## Typology of Orthography

## David R．Mortensen

March 19， 2024

## Logographic Writing Systems

The oldest type of writing systems，exemplified by Chinese characters， Mesopotamian cuneiform，and Egyptian hieroglyphics，among others，uses symbols to represent whole morphemes or words without respect，necessar－ ily，to their pronunciation．This is called a Logographic writing system． The symbols in logographic writing systems typically start of as ICONIC ${ }^{1}$ representations of the meaning they are intended to convey．

|  | 为为 | （1axio | （4） | \％ | maint | （1at） | \％umin |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 馬 | J | 岸 | 第 | 馬 | 馬 | － | 亏 |
| 日 | $\square$ | $\bigcirc$ | $\theta$ | 日 | 日 | 日 |  |
| 月 | D | 1） | 1 | 月 | 月 | 月 | 月 |
| 鳥 | \％ | 丞 | 皆 | 鳥 | 鳥 | 鳥 | 多 |
|  | M | $\downarrow$ | W | 山 | 山 | $山$ |  |

In Figure1，we can see that the character for＇mountain＇（山 shān in Mod－ ern Chinese）started its life as a drawing of a literal mountain（or，rather，a carving of a literal mountain into tortoise shells used for divination）．Over time，the characters became more and more abstracted from their original iconic representation．

Note that the characters never stood directly for meanings．They stood for particular words which stood for meanings．In other words，words（or really， morphemes）in ancient Chinese were three tuples of
$\langle$ semantics，orthography，phonology〉

## Abjads

Around 1800 BCE，in the Levant，people developed a different idea：just write the sounds．Specifically，the decided to just write consonants．This kind of system—used today by Hebrew，Arabic，and Aramaic—is called an abjad．
${ }^{1}$ Iconicity refers to the situation in which the signifier resembles the signified in some way．

Figure 1：Historical origins of some Han characters（＇horse＇，＇sun＇，＇moon＇，＇bird＇， ＇mountain＇）．


Figure 2：A sample of Proto－Sinaitic script．

Hebrew and Arabic are not pure abjads in that certain long vowels are written. ${ }^{2}$

Consider Table 1 of Hebrew vowels (simplified): The short vowels are,

orthographically, invisible. However, the long vowels are marked using consonant symbols ( $\Pi$ alef, ' yod, and 1 vav, which represent $/ \mathrm{R} / \mathrm{/} / \mathrm{j} /$, and $/ \mathrm{v} /$, respectively). This kind of writing system is sometimes called an "impure" abjad, because some of the vowels are written.

## Syllabaries

| Consonant | a |  |  | e |  |  | i |  |  | - |  | u |  | v [⿰๊๊] |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\varnothing$ | D a |  |  |  |  |  |  | i |  |  |  | $\mathrm{O}^{\prime}$ | u |  | $v$ |
| g/k | \& ga | (e) ka |  |  |  |  |  | gi |  |  | go | J | gu |  | gv |
| h | of ha |  |  |  |  |  |  | hi |  |  | ho | $\Gamma$ | hu |  | h hv |
| 1 | W la |  |  | $\delta$ | le |  |  | li |  |  |  | M | lu | ฯ | Iv |
| m | $\chi^{\text {\% }}$ ma |  |  | 01 |  |  |  | mi |  |  | mo | $y$ | mu |  | * mv |
| $\mathrm{n} / \mathrm{hn}$ | $\boldsymbol{\theta}$ na | t, hna | G nah | $\Omega$ | ne |  |  | ni |  |  | no | ๆ | nu |  | nv |
| $\begin{gathered} \text { qu } \\ {\left[\mathbf{k}^{w}\right]} \end{gathered}$ | U qua |  |  |  | que |  |  | qui |  |  | quo | C | quu |  | quv |
| $s$ | 00 s | U sa |  | 4 |  |  |  | si |  |  | so | $\mathscr{8}$ | su |  | sv |
| d/t | L da | W ta |  | S |  | て te |  | di | J ti | V | do | S | du |  | dv |
| $\begin{gathered} \mathrm{dl} / \mathrm{tl} \\ {\left[\mathrm{~d}_{3}\right] /[\mathrm{tt]}]} \end{gathered}$ | 8 dla | $\mathcal{L}$ tla |  | L | tle |  |  | ti |  |  | tlo | $\gamma$ | tlu | P | tlv |
| $\begin{gathered} \mathrm{ts} \\ {[\mathrm{ts}]} \end{gathered}$ | G tsa |  |  | V |  |  |  | tsi |  |  | tso | d | tsu |  | m tsv |
| $\begin{gathered} \mathbf{w} \\ {[\underline{[w]}} \end{gathered}$ | G wa |  |  |  | we |  |  | wi |  |  | wo | 9 | wu | 6 | wv |
| $\begin{gathered} \mathbf{y} \\ {[j]} \end{gathered}$ | $\propto$ ya |  |  | $\beta$ |  |  |  |  |  |  | yo | G | yu | B | yv |
| * The character G was previously used to represent the syllable mv, but is no longer used. ${ }^{\text {[note 1] }}$ |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

## Abugidas

Daniels and Bright defined Abugidas as "a type of writing system whose basic characters denote consonants followed by a particular vowel, and in which
${ }^{2}$ Many languages make a distinction between vowels based on their duration. This is called "vowel length."

Table 1: Hebrew short and long vowels
diacritics denote other vowels."

| प | pa |
| :--- | :--- |
| पि | pi |
| पु | pu |
| पृ | pr |
| पे | pe |
| पो | po |
| पू | $\mathrm{pæ}$ |

## क का कॅ कॉ कि की कु कू के के कै को को कौ कृ कृ का क्ष कं क: क् क़

## Alphabets

Ancient Greeks, on learning the Phonecian Abjad, had an insight, namely that one could write the vowels as well. This gave rise the to Greek alphabetthe first alphabet of which we are aware. And an alphabet is distinguished from an abjad in precisely this way: there is roughly one symbol ("letter") for each phoneme, vowels and consonants. Most modern alphabets, as shown in Figure 6, are descents of the Greek alphabet and (in the great majority of cases, the Latin alphabet).

One important exception is the Korean Hangul alphabet, invented by the fourth kind of the Joseon Dynasty, Sejong the Great. At first, it may look like a syllabary: it is written using symbols that represent syllables. However, as shown in Figure 5, each of these syllable blocks can be decomposed into 1-5 subcomponents which are essentially letters. They are not written linearly, as in most alphabets, but-just like letters in ideal alphabets-there is one of these letters for each phoneme and a phoneme for each letter. It has the added feature of systematicity-the shape of the letters follows from the place and manner of articulation. For these reasons, Korean Hangul is considered by many linguists to be the most scientifically insightful writing systems in common use.

Table 2: Some Devanagari vowels with the consonant $/ \mathrm{p} /$.


Figure 5: The order and position of letters in Hangul.


Figure 6: A family tree of writing systems descended from Prot-Sinaitic

## A Note on Encoding

We have seen, that the graphical order of symbols in the writing systems of the world is not consistent. Arabic and Hebrew are written from write to left; Hindi and English are written from left to right. Chinese was once often written top to bottom. Hindi vowels may be written left of, right of, above, and beneath the consonant symbols they modify. Hangul letters are stacked non-linearly. However, when we encode these writing systems using Unicode, they have to be represented as sequences of bytes.

The general policy is to represent the characters so that code-point order is monotonic with pronunciation (where this makes sense). There is, as we will see, a separation between PRESENTATION (how written language is RENDERED through the use of fonts) and the logical structure. For example, in Devanagari orthographies, the consonant symbols is always represented first, followed by the vowel symbol. In general, too, when one symbols combines with another, the "base" comes first in the Unicode stream and the combining modifier comes after it.

## References

