
                                    </phr>
                                    <w type="verb"
inflection="infinitive" ana="#L031047">
                                        <c type="phon">tadune</c>
                                    </w>
                                </cl>
                                <lb/>
                                <w type="verb" inflection="infinitive"
ana="#L031722">
                                    <c type="logo">mukape</c>
                                </w>
                            </cl>
                            <w type="particle" subtype="foc"
ana="#L000506">
                                <c type="phon">ka</c>
                            </w>
                        </phr>
                        <w>
                            <w type="verb" inflection="stem"
ana="#L031840">
                                <c type="logo">yuka</c>
                            </w>
                            <m type="auxiliary" function="conjectural"
inflection="adnconc"
ana="#L000002">
                                <c type="logo">mu</c>
                            </m>
                        </w>
                    </cl>
                </s>
                <lb/>
                <s>
                    <cl>
                        <phr>
                            <cl>
                                <w type="verb" inflection="stem"
ana="#L031644">
                                    <c type="logo">mati</c>
                                </w>
                            </cl>
                            <w type="particle" subtype="case"
function="dat" ana="#L000519">
                                <c type="phon">ni</c>
                            </w>
                            <w type="particle" subtype="foc"
```

```
ana="#L000506">
                                <c type="phon">ka</c>
                            </w>
                        </phr>
                        <w>
                            <w type="verb" inflection="stem"
ana="#L031644">
                                <c type="logo">mata</c>
                            </w>
                            <m type="auxiliary" function="conjectural"
ana="#L000002"
inflection="adnconc">
                                <c type="logo">mu</c>
                            </m>
                        </w>
                    </cl>
                </s>
            </ab>
```

Conclusion

This small inventory of syntactic elements and conventions for their use, combined with the material they can contain, will allow unique identification of at least all of these elements or properties in the corpus: topics, right dislocated elements, focused elements, noun phrase heads, particle scope, clause predicates (including analytic predicates), zero marked arguments, topicalized arguments, relative order of case marked and zero marked arguments (including ordering relative to focused elements), and clause types (main, subordinate, adnominal, nominalized). Furthermore, all such elements and properties, as well as combinations of them, and combinations with other items and properties coded in the corpus will be searchable and extractable from the corpus. For example, we will be able to use the corpus to extract all attested syntactic frames for individual verbs, within individual stages of the language as well as across different stages. All of this is highly relevant, not just to the VSARPJ research project, but also more generally and widely to investigation of most features of pre-modern Japanese syntax.[11]

[11] Needless to say, these coding conventions easily lend themselves to the creation of equally powerful corpora of modern Japanese.

Abbreviations

General

TEI	Text Encoding Initiative
VSARPJ	Verb Semantics and Argument Realization in Pre-modern Japanese

Grammatical Terms

AND	Adnominal
CONC	Concessive
CONCL	Conclusive
DAT	Dative
EMPH	Emphatic
ETOP	Emphatic topic
HON	Honorific
NEG	Negative
OPT	Optative
RESP	Respect
TOP	Topic

Languages

EMJ	Early Middle Japanese
LMJ	Late Middle Japanese
MJ	Middle Japanese
NJ	Modern Japanese
OJ	Old Japanese

Texts

MYS	*Man'yōshū*

References

Frellesvig, Bjarke and Whitman, John. eds. 2008. *Proto-Japanese: Issues and Prospects*. Amsterdam: John Benjamins.

Frellesvig, Bjarke. 2010. *A History of the Japanese Language*. Cambridge: Cambridge University Press.

Frellesvig, Bjarke; Hom, Stephen Wright; Russell, Kerri L.; Sells, Peter. *The Oxford Corpus of Old Japanese. http://vsarpj.orinst.ox.ac.uk/corpus/corpus.html*.

Levin, Beth and Hovav, Malka Rappaport. 2005. *Argument Realization.* Cambridge: Cambridge University Press.

Text Encoding Initiative. (n.d.) *P5: Guidelines for Electronic Text Encoding and Interchange. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index-toc.html.*