A FORMAL COMPUTATIONAL ANALYSIS OF INDIC SCRIPTS

Richard Sproat
University of Illinois at Urbana-Champaign

Introduction

The Brahmi-derived Indic scripts occupy a special place in the study of writing systems. They are *alphasyllabic* scripts (Bright, 1996a) (though Daniels (1996) prefers the term *abugida*), meaning that they are basically segmental in that almost all segments are represented in the script, yet the fundamental organizing principle of the script is the (orthographic) syllable: crudely, an orthographic syllable — *akṣara* — in an Indic script consists of a consonantal core consisting of one or more consonant symbols, with the vowels variously arranged around the core. The one exception to the principle that segments are represented is the *inherent vowel*, usually a schwa or /a/-like phoneme: when the consonantal core is marked with no vowel, it is interpreted as having the inherent vowel (though phonological rules such as Hindi schwa-deletion (Ohala, 1983), can complicate this characterization).

Brahmi-derived scripts are by no means the only scripts that are basically segmental but have the syllable as an organizing principle: notable other scripts are Ethiopic (Haile, 1996) and Korean Hankul (King, 1996); there have been claims that the latter is derived from Indic scripts, though King notes (page 219) that there have been “no less than ten different ‘origin theories’.”[1] But Brahmi scripts are indubitably the most successful, spawning not only several scripts in the Indian subcontinent, but also spreading into Southeast Asia and spawning scripts such as Thai and Lao (Diller, 1996), and even arguably influencing the development in modern times of the Pahawh Hmong messianic script (Smalley, Vang, and Yang, 1990; Ratliff, 1996).

The Indic scripts have some properties that make them interesting from a number of perspectives. For instance, while the individual scripts differ in details, they share a lot of properties; for example, many scripts have encodings for the full set of Sanskrit-derived consonants, and in all the scripts, vowel-initial *akṣara* are written with a full vowel form,[2] whereas in consonant-initial *akṣara* are written with diacritic vowels around the consonant. This is convenient from the point of view of text encoding since it means that to a large extent to transliterate between different Indic scripts, one merely needs to shift the codespace of the text; this is what is done in Unicode (Aliprand, 2003, and `www.unicode.org`). The parallels between scripts can be quite strong indeed: as I shall argue in a later section, the encoding of diacritic vowels in Devanagari and Oriya are identical at the right level of abstraction. Indic scripts also make a much richer use of the two-dimensional layout possibilities afforded by the written medium than do Western alphabetic scripts, since vowels may be written above, below or even before the consonants they logically follow; and in consonant sequences in many scripts—Tamil is a notable exception (Steever, 1996)— all but one of the consonants takes on a reduced form, which may be written either in linear sequence with, or above or below the "primary" consonant (usually either the first or last in the logical sequence).[3] In this paper I will express these two-dimensional layout possibilities in terms of a planar regular calculus.

Finally Indic scripts have some interesting psycholinguistic properties. With the exception of the inherent vowel, vowels are always represented explicitly in the Indic scripts, so in that sense the scripts are segmental. But vowels are clearly subordinate to consonants (except in vowel-initial *akṣara*) since they are written as diacritics on the consonant. As a result, various studies have shown that readers of Indian languages written in Indic scripts have less overall segmental awareness than readers of a language like English. Other studies have also demonstrated psycholinguistic consequences of certain properties of Indic scripts; for example, /i/ in Devanagari, which is written before the consonant it logically follows, causes processing problems in that words that contain /i/ take longer to identify and longer to name (i.e., pronounce aloud) in controlled experiments. I will discuss these issues and cite the relevant studies in a later section.

The goal in this paper is threefold. First, I will introduce a formalism for describing planar arrangements of symbols and apply that formalism to selected phenomena in Indic scripts. Second, I will present in some detail a computational model of Devanagari that was developed as part of a Hindi text-to-speech system. Finally, I will explore some of the results of psycholinguistic research on Indic scripts, and relate them to the formal model previously developed: since the formal model as stated treats Indic scripts as segmental at an abstract level, this potentially has psycholinguistic implications, which we will explore.

Formal Preliminaries

The relation between writing and the linguistic information it encodes can naturally be thought of as a mapping in the algebraic sense. Consider that we have a relation γ, that takes a string of linguistic units σ (say, a sequence of phonemes), and maps it to a graphemic representation,

$\gamma(\alpha)$; we may refer to $\gamma(\alpha)$ as the *image* of α under γ. One question that immediately arises about γ is what kind of relation it is. In (Sproat, 2000) I argue that it is a *regular* relation (Kaplan and Kay, 1994) in the formal sense of regular languages. Regular relations are those relations that can be implemented using *finite-state transducers* (FSTs), for which efficient algorithms are known, (Mohri, 1997; Mohri, Pereira, and Riley, 1998). This conclusion has implications for implementations of systems, such as text-to-speech (TTS) systems that compute the mapping between text and linguistic structure (Sproat, 1997a).

Regular relations, like regular languages, are definable in terms of an alphabet, and a set of operations on that alphabet; see (Hopcroft and Ullman, 1979) for these notions for regular languages. Specifically, regular relations are those relations that can be constructed out of a finite alphabet of *pairs* of symbols using only the operations of *(con)catenation* ($\cdot$), *union* ($\cup$) and *closure* (*). Union and closure are to be understood in their normal set-theoretic sense, and catenation means string concatenation such that for two strings α and β, their catenation ($\alpha \cdot \beta$) is simply the combined string ($\alpha\beta$).

In the analysis of writing systems it is desirable to extend the definition of catenation to include two-dimensional operations. Specifically, I propose five planar catenation operators:

- left catenation $\overset{\rightarrow}{\cdot}$
- right catenation $\overset{\leftarrow}{\cdot}$
- downwards catenation $\overset{\downarrow}{\cdot}$
- upwards catenation $\overset{\uparrow}{\cdot}$
- surrounding catenation $\odot$

The semantics of each of these is illustrated in Figure 1.

Surrounding catenation seems not to be needed for Indic scripts, though it is needed for Chinese; see (Sproat, 2000). However most Indic scripts employ all the other catenators; I will give illustrations in the next section. Note that catenators are associative within the same type so that $\alpha \overset{\downarrow}{\cdot} (\beta \overset{\downarrow}{\cdot} \gamma) = (\alpha \overset{\downarrow}{\cdot} \beta) \overset{\downarrow}{\cdot} \gamma$; however across types this is not the case so that in general $\alpha \overset{\downarrow}{\cdot} (\beta \overset{\rightarrow}{\cdot} \gamma) \neq (\alpha \overset{\downarrow}{\cdot} \beta) \overset{\rightarrow}{\cdot} \gamma$.

The planar catenators are, of course, similar to the traditional way of describing the placement of vowel symbols in Indic scripts, by using a dash to represent the consonant, and by placing the vowel symbol in

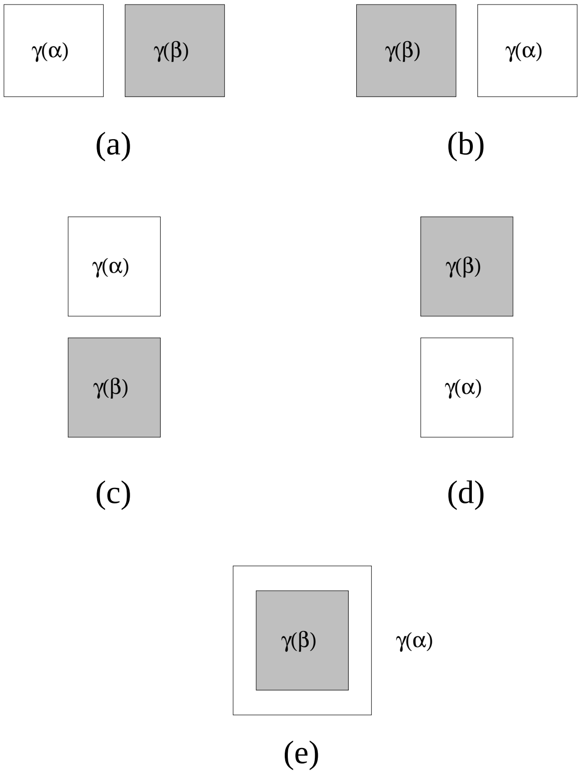

Figure 1: the five planar concatenation operations:
(a) $\gamma(\alpha) \overset{\rightarrow}{\cdot} \gamma(\beta)$; (b) $\gamma(\alpha) \overset{\leftarrow}{\cdot} \gamma(\beta)$; (c) $\gamma(\alpha) \overset{\downarrow}{\cdot} \gamma(\beta)$; (d) $\gamma(\alpha) \overset{\uparrow}{\cdot} \gamma(\beta)$; (e) $\gamma(\alpha) \bigodot \gamma(\beta)$.

the appropriate geometric position relative to the dash, representing the normal position of that vowel when written with a real consonant symbol. There is a difference, however: the catenators are more abstract in that they distinguish only four relative positions, namely left-of, right-of, above and below. Indeed, the specification of the combinatorics of glyphs in terms of the planar catenators is at an intermediate level of abstractness: it is more abstract than the traditional description since it does not give the *exact* placement of glyphs only the general placement relative to other glyphs; on the other hand it is less abstract than a representation that merely says that two symbols are combined (catenated) in a logical order, as is the case in the standard Unicode representation. One may legitimately ask if there is any justification for having such an intermediate level of representation. I believe there is, and I will attempt to justify this later in the paper.

All scripts have an overall, or macroscopic, direction.[4] In Indic scripts the macroscopic direction is left-to-right. In terms of the planar catenators, the catenator encoding the macroscopic direction in Indic scripts is $\overset{\rightarrow}{\cdot}$. Let us term $\overset{\rightarrow}{\cdot}$ the *macroscopic catenator* for Indic scripts. Deviation from the macroscopic direction—choice of a catenator other than the macroscopic catenator—is always local and occurs within what I term the *small linguistic unit* (SLU) in (Sproat, 2000):

> LOCALITY HYPOTHESIS: Changes from the macroscopic catenation type can only occur within a graphic unit that corresponds to a small linguistic unit.

The SLU seems to vary across writing systems, though in many writing systems the SLU is the syllable. In most Indic scripts, the SLU is the orthographic syllable, or *akṣara*.

One other assumption is that when two elements are adjacent in the linguistic representation, their graphical expressions must be adjacent, according to whatever catenator is used. So $\gamma(ab)$ is $\gamma(a) \cdot \gamma(b)$, for some catenator '$\cdot$'. In particular, if a is a sequence, and '$\cdot$' is $\overset{\leftarrow}{\cdot}$, we expect $\gamma(b)$ to be written *before* whatever is written first in $\gamma(a)$. Let us refer to this as the ADJACENCY AXIOM. We can express this formally as follows, where $\alpha \prec_p \beta$ (read 'α immediately path-precedes β') denotes the situation where α precedes and is adjacent to β (Sproat, 2000):

> ADJACENCY AXIOM: If $\alpha \prec_p \beta$ and if $\gamma(\alpha\beta)$ is not idiosyncratically specified to be written with a single symbol or ligature, then
> $\gamma(\alpha\beta) = \gamma(\alpha) \cdot \gamma(\beta)$.

In what follows I will develop some additional formalisms beyond what was introduced above.

Illustrations

I will illustrate the formalism by describing some phenomena in a variety of Indic scripts, starting with Devanagari.

DEVANAGARI

Devanagari (Bright, 1996a) is used to represent several Northern languages, including Hindi, Nepali and Marathi, and is the script most commonly used to represent Sanskrit. For the sake of the present discussion, we will assume its use for Hindi. In the transliterations throughout this paper, I will mostly avoid phonetic symbols. Digraphs representing single phonemes will be enclosed in square brackets; thus '[sh]'. Long vowels will be shown with doubling; thus ‘[ii]’ versus ‘i’. I will use ‘M’ after vowels to represent vowel nasalization.

The script has several features in common with most other Indic scripts:

- Syllable-initial vowels are represented as full symbols, but post-consonantal vowels appear as diacritics; see Table 1. /a/ (phonetically /ə/ in Hindi) is the inherent vowel, and thus has no diacritic mark.

- The orthographic syllable or *akṣara* starts at the left edge of a sequence of consonants, and ends at the right edge of the syllable’s vowel. Consonant sequences are represented orthographically as ligatured groups. Thus स <s> + क <k> yields स्क <sk>; क <k> + ष <[.s]> + म <m> yields क्ष्म <k[.s]m>. Ligatures may involve leftwards catenation or downwards catenation; some ligatures, such as the क्ष <k[.s]> ligature in the previous example are synchronically unanalyzable or ‘suppletive’ (on analogy with morphological suppletion). Diacritic vowels that combine with $\overleftarrow{\cdot}$ (in Devanagari <i>) appear before the entire sequence: /ski/ is represented स्कि <i+sk>

- The *anusvara* symbol, representing a nasalized vowel, or a postvocalic homorganic nasal stop, appears above the rightmost symbol of the *akṣara*: हां <h[aa]M>. Strictly speaking, a purely nasalized vowel should be written with *candrabindu* (thus हाँ <h[aa]M>, which has the same placement as *anusvara*, but this is often not observed.

- Orthographic-syllable-initial /r/ is represented by a superscript symbol (*repha* or *arka*) occurring above the last symbol of the *akṣara*: thus /varm[aa]/ is represented as वर्मा <vm[aa]+r>.

The procedure for marking orthographic syllables, given a string of phonemes, that applies to most Indic scripts (Tamil is an exception) can be stated as follows:

> Divide the phonological string into orthographic syllables by placing a syllable boundary $_{syl}$:
>
> - At the beginning of the word;
> - Between each pair of adjacent (non-tautosyllabic) vowels (as between the /[aa]/ and /[ii]/ in *bhaaii* ‘brother’)

Catenator	Full form	Diacritic form
(Null	अ <a>	क <ka>)
$\overset{\rightarrow}{\cdot}$	आ <[aa]> ओ <o> औ <[au]> ई <[ii]>	का <k[aa]> को <ko> कौ <k[au]> की <k[ii]>
$\overset{\uparrow}{\cdot}$	ए <e> ऐ <[ai]>	के <ke> कै <k[ai]>
$\overset{\downarrow}{\cdot}$	उ <u> ऊ <[uu]> ऋ <[ri]>	कु <ku> कू <k[uu]> कृ <k[ri]>
$\overset{\leftarrow}{\cdot}$	इ <i>	कि <ki>

Table 1: Full and diacritic forms for Devanagari vowels, classified by catenator inherent to the diacritic forms.

- Before the first consonant of a consonant sequence, with the exception of a homorganic nasal, which may optionally be syllabified with the preceding *akṣara*.

To get the ligatured consonants we assume a function Lig, which takes a sequence of consonant symbols and forms a ligatured consonant sequence:

$$\gamma(C_1C_2\ldots C_n) = Lig(\gamma(C_1)\cdot\gamma(C_2)\cdot\ldots\cdot\gamma(C_n))$$

Lexical specifications for certain ligatures, such as क+स = क्ष <k[.s]> will override the general ligature. By default the catenation operator is $\overset{\rightarrow}{\cdot}$, but in certain instances, especially when the full form of the consonant does not have a vertical bar (Bright, 1996a)—e.g., द <d>—the catenation operator may be $\overset{\downarrow}{\cdot}$: द+व becomes द्व <dv>. Unlike the diacritic vowels (see below) where catenation type is purely a property of the vowel, this choice of $\overset{\downarrow}{\cdot}$ seems to be a choice of the pair of consonants. It also seems to be subject to some stylistic variation: one can sometimes combine a pair of consonants in more than one way: ल+ल can be ल्ल or ल्ल <ll>. Sequence-final /r/, is always produced by a reduced form, which we will notate as $\gamma_{red}(\mathrm{r})$, subscripted (or in some cases idiosyncratically fused with) the previous consonant: $\gamma(C_1\mathrm{r}) = Lig(\gamma(C_1)\overset{\downarrow}{\cdot}\gamma_{red}(\mathrm{r}))$ so that, e.g., प+र becomes प्र <pr>.

If $full(V)$ represents the full (syllable-initial) form of V, then the rule for graphically encoding syllable-initial vowels can be given as follows:

$$\gamma(V) \rightarrow full(V)/\ {}_{syl}\underline{\quad}.$$

For postconsonantal vowels, the following general rules apply, where γ_{red} represents the reduced diacritic vowel:

$$\gamma(\mathrm{a}) \rightarrow \emptyset$$
$$\gamma(V) \rightarrow \gamma_{red}(V)$$

That is, /a/ deletes, and all other vowels are represented by their diacritic form. The catenation operator depends upon the vowel: For a consonant sequence κ and vowel υ, the catenation operators are chosen as follows:

$$\begin{array}{ll}
\gamma(\kappa) \overset{\rightarrow}{\cdot} \gamma_{red}(\upsilon) & \text{if } \upsilon = \text{/a,o,[au],[ii]/} \\
\gamma(\kappa) \overset{\uparrow}{\cdot} \gamma_{red}(\upsilon) & \text{if } \upsilon = \text{/e,[ai]/} \\
\gamma(\kappa) \overset{\downarrow}{\cdot} \gamma_{red}(\upsilon) & \text{if } \upsilon = \text{/[uu],u,[ri]/} \\
\gamma(\kappa) \overset{\leftarrow}{\cdot} \gamma_{red}(\upsilon) & \text{if } \upsilon = \text{/i/}
\end{array}$$

The vowels /e,[ai],o,[au]/ allow a deeper analysis, to which we will return in the next section.

Anusvara and *candrabindu* are simply conjoined with their logical predecessor using the $\overset{\uparrow}{\cdot}$ operator:

$$\gamma(xM) \rightarrow \gamma(x) \overset{\uparrow}{\cdot} \gamma(M)$$

Finally, *repha*/*arka* can be described as follows: Given an orthographic syllable starting with /r/ and remainder κ starting with a *non-null* consonant sequence:

$$\gamma(\mathrm{r}\kappa) = arka \overset{\downarrow}{\cdot} \gamma(\kappa).$$

This places the *arka* above the remainder of the *akṣara*; to place it on the righthand side of the *akṣara* we assume a general stylistic principle of Devanagari that prefers to place superscripted materials to the righthand edge of the *akṣara*. We will see in the next section that this is useful beyond this one case.

Catenator	Example				
$\overset{\downarrow}{\cdot}$	ଷ <[sh]>	+	ପ <p>	=	ଷ୍ପ <[sh]p>
	ଣ <[.n]>	+	ଠ <[.th]>	=	ଣ୍ଠ <[.n][.th]>
	ସ <s>	+	ତ <t>	=	ସ୍ତ <st>
	କ <k>	+	ର <r>	=	କ୍ର <kr>
$\overset{\uparrow}{\cdot}$	ଦ <d>	+	ବ <b>	=	ଦ୍ବ <bd>
	ଙ <[ng]>	+	କ <k>	=	ଙ୍କ <[ng]k>
	ତ <t>	+	ସ <s>	=	ତ୍ସ <ts>
	ର <r>	+	କ <k>	=	ର୍କ <rk>
$\overset{\rightarrow}{\cdot}$	ପ <p>	+	ଯ <p>	=	ପ୍ୟ <py>
(fused)	କ <k>	+	ଷ <[sh]>	=	କ୍ଷ <k[sh]>

Table 2: Catenators used for Oriya consonant sequences. In many of the above cases, the glyphs take on a reduced form, and ligaturing applies. Note the representation of <rk> with a superscript reduced <r>, as in Devanagari. <k[sh]> is represented with a fused glyph, also as in Devanagari.

ORIYA

Oriya (Mahapatra, 1996) is a Northern Indic script distantly related to Devanagari (Salomon, 1996), though quite different in appearance. The basic design of the Oriya script is of course identical to that of Devanagari: syllable-initial vowels are represented with their own symbols, otherwise dependent vowel symbols are used, so that the orthographic syllable parsing procedure given above for Devanagari also works for Oriya. Consonant sequences are represented by combining the single consonants using various catenators and often involving ligaturing the consonants together.

Oriya consonant sequences are complex and there is much sequence-specificity in the choice of catenators (often $\overset{\downarrow}{\cdot}$ or $\overset{\uparrow}{\cdot}$, but sometimes $\overset{\rightarrow}{\cdot}$), as well as the degree of fusion. Simple cases include ଷ <[sh]> plus ପ <p>, which becomes ଷ୍ପ <[sh]p>; and ଣ <[.n]> plus ଠ <[.th]> which becomes ଣ୍ଠ <[.n][.th]>; both using the $\overset{\downarrow}{\cdot}$ operator. Various examples are given in Table 2.

The diacritic vowels are more orderly. Oriya simplex diacritic vowels are shown in Table 3. The grammar for these vowels is similar to that for Devanagari, except that /i/ uses $\overset{\uparrow}{\cdot}$ rather than $\overset{\leftarrow}{\cdot}$, and /e/ uses $\overset{\leftarrow}{\cdot}$ rather than $\overset{\uparrow}{\cdot}$:

Catenator	Diacritic form	
(Null	କ	<ka>)
$\overset{\rightarrow}{\cdot}$	କା କୀ	<k[aa]> <k[ii]>
$\overset{\uparrow}{\cdot}$	କି	<ki>
$\overset{\downarrow}{\cdot}$	କୂ କୁ କୃ	<k[uu]> <ku> <k[ri]>
$\overset{\leftarrow}{\cdot}$	କେ	<ke>

Table 3: Forms for simplex diacritic Oriya vowels, classified by catenator.

$$\begin{array}{ll}
\gamma(\kappa) \overset{\rightarrow}{\cdot} \gamma_{red}(\upsilon) & \text{if } \upsilon = \text{/a,[ii]/} \\
\gamma(\kappa) \overset{\uparrow}{\cdot} \gamma_{red}(\upsilon) & \text{if } \upsilon = \text{/i/} \\
\gamma(\kappa) \overset{\downarrow}{\cdot} \gamma_{red}(\upsilon) & \text{if } \upsilon = \text{/[uu],u,[ri]/} \\
\gamma(\kappa) \overset{\leftarrow}{\cdot} \gamma_{red}(\upsilon) & \text{if } \upsilon = \text{/e/}
\end{array}$$

The three vowels, /o, [ai], [au]/ are expressed using combinations of symbols requiring more than one catenator. Thus <ko> is a combination of diacritic <e> and diacritic <[aa]>, thus କୋ ; <k[ai]> is a combination of diacritic <e> and a new superscript diacritic, thus କୈ; and <k[au]> is a combination of all three, thus କୌ.[5] Such complex vowels are common in Indic scripts (as well as Indic-derived scripts of Southeast Asia, such as Thai), and Devanagari is more the exception than the rule in seemingly lacking them.

Let us assume that the complex vowels are represented as a set of symbols each with their own catenator; furthermore let us assume that the glyph ୈ for <[ai]> and <[au]> represents a 'diphthong' feature DIPH (since historically at least, these vowels were diphthongs). Then:

$$\begin{array}{lcl}
\text{<[ai]>} & = & \{\gamma_{red}(\text{e}),\gamma(\text{DIPH})\} \\
\text{<o>} & = & \{\gamma_{red}(\text{e}), \gamma_{red}([\text{aa}])\} \\
\text{<[au]>} & = & \{\gamma_{red}(\text{e}),\gamma(\text{DIPH}),\gamma_{red}([\text{aa}])\}
\end{array}$$

Let us further assume that the catenation of the elements of a symbol set is defined according to the following axiom, where '$\cdot_i$' is the catenator for the ith set member:

$$\boxed{\kappa \cdot \{w_1, w_2 \ldots w_n\} = \kappa \cdot_1 w_1 \wedge \kappa \cdot_2 w_2 \ldots \wedge \kappa \cdot_n w_n}$$

In other words, the catenation of a symbol κ with a symbol set is simply the conjunction of the catenation of that symbol with the individual symbols of the set. Then following this axiom we have, for <k[au]>:

$$\begin{array}{lcl}
\gamma(\mathrm{k}) \cdot \gamma([\mathrm{au}]) & = & \\
\gamma(\mathrm{k}) \cdot \gamma_{\mathrm{red}}([\mathrm{au}]) & = & \\
\gamma(\mathrm{k}) \cdot \{\gamma_{\mathrm{red}}(\mathrm{e}), \gamma(\textsc{diph}), \gamma_{\mathrm{red}}([\mathrm{aa}])\} & = & \\
\gamma(\mathrm{k}) \overset{\leftarrow}{\cdot} \gamma_{\mathrm{red}}(\mathrm{e}) \wedge \gamma(\mathrm{k}) \overset{\uparrow}{\cdot} \gamma(\textsc{diph}) \wedge \gamma(\mathrm{k}) \overset{\rightarrow}{\cdot} \gamma_{\mathrm{red}}([\mathrm{aa}]) & = & \text{କୌ}
\end{array}$$

If we return to Devanagari, we can see that exactly the same grammar works for the vowels /o,[ai],[au]/ as it does in Oriya. Starting with <ke> के we can suggest that the second stroke in <k[ai]> कै encodes DIPH, so that <k[ai]> involves a combination of <e> and DIPH. Similarly, <ko> is just a combination of <e> and <a>. Here, the general stylistic principle of Devanagari that we appealed to above that places superscripts at the right edge of the *akṣara* will guarantee that the <e> stroke will appear above the <[aa]> rather than the <k>: को. Finally, <au> is just a combination of <e>, <[aa]> and DIPH, with the placement of the <e>+DIPH complex above the <[aa]> being again due to the stylistic principle: कौ. Thus, in detail:

$$\begin{array}{lcl}
\gamma(\mathrm{k}) \cdot \gamma([\mathrm{au}]) & = & \\
\gamma(\mathrm{k}) \cdot \gamma_{\mathrm{red}}([\mathrm{au}]) & = & \\
\gamma(\mathrm{k}) \cdot \{\gamma_{\mathrm{red}}(\mathrm{e}), \gamma(\textsc{diph}), \gamma_{\mathrm{red}}([\mathrm{aa}])\} & = & \\
\gamma(\mathrm{k}) \overset{\uparrow}{\cdot} \gamma_{\mathrm{red}}(\mathrm{e}) \wedge \gamma(\mathrm{k}) \overset{\uparrow}{\cdot} \gamma(\textsc{diph}) \wedge \gamma(\mathrm{k}) \overset{\rightarrow}{\cdot} \gamma_{\mathrm{red}}([\mathrm{aa}]) & = & \text{कौ}
\end{array}$$

This decomposition also carries over to the Devanagari full form vowels. Thus <o> is (full) <[aa]> plus diacritic <e>: ओ. <[ai]> is (full) <e> plus DIPH: ऐ. And <[au]> is (full) <[aa]> plus diacritic <e> plus DIPH: औ.[6] Note that this explains the otherwise puzzling fact that for full <[ai]> ऐ, there is only one stroke above the ए, whereas for all other 'diphthongs' there are two: in this case the single stroke represents the DIPH feature, and the second stroke, which would normally represent diacritic <e> is unnecessary since we already have the full form <e>. This system for representing <e>, <o>, <ai> and <au>, would seem to have been inherited from Brahmi; see (Salomon, 1996, Table 30.2, page 374).

KANNADA

Kannada is a Southern Brahmi-derived script used to write the Kannada language (Spencer, 1950; Bright, 1996b). The underlying structure of the script is again, essentially the same as other Indic scripts, though the surface details differ. Consonant clusters are written in a much more regular fashion than Devanagari or Oriya, with the typical mode of expression being to use the $\overset{\downarrow}{\cdot}$ catenator to place a reduced form of a non-initial consonant underneath the consonant it logically follows: thus ಲ <l> becomes ಲ್ಲ <l>, so that $\gamma(\mathrm{ll}) = \gamma(\mathrm{l}) \overset{\downarrow}{\cdot} \gamma_{\mathrm{red}}(\mathrm{l})$. If there are more than two consonants in the sequence, then the subsequent consonants are not stacked, but are written in a left-to-right sequence following the reduced and subscripted second consonant: ಲಕ್ಷ್ಮಣ <lak[.s]ma[.n]a> 'Lakshmana' (Spencer, 1950, page 99). This motivates a general rule for subscripted sequences whereby in a sequence of $\overset{\downarrow}{\cdot}$ catenations, all but the initial catenator are changed to $\overset{\rightarrow}{\cdot}$:

KANNADA LINEARIZATION: $\overset{\downarrow}{\cdot} \rightarrow \overset{\rightarrow}{\cdot}$ / $\overset{\downarrow}{\cdot}$ __

(Note that this rule must apply right-to-left.) Thus the $k[.s]m$ sequence is analyzed as $\gamma(\mathrm{k}) \overset{\downarrow}{\cdot} (\gamma([.\mathrm{s}]) \overset{\rightarrow}{\cdot} \gamma(\mathrm{m}))$.

Anusvara and *arka* are written at the end of their *akṣaras* but are written inline; thus in Kannada *anusvara* uses $\overset{\rightarrow}{\cdot}$ since it is logically at the end; and *arka* uses $\overset{\leftarrow}{\cdot}$ since it is logically at the beginning:[7]

ತ <t> becomes ರ್ತ <rt>:	$\gamma(\mathrm{rta}) = \mathrm{arka} \overset{\leftarrow}{\cdot} \gamma(\mathrm{ta})$
ತ <t> becomes ತಂ <t(a)M>:	$\gamma(\mathrm{taM}) = \gamma(\mathrm{ta}) \overset{\rightarrow}{\cdot} \gamma(\mathrm{M})$

The diacritic vowel system is also simplified. The catenator $\overset{\leftarrow}{\cdot}$ is not used for vowels, which are limited to $\overset{\uparrow}{\cdot}$ and $\overset{\rightarrow}{\cdot}$, with $\overset{\downarrow}{\cdot}$ being used only for <[ri]>, <[rii]> and the lower symbol of <ai> (a complex symbol). Table 4 gives the analysis of Kannada diacritic vowels in terms of these three catenators. I also include the symbol sets for complex vowels, where LONG is used for a glyph that seems to be associated only with long vowels.[8] With consonant sequences the vowels are standardly described as being associated with the first (inline) consonant (Bright, 1996b). According to the ADJACENCY AXIOM, vowels that are written above ($\overset{\uparrow}{\cdot}$) the *akṣara* must appear above the first consonant, which they do, consistent with the standard description. For vowels that have an $\overset{\rightarrow}{\cdot}$ or $\overset{\downarrow}{\cdot}$ component, these components must appear to the right or below the vertically arranged consonant cluster: so either $\gamma(C_1C_2\ldots C_n) \overset{\rightarrow}{\cdot} \gamma(V)$ or $\gamma(C_1C_2\ldots C_n) \overset{\downarrow}{\cdot} \gamma(V)$. Thus we have (with

Catenator	Diacritic form	
(Null	ಕ	<ka>)
$\overset{\rightarrow}{\cdot}$	ಕು	<ku>
	ಕಾ	<k[aa]>
	ಕೂ	<k[uu]>
$\overset{\uparrow}{\cdot}$	ಕಿ	<ki>
	ಕೆ	<ke>
	ಕೌ	<k[au]>
$\overset{\downarrow}{\cdot}$	ಕೃ	<k[ri]>
	ಕೄ	<k[rii]>
{i ($\overset{\uparrow}{\cdot}$), LONG ($\overset{\rightarrow}{\cdot}$)}	ಕೀ	<k[ii]>
{e ($\overset{\uparrow}{\cdot}$), [uu] ($\overset{\rightarrow}{\cdot}$)}	ಕೊ	<ko>
{e ($\overset{\uparrow}{\cdot}$), [ai] ($\overset{\downarrow}{\cdot}$)}	ಕೈ	<k[ai]>
{e ($\overset{\uparrow}{\cdot}$), LONG ($\overset{\rightarrow}{\cdot}$)}	ಕೇ	<k[ee]>
{e ($\overset{\uparrow}{\cdot}$), [uu] ($\overset{\rightarrow}{\cdot}$), LONG ($\overset{\rightarrow}{\cdot}$)}	ಕೋ	<k[oo]>

Table 4: Forms for diacritic Kannada vowels, classified by catenator. Complex vowels are given at the bottom of the table. Note that <[rii]> only occurs in Kannada transliterations of Sanskrit (Bill Bright, p.c.).

$\overset{\rightarrow}{\cdot}$) ಕಣ್ಣೀರು <ka[.n][.n][ii]ru> ‘tears’ and ಲಕ್ಷ್ಮೀಶ <lak[.s]m[ii][.s]a> ‘husband of Laxmi (Vishnu)’ (Spencer, 1950, pages 29, 341); here the right component of <[ii]> appears after the consonant sequence. Similarly (with $\overset{\downarrow}{\cdot}$) we have ಕ್ರೈಸ್ತ <kr[ai]sta> ‘Christian’, ಏಶ್ವೈಕ್ಯ <[ee][sh]v[ai]kya> ‘universal unity’ (Spencer, 1950, page 156, 343); where the lower component of <[ai]> appears after the sequence of lowered consonants. Note in particular the application of KANNADA LINEARIZATION to the righthand component of <[ai]>. It is worth observing that while this arrangement is consistent with the ADJACENCY AXIOM, it contradicts the standard description of the arrangement of vowels with consonant sequences.

TAMIL

From a formal point of view, Tamil (Steever, 1996; Radhakrishnan, 2002) is probably the simplest Indic script since it has eliminated the *akṣara* as a fundamental unit of the script. Virtually all consonant conjuncts have been eliminated, and consonant sequences are written left-to-right with unligatured consonant symbols. Steever (1996, page 426) suggests that this trend may have been related to the introduction of typography from the West. However, the vowels behave as they do in other Indic scripts: syllable-initial vowels are written in a full form, whereas postconsonantal vowels are written with a diacritic form, with various of the four catenators being used, depending upon the vowel.

Given that consonant sequences have been completely linearized in Tamil, with all consonant sequences being constructed using the macroscopic catenator $\overset{\rightarrow}{\cdot}$, there would seem to be no motivation for maintaining that the SLU is the orthographic syllable. Rather the SLU in Tamil is a maximally CV(V) unit, with the possibilities being V(V) (syllable-initial short or long vowel), C (consonant-sequence-medial consonant) and CV(V) (consonant with diacritic short or long vowels). Ignoring for the moment the decomposability of CV(V) elements and the fact that there are simplex vowel glyphs representing both long and short vowels, Tamil orthography is almost what I termed in (Sproat, 2000) a *core syllabary*, like Japanese *kana*. Tamil is moving towards being an alphasyllabic version of *kana*.

If the SLU is the CV(V) unit, not an orthographic syllable, the theory predicts that vowel diacritics in Tamil can only attach to the final consonant in a consonant sequence. In particular, a vowel that uses the $\overset{\leftarrow}{\cdot}$ catenator must appear before the last consonant of a sequence, and cannot, for example appear before the initial consonant, as it would in Devanagari or Oriya. This is because appearance before a non-final consonant would require deviation from the macroscopic order in a unit larger than the SLU. This prediction is correct: thus ஹெக்டே <hek[.t][ee]> is written with the [ee] symbol appearing before the ட <[.t]>.

A procedure that would be appropriate for dividing Tamil words into SLU's would be as follows:

> Divide the phonological string into orthographic units by placing a unit boundary ']':
>
> - At the beginning of the word;
> - Between each pair of adjacent vowels;
> - Before every consonant of a consonant sequence.

The macroscropic catenator in Tamil is $\overset{\rightarrow}{\cdot}$, as in other Indic scripts.

The simplex Tamil vowel diacritics are shown in Table 5. It should be noted that while the forms with <p> are fairly transparent, many of the consonant-vowel combinations involve more or less opaque ligatures so

Catenator	Diacritic form
(Null	ப <pa>)
$\overset{\rightarrow}{\cdot}$	பா <p[aa]> பி <pi>
$\overset{\uparrow}{\cdot}$	பீ <p[ii]>
$\overset{\downarrow}{\cdot}$	பு <pu> பூ <p[uu]>
$\overset{\leftarrow}{\cdot}$	பெ <pe> பே <p[ee]> பை <p[ai]>

Table 5: Forms for simplex diacritic Tamil vowels, classified by catenator.

that, for instance <[.n]> is ண, whereas <[.n][uu]> is ணூ, which is quite different from what one might expect on the basis of பூ <puu>. Tamil seems to be evolving towards *kana* in another regard, namely that the CV(V) units are becoming unanalyzable.

The diacritic forms of <o>, and <[au]>, involve the same symbol sets as in Devanagari and Oriya, and in addition Tamil has <[oo]>, which is formed on the basis of <[ee]> (these latter two invented by missionaries, Bill Bright, p.c.):

<o>	=	$\{\gamma_{red}(\mathrm{e}), \gamma_{red}([\mathrm{aa}])\}$	பொ	<po>
<[au]>	=	$\{\gamma_{red}(\mathrm{e}),\gamma(\textsc{diph}),\gamma_{red}([\mathrm{aa}])\}$	பௌ	<p[au]>
<[oo]>	=	$\{\gamma_{red}([\mathrm{ee}]), \gamma_{red}([\mathrm{aa}])\}$	போ	<p[oo]>

Note that in Tamil the DIPH glyph is ligatured to the beginning of the <[aa]> glyph, though we assume that it is specified as catenating with $\overset{\rightarrow}{\cdot}$ to the preceding consonant.

Sequence-medial consonants are marked with a dot—*puḷḷi*—that ‘cancels’ the inherent vowel; this has the same function as, say, *virāma*[9] in Devanagari, but is more common in Tamil since vowellessness cannot be encoded by membership in a conjunct. Thus note the dot above the <k> (க) in ஹெக்டே <hek[.t][ee]>. The *puḷḷi* can be placed by rule as follows:

$$\gamma(\mathrm{C}) \rightarrow \gamma(\mathrm{C}) \overset{\uparrow}{\cdot} \cdot / \_\_]$$

A similar rule (replacing $]$ with $_{syl}$) would be required to encode cancellation signs in other Indic scripts.

Computational analysis of Devanagari

As part of a Hindi text-to-speech system that we were developing at Bell Labs (Narasimhan, Sproat, and Kiraz, 2003; Sproat, 1997b) (and see (Bhaskararao and Mathew, 1992; Bhaskararao, Peri, and Updikar, 1994) for other work on Hindi TTS), we developed a practical implementation of Hindi orthography as part of the text analysis component of the system. The task of a text-analysis component of a TTS system is to map from text into a linguistic representation; it can thus be thought of as a computational model of what humans do when they read aloud. The Hindi text-analysis component transformed raw text into a phonological representation, including as part of that process the expansion of abbreviations, numbers and time expressions, as well as phonological phenomena such as schwa deletion. All of this was implemented in terms of the FST-based approach to text-analysis outlined in (Sproat, 1997a; Sproat, 2000).

The first component of the text-analysis module converted from Devanagari text into an abstract orthographic representation, where symbols are left-catenated together in their logical order. Thus वर्मा would be represented as <varm[aa]>, and हिंदी would be represented as <hiNdii>. Note that while this representation could also be thought of as phonological (Hindi orthography being fairly close to a phonemic representation), phonological phenomena such as schwa-deletion are not represented at this level.

Input text was represented using the mock ISO-8859 *xdvng* encoding, for use with the X window system under Unix.[10] The fonts were designed to work under an ordinary browser with no special Devanagari-specific rendering. Super- and subscripted vowels are handled in this scheme by having the corresponding character codes appear in their logical order (as in Unicode), but with the glyphs so designed that they will over- or understrike the previous consonant. The placement of super- or subscripts was thus technically correct if not always esthetically pleasing. For symbols that are massively 'out of order', such as short <i> or preconsonantal <r> (*arka*), these would be coded in their surface orthographic order rather than their logical order. The consonant symbols are mainly represented in the font without the vertical bar, so that consonant sequences can be (mostly) represented by simply catenating the symbols together, reserving a vertical bar (which is a separate glyph with its own code) for the last consonant.

We adopted a generative approach to the problem in that we started with the abstract orthographic representation and built FSTs that transduced from that level to the surface orthographic realization. Since FSTs are closed under *inversion* (Kaplan and Kay, 1994), we merely need to invert this model to get something that will map from the surface orthography into the abstract representation from which subsequent linguistic processing can then take place.

The first stage of the mapping involved parsing the input string into orthographic syllables (*akṣara*). This was handled using the following rewrite rule, which simply introduces a syllable boundary $[_\sigma$ between a vowel and any following consonant:[11]

ORTHSYLL: $\epsilon \rightarrow [_\sigma / V\_\_C$

Here ϵ represents the empty string per convention. Note that context-dependent rewrite rules have been known since the 1970's to be implementable by FSTs (Johnson, 1972), and practical algorithms have been developed in the last decade or so (Kaplan and Kay, 1994; Mohri and Sproat, 1996).

The orthographic syllable boundary introduced by ORTHSYLL plays an important role in various processes. For example, the placement of <i> is handled by a pair of rules, one which converts a postconsonantal <i> into a 'trace' (i), and a second that inserts a precursor for the diacritic (non-full-form) <i> (i_{diacr}) after an orthographic syllable boundary, before a consonant sequence that is followed by the <i> trace:

I-TRACE: $\mathrm{i} \rightarrow (\mathrm{i}) / C\_\_$
ORTH-I: $\epsilon \rightarrow \mathrm{i}_{diacr} / [_\sigma\_\_C^+(\mathrm{i})$

Note that the I-TRACE is needed since even though the <i> is written before the consonant sequence, the consonant sequence still behaves as if there is a vowel after it: the rule VBAR that later inserts a vertical bar on the final consonant of the sequence is sensitive to the presence of a vowel in order to find the end of the sequence.

In a similar way, preconsonantal <r> is positioned by a pair of rules to the end of the orthographic syllable, as follows:

R-TRACE: $\mathrm{r} \rightarrow (\mathrm{r}) / [_\sigma\_\_C$
ORTH-R: $\epsilon \rightarrow \mathrm{r}_{diacr} / (\mathrm{r})C^+V^*\_\_[_\sigma$

In the case of <r> there is no reason to preserve the trace, so this can be automatically deleted.

The two-stage process involving deletion and insertion can be viewed as a poor-man's implementation of the directional catenators introduced in the previous discussion, necessitated by the fact that the FST toolkit used in the implementation effectively encodes only one catenation operator. Thus a more faithful implementation would select $\overleftarrow{\cdot}$ for <i> and leave the relative order of the glyphs up to the final rendering.

Postconsonantal vowels other than /i/ are spelled out in their diacritic form:

$$
\begin{array}{llllll}
\text{a} & \rightarrow & \text{a}_{diacr} & / & [_{\sigma}C^{+}\_\_ \\
\text{[aa]} & \rightarrow & \text{[aa]}_{diacr} & / & [_{\sigma}C^{+}\_\_ \\
\text{[ii]} & \rightarrow & \text{[ii]}_{diacr} & / & [_{\sigma}C^{+}\_\_ \\
\text{u} & \rightarrow & \text{u}_{diacr} & / & [_{\sigma}C^{+}\_\_ \\
\text{[uu]} & \rightarrow & \text{[uu]}_{diacr} & / & [_{\sigma}C^{+}\_\_ \\
\text{e} & \rightarrow & \text{e}_{diacr} & / & [_{\sigma}C^{+}\_\_ \\
\text{[ai]} & \rightarrow & \text{[ai]}_{diacr} & / & [_{\sigma}C^{+}\_\_ \\
\text{o} & \rightarrow & \text{o}_{diacr} & / & [_{\sigma}C^{+}\_\_ \\
\text{[au]} & \rightarrow & \text{[au]}_{diacr} & / & [_{\sigma}C^{+}\_\_ \\
\text{[ri]} & \rightarrow & \text{[ri]}_{diacr} & / & [_{\sigma}C^{+}\_\_
\end{array}
$$

Nasalized vowels are marked with *candrabindu*, though this is supposed to be replaced by *anusvara* if the vowel diacritic is a superscript. Thus we have पुँ for <puM> with ँ *candrabindu*, but पें with ं *anusvara* for <peM>.[12] We had the following rules for nasalized vowels:

$$
\begin{array}{llllll}
\text{aM} & \rightarrow & \text{a}_{diacr}\text{[candrabindu]} & / & [_{\sigma}C^{+}\_\_ \\
\text{[aa]M} & \rightarrow & \text{[aa]}_{diacr}\text{[candrabindu]} & / & [_{\sigma}C^{+}\_\_ \\
\text{[ii]M} & \rightarrow & \text{[ii]}_{diacr}\text{[anusvara]} & / & [_{\sigma}C^{+}\_\_ \\
\text{uM} & \rightarrow & \text{u}_{diacr}\text{[candrabindu]} & / & [_{\sigma}C^{+}\_\_ \\
\text{[uu]M} & \rightarrow & \text{[uu]}_{diacr}\text{[candrabindu]} & / & [_{\sigma}C^{+}\_\_ \\
\text{eM} & \rightarrow & \text{e}_{diacr}\text{[anusvara]} & / & [_{\sigma}C^{+}\_\_ \\
\text{[ai]M} & \rightarrow & \text{[ai]}_{diacr}\text{[anusvara]} & / & [_{\sigma}C^{+}\_\_ \\
\text{oM} & \rightarrow & \text{o}_{diacr}\text{[anusvara]} & / & [_{\sigma}C^{+}\_\_ \\
\text{[au]M} & \rightarrow & \text{[au]}_{diacr}\text{[anusvara]} & / & [_{\sigma}C^{+}\_\_
\end{array}
$$

Full vowels are encoded using a set of rules similar to the above rules for diacritic vowels, except that the rules are context independent (the post-consonantal vowels having already been taken care of by the rules above), and the full rather than the diacritic form is used. Rules were also included for English open-o /ɔ/ (full form ऑ).

Additional rules translated consonants into their orthographic encodings, deleted the orthographic syllable boundary, and optionally encoded homorganic nasals as *anusvara* so that पेन्सिल <pensil> 'pencil' could also be encoded as पेंसिल <peMsil>.

A final set of rules converts the abstract specifications of glyphs as output by the rules above into actual *xdvng* glyphs. Consonants are divided into two classes, those that need a vertical bar to be complete in final position (e.g. प <p>, ब <b>, भ <[bh]>, न <n>), and those that do not (e.g. ट <ṭ>, ठ <[ṭh]>)

Non-VBAR consonants, including consonant sequences ending in <r> or <y> (which cannot include any following consonant) can be translated directly into glyphs. Thus, for example:

pr	→	प्र
sr	→	स्र
ṭk	→	ट्क
ḍy	→	ड्य

A vertical bar is inserted between VBAR consonants and a following vowel, which includes I-TRACE. Within consonant sequences, no vertical bar is inserted, which produces the appropriate ligatured forms in most cases. Thus <skr> would appear as स्क्र and <st> as स्त, with the vertical bar attached to the <t> in the latter case. Special ligatures such as <kṣ> क्ष are also produced by rule.

Vowels are either spelled out in their full form or diacritic form depending upon the output of the abstract rules above; the same is true for pre-consonantal <r>.

The complete set of rules in *lextools* format[13] can be made available upon request.

Psycholinguistic Implications

The formal model developed above explicitly treats Indic scripts as segmental at some level of analysis. Nevertheless, they are interestingly different from Western segmental scripts in that they make far greater use of two-dimensional layout in arranging their symbols, the vowel symbols are largely subordinate to the consonant symbols, and there is often a great deal of fusion between different segments which tends to render the *akṣara*, or various portions thereof, rather opaque. What do these similarities and differences with Western segmental scripts imply about how Indic scripts are understood and processed by native readers?

Much has been made in the literature on writing and language on the issue of “segmental awareness” and its relation to the use of alphabetic scripts. It has been claimed that people who learn writing systems based on non-alphabetic scripts, including morphosyllabic systems like Chinese, or quasi-syllabic systems like Japanese kana, are not able to perform certain tasks that require awareness of phonological structure at the level of segments. For example, *phoneme reversal*, whereby one takes an input like /poki/ and one is required to reverse the vowels to produce /piko/, is very hard for even literate speakers of Japanese to do, though it is quite easy for literate speakers of English. In contrast, the ability of speakers to manipulate syllables — e.g. *syllable reversal* whereby /poki/ becomes /kipo/ — is unaffected by the writing system one learns, and can even be handled by illiterates; see, e.g., (Prakash et al., 1993).

An obvious question that comes up in this context is where one draws the line between what one considers ‘alphabetic’ and what is ‘non-alphabetic’. Indic scripts are particularly interesting in this regard since they are clearly segmental in their abstract design, and yet the (orthographic) syllable plays an important role; hence the commonly used term *alphasyllabary* (Bright, 1996a). More importantly, perhaps, they are frequently taught as syllabaries (Karanth, 2003)[14] something that would surely affect literate speakers’ conscious phonological awareness.

The strongest position on this issue is perhaps the one taken by Faber (Faber, 1992). Faber explicitly argues that phonemic segmentation is an epiphenomenon due to alphabetic writing (page 111):

> . . . investigations of language use suggest that many speakers do not divide words into phonological segments unless they have received explicit instruction in such segmentation comparable to that involved in teaching an alphabetic writing system.

She ties this claim in with a point made earlier on the same page to the effect that recent phonological theories have argued against a primitive status for segments:

> There is by now a large and convincing body of evidence that linguistic units representing acoustic or articulatory steady states need not be included as primitives in linguistic representations of phonological structure.[15]

We first need to observe that Faber is raising two quite distinct issues here. The first relates to the issue of speakers’ *conscious* awareness of their language, and their ability to consciously manipulate phonological units of a particular kind. The second relates to speakers *unconscious* knowledge of language, and phonological and phonetic theories of that knowledge. The two are not necessarily connected. It is perfectly conceivable, for example, that segments in the traditional structuralist linguistic sense are a primitive of people's unconscious phonological system, yet they do not become consciously aware of segments unless explicitly trained (any more than untrained people are conscious of how their articulators make sounds).

Yet, one activity that would surely seem to require a conscious awareness of segments is the design of a segmental writing system, so if conscious awareness of segments is an epiphenomenon we would seem to have a classic chicken-and-egg problem: how were segmental systems invented in the first place? For the Greek alphabet, the main question is how the vowel symbols developed, since the consonants were already provided by the Phoenician precursor. Faber's explanation (Faber, 1992, page 126) is that the vowels were a *mis*interpretation (emphasis hers) of several of the Phoenician consonant symbols as vowel symbols, due to the misperception of consonant sounds not found in Greek.[16] According to this story, then, the Greek alphabet is the accident that happened just once. But what about various other scripts including the alphasyllabic scripts of South Asia and Ethiopia, as well as Hankul? In all of these scripts the (orthographic) syllable plays a key role in the arrangement of symbols, yet the design of the system is basically segmental in nature. Faber defines these cases away by claiming that only scripts where consonants and vowels are both represented (thus Semitic scripts fail), are on a par (thus Indic scripts fail), and are linearly arranged (thus Hankul fails), count as alphabetic and thus segmental. So, in Faber's classification, Indic scripts do not count as alphabetic.

Now, there is certainly evidence that people who were taught to read Indic scripts (and are not literate in a purely alphabetic script such as English), are less segmentally aware than their counterparts in places where alphabetic scripts are used. This in turn is in keeping with Faber's expectations.

Consider, for example, work by Padakannaya (2000) on Kannada, where he experimented with a variety of tasks that tested the abilities of children of different ages to manipulate the sounds of words. These tasks included recognition or manipulation of syllables:

- Rhyme recognition: determining whether two words rhyme or not;
- Syllable deletion;
- Syllable reversal;

as well as tasks that involved recognition or manipulation of segments:

- Phoneme oddity: given four stimuli, determine the one stimulus for which there is a different phoneme at a specified position. Thus for nonsense words *chota*, *beti*, *kale*, *mito*, the odd one out is *kale* since there is an /l/ rather than a /t/ in the third position;
- Phoneme deletion;
- Phoneme reversal.

Padakannaya compared two sets of children in school grades I through VII. The first set were sighted children learning the standard Kannada alphasyllabary, and the second were blind children who were learning a purely alphabetic Kannada braille.

On syllable manipulation tasks there was no significant difference found between the two sets of children. So for example, except for first grade, where blind children outperformed sighted children, blind children and sighted children were identical in their performance on syllable deletion. On the other hand, for phoneme manipulation, blind children, exposed to the alphabetic script, consistently outperformed sighted children. For example, Figure 2 shows the performance in terms of percent correct for the phoneme reversal task for both populations of children from grades I through VII. For the first three grades, where the blind children steadily improve in performance, the sighted children are not able to do the task at all. In fourth grade sighted children start to be able to do the task, but it is not until fifth grade that they start to catch up with the blind children's ability; not coincidentally, it is in the fifth grade that English is first introduced into the curriculum. Results showing similarly lower performance for segment manipulation tasks by readers of Indic scripts are reported in (Prakash et al., 1993; Padakannaya, 1999; Karanth, 2002), inter alia.

On the other hand, as we can see even in Figure 2, segmental manipulation is not a categorical ability: by fourth grade, before they started learning English, the sighted children had developed some, albeit weak, ability to reverse phonemes. One factor that seems to affect phonemic awareness with readers of Indic scripts is how "separable" or "noticeable" particular glyphs are are (Prakash et al., 1993, and Prakash Padakannaya, p.c.). For example, Padakannaya and his colleagues (1993, page 65) note that their Hindi subjects were 95% successful in a phoneme deletion task at deleting the /d/ in *doshii* दोशी <do[sh][ii]> to yield *osh[ii]*. Here, the द <d> forms a separate glyph with an additional symbol to represent the vowel. On the other hand,

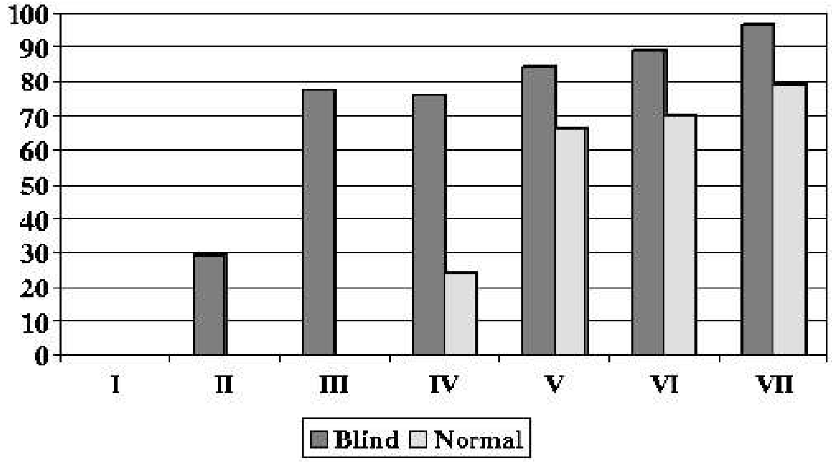

Figure 2: Phoneme reversal in sighted and blind Kannada children (the blind children learning an alphabetic Braille), from (Padakannaya, 2000), used with permission of the author. The horizontal axis represents the breakdown by grade in school of the subjects; the vertical axis is percent correct on the phoneme reversal task. Note in particular that the normals only begin to catch up with their blind counterparts in fifth grade, which is when they start to learn English.

they were not able to correctly delete the /n/ in *nadii* नदी <nd[ii]>; in this case the first syllable has the inherent vowel /a/, which is not written with a separate glyph from the न <n>. Similarly, literate Hindi speakers found it easier to delete the /y/ in प्य <py> which, while fused with the preceding <p> is still in more its full form; than they did with the /r/ in प्र <pr>, which is written as a small stroke at the bottom of the <p>.

Cross-scriptal variation also results in differences in the ability to treat certain sounds as separate units. Thus as Padakannaya and his colleagues (Prakash et al., 1993, page 66) note, Hindi speakers find it hard to treat *anusvara* and *arka* as separate segments, since they are written as diacritics above the end of the orthographic syllable: पेंसिल <peMsil>; अर्क <aka+r> (<arka>). On the other hand, this is easy for Kannada speakers since in each case these symbols are written inline in Kannada script (though in the case of *arka*, out of its "logical" order): ಪೆಂಸಿಲ <peMsil>; ಅರ್ಕ <aka+r>.[17]

So some conscious segment-level manipulation is clearly possible, but one might wonder how the segmental aspects of Indic scripts affect the unconscious processing that is presumably involved when fluent readers read words. A study by Vaid and Gupta (Vaid and Gupta, 2002) addresses this question. Vaid and Gupta investigated naming latencies (the speed with which readers are able to vocalized printed words) in 40 Hindi-speaking adults, and naming errors (the numbers of errors made when vocalizing printed words) among 10 Hindi-speaking children (7.5–10.0 years old) with words containing short /i/, which is written before the consonants it logically follows. In particular, they considered cases where there was a single consonant (तिलक <i+tlk> /tilak/ ‘sectarian mark’, versus cases where there was a phonologically *heterosyllabic* consonant sequence (मस्जिद <m i+sjd> /masjid/ ‘mosque’). If Hindi readers of the Devanagari script processed things syllable-by-syllable (as one might expect if Devanagari is really a syllabary) then one would expect that the misorder of the <i> would only cause problems if a phonological syllable boundary intervened. In particular, in cases like तिलक <i+tlk> /tilak/, the first syllable would simply be processed as a unit, and it should not matter that the <i> is out of its logical order; such cases should be as readily processed as cases where the vowel symbol is written after the consonant it logically follows. On the other hand, with cases like मस्जिद <m i+sjd> /masjid/, there would be a processing cost since the <i> is separated from the syllable it logically belongs with by the intervening <s>. The results are consistent with the view that at least some segmental processing is going on. In particular, while *masjid*-type cases resulted in slower naming than *tilak*-type cases, *tilak*-type cases were also significantly slower than cases not involving short <i>.[18] Similarly, children produced more naming errors for *masjid*-type cases than for *tilak*-type cases, they also produced more errors for *tilak*-type than for cases not involving short <i>. Thus, it cannot be the case that Hindi readers simply process the script syllable-by-syllable: at least for glyphs that occur ‘inline’ (Vaid and Gupta (2002) present no evidence of what happens with super- or subscripted vowels), there is clear evidence that readers also process segment-by-segment.

What the cases cited from (Prakash et al., 1993) and (Vaid and Gupta, 2002) — Devanagari <d>, <y>, short <i>, and Kannada *arka* and *anusvara* — all have in common is that each of these involves a glyph that is written inline. That is, they involve a catenation operator (either $\overrightarrow{\cdot}$ or $\overleftarrow{\cdot}$) that is horizontal to the overall script direction ($\overrightarrow{\cdot}$). This leads one to the hypothesis that the ability to process a symbol as a separate segment is enhanced when that symbol is written horizontal to the overall script direction. This view is consistent with other research that shows that superscripted diacritics are less noticeable than inline glyphs; thus van Heuven (2002) shows for Dutch that errors in the placement of dieresis (required by the rules of Dutch orthography in cases where adjacent vowels constitute separate syllables) have hardly any effect on lexical decision tasks, whereas an error in a letter has a substantial effect.

On the other hand, this cannot be the whole story. Padakannaya (2001) notes that in Kannada consonant sequences C_1C_2, C_1 had a much lower success rate in a phoneme deletion task than C_2. Thus in ರಕ್ತ <rakta> ‘blood’, the /t/ should be easier to delete than the /k/. Padakannaya suggests that this is because the first consonant, while written inline, is generally ligatured with vowel symbols (though not in this particular case), whereas the second position (though written underneath the preceding consonant and though often reduced, as in the case of <t> here) is not adorned with vowel symbols. So there must be an additional factor of visual distinctiveness: if it is easier to parse out the segment in question, it will be processed more like a segment. The use of ligature is clearly a factor in making a segmental analysis less accessible.

So, there is clear evidence that properties of Indic scripts have an effect on processing at the segmental level. But while it is clearly the case that learners of Indic scripts are not able to perform segmental manipulations as well as learners of purely alphabetic scripts, it is also clear that segmental abilities do show up in cases where the script supports it. In particular, segmental glyphs that are written inline, and glyphs that are readily separable from their surroundings have a better chance of engendering segmental awareness than glyphs that do not have these properties.

Still, no matter what the psycholinguistic effects of learning an Indic script may be, one must still deal with the point that Indic scripts are segmental in their basic design, and this raises the question of how the Brahmi script was invented in the first place. Assuming it was not the same ‘accident’ that has been ascribed to the development of the Greek alphabet from Phoenician, and perhaps even if it was, this would seem to presume a fair level of segmental awareness and general phonological sophistication on the part of the previously illiterate inventors of the script (Patel, 1993). One can speculate that such phonological sophistication may have arisen out of the language games that were played by Vedic pracitioners for the purpose of preserving the sacred texts in an oral tradition. These games were quite complex: a particularly baroque instance was the *ghana-pāṭha* (Macdonell, 1900, page 42) which required one to step through a sequence of words *a b c d* as *ab, ba, abc, cba, abc; bc, cb, bcd* . . . . Since this process required untying segmental sandhi phenomena at each word juncture, one can imagine that this exercise would enhance the practitioner's awareness of segmental phenomena. It is an open question whether the most elaborate of these games predated literacy, but it seems to be widely believed that Vedic practitioners did use such techniques even before literacy (and perhaps even as an antidote to the threat of literacy) in order to guarantee that the sacred texts were preserved in their correct form; Steve Farmer, p.c., and see also (Farmer, Henderson, and Witzel, 2002, page 68).[19]

In principle, there is no reason to assume that the ability to manipulate segments (or phonological units of any size) depends solely upon learning a script with a particular set of properties. Rather, it is reasonable to assume that the propensity for this ability is there in all of us, but that it takes certain external factors, such as learning a script, or participating in certain kinds of language games, for the ability to fully develop.

Justification for the Level of Abstraction

One issue that we have left hanging is the justification for the particular level of abstraction I have chosen for the formal description. As we noted before, there are a number of levels of abstraction that one could choose in describing a writing system, ranging from a very abstract level at which elements are catenated in their logical order, to the surface level where the exact placement and rendering of glyphs is accomplished. We have chosen a level intermediate to that, where directional catenators allow one to broadly describe the arrangement of symbols, but we still abstract away from the exact placement. In this scheme it is sufficient to say that the diacritic <i> in Devanagari occurs before the consonant sequence it logically follows, abstracting away from the detail that the upper stroke of the <i> is actually written over the following consonant sequence as in कि <ki>. Similarly one can describe the diacritic vowel <[uu]> in Oriya as occurring under the consonant it logically follows as in କୂ <k[uu]>, without the finer grained description that it actually occurs on the right at the bottom.

The justification for this is that it seems to be the right level of granularity to describe some of the psycholinguistic results we have discussed. So, it seems to be a factor that certain glyphs are written above or below the main glyphs, but there is no justification for making a finer-grained distinction than that. In Vaid and Gupta's (2002) experiment, the placement of diacritic <i> before the consonant sequence was clearly relevant, whereas the fact that it also extends above the consonant sequence was not, so far as one knows, a factor.

Of course, this is all extremely sketchy at this point since no psycholinguistic work has yet been done that addresses the question of whether, e.g., placement of a glyph on top to the right, or on top to the left, makes a significant difference in some measurable behavior. My expectation is that no such results will be forthcoming, but that it will be possible to measure more effects of choice among the four basic catenators used here.

Conclusions

The formal model presented in this paper allows for the succinct description of some quite intricate writing systems, such as the writing systems of India. The theory is able to make predictions, as we saw in the case of the placement of vowels in Tamil that deviate in their catenation from the macroscopic direction, and the placement of the secondary $\overset{\rightarrow}{\cdot}$- and $\overset{\downarrow}{\cdot}$- catenated vowel components in Kannada, and it is able to provide some insights into the commonalities among Indic scripts, as in the case of the encoding of <o> in diphthongs in Devanagari and Oriya.

The model also has implications for human processing of scripts. Since it treats Indic scripts as fundamentally segmental, despite their obvious differences from Western alphabets, it implies that there should be some evidence that readers of Indic scripts are able to develop some segmental awareness, and that psycholinguistic studies of reading should show some processing at the segment level rather than just at the syllable level. These expectations are borne out: although it is very clear that readers of Indic scripts are much less able to manipulate segments than readers of alphabetic scripts, there are clearly also cases where they are able to do so. And the naming studies reported in (Vaid and Gupta, 2002) show that processing of Devanagari at least partly occurs at the level of the segment. Through the work of Padakannaya and others, there is much that is now understood about the psycholinguistic effects of learning Indic scripts, but there is also much that is not known. To my mind this will be one of the most fruitful areas of future research on Indian writing systems.

Acknowledgments

I wish to thank Bill Bright, Steve Farmer, Prakash Padakannaya, Pramod Pandey, and George Thompson, for helpful comments on previous versions of this paper.

References

Aliprand, Joan, editor. 2003. *The Unicode Standard, Version 4.0*. Addison Wesley Professional, Reading, MA.

Bhaskararao, Peri and Suresh Mathew. 1992. Phonemic transcription rules for text-to-speech synthesis of Hindi. In R.M.K. Sinha, editor, *Computer Processing of Asian Languages*. TATA McGraw Hill, New Delhi.

Bhaskararao, Peri, Venkata Peri, and Vishwas Updikar. 1994. A text-to-speech system for application by visually handicapped and illiterate. In *Proceedings of ICSLP 94*, pages 1239–1242, Yokohama.

Blevins, Juliette. 2002. *Evolutionary Phonology*. Ms., University of California, Berkeley, Submitted to Cambridge University Press.

Bright, William. 1996a. The Devanagari script. In Peter Daniels and William Bright, editors, *The World's Writing Systems*. Oxford University Press, New York, NY, pages 384–390.

Bright, William. 1996b. Kannada and Telugu writing. In Peter Daniels and William Bright, editors, *The World's Writing Systems*. Oxford University Press, New York, NY, pages 413–419.

Browman, Cathe and Louis Goldstein. 1989. Articulatory gestures as phonological units. *Phonology Yearbook*, 6:201–251.

Daniels, Peter. 1996. The study of writing systems. In Peter Daniels and William Bright, editors, *The World's Writing Systems*. Oxford University Press, New York, NY, pages 3–17.

DeFrancis, John. 1989. *Visible Speech: The Diverse Oneness of Writing Systems*. University of Hawaii Press, Honolulu, HI.

Diller, Anthony. 1996. Thai and Lao writing. In Peter Daniels and William Bright, editors, *The World's Writing Systems*. Oxford University Press, New York, NY, pages 457–466.

Faber, Alice. 1992. Phonemic segmentation as epiphenomenon. evidence from the history of alphabetic writing. In Pamela Downing, Susan Lima, and Michael Noonan, editors, *The Linguistics of Literacy*. John Benjamins, Amsterdam, pages 111–34.

Farmer, Steve, J.B. Henderson, and Michael Witzel. 2002. Neurobiology, layered texts, and correlative cosmologies: A cross-cultural framework for premodern history. *Bulletin of the Museum of Far Eastern Antiquities*, 72:48–90.

Fujimura, Osamu. 1994. The C/D model: A computational model of phonetic implementation. In *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, volume 17, pages 1–20. American Mathematical Socieaty.

Gill, Harjeet Singh. 1996. The gurmukhi script. In Peter Daniels and William Bright, editors, *The World's Writing Systems*. Oxford University Press, New York, NY, pages 395–398.

Haile, Getatchew. 1996. Ethiopic writing. In Peter Daniels and William Bright, editors, *The World's Writing Systems*. Oxford University Press, New York, NY, pages 569–576.

Hopcroft, John and Jeffrey Ullman. 1979. *Introduction to Automata Theory, Languages and Computation*. Addison-Wesley, Reading, MA.

Johnson, C. Douglas. 1972. *Formal Aspects of Phonological Description*. Mouton, Mouton, The Hague.

Kaplan, Ronald and Martin Kay. 1994. Regular models of phonological rule systems. *Computational Linguistics*, 20:331–378.

Karanth, Prathibha. 2002. Reading into reading research through nonalphabetic lenses: Evidence from Indian languages. *Topics in Language Disorders*, 22(5):20–31.

Karanth, Prathibha. 2003. Literacy and language processes —orthographic and structural effects. In Prathibha Karanth and Joe Rozario, editors, *Learning Disability in India: Willing the Mind to Learn*. Sage, chapter 6, pages 145–160.

King, Ross. 1996. Korean hankul. In Peter Daniels and William Bright, editors, *The World's Writing Systems*. Oxford University Press, New York, NY, pages 218–227.

Macdonell, Arthur. 1900. *A History of Sanskrit Literature*. Heinemann, London.

Mahapatra, B. P. 1996. Oriya writing. In Peter Daniels and William Bright, editors, *The World's Writing Systems*. Oxford University Press, New York, NY, pages 404–407.

Mohri, Mehryar. 1997. Finite-state transducers in language and speech processing. *Computational Linguistics*, 23(2).

Mohri, Mehryar, Fernando Pereira, and Michael Riley. 1998. A rational design for a weighted finite-state transducer library. *Lecture Notes in Computer Science*, (1436).

Mohri, Mehryar and Richard Sproat. 1996. An efficient compiler for weighted rewrite rules. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 231–238, Santa Cruz, CA. ACL.

Narasimhan, Bhuvana, Richard Sproat, and George Kiraz. 2003. Schwa-deletion in Hindi text-to-speech synthesis. *International Journal of Speech Technology*. forthcoming.

Ohala, Manjari. 1983. *Aspects of Hindi Phonology*. Motilal Banarsidass, Delhi.

Padakannaya, Prakash. 1999. Reading disability and knowledge of orthographic principles. *Psychological Studies*, 44:59–64.

Padakannaya, Prakash. 2000. Is phonemic awareness an artefact of alphabetic literacy?! Presented at ARMADILLO 11, Texas A&M, October 13–14, 2000.

Padakannaya, Prakash. 2001. Syllabic and phonemic awareness in children acquiring literacy in a semisyllabic script. Department of Psychology, University of Mysore.

Patel, P. G. 1993. Ancient India and the orality-literacy divide theory. In Robert Scholes, editor, *Literacy and Language Analysis*. Lawrence Erlbaum Associates, Hillsdale, NJ, pages 199–208.

Prakash, P., D. Rekha, R. Nigam, and P. Karanth. 1993. Phonological awareness, orthography and literacy. In Robert Scholes, editor, *Literacy and Language Analysis*. Lawrence Erlbaum Associates, Hillsdale, NJ, pages 55–70.

Radhakrishnan, Sankaran. 2002. *Tamil Script Learners Manual*. University of Texas at Austin, Department of Asian Studies, Austin, TX. Available at `inic.utexas.edu/asnic/radhakrishnan/pages/tamilscript.html`.

Ratliff, Martha. 1996. The Pahawh Hmong script. In Peter Daniels and William Bright, editors, *The World's Writing Systems*. Oxford University Press, New York, NY, pages 619–624.

Salomon, Richard. 1996. Brahmi and Kharoshthi. In Peter Daniels and William Bright, editors, *The World's Writing Systems*. Oxford University Press, New York, NY, pages 373–383.

Sampson, Geoffrey. 1985. *Writing Systems*. Stanford University Press, Stanford, CA.

Smalley, William, Chia Koua Vang, and Gnia Yee Yang. 1990. *Mother of Writing: The Origin and Development of a Hmong Messianic Script*. University of Chicago Press, Chicago, IL.

Spencer, Harold. 1950. *Kanarese Grammar*. Wesley Press, Mysore. Revised edition by W. Perston.

Sproat, Richard. 1997a. Multilingual text analysis for text-to-speech synthesis. *Journal of Natural Language Engineering*, 2(4):369–380.

Sproat, Richard, editor. 1997b. *Multilingual Text to Speech Synthesis: The Bell Labs Approach*. Kluwer Academic Publishers, Boston, MA.

Sproat, Richard. 2000. *A Computational Theory of Writing Systems*. Cambridge University Press, Cambridge.

Steever, Sanford. 1996. Tamil writing. In Peter Daniels and William Bright, editors, *The World's Writing Systems*. Oxford University Press, New York, NY, pages 426–430.

Swiggers, Pierre. 1996. Transmission of the Phoenician writing system to the west. In Peter Daniels and William Bright, editors, *The World's Writing Systems*. Oxford University Press, New York, NY, pages 261–270.

Vaid, Jyotsna and Ashum Gupta. 2002. Exploring word recognition in a semi-alphabetic script: The case of Devanagari. *Brain and Language*, 81:679–690.

van Heuven, Vincent. 2002. The effects of diaeresis on visual word recognition in Dutch. In Martin Neef, Anneke Neijt, and Richard Sproat, editors, *The Relation of Writing to Spoken Language*, number 460 in Linguistische Arbeiten. Max Niemeyer, Tübingen, pages 99–114.

Notes

[1] As is well-known, Sampson (1985), has argued that Hankul is a featural script. This has been argued against by DeFrancis (1989), and see also (Sproat, 2000).

[2] Though in Gurmukhi the full-vowel forms are replaced with vowel-specific "vowel bearers", along with the diacritic form (Gill, 1996), and this trend of using vowel bearers has developed further in Southeast Asian Brahmi derivatives.

[3] I will use the term *consonant sequence* throughout to refer to a sequence of consonants within an *akṣara*. By so doing, I wish to emphasize the fact that these do not generally represent phonological consonant clusters.

[4] Some, like Chinese, have several different overall directions to choose from.

[5] In Mahapatra’s Table 35.1 (1996, page 405), <k[au]> appears as କୌ with the ୗ and ା glyphs unfused.

[6] In Oriya, where the decomposition of /o,[ai],[au]/ is less controversial than the proposal for Devanagari, the full vowel forms are more idiosyncratic.

[7] Use of *arka* is not obligatory in Kannada: one can also write the sequence *rC* in its logical way, with the full form of the <r> glyph and the second consonant in a reduced form underneath it: e.g., ರ್ತ <rta>.

[8] Interestingly Spencer (1950) suggests that the inherent vowel /a/ is actually represented in the orthography by the headstroke that is included with many of the consonant symbols—e.g. ಕ <k>—and which is lost in the reduced forms: consider the second <k> of ಕ್ಕ <kk>. His argument is that the headstroke is lost with some of the vowels, as with ಕಿ <ki>. This seems doubtful, however: the headstroke is only lost in the reduced forms, and with vowels that would conflict with the headstroke. Thus the headstroke is not lost with ಕು <ku>.

[9] Also used to ‘cancel’ inherent vowels and occasionally used in Devanagari as a substitute for a ligatured consonant.

[10] See `http://www.sibal.com/sandeep/jtrans`.

[11] This rule does not handle the case of a homorganic nasal being written with *anusvara*, and hence being syllabified with the preceding *akṣara*.

[12] In practice, however, we found that the conventions for candrabindu usage are often not adhered to; a later ‘stylistic’ rule allows *anusvara* to substitute for *candrabindu*.

[13] `http://www.research.att.com/sw/tools/lextools/`

[14] Though there have been recent attempts to decompose the complex symbols when they are taught (Karanth, 2002).

[15] The work in phonology that argues against the segment as a primitive of phonological representation relies heavily on the concept that phonetic features are features of syllables rather than segments. Segments appear phonetically as an epiphenomenon of coordinated timing of specific features. This view has been expressed in various phonological theories, but is perhaps most strongly associated with “articulatory phonology”(Browman and Goldstein, 1989), and see also (Fujimura, 1994).

Problematic for that view is recent work by Blevins (2002) on historical metathesis, which involves shifting articulatory features from one (segmental) position within a syllable to another, typically flipping between pre- and post-rime position. The historical evidence suggests that language learners have a strong propensity to associate those features with one or another position, with change occurring when the position decided on by learners of one generation differs from the position assumed by learners of the preceding generation. But if the anti-segmental view favored by Faber is right, one has to wonder why learners would have any propensity at all to associate the features with a particular segmental position. In particular, why could the features not just remain associated with the syllable, with no preferred timing with respect to other segments?

[16] On the other hand, Swiggers (1996) seems to imply a rather more conscious design on the part of the Greeks than is suggested by Faber's explanation.

[17] While *arka* and *anusvara* make segmental manipulation easier, interestingly they also make syllable deletion harder (Padakannaya, 2001).

[18] Note that the naming latency difference between *tilak* and *masjid* probably cannot be explained solely by the fact that *masjid* has more letters than *tilak*: both were slower than cases like *tasv[ii]r* 'picture', which has the same number of letters as *masjid*.

[19] George Thompson (p.c.) also notes that the notion of *akṣara* itself precedes literacy.