The ‘end of word’ problem in Sanskrit: preliminary conclusions of the workgroup

The workgroup was established to consider the problem that in Sanskrit word junctions are often obscured or eliminated, and to propose XML/TEI markup to address this problem. Discussion has revealed that the problem is in fact only one manifestation of a more general issue.

For the convenience of non-Sanskritists the discussion here uses Roman transliteration, but the difficulty arises within the Devanāgarī script in which Sanskrit is normally written. Devanāgarī is a syllabary, in which one syllable consists of any number of consonants followed by one vowel followed optionally by *ṃ* or *ḥ* (*anusvāra* or *visarga*).[1] If a word ends in a consonant, it therefore has to share a syllable with the next word, so that *āsīd rājā* (‘there was a king’) is written **ā - sī - drā - jā**. To make matters worse, sandhi (phonological change at word boundaries) may fuse two consecutive vowels together, so that, even ignoring orthography, the words can no longer be divided — for example *tathā api* (‘even so’) becomes *tathāpi*, where the single vowel *ā* is shared by two inseparable words.

This presents no problems for display, but it poses very serious problems for any kind of analysis, and it has long been clear that we need to find some way round it. Various ad hoc ‘solutions’ have been used by people who have typed up Sanskrit texts, but none is at all satisfactory. It is clear that the most appropriate solution would be a form of specialised markup — hence XML and the Text Encoding Initiative.

The workgroup considered a form of markup which addresses the problem at the level of the syllable affected. This involves usages along the lines of

āsī<foo bar="d rā">drā</foo>jā

and

ta<foo bar="thā a">thā</foo>pi

This style of markup was eventually rejected on the grounds that, by intruding into words, it effectively prevents the use of other markup at the level of the word itself. Anyone who wants to tag the word *āsīd* in the first example above will be prevented from doing so by the very markup which seeks to establish that *āsīd* is indeed a word. This is clearly unacceptable.

Instead, a solution was agreed in principle which operates on words and groups of words:

```
<wordGroup>āsīdrājā
    <word>āsīd</word>
    <word>rājā</word>
</wordGroup>
```

and

```
<wordGroup>tathāpi
    <word>tathā</word>
    <word>api</word>
</wordGroup>
```

This has a number of advantages:

[1] This is a slight simplification, deliberately omitting certain ‘optional’ features such as Vedic accents. These do not, however, affect the principle as stated here.

- Such markup isolates individual words without intruding into them, allowing the possibility of further markup on the words themselves.
- Word-based markup, though slightly more verbose than syllable-based markup, is very much more readable.
- Word-based markup, unlike syllable-based markup, can easily be generalised for use with languages other than Sanskrit.

This last point is an especially important one. Members of the workgroup were aware from the start that Japanese presents problems that are comparable in some ways with those of Sanskrit. During the course of the discussion it has emerged that Avestan and Old High German could also benefit from the sort of markup we are now proposing. It seems likely that other languages too may contain features that could be usefully addressed in the same general way. We therefore regard it as particularly important that the solution adopted should be amenable to use in principle with *any* language. The markup should permit simple occasional use:

```
We’re all <wordGroup>gonna
    <word>going</word>
    <word>to</word>
</wordGroup> die!
```

However, the markup must also permit sophisticated analysis of word groups in languages where such groups are an established feature. This would necessitate the definition of further, language-specific markup. In any given document, language-specific markup would default to the language specified for the document as a whole with the **lang** attribute of the **<text>** tag (or whatever other mechanism may replace this); occasional citations in other languages would make use of the **lang** attribute of each **<wordGroup>** tag.

In some cases, it may be necessary to give alternative analyses: this can be achieved using the **<choice>...</choice>** mechanism. For example, we envisage that Sanskrit-specific markup will permit different ‘levels’ of analysis; this is because many existing electronic texts could easily be converted to incorporate partial (orthographic only) analysis. It is possible for different levels to coexist:

```
<wordGroup>āsīdrājā
    <choice>
        <word level="1">āsīd</word>
        <word level="3">āsīt</word>
    </choice>
    <word>rājā</word>
</wordGroup>
```

Sometimes it makes more sense for **<choice>...</choice>** to bracket alternative **wordGroup**s rather than alternative **word**s. In Japanese the delimitation of words is itself often uncertain:

```
<choice>
    <wordGroup type="x">
        <word>iwate</word>
        <word>no</word>
        <word>kai</word>
    </wordGroup>
    <wordGroup type="y">
        <word>iwateno</word>
        <word>kai</word>
    </wordGroup>
    ...
</choice>
```

The names used for the tags are provisional. It has been suggested that the existing TEI tag **<w>** could be used in place of **<word>**, but that is probably not a good idea: the point of a **<word>** is that it is a component of a **<wordGroup>**, not that it is necessarily a grammatical word. Partially analysed (level 1) Sanskrit texts will contain **word**s that are actually sequences of words.

The workgroup would be happy to receive feedback from the Council on the general shape of their proposal as outlined above. They would then proceed towards a more formal definition of the proposed tags **<word>** and **<wordGroup>**, and would also begin work on the detailed language-specific markup for Sanskrit and Avestan. The other known language for which detailed markup will be required is Japanese, but this workgroup would not be an appropriate body to deal with this topic, since it contains only one member who knows the language.

John D. Smith
May 5, 2004