Orthography, Phonology, Morphology, and Meaning
R. Frost and L. Katz (Editors)

CHAPTER 1

839

Linguistic Awareness and Orthographic Form

Ignatius G. Mattingly
Department of Linguistics, University of Connecticut, Storrs

Introduction: The taxonomy of writing systems

To impose some pattern on the vast array of writing systems, present and past,[1] several investigators have proposed typologies of writing (Gelb, 1963; Hill, 1967; Sampson, 1985; DeFrancis, 1989; see DeFrancis for a review). While typology for its own sake may seem a dubious goal, these proposals bring to notice certain interesting questions.

Consider first the problem posed by logograms. It is generally recognized that the signs found in writing fall into two broad categories: logographic and phonographic. Logograms stand for words, or more precisely, morphemes. Thus, in Sumerian writing, there is a logogram that stands for the morpheme ti, 'arrow.' Phonographic signs stand for something phonological: syllables or phonemic segments. Thus, in Old Persian, there is a sign for the syllable da, and in Greek alphabetic writing, a sign for the vowel a. This distinction suggests that writing systems might be classified according to whether they are logographic or phonographic. But the attempt to impose such a classification is embarrassed by the fact that while the many systems in the West Semitic tradition are indeed essentially phonographic and have no logograms, writing systems of all other traditions use both logograms and phonograms. There have been no purely logographic systems: phonographic signs are found in all traditions.

In these circumstances, Gelb sets up a hybrid category "word-syllabic," in which he includes Sumerian, Egyptian (whose phonographic signs he takes to be syllabic[2]), and

[1] It will be assumed here, following Gelb (1963), Jensen (1970), DeFrancis (1989) and others, that there are six major orthographic traditions: (1) Mesopotamian cuneiform, beginning with Sumerian (c. 3100 B.C.) and including Akkadian, cuneiform Hittite, Urartian, Hurrian, Elamite, Old Persian; (2) Cretan, including Minoan Linear A, Mycenaean Greek Linear B, Cypriote, and Hittite hieroglyphics, all probably derived from a common source (c. 2000 B.C.); (3) Chinese, beginning with Chinese itself (c. 1300 B.C.) and including Korean nonalphabetic writing and Japanese; (4) Mayan (c. 300 A.D.); (5) Egyptian (c. 3000 B.C); (6) West Semitic, beginning with Phoenician (c. 1600 B.C.) and including Ras Shamrah cuneiform, Old Hebrew, South Arabic, Aramaic, and Greek alphabetic writing. From Aramaic derive Hebrew, Arabic, and many others; from Greek derive Etruscan, Latin, and many others. Germanic runes and Korean alphabetic writing probably belong in this tradition also, though the derivations are not clear. All but the most dogmatic monogeneticists would agree that the Mesopotamian, Cretan, Chinese, and Minoan traditions are probably independent developments. But some scholars (e.g., Driver, 1976; Ray, 1986) would derive Egyptian writing from Mesopotamian, and some (e.g., Driver, 1976), with somewhat greater plausibility, would derive West Semitic from Egyptian.

[2] Egyptologists and most other students of writing believe that Egyptian phonographic signs stand for consonants, the vowels not being regularly transcribed. But according to Gelb, they stand instead for generalized syllables, e.g., the Egyptian sign usually interpreted as consonantal w actually stands for **wa**,

Chinese. Other orthographic taxonomists allow a writing system to belong to two different categories. Thus for Hill, Egyptian is both "phonemic" and "morphemic" and for Sampson, Japanese is both "phonographic" and "logographic." DeFrancis, recognizing that logograms are neither necessary nor sufficient for an orthography, more sensibly treats logography as an optional accompaniment to various phonographic categories. But the question of interest is *why* logograms should play only this secondary role, why there have been no pure logographies.

A second problem arises in sorting out the phonographic categories. Here one might recognize, with DeFrancis, systems like Sumerian or Linear B, in which the phonographic signs stands for syllables; systems like Egyptian or Phoenician, in which they stand for consonants; and systems like Greek or English, in which they stand for both consonants and vowels (*plene* systems).

The distinction between consonantal and *plene* systems, however, proves to be less than rigid. In Egyptian, the letters for **j**, **w**, and **?** are used to write **i**, **u** and **a**, respectively, in foreign names (Gelb, 1963). Phoenician, indeed, is a strictly consonantal, but the other "consonantal" systems deriving from it all have some convention for transcribing vowels when necessary. For example, in Aramaic, the letters *yodh*, *waw*, and *he* (or *aleph*) were used to write final **i**, **u**, and **a**, respectively, and to render vowels in foreign names (Cross & Freedman, 1952). In Masoretic Hebrew, Arabic, and various Indic systems, vowels are regularly indicated by diacritic marks on consonant letters. And, of course, the first clearly *plene* system, the Greek alphabet, is a development from the Phoenician consonantal system. The taxonomist thus has to decide where to draw the line between essentially consonantal systems, hybrid systems, and undoubted *plene* systems. Perhaps the wisest course is the one followed by Sampson: simply to classify all these systems as "segmental."

Syllabic systems, in contrast, are clearly a separate category and present no problem to the taxonomist. There is no writing system that must be regarded as a hybrid between a syllabic and a segmental system. Syllabic systems show no tendency to analyze syllables into segments. What is found, rather, is that when analysis becomes necessary, complex syllables are analyzed into simpler syllables. Thus, neither the Mesopotamian nor the Mayan syllabaries had signs for all possible $C_1V_1C_2$ syllables in their respective languages. Instead, such syllables were written in Mesopotamian as if they were $C_1V_1 + V_1C_2$ (Driver, 1976) and in Mayan as if they were $C_1V_1 + C_2V_1$ (Kelley, 1976)). Similarly, Greek $C_1C_2V_1$...syllables were written in Linear B as $C_1V_1 + C_2V_1 + ...$(Ventris & Chadwick, 1973). Nor, despite suggestions to the contrary by Gelb and DeFrancis, has a syllabic system ever developed into a segmental system, or conversely.[3] It cannot be excluded that the Egyptians may, as DeFrancis says (following Ray, 1986), have gotten the idea of writing from the Sumerians. But there is certainly no reason to believe that they borrowed the idea of *syllabic* writing from the Sumerians and then adapted it to consonantal writing, in the way that the Greeks may be said to have borrowed the idea of consonantal writing

---

wi, **we**, **wu**, or **wo**, according to context. It is obviously difficult to distinguish these two accounts empirically. The only support Gelb offers for his position is that "the development from a logographic to a consonantal writing, as generally accepted by the Egyptologists, is unknown and unthinkable in the history of writing" (Gelb 1963, p. 78). But this argument is clearly circular (Edgerton, 1952; Mattingly,1985).

[3]Gelb (1952, 1963) proposed some cases in which syllabic systems are supposed to have developed into segmental systems ; but see Edgerton (1952). Ethiopic writing, derived from the West Semitic consonantal tradition, might be viewed as a syllabic system derived from a segmental system, because the signs do correspond to syllables. But, with a few exceptions, each sign actually consists of a consonant letter plus a vowel mark, except that **a** is left unmarked. As in the case of Indic systems, one could argue about whether this is a consonantal or a *plene* system, but it is certainly not a syllabic system (Sampson, 1985).

from the Phoenicians and adapted it to *plene* writing. The various orthographic traditions are remarkably self-consistent in this matter. The Mesopotamian, Chinese, Cretan and Mayan traditions began and remained syllabic; the Egyptian and West Semitic traditions began and remained segmental.

If the main purpose here were to arrive at a taxonomy of writing systems, the conclusion would have to be that there are two primary categories: syllabic and segmental. Either of these may or may not be accompanied by logograms. Transcription of vowels in segmental systems is a matter of degree, with Phoenician at one end of the scale and Greek at the other. The interesting question, however, particularly given the degree of overlap or hybridization that is found between logographic and phonographic categories, and between consonantal and *plene* categories, is why the syllabic and segmental categories have remained so distinct.

In an attempt to answer the questions just posed, it is necessary to consider why an orthography can make reading and writing possible, what constraints there are on the form of orthographies, how orthographies could have been invented, and what happens when orthographies are transmitted from one culture to another.

Why reading and writing are possible[4]

When a listener has just heard an utterance in a language he knows, he has available for a brief time not only his understanding of the semantic and pragmatic content of the utterance (the speaker's *message*), but also a mental representation of its linguistic structure. The basis for this claim is that a linguist, by analyzing the intuitions of informants about utterances in their native language (such as that two utterances are or are not the same word, or that a certain word is the subject of a sentence), can formulate a coherent grammar, consistent with grammars that would be formulated by other linguists working with other informants on the same language. This holds true even if, as is typically the case for a language with no writing system, the informants are quite unaware of the linguistic units into which utterances in their language can be analyzed. Because the informants' intuitions are apparently valid, they must be based on linguistic representations of some kind.

While linguists are not in total agreement about the nature of the linguistic representation of an utterance, it seems reasonably clear that such a representation must include the syntactic structure, the selection of lexical items and their component morphemes, the phonological structure, and the phonetic structure. The linguist's syntactic diagrams and phonological and phonetic transcriptions are formal reconstructions of different levels of the representation. These levels are not independent of one another. Syntax constrains lexical choice, lexical choice determines morphology and phonology, syntax and phonology determine phonetic structure. The representation thus has extensive inherent redundancy.

The linguistic representation is strictly structural rather than procedural. The listener has no access to the many intermediate steps he must presumably go through in the course of parsing the utterance, so that these steps are not represented. Acoustic details such as formant trajectories are not part of the linguistic representation, simply because the listener does not perceive them as such, but only the phonetic events they reflect. Other aspects of the utterance, such as individual voice quality, speaking rate, and loudness, which the listener can hear, must be presumed to be excluded because they are not linguistic at all and never serve to mark a linguistic difference between two utterances.

[4]The proposals in this section are developed in more detail in Mattingly (1991).

Access must be distinguished from awareness. All normal language users, it has been claimed, have access to the contents of linguistic representations. This means that they have a potential ability to introspect and report on significant details of the representation, and to regard it as a structure of phrases, words, and segments, not that they can actually do so. The representation is a complicated affair, and a person who is not "linguistically aware" can no more be expected to notice its characteristic units and structure than an electronically naive person can be expected to appreciate the units and structure of a circuit diagram (Mattingly, 1972). Linguistic awareness must in large part be acquired. The principal stimulus for linguistic awareness in modern cultures is literacy (Morais, Cary, Alegria, & Bertelson, 1979). Unlike illiterate adults or preliterate children, those who have learned to read can readily report on and manipulate at least those units of the linguistic representations of spoken utterances to which units of the orthography correspond (Read, Zhang, Nie, & Ding, 1986). However, there must certainly be other sources of linguistic awareness: Long before writing was known, poets composed verse in meters requiring strict attention to subtle phonological details.

It is not agreed how linguistic representations are created. On one view, they are a byproduct of the cognitive processes by which utterances are analyzed. Linguistic information, recovered step by step from the auditory image of the input signal, is temporarily represented in memory until, at a later stage, the speaker's message can be computed (Baddeley, 1986). The difficulty with this view is that, as has been noted, the language user seems to have no access to the supposedly cognitive analytic steps that must precede the formation of the representation or to the subsequent steps by which the message is derived from this representation. An alternative view is that the representation, as well as the message itself, is not a byproduct but a true output of a specialized, low-level processor (the "language module") whose internal operations, being inaccessible to cognition, have no cognitive byproducts (Fodor, 1983). This view implies that the linguistic representation must have some biological function other than communication, for which the message alone would suffice. What this function might be is unclear (but see Mattingly, 1991, for some speculations).

So far, the cognitive linguistic representation has been considered just as the product of the perception of utterances. But such representations are produced in the course of other modes of linguistic processing as well. Thus, a linguistic representation is formed in the production of an utterance, so that the speaker knows what it is he has just said. And when one rehearses an utterance in order to keep it in mind verbatim, what presumably happens is that the linguistic processor uses a decaying linguistic representation to construct a fresh version of the representation, and incidentally, of the message. This seeming defiance of entropy is possible for linguistic representations (as it may not be for mental representations in general) because of their high inherent redundancy.

Consideration of rehearsal also shows that the linguistic representation can be an input to as well as an output from the linguistic processor. Even more significantly, for the present purposes, a representation not originally produced by primary processes of perception or production can be such an input. An introspective, linguistically aware person can readily compose a "synthetic" linguistic representation according to some arbitrary criterion: the first five words he can think of that begin with /b/, for example. This is obviously a very partial representation: just a sequence of phonological forms drawn from the lexicon, without explicit phonetics or syntax. But if this sequence is rehearsed, the phonetic level, together with whatever syntactic structure or traces of meaning may be accidentally implicit

in the sequence, will be computed, just as if the sequence were what remained of a natural representation resulting from an earlier act of production, perception, or rehearsal. All that is required for a synthetic representation to serve as input for computing a natural one is that it contain enough information so that the rest of the structure of the utterance is more or less determined.

These various considerations suggest how it is that one linguistically aware language user can communicate with another, not by means of speech, but by means of synthetic representations, provided a way of transcribing such representations, that is, an orthography, is available. The writer speaks some utterance (at least to himself), creating a linguistic representation. The orthography enables him to transcribe this representation in some very partial fashion. From this transcription, the reader constructs a partial, synthetic linguistic representation. Such a representation is enough to enable the reader's linguistic processor to compute a complete, natural representation, as well as the writer's intended message.

If we compare what happens between writer and reader with what happens between speaker and hearer, it can be seen that the difference is much more than merely a matter of sensory modality. In speech perception, there is a natural and unique set of "signs"—the acoustic events that the human vocal tract can produce—and they are already in a form suitable for immediate linguistic processing (Liberman, this volume). Only the output of this processing is a linguistic representation. The input speech signal is in no sense a partial linguistic representation, but rather a complete representation of a very different kind. Moreover, the specification of the complex relation between the phonetically significant events in the signal and the units of the linguistic representation is acquired precognitively (Liberman & Mattingly, 1991); it does not have to be learned. Indeed, as has been remarked, the hearer has no access to the acoustic events, and may have little or no awareness of the units of the linguistic representation. In reading, on the other hand, there is no one, natural set of input symbols. Linguistic processing must therefore be preceded by a stage having no counterpart in speech perception: a cognitive translation from the orthographic signs to the units of the synthetic linguistic representation. The beginning reader must therefore deliberately master the mapping between the signs and the units, and for this he must have an awareness of the appropriate aspects of the linguistic representation.

Constraints on orthographic form

What psychological factors constrain the form of an orthography? Gelb (1963) makes a useful distinction between "outer form"—the shape of the visible symbols and their arrangement in a text—and "inner form"—the nature of the correspondence of the symbols to linguistic units. Beyond the trivial requirement that the symbols be visually discriminable, there appear to be no particular psychological constraints on outer form. The shapes of the signs in the writing systems of the world and the way they are arranged are extremely various, and such limitations as exist are to be accounted for not by cognitive or linguistic factors but by practical ones, such as the nature of the writing materials available and what patterns are easily written by hand, or by esthetic ones, such as the beauty of particular stroke patterns. This variety is possible because, as has just been seen, a cognitive translation is required for reading and writing in any event. This price having been paid, outer form can vary almost without limit.

Inner form, on the other hand, is highly constrained. In the first place, the orthography must correspond to the linguistic representation, because there is no other cognitive path to linguistic processes. This is the reason that proposals to treat spectrographic displays of

speech as, in effect, an orthography the deaf could learn to read (Potter, Kopp, & Kopp, 1966) are not likely to succeed. On the one hand, the reader of spectrograms cannot process the visually-presented spectral information as a listener can process the same information in the auditorially-presented and biologically-privileged speech signal. On the other hand, the spectrogram reader has no natural cognitive access to raw spectral events, and, a fortiori, no awareness of them. Therefore, even if he could somehow synthesize a cognitive spectral representation from the visible one, there is no reason to believe it could be an input to linguistic processes. All he can do is to apply his cognitive knowledge of acoustic phonetics to the task of inferring the linguistic representation from the spectrogram. Because the relation between spectral patterns and even the most concrete level of this representation, the phonetic level, is extremely complex, and a great deal of extraneous information is present, "reading" spectrograms is a slow and unreliable process. Analogous observations, obviously, could be made with respect to other records of physical activity in which linguistic information is implicit, such as the speech waveform or traces of articulatory movements. What has to be transcribed, then, is some level or levels of the linguistic representation itself.

However, certain levels of the linguistic representation are seldom or never transcribed in traditional orthographies. For example, syntactic structure is never transcribed. The few features of orthography that might be considered syntactic, such as punctuation and sentence-initial capitalization, are more reasonably regarded as transcriptions of prosodic elements. Why is syntax thus avoided? It is not just that tree diagrams are cumbersome to draw and nested brackets difficult to keep track of, but that the syntactic structure alone would be insufficient to specify a particular sentence: Each possible phrase marker is shared by an indefinitely large number of sentences. It would therefore be necessary that a syntactic orthography also transcribe in some way the particular lexical choices. But if this is to be done, the phrase-marker itself becomes redundant, because (barring some well-known types of structural ambiguity, such as those discussed by Chomsky, 1957) the words, and the order in which they occur, are themselves sufficient to specify syntactic structure.

Again, someone who supposed that speech and writing converged at the lowest conceivable level, given the difference of modality, might expect that the most efficient form of writing would be a narrow phonetic transcription (see Edfeldt, 1960). This transcription would correspond to the output of the phonological component of the grammar, presumably the level of the linguistic representation closest to the speech signal itself. Owing to contextual variation, higher-level units such as phonemes, syllables, morphemes, or words are not consistently transcribed or explicitly demarcated in such a transcription. But, in contrast to the syntactic orthography just considered, more than enough linguistic information to specify the linguistic representation would nevertheless be implicit. Why is such an orthography not found? A partial answer is that because, as has been suggested, writing and speech are not, in fact, so simply related, there is no particular advantage to a low-level, phonetically veridical representation. Moreover, it seems more difficult to attain awareness of phonetic details insofar as they are predictable. Once the language-learner is able to represent words phonemically, the phonetic level seems to sink below awareness. But as will be seen, there is a still more fundamental reason why a narrow phonetic transcription would be impractical.

It is important to distinguish between the linguistic unit used for the actual processing of an utterance by writer and reader, and the linguistic units to which the various graphemic units correspond. Elementary graphemic units correspond to phonemes (English letters or

digraphs), syllables (Japanese kana[5]), or morphemes (simple Chinese characters). These are usually organized into complex units that have been called "frames" (Wang, 1981). A spelled word in English, a complex Chinese character, a grouping of Egyptian hieroglyphics are examples. Frames are usually demarcated by spaces in modern writing, but other demarcative symbols have been used. Sometimes the frame is implicit: The structure of the frame itself may be sufficient to demarcate it from adjacent frames, as in Japanese, where a kanji logogram or logograms is regularly followed by kana syllable signs specifying affixes. Some orthographies, such as those early alphabetic orthographies in which there is no demarcative information of any kind, have no frames larger than their elementary signs. Frames often correspond to linguistic words, but not always: In Chinese and Sumerian, they correspond to morphemes.

By "unit of transcription" is meant the linguistic unit that the writer actually transcribes and the reader cognitively translates to form the synthetic linguistic representation. One might expect that the units of transcription for a particular orthography would be those to which its frames corresponded. Thus, in English, the frames are consistent spellings of words, and the experienced reader's intuition is surely that he reads word by word and not letter by letter, as he would if the transcription unit were the segment. This intuition is borne out by demonstrations of "word superiority." In these experiments, it is found, for example, that subjects can recognize a letter faster and more accurately when it is part of a real written word than when it appears alone or in a nonword (Reicher, 1969). This result suggests that in the case of a real word, subjects can use the orthographic information to recognize the word very rapidly, and then report the letters it contains. If the segment were the transcription unit, the letters corresponding to the segments should be recognized and reported faster than the words.

However, it is possible that the unit of transcription does not really depend on the frame used in a particular orthography, but is in fact *always* the word. One reason for believing this is that the word has to be the most efficient unit of transcription, because words are the largest lexical structures. Anything smaller would require processing more units per utterance; anything larger could not be readily coded orthographically.

Chinese writing allows a test of this possibility. A Chinese word consists of one or more monosyllabic morphemes. In the writing, characters are the frames and correspond to these morphemes. Words as such are not demarcated. There is some evidence, however, that the unit of transcription is nonetheless the word. In a recent experiment (Mattingly& Xu, in preparation), Chinese speakers were shown sequences of two characters on a CRT. In half the sequences, one of the characters was actually a pseudocharacter, consisting of two graphic components that in actual writing occur separately as components of other characters, but not together in the same character. Of the sequences in which both characters were real, half were real bimorphemic words and half were pseudowords. The subject's task was to respond "Yes," if both characters in a sequence were genuine and "No," if either was a pseudocharacter. Subjects performed this task faster for words than for pseudowords, and it was possible to show that this was not simply an effect of the higher transitional probabilities of the word sequences, but rather a valid "word superiority" effect. This result, like that of an earlier experiment by C. M. Cheng (1981, summarized in Hoosain, 1991) suggests that despite morphemic framing and the absence of word

[5]Japanese kana correspond, strictly speaking, to moras, which are not equivalent to English syllables. But they do belong to a general class of phonological units that can be called "syllables" (see, e. g., Hyman, 1975).

boundaries, the word is the transcription unit for Chinese readers. Other writing systems in which words are not framed remain to be investigated.

But if word-size frames are not essential for reading word by word, why is a narrow phonetic transcription an unlikely orthography? The reason must be that the shapes of words in such a transcription are context-sensitive and thus difficult to recognize. (Notice what happens to /hænd/, *hand*, in [hæntuwlz], *hand tools*, [hæŋgrənejd], *hand grenade*, [hæmpɪkt], *hand picked*, etc.). The reader is therefore forced to process the transcription symbol by symbol, a slow and arduous procedure. In Chinese, on the other hand, though word-boundaries are absent, the form of an orthographic word is constant, or at least not subject to contextual variation. It is suggested that this is a minimal constraint that all writing systems must meet, so that words can serve as units of transcription.

Although words are the transcription units, writing always employs graphemic units corresponding to linguistic units smaller than the word. It might seem possible, in principle, to have a pure logographic system, consisting simply of one monolithic symbol for each word. But the difficulty with such a system is that while the lexicon of a language is, in principle, finite, it is in practice, indefinite: New words are continually being coined or borrowed. In some cases—a nonce word or an unusual foreign name, for example—it would make little sense to provide a special logogram. A writer could thus find himself with no means of writing a particular word because no logogram for it existed. Or, of course, he could be stuck simply because he did not know the correct logogram. An actual writing system insures that the writer will never be in this situation by providing a system of spelling units. The availability of the spelling system guarantees that the orthography will be "productive," that is, that the writer who has mastered the spelling rules will always have some way (though it may not be the "correct" or standard way) to write every word in the language (Mattingly, 1985).

The only linguistic units that have served as the basis for spelling units are syllables and phonemes. It might be thought that morphemes could be the basis of a spelling system and some (e.g., Sampson, 1985) have argued that Chinese has such a system, because the characters correspond to morphemes. This is true, but, as has already been noted, these morphemic units are frames: Relatively few of the characters in the inventory are simple logograms. Over 90% are phonetic compounds, each consisting of two graphic components that (in general) occur also as separate logographic characters. One of these, the "phonetic" stands, in principle, for a particular phonological syllable, and the set of phonetics thus constitutes a syllabary. The other, the "semantic," is one of 214 determiners that serve to mitigate the extensive homophony of Chinese: The number of monosyllabic morphemes far exceeds the number of phonologically distinct syllables. The situation is complicated, however, because there is usually more than one phonetic corresponding to a particular phonological syllable (there are about 4000 in all for about 1300 phonologically distinct syllables), and because, through various accidents of linguistic history, a phonetic often has different phonological values in different characters. But these circumstances should not obscure the highly systematic, syllabographic nature of the spelling, any more than the existence of several spelling patterns for one sound, and numerous inconsistencies in letter-to-sound correspondence, should obscure the systematic, alphabetic nature of English spelling (DeFrancis, 1989).

Words can indeed be analyzed into morphemes as well as segments and syllables, but the inventory of morphemes in a language, like the inventory of words itself, is indefinitely large and subject to continual change. While logograms that are morphemic signs can have a

valuable supplementary function in orthography, they could not constitute a productive spelling system, and there is no orthography in which they play this role.

Syllables and segments, on the other hand, have several properties that make them suitable as a basis for spelling units. First, a word can always be analyzed as a sequence of phonological elements of either type. Second, the inventory of syllables may be small (and indeed was small in all the languages for which syllabic spelling developed independently) and the inventory of segments is *always* small. Third, the membership of these inventories changes only very slowly. No other linguistic units have these convenient properties, save perhaps phonological distinctive features (Because a diacritic is used to indicate voicing, it could be maintained that features have a marginal role in Japanese spelling).

In sum, every orthography needs to have a spelling system and a spelling system is necessarily phonographic. It is not accidental that all orthographies spell either syllabically or segmentally: there is probably no other way to spell.

The invention of writing[6]

Writing was invented, probably several times, by illiterates. From what has been said already, it follows that what had to be discovered was one or the other of the two possible spelling principles, the syllabic or the segmental, and that this must have required awareness of these units of the linguistic representation. How could the inventors have arrived at such awareness?

Some linguistic units seem to be more obvious than others. Awareness of words can perhaps be assumed for most speakers, even if they are preliterate or illiterate. It probably requires only a very modest degree of awareness to appreciate that an utterance is analyzable as a sequence of syntactically functional phonological strings, if only because sequences consisting of just one such string are quite frequent: Words may occur in isolation. Certainly preliterate children have no difficulty in understanding a task in which they are to complete a sentence with some word, and a linguist's naive informant readily supplies the names of objects. Awareness of syllables as countable units may also be fairly widespread. The syllable is the basis for verse in many cultures; preliterate children can count the number of syllables in a word. This kind of syllabic awareness, however, is probably not the same thing as being aware (if such is indeed the case) that the syllables of one's language constitute a small inventory of readily demarcatable units.

These limited degrees of linguistic awareness are probably readily available to speakers of all languages. But more subtle forms of awareness may well have arisen only because they were facilitated by specific properties of certain languages, including, in particular, those for which writing was originally invented.

Consider, first, Chinese. In the Ancient Chinese language, words were in general monomorphemic, there being neither compounding nor affixation. Morphemes were monosyllabic and a particular morpheme was invariant in phonological form. Because of restrictions on syllable structure, the inventory of syllables was small. Homophony was therefore very extensive, one syllable corresponding to many morphemes (Chao, 1968).[7]

[6]An earlier formulation of some of the proposals in this section can be found in Mattingly (1987).

[7]DeFrancis (1950), protesting against the "monosyllabic myth," has suggested that there actually were many polysyllabic words in Ancient Chinese, just as in Modern Chinese, but that only one of the syllables in a word was transcribed in the writing. Thus, morphemes that appear from the writing to be monosyllabic homophones may actually have been polysyllabic morphemes with common homophonous syllables. Y.-R. Chao's (1968) response was that "so far as Classical Chinese and its writing system is concerned, the monosyllabic myth is one of the truest myths in Chinese mythology" (p. 103). For the present purpose,

The number of different characters in the Chinese writing system sharing a particular phonetic component gives some notion of the degree of homophony in Ancient Chinese, and this number often exceeds twenty. Chinese thus contrasts sharply with English and other Indo-European languages, in which morphemes vary in phonological form, may be polysyllabic, and may not even consist of an integral number of syllables; syllable structure is complex; the number of possible syllables is relatively large; and homophony is therefore a marginal phenomenon.

Since words coincided with morphemes in Chinese, awareness of morphemes required no analysis, and the use of logograms, i.e., morphemic signs, was an obvious move. The extensive homophony made "phonetic borrowing"—using the sign for one morpheme to write another morpheme with the same syllabic form[8]—a strategy that was both obvious and productive; when a writer needed to write a morpheme, a sign with the required sound was very likely to be available. It thus became obvious that the number of different sounds was in fact small, yet every morpheme corresponded to one of them. Awareness of demarcatable syllable units thus developed. Of course, the same extensive homophony that fostered the discovery of these units also meant that their signs had to be disambiguated by the use of logograms as determiners, as in the large class of characters called "phonetic compounds," described earlier.

Chinese morphophonological structure thus encouraged the discovery of the syllable; on the other hand, it did not encourage the discovery of the phonemic segment. There was nothing about this structure that would have served to isolate phonemes from syllables or morphemes.

Sumerian was an agglutinative language. A word consisted of one or two monosyllabic CVC morphemes and various inflectional and derivational affixes. Its phonology had certain properties that imply a preference for a CVCVC...VC syllabification. There were no intrasyllabic consonant clusters; a cluster simplification process deleted the first of two successive consonants across syllable boundaries, resulting in such alternations as **til**, **ti**, 'life'; and final vowels were deleted (Driver, 1976; Kramer, 1963). In other relevant respects, however, Sumerian resembled Chinese and, like Chinese, favored awareness of morphemes and of syllables as demarcatable units. Aside from the effects of the syllable-forming processes just mentioned, a root maintained an invariant phonological form. A root could be repeated to indicate plurality. Because the morphemes were monosyllabic, and because of the restricted syllable structure, the number of possible distinct syllables was small. These circumstances, resulted, again, in extensive homophony.

For a speaker of Sumerian to become aware of morphemes was perhaps not quite as easy as for a speaker of Chinese. He would have had to notice that words with similar meanings often had common components, for the most part corresponding to syllables. This stage of awareness having been achieved, morphemic writing is possible. From this point on, the story is quite similar to that for Chinese, homophony leading to phonetic borrowing, and then to syllable writing supplemented with determiners.

There is, however, one striking difference between the Sumerian and the Chinese writing systems. While Chinese makes no internal analysis of syllables, Sumerian does. A sign for a $C_1V_1C_2$ morpheme could be borrowed to write a $C_1V_1C_3$ morpheme, e.g., the RIM sign was used to write **rin**. A VC syllable sign could be used as a partial phonetic indicator after

however, it does not matter whether the myth is true or false. DeFrancis's partial homophony will serve as well as the total homophony more usually attributed to Ancient Chinese.

[8]Or, on DeFrancis' (1950) view, another morpheme having a syllable in common.

a logogram, e.g., GUL + UL. For many of the $C_1V_1C_2$ syllables, as has been mentioned, there was no special sign; instead, such a syllable was written with the sign for the C_1V_1 followed by the sign for V_1C_2. Thus the syllable ral is written RA AL (examples from Gelb, 1963). A possible explanation of these various practices is that in spoken Sumerian, consistent with its preference for CVCVC...VC structure, some form of vowel coalescence took place when two similar vowels came together, so that $C_1V_1 + V_1C_2$ sequences became phonetically $C_1V_1C_2$, and thus homophonous with original $C_1V_1C_2$ syllables. Such homophony could have suggested analyzing and so writing the latter as $C_1V_1 + V_1C_2$. Again CV signs as well as VC signs were used to indicate the endings of $C_1V_1C_2$ morphemes. For example, because of multiple semantic borrowing, the logogram DU could stand not only for **du**, 'leg,' but also for **gin**, 'go,' **gub**, 'stand,' and **tum**, 'bring'. Which of the latter three was intended was indicated by writing DU NA for gin, DU BA for gub, and DU MA for tum (Driver, 1976). This practice perhaps arose because the phonological final vowel deletion made $C_1V_1C_2$ and $C_1V_1C_2V_2$ sequences homophonous, suggesting that what followed C_1V_1 could be written in either case as if it were C_1V_2. Thus the Sumerians may have viewed $C_1V_1C_2$ morphemes either as $C_1V_1 + V_1C_2$ or as $C_1V_1 + C_2V_2$, either of which was entirely consistent with their syllabic phonological awareness.

With Egyptian, in contrast to Chinese and Sumerian, the morphology and phonology of the language of the language favored segmental awareness. In Afro-Asiatic languages, the roots are biconsonantal and triconsonantal patterns into which different vowels or zero (that is no vowel at all) are inserted to generate a large number of inflected forms. Because the vowels of Egyptian are unknown, it is easier to illustrate this point with an example from another Afro-Asiatic language, e.g., Hebrew. From the Hebrew root k-t-b are derived **kātab**, 'he wrote'; **yikkātēb**, 'he will be inscribed'; **kātob** 'to write'; **kātub**, 'written'; **miktab**, 'letter; and many other forms. Because of phonological restrictions, the number of different consonantal patterns in Egyptian was relatively small, and there were consequently numerous homophonous roots, e.g., **n-f-r**, 'good'; **n-f-r**, 'lute' (Jensen, 1970).

It is not difficult to imagine an Egyptian noticing that many sets of semantically similar words in his language had a common consonantal ground and a varying vocalic figure, though at first he may not have individuated the consonants. Accordingly, signs for root morphemes were devised. The homophony of Egyptian then did for phonetic segments what homophony in Chinese and Sumerian did for syllables. A morphemic sign was frequently borrowed to write a homophonous morpheme, e.g., NFR, the sign for **n-f-r**, 'lute', used to write **n-f-r**, 'good,' or WR, 'swallow,' used to write **w-r**, 'big.' The signs were now generalized to stand for consonantal sequences that were not morphemes, e.g., WR < WR was used to write the first part of **w-r-d**, 'weary.' And because in some cases roots were actually uniconsonantal, and in other cases the second consonant had become silent, some signs came to stand for single consonants, and constituted a consonantal alphabet. Thus the **d** in **w-r-d** could written with the sign D < DT, the final consonant in **d-t**, 'hand,' being actually the feminine suffix, not part of the root. Finally, logograms were employed as determiners to clarify ambiguous transcriptions: the spelling MN N H for the word **m-n-h** being followed by the determiner for 'plants' when this word had the sense 'papyrus plant,' the determiner for 'men' when it had the sense 'youth,' and the determiner for 'minerals' when it had the sense 'wax' (examples from Jensen, 1970). In this fashion, the Egyptians arrived at a consonantal spelling system.

If the Egyptians had thus achieved segmental awareness, why did they not transcribe the vowels as well as the consonants? It is not likely that they were unable to hear the different

vowels. The explanation is rather that because the vowels ordinarily conveyed only inflectional information, the writing was sufficiently unambiguous without such indications, just as English writing is sufficiently unambiguous without stress marking. But as has already been noted, there was a convention for writing vowels when necessary. Such writing is found very early in the history of Egyptian writing (Gelb, 1963).

The Egyptians could hardly have arrived at a syllabic system instead. Because zero alternated with vowels in the generation of words, there was no obvious correspondence between morphemes and syllables or syllable sequences. And because of such alternations, a syllabic orthography would have resulted in a number of dissimilar spellings for the same morpheme.

These examples suggest that the phonological awareness required for the invention of writing develops when morphemes have a highly restricted phonological structure—monosyllabic, in the case of Sumerian and Chinese; consonantal in the case of Egyptian—that results in pervasive homophony. Speakers of such languages are naturally guided to the invention of writing by these special conditions. (A corollary is that it is not necessary to propose a derivation of Egyptian from Sumerian to account for parallels in the development of the two systems.) On the other hand, Indo-European languages and many others lack any such restrictions, and would not have favored phonological awareness in this way. Indeed, one has to wonder whether, for such languages, writing could have been invented at all.

In the early discussion of the psychology of reading, the precise role of phonological awareness in learning to read appeared equivocal. Is phonological awareness a prerequisite for reading? Or, on the other hand, does the experience of reading engender phonological awareness (Liberman, Shankweiler, Liberman, Fowler, & Fischer, 1977)? It was later seen, however, that both statements must be true: The beginning reader must, indeed, have some degree of awareness, but this awareness is increased and diversified in appropriate directions as a result of his encounter with the orthography (Morais, Alegria & Content, 1987). In the same way, the invention of writing must have been an incremental process, beginning with an initial awareness of morphemic structure. The experience of working out ways to transcribe morphemes for which there were no logograms led to awareness of the syllabic or phonemic structure of these morphemes, and then to awareness of such structure generally.

To say that the process was incremental is not to say that it was not quite rapid. It is noteworthy that in all three of the writing traditions just considered, evidence of spelling is found very early: in Sumerian writing from the Uruk IV stratum (Gelb, 1963); in Chinese writing of the Shang dynasty (DeFrancis, 1989); in Egyptian writing of the First Dynasty (Gelb, 1963). These facts are consistent with the proposal that for general-purpose writing, a purely logographic system is impractical. As has been argued, an orthography is not productive without a spelling system: The invention of the one requires the invention of the other.

To the extent that this account of the invention of writing is plausible, it supports the dichotomy between syllabic and segmental spelling proposed earlier, for what had to be invented was one or the other of the two spelling principles that provide the basis for the classification. It should also be noted that the segmental principle did not develop in Egypt by elaborating on the syllabic principle, but rather by generalizing from the segmental transcription of morphemes: The syllable played no role. And, conversely, when Sumerians analyzed complex syllables, they did not resolve them into their constituent phonemes, but rather into simpler syllables. The discovery of one method almost seems to have guaranteed

that the other would not be discovered. In effect, speakers of these languages come to regard them as as essentially syllabic or as essentially segmental, and their writing systems reflect one of these two phonological theories.

Transmission of writing systems

It has already been noted that orthographic traditions are either consistently syllabic or consistently segmental. Some explanation for this consistency is required. It seems natural enough, perhaps, that a segmental tradition should not become syllabic, for this would appear to be a backward step. But that no syllabic tradition should have become segmental is puzzling, the more so because there have been at least two occasions when such a development might reasonably have been expected. The first was when speakers of Akkadian, an Afro-Asiatic language with consonantal root structure similar to that of Egyptian and Hebrew, borrowed Sumerian syllabic writing. A proper awareness of the morphophonology of their language would have suggested that they convert the Sumerian system into a consonantal system. But instead, the Akkadians preserved the syllabic character of the borrowed writing, even though to write the same triconsonantal pattern in different ways depending on the particular inflectional vowels obscured the roots of native words. Similarly, the Mycenaean Greeks borrowed Minoan syllable writing, and instead of making an alphabet out of it, as would have been sensible, given the extensive consonant clustering in Greek, they continued to write with signs that stood for CV syllables, either ignoring the "extra" consonants or pretending that they were syllables. This resulted in such bizarre transcriptions such as A RE KU TU RU WO for **alektruōn**, 'cock' (Ventris & Chadwick, 1973). What can have happened to linguistic awareness in these cases?

The explanation begins with the observation that the mismatches between language and writing observed for Akkadian and Mycenean Greek are not unparalleled; they are simply fairly extreme cases. While an originally invented writing system clearly reflects the morphophonological structure of the language it was invented to write, this situation is obviously exceptional. In general, the system used at a particular time to write a particular language has been inherited from an earlier stage in the history of that language, or has been adapted from a system (itself perhaps an adaptation) used for some other language, or, most commonly, both. The consequence, in many cases, is that the writing often seems very poorly suited to the spoken language. If Akkadian and Mycenaean Greek illustrate the risks of borrowing, the English writing system is a good illustration of the effects of orthographic inheritance. The phonology of English has changed considerably since the fifteenth century, most notably in consequence of the Great Vowel Shift, but the writing system has remained very much as it was then (Pyles, 1971). As a consequence, the system has a number of features that must seem very peculiar to the foreigner learning English: For example, the same letter is used to write phonetically dissimilar vowels, a tense vowel is denoted by an E after the following consonant, and a lax vowel is denoted by the doubling of this consonant. A similar account could be given for Chinese writing, which corresponds more closely to Classical Chinese than to any modern dialect.

It cannot be doubted, given what has been learned in recent years about the relation between orthographic structure and learning to read in modern languages, that such complications place a heavy burden on the learner (Liberman, Liberman, Mattingly, & Shankweiler (1980). What is surprising, given the close connection between literacy and awareness of linguistic representations, a connection clearly essential in the invention of writing, is that readers and writers have so often happily accepted (once they have learned

it) an orthography that seems poorly matched to their language. It might have been expected that Akkadian cuneiform would have been rejected as soon as it was proposed, and that English orthography would by now have been abandoned as obsolete. But, instead, it is reported that the Akkadians believed their writing system to be of divine origin (Driver, 1976), and Chomsky and Halle (1968) say that "conventional [English] orthography is...a near optimal system for the lexical representation of English words" (p. 49).

In the case of inherited orthographies, the explanation may be that the orthography itself may determine not only which aspects of linguistic representations are singled out for awareness, but perhaps, indirectly, the character of these representations themselves. This could come about if the orthographically based, synthetic input representations were taken seriously by the language processor as evidence about the structure of the language, and thus led to adjustments in the beginning reader's morphophonology. It will be recalled that according to the sketch of the reading and writing process given earlier, the processor does not distinguish synthetic representations from natural ones. Consistent with this possibility is the fact that orthographic conventions sometimes mimic phonology: The conventions for marking English tense and lax vowels invite the reader to assume that underlying lax vowels become tense in open syllables and underlying tense vowels become lax before underlying geminate consonants. Such pseudophonological rules, as well as derivational morphological relations as those between *heal, health* or *telegraph, telegraphy,* though at first having merely orthographic status, may acquire linguistic reality for the experienced reader.[9] For such a reader, the orthography corresponds to linguistic representations because the representations themselves have been appropriately modified, and English orthography now indeed seems "near optimal."

In the case of borrowed orthographies, a similar explanation may apply. The phonological awareness of a borrowing group, such as the Akkadians or the Greeks, was not guided by peculiarities of their own spoken language, as was the awareness of the original inventors of writing, but by the writing system they were borrowing. This is hardly surprising: The borrowers were not sophisticated consumers, comparing competing technologies to decide which was better for their particular needs. They did not realize that there was a choice that could be made between the two different spelling principles and the theories of phonology implicit in each. They simply embraced unquestioningly the spelling principle—syllabic in the cases considered above—used by the culture under whose influence they had come, just as beginning readers accept the principle of the writing system they inherit. This principle having been accepted, the morphophonologies of the borrowers adjusted so that their linguistic representations became, in fact, a good match to their syllabic orthographies.

If this account is correct, it has to apply to the transmission of segmental systems, as well. A segmental system has obvious advantages over a syllabary for languages with complex syllable structure. But the spread of the alphabet is perhaps to be explained by an appeal to the forces of tradition rather than to those of reason.

An orthographic tradition can perpetuate itself because it offers a particular brand of morphophonological awareness ready-made. The processes of introspection needed to invent writing in the first place are not demanded. The kind of awareness offered may be poorly matched to a particular language, but this does not impede the process. Whether the

[9]These changes in the morphophonologies of individual readers have, by hypothesis, no basis in the spoken language and are transmitted only from writer to reader, and not from mother to child. Thus, though psychologically real, they are not part of the grammar of the language as usually conceived of.

writing system is borrowed or inherited, the morphophonology of the new reader adjusts to meet the presuppositions of the system.

Conclusions

It has for some time been widely agreed that the notion of linguistic awareness is essential for an understanding of the reading process, the acquisition of reading and reading disability. This notion is likewise essential for an understanding of the invention and dissemination of orthographies. There are really only two possible ways to write, the syllabic method and the segmental method, because only by using one of these two methods is the writer assured of being able to write any word in his language. But for an illiterate to discover either of these methods, and thus be in a position to invent writing, requires awareness of the appropriate unit of linguistic representations. Awareness of syllables, or, on the other hand, of segments, is fostered by special morphophonological properties found in those languages for which writing systems were invented, though by no means in all languages. But once it has become established, the writing system itself shapes the linguistic awareness, and even the phonology, both of those who inherit the system and of those who borrow it to transcribe some other language. Thus, in the history of writing, syllabic and segmental traditions are clearly distinguished.

Acknowledgment

Preparation of this paper was supported in part by NICHD grant HD-01994 to Haskins Laboratories. Alice Faber, Leonard Katz, Alvin Liberman, and Yi Xu gave helpful comment and criticism on an earlier draft.

References

Baddeley, A. D. (1986). *Working memory.* Oxford: Clarendon Press.

Chao, Y.-R. (1968). *Language and symbolic systems.* London: Cambridge University Press.

Cheng, C. M. (1981). Perception of Chinese characters. *Acta Psychologica Taiwanica, 23,* 137-153.

Chomsky, N. (1957). *Syntactic structures.* The Hague: Mouton.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English.* New York: Harper & Row.

Cross, F. M., & Freedman, D. N. (1952). *Early Hebrew orthography: A study of the epigraphic evidence.* New Haven: American Oriental Society.

DeFrancis, J. (1950). *Nationalism and language reform in China.* Princeton, NJ: Princeton University Press.

DeFrancis, J. (1989). *Visible speech: The diverse oneness of writing systems.* Honolulu: University of Hawaii Press.

Driver, G. R. (1976). *Semitic writing: From pictograph to alphabet.* London: Oxford University Press.

Edfeldt, A. W. (1960). *Silent speech and silent reading.* Chicago: University of Chicago Press.

Edgerton, W. F. (1952). On the theory of writing. *Journal of Near Eastern Studies, 11,* 287-290.

Fodor, J. A. (1983). *The modularity of mind.* Cambridge, MA: MIT Press.

Gelb, I. J. (1952). *A study of writing: The foundations of grammatology.* Chicago: University of Chicago Press.

Gelb, I. J. (1963). *A study of writing* (rev. ed.). Chicago: University of Chicago Press.

Hill, A. A. (1967). The typology of writing systems. In W. M. Austin (Ed.), *Papers in linguistics in honor of Leon Dostert* (pp. 92-99). The Hague: Mouton.
Hoosain, R. (1991). *Psycholinguistic implications for linguistic relativity: A case study of Chinese*. Hillsdale, NJ: Lawrence Erlbaum.
Hyman, L. (1975). *Phonology theory and analysis.* New York: Holt-Rinehart-Winston.
Jensen, H. (1970). *Sign, symbol and script: An account of man's efforts to write* (3rd ed.). Tr. G. Unwin. London: G. Allen & Unwin.
Kelley, D. H. (1976). *Deciphering the Mayan script.* Austin. TX: University of Texas Press.
Kramer, S. N. (1963). *The Sumerians: Their history, culture, and character.* Chicago: University of Chicago Press.
Liberman, A. (1992). The relation of speech to reading and writing.
Liberman, A. M., & Mattingly, I. G. (1991). Modularity and the effects of experience. In R. R. Hoffman & D. S. Palermo (Eds.), *Cognition and the symbolic processes: applied and ecological perspectives* (pp. 33-38). Hillsdale, NJ: Lawrence Erlbaum.
Liberman, I. Y., Liberman, A. M., Mattingly, I. G., & Shankweiler, D. (1980). Orthography and the beginning reader. In J. F. Kavanagh & R. Venezky (Eds.), *Orthography, reading, and dyslexia* (pp. 137-153). Baltimore: University Park Press.
Liberman, I. Y., Shankweiler, D., Liberman, A. M., Fowler, C., & Fischer, W. F. (1977). Phonetic segmentation and recoding in the beginning reader. In A. S. Reber & D. A. Scarborough (Eds.), *Towards a psychology of reading* (pp. 207-225). Hillsdale, NJ: Lawrence Erlbaum.
Mattingly, I. G. (1972). Reading, the linguistic process, and linguistic awareness. In J. F. Kavanagh & I. G. Mattingly, *Language by ear and by eye: The relationships between speech and reading* (pp. 133-147). Cambridge, MA: MIT Press.
Mattingly, I. G. (1985). Did orthographies evolve? *RASE remedial and special education, 6(6),* 18-23.
Mattingly, I. G. (1987). Morphological structure and segmental awareness. *CPC cahiers de psychologie cognitive, 7,* 488-493.
Mattingly, I. G. (1991). Reading and the biological function of linguistic representations. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the Motor Theory of Speech Perception* (pp. 339-346). Hillsdale, NJ: Lawrence Erlbaum.
Mattingly, I. G., & Xu, Yi (in preparation). *Word superiority in Chinese.*
Morais, J., Alegria, J., & Content, A. (1987). The relationships between segmental analysis and alphabetic literacy: An interactive view. *CPC Cahiers de Psychologie Cognitive, 7,* 415-438.
Morais, J., Cary, L., Alegria, J., & Bertelson, P. (1979). Does awareness of speech as a sequence of phones arise spontaneously? *Cognition, 7,* 323-331.
Potter, R. K., Kopp, G. A., & Kopp, H. G. (1966). *Visible speech* (2nd ed.). New York: Dover.
Pyles, T. (1971). *The origins and development of the English language* (2nd ed.). New York: Harcourt Brace Jovanovich.
Ray, J. D. (1986). The emergence of writing in Egypt. *World Archaeology, 17(3),* 308-316.
Read, C., Zhang, Y.-F., Nie, H.-Y., & Ding, B.-Q. (1986). The ability to manipulate speech sounds depends on knowing alphabetic reading. *Cognition, 24,* 31-44.
Reicher, G. M. (1969). Perceptual recognition as function of the meaningfulness of the stimulus material. *Journal of Experimental Psychology, 81,* 275-280.
Sampson, G. (1985). *Writing systems A linguistic introduction.* Stanford, CA: Stanford University Press.
Ventris, M., & Chadwick, J. (1973). *Documents in Mycenaean Greek* (2nd ed.) Cambridge: Cambridge University Press.
Wang, W. S.-Y. (1981). Language structure and optimal orthography. In O. J. L. Tzeng & H. Singer (Eds.), *Perception of print: Reading research in experimental psychology* (pp. 223-236). Hillsdale, NJ: Lawrence Erlbaum.