The ‘end of word’ problem in Sanskrit: addendum to the preliminary report

The Council considered the preliminary report of the workgroup at their meeting in early May of this year, and their responses were passed on to us. Their chief concern appears to have been that the form of markup proposed in our report was too specific, and that if possible a more generic solution should be found. It was even suggested that the **<choice>** tag might provide a suitable approach.

The Indianists on the workgroup accepted the Council’s concern that the ‘end of word’ problem should be tackled in a suitably generic way; however, they felt that **<choice>** was both *too* generic and not appropriate for this purpose. The markup that is finally agreed must accurately reflect the specific realities we are dealing with in our different languages, and those realities do not involve a ‘choice’ between a grammatically valid form and an underlying abstraction.

All the cases we have considered so far involve a ‘chunk’ of text representing a sequence of isolable segments, where those segments do not appear in their normal identifiable form. Examples include:

- a group of words in which, for orthographic or other reasons, the word junctions are not indicated (Sanskrit, Japanese)
- a group of words whose forms have been affected by ‘euphonic’ sandhi changes (Sanskrit, Breton)
- a compound word (Sanskrit, Avestan, many other languages)
- a contraction (Gothic, English, many other languages)

This suggests that an approach using generic markup needs to be based round a pair of tags, one for the ‘chunk’ as a whole, the other for each constituent of it. Here I am using **<sequence>** and **<segment>** for these.

Returning to our old friend ‘there was a king’, one might envisage something like

A **<sequence type="wordGroup">āsīdrājā**
<segment type="word">āsīt</segment>
<segment type="word">rājā</segment>
</sequence>

A compound word in Sanskrit could be dealt with as follows:

B **<sequence type="compound">dharmakṣetre**
<segment type="compoundMember">dharma</segment>
<segment type="compoundMember">kṣetre</segment>
</sequence>

However, it has been suggested that it would be preferable not to rely purely on the nesting of tags to indicate the relationship between the sequence and its constituent segments. If this view were to prevail, the two examples above could be rewritten as follows:

C **<sequence>**
<sequenceText type="wordGroup">āsīdrājā</sequenceText>
<sequenceAnalysis>
<segment type="word">āsīt</segment>
<segment type="word">rājā</segment>
</sequenceAnalysis>
</sequence>

and

D **<sequence>**
<sequenceText type="compound">dharmakṣetre</sequenceText>
<sequenceAnalysis>
dharma
kṣetre
</sequenceAnalysis>
</sequence>

The approach adopted in ***A*** and ***B*** has the merit of brevity, and some members of the workgroup consider that the use of nested tags adequately expresses the subordinate relationship between actual text and proposed analysis. Other members favour the more prolix but more explicit approach of ***C*** and ***D***. We would welcome feedback from the Council on this.

Whether the final outcome is ***AB*** or ***CD***, we feel that the general approach outlined in this document gets the balance between the desirably generic and the necessarily specific about right, and we hope that the Council will feel able to endorse it.

John D. Smith
June 28, 2004