A Brief History of Documents and Writing Systems

1

Henry S. Baird

Contents

Abstract

This chapter provides a review of the history of written language, with emphasis on the origins and evolution of characteristics which have been found to affect – and in some cases continue to challenge – the automated recognition and processing of document images.

Keywords

Alphabet • Document • Font • Glyph • Graph • Grapheme • Image quality • Language • Layout style • Punctuation • Reading order • Script • Style • Syllabary • Text blocks • Text lines • Typeface • Writing system • Words

H.S. Baird
Computer Science and Engineering Department, Lehigh University, Bethlehem, PA, USA
e-mail: baird@cse.lehigh.edu

D. Doermann, K. Tombre (eds.), *Handbook of Document Image Processing and Recognition*, DOI 10.1007/978-0-85729-859-1_2,

Introduction

Gaur's profusely illustrated history of writing [2] makes clear that in addition to characters handwritten or machine-printed on flat surfaces, human communication embraces rock paintings (by many prehistoric people), message sticks (Aboriginal Australians), beans marked with dots and lines (Incas), *quipu* (knotted cords of hair or cotton; Inca plus China, Africa, Polynesia, etc.), wampum belts (North American Natives), sets of cowrie shells (Yoruba of Nigeria), and strings of tally sticks (Torres Strait Islands). Although none of the above has yet to be read automatically by computers, it is conceivable that the document image analysis R&D community will someday attempt them.

The Origins of Writing

Gaur distinguishes between "thought writing" (which "transmits an idea directly," e.g., "the drawing of a tree means "tree") and "sound writing" in which the "phonetics" of speech is made visible by a conventional set of signs (Sampson calls this a "glottographic" system [7]). Parkes [6] describes the early relationship between speech and writing in the West as follows:

> In Antiquity the written word was regarded as a record of the spoken word, and texts were usually read aloud. But from the sixth century onwards attitudes to the written word changed: writing came to be regarded as conveying information directly to the mind through the eye, and Isidore of Seville (c. 560–636 CE) could state a preference for silent reading which subsequently became established as the norm.

Daniels and Bright's [1] 1996 survey of the world's writing systems lists over 50 major families, some with a dozen or more subfamilies. The vast diversity of (especially phonetic) writing conventions suggests that many are largely arbitrary cultural inventions: this variety remains of course the single most confounding technical challenge to automatic recognition. The key stages of the evolution of writing systems are also imperfectly understood and may always remain so. Perhaps as a result, the present state of writing systems appears largely chaotic: few broadly applicable rules are evident. Even systems with long historical records – notably Chinese – tend to defy completely systematic analysis. Similar pessimism, qualified no doubt by progress in modern linguistics, might be extended to the thousands of languages known, only a small fraction of which enjoy a writing system. (In 2009, the SIL Ethnologue [4] listed 6,909 living human languages and estimated that 7,000–10,000 more or less distinct living languages exist.)

Writing System Terminology

First, a brief review of terms used to describe the appearance of writing systems, starting, as modern document image analysis systems usually do, with an image of a single sheet of paper (a "page") with a message inked on it. This image may contain

a mixture of text and non-text regions. Textual regions typically contain blocks (or "columns") of text organized into text "lines," which (according to the language) may run horizontally or vertically (very rarely, in a spiral). Within a block, text lines are read typically top to bottom (for horizontal lines) and left to right (for vertical lines); this choice, which may seem arbitrary, nevertheless holds, interestingly, for many ancient texts as well as for virtually all modern texts. The order of reading within text lines also varies by language; in some ancient texts, the order switches from line to line so that if one line was read left to right, the next was read right to left (a technical term for this is *boustrophedonic*, from Greek "as the ox plows"). Text lines contain symbolic images of words from the language (and punctuation, discussed below). Almost universally, the words are written in the same order that they would be spoken.

Resuming the review of basic terminology, into what smaller elements should a text-line image be segmented? In all Western European (and many other) writing systems, a "word-space" convention assists breaking text lines into "word" images; although, these may contain punctuation and thus not map directly onto linguistic words. Even in these systems, automatic segmentation can be difficult to achieve reliably using purely "geometric" cues (such as the distribution of horizontal spaces separating characters, scaled by estimates of local type size): ambiguities often require assistance from symbol recognition and even from higher levels of interpretation.

Reading Order and Segmentation

Since in spoken language words occur in time, in sequence one after the other, almost all phonetic writing is also arranged linearly in space. By contrast, most writing media are two dimensional (at least), but the linear convention copied from speech seldom exploits these extra dimensions. Some "primitive" writing such as the Yukaghir messages discussed by Sampson does not encode any fixed word order and so can be "read out loud" in a multiplicity of narratives; and some modern "super-textual" writing, such as mathematics and music, expands beyond one dimension (more on this later in the book). The fact that once a page of text has been decomposed ("segmented") into blocks and text lines and the intended reading order inferred, recognition faces a decisively simpler class of linearized problems.

Independently of the document analysis community, the speech recognition (and more generally the computational linguistics) R&D community had, beginning in the 1970s, discovered the power of a class of dynamic programming optimization algorithms in analyzing times series problems. Methods that depend for efficiency on linear ordering include grammars, Markov models, Hidden Markov models, dynamic time warping, finite state transducers, and so forth. These algorithmic advances, which revolutionized other fields, were slow to penetrate the document analysis community until the early 1990s; but this process has now taken hold. It is not widely appreciated that most of these dynamic programming methods depend for their efficiency on special properties (often called "optimal substructure," also known, earlier, as "the principle of optimality") that apply to many one-dimensional

problems but which rarely generalize to higher dimensions. Many two (and higher)-dimensional optimization problems seem to be intrinsically harder in this sense: this may account, in part, for the relatively slower evolution of layout analysis methods compared to text recognition methods.

Other languages, notably the major modern East Asian languages, lack a word-space convention, and so the next level of segmentation must be directly related to individual symbols. In some writing systems, such as Arabic, a linguistic word is written as a sequence of spaced-out connected groups of symbols ("sub-words"): true word spaces exist alongside interword breaks and thus complicate segmentation. Much handwriting is cursive, in which many (or all) symbols within a word are connected. Even in some machine-printed systems, such as Arabic, calligraphic influences strongly survive, and typefaces are designed to imitate careful but still cursive handwriting.

It is natural to suppose that all these language and writing system dependent policies have been modeled and implemented as a segmentation algorithm able to detect and isolate, from an image of a text line, each individual *symbol*, which normally will be a fundamental unit of the written language, such as a character shape permitted by the alphabet or syllabary.

However, exceptions to a straightforward one-to-one mapping between linguistic symbols and images of symbols are surprisingly frequent. Ligatures (and digraphs, and contractions generally) merge two or more linguistic symbols into a single written character. In some writing systems, e.g., Medieval manuscripts, the number of contractions permitted can outnumber the letters of the underlying alphabet. The impact on document recognition engineering can be daunting: in effect, for purposes of image recognition, the alphabet has expanded, perhaps by a large factor, increasing the effort of collecting labeled samples for each class. More seriously perhaps, the set of "characters" to be found in the document image may be unknown at the outset; variations may be discovered on the fly; what is a legitimate variation – not a typo or misinterpretation – may not be clear; and professional historians may need to be consulted. In this sense, many calligraphic writing systems, even in the West, are "open," lacking a fixed set of conventional letterforms.

An image of a correctly isolated symbol is called a *graph* (some authorities prefer the term *graphemes* for what are here called symbols). Now consider the set of all graphs segmented from a document image; the task of a *character classifier* is to assign to each graph its correct linguistic character label (in the case of contractions, the correct output is a string of linguistic labels). Now graphs for the same symbol can be expected to differ in detail, due to the vagaries of printing (e.g., text size, inking, paper quality), handwriting, imaging (point-spread function, scanning resolution, etc.), and variations even in segmentation style. Of course such variations are a principal technical challenge of text-image classifier design.

But there are deeper challenges, due to variations of other kinds. In some writing systems, more than one shape is allowed to be used to represent a single symbol: such a set of visually dissimilar but linguistically identical character shapes are sometimes called *allographs*. The underlying shapes may be so different that they must, as a practical matter of classifier training, be separated into distinct classes:

in this case, once again, the classes required for image recognition cannot be mapped one to one onto linguistic classes. But, from another point of view, an inability to generalize, during training, across dissimilar allographs can be judged a symptom of inadequacy in the trainable classifier technology, and if this criticism is justified, then couldn't the technology also fail to generalize across other variations such as extremes of image quality? Indeed, document recognition engineers often feel the necessity of making manual adjustments – which a linguist might regard as irrelevant and distracting interventions – to the labeling of training sets, splitting and combining classes, or organizing them into tree structures. The hope of minimizing this potentially open-ended manual "tuning" is one motive for trying classification trees (CARTs); unfortunately, training good trees has always been either computationally prohibitive or weakly heuristic. Note that these problems, due to certain "open-ended" characteristics of writing systems and typographic conventions, can arise even for modern languages in high-technology cultures.

Furthermore, there is the problem (and promise) of *style*: an individual's writing personality is an example, as are typefaces in machine-print; also image quality can be thought of helpfully as a kind of style (much more on this later).

Types of Writing Systems

Harris' 1986 history of writing systems [3] attempted to classify the various types of symbols (he says "signs") used in the writing systems of the world as follows:

> *Alphabetic*: a set of symbols representing the complete set of consonants (e.g., "s") and vowels (e.g., "a") occurring in speech, as in English and most classical and modern Western scripts (perhaps "ultimately from the North Semitic alphabet of the 2nd half of the 2nd millennium B.C.")
> *Syllabic*: a set of symbols, one for each syllable (short consonant-vowel or consonant-vowel-consonant combination), e.g., "ka" (Japanese)
> *Logographic*: a set of symbols "representing words but giving no indication of pronunciation," as in "$" to indicate "dollar," and frequently throughout the Chinese Han system (used also in Japan and Korea)
> *Pictographic*: symbols "which take the form of a simplified picture of what they represent," as in "a circle with rays" to indicate the sun and arguably in certain Egyptian hieroglyphs
> *Ideographic*: symbols "representing an idea of a message as a whole, rather than any particular formulation of it," as in "an arrow sign" to indicate direction

Although this taxonomy is simplistic (and still somewhat controversial), it should be clear enough for the purposes of this chapter. The principal implications for document recognition are that (a) alphabetic, syllabic, and logographic systems dominate almost all modern (and many ancient) scripts; (b) the recognition of pictographic and ideographic writing systems has been relatively neglected by the OCR community (except for "logo" recognition in business documents), though this may change radically as the challenges of "cityscape scenes" are taken more seriously, including the problems of detection, isolation, recognition, and interpretation of traffic signs and the rapidly growing set of "international" signs and symbols; (c) alphabets tend to be much smaller than syllabaries which in turn

are very much smaller than logographic sets, with important implications for the engineering costs of supervised training; and (d) while alphabets and syllabaries are typically "closed" (complete and fixed), logographic systems tend to be "open" (incomplete, freely extensible).

It is difficult to generalize across all the variations exhibited in writing systems. However, one strong tendency, in virtually all phonetic writing systems is towards the use of compact "physical support" for individual symbol images: that is, they tend all to fit within small non-overlapping cells of approximately equal size.

The implications for document image recognition are daunting: in order to process a new language, several hurdles must be overcome, including: descriptions of all the graphemes used, collection of samples of glyphs (many for each grapheme, and more for each distinct style), analysis of page layout conventions, amassing dictionaries (lexica or morphological analyzers), at least. Some of these hurdles may require assistance from professional linguists.

Origins of Writing Media

A wide range of materials have served for early writing: Gaur highlights stone, leaves, bark, wood generally, clay, skin, animal bones, ivory, bamboo, tortoiseshell, and many metals notably copper and bronze. Although relatively perishable, one Egyptian wooden writing board survives from about 2000 BCE. "Some of the earliest Chinese writing" survives on "oracle" bones from about 1700 BCE. Waxed writing tablets, conveniently reusable, originated as early as the 8th C. BCE, and were used pervasively by the ancient Greeks and Romans; Roman laws, however, were published by inscription on bronze tablets which were displayed on doors.

The scale of production of certain writing media grew remarkably, even in ancient times, starting with clay tablets in Mesopotamia and continuing with papyri in Egypt. In South and Southeast Asia, palm leaves were the dominant medium until modern times. Vast corpora of palm leaves survive, many containing Jain, Buddhist, and Hindu scriptures: these have already been the object of serious document image recognition research. The rapid growth of interest worldwide in preservation of and access to historical documents seems likely to leave few of these arcane document types untouched, and reveal many new technical challenges.

Note that each of the three writing cultures above amassed huge collections of documents which were apparently intended to be highly uniform in material, size, and appearance, including the order in which symbols were written and in the shapes of symbols. The evidence is strong for large cadres of professional scribes trained in uniform practices. The wide diversity – indeed profusion of creative variations – in modern writing styles, which one may be tempted today to take for granted, was not the norm in early societies, and it accelerated only with the industrial age. A significant technical trend in today's document recognition research is interest in *style-conscious* methods which can exploit known (or merely guessed) uniformity on the input images. The older a written corpus is, the more likely it is to be

crafted in a uniform style: thus modern style-conscious methods may turn out to be particularly (even surprisingly) effective when applied to pre-modern documents.

Another important implication is that each medium could, and often did, affect the evolution of writing styles. For example, the introduction of serifs occurred in monumental classical inscriptions (such as a highly influential Trajan column), driven by technical constraints peculiar to carving (chiseling) marble. The survival of serifs into modern times has ostensibly been due to aesthetics, though one could argue that they also aid in legibility.

Some writing materials were (and remain) far most costly than others. The expense of relatively long-lasting media such as vellum drove the development of elaborate Medieval scribal conventions to save space, including an explosion of concise contractions and diacritical marks.

Punctuation

Parkes' profusely illustrated 1993 study [6] showed that, in the West at least, by the Middle Ages,

> punctuation became an essential component of written language. Its primary function is to resolve structural uncertainties in a text, and to signal nuances of semantic significance . . . .

Nevertheless the functions of punctuation have received relatively little attention by classical and even modern computational linguists. One exception is Sproat's 2000 formal theory of orthography [8] embracing several modern writing systems including Russian, Chinese, and Hangul (Korean): his principal aim is to analyze encoded text corpora in order to drive (control) an intelligible text-to-speech synthesis system; he shows that this requires finite-state models at both "shallow" and "deep" levels; and he suggests that complete models of this kind are unlikely to be learnable from training data using purely statistical inference. Nunberg's 1990 thoughtful study [5] revealed that punctuation rules in English are more complex than the regular expressions used in state-of-the-art OCR machines would suggest.

Conclusion

Some clear trends in the history of writing systems that are potentially important to the document image analysis R&D community have not, as far as is known, received sustained scholarly attention of any kind. Careful studies of the reasons for the early and persistent dominance of monochrome (bilevel) documents are not known to the present writer. Although much is known about the evolution of certain, mostly Western and Asian, alphabets (and syllabaries, ideographic systems, etc.), details are often missing concerning the crucial shifts from open-ended symbol sets to finite and fixed sets. (It is interesting to contrast this fact with the persistent open-endedness of dictionaries in all living languages.) In most writing systems with long histories, the graphs of symbols have evolved steadily from complex

to relatively simplified forms. Within living memory, the Han writing system underwent a dramatic refinement to smaller symbol sets and simplified glyph forms. One such event, which now seems anomalous and even embarrassing, occurred when manufacturers of early OCR systems despaired of coping with naturally occurring printed text and invented "OCR fonts" such as OCR-A and OCR-B to make their problem simpler and then seriously (if ineffectively) proposed them for widespread commercial use.

Perhaps most of these events, all of which have tended to simplify the task of recognition, have been driven by a slow but inexorable broader impulse towards standardization and automation. But this explanation begs the question of whether or not simplification will continue unabated.

Cross-References

▸Document Creation, Image Acquisition and Document Quality
▸The Evolution of Document Image Analysis

References

1. Daniels PT, Bright W (eds) (1996) The world's writing systems. Oxford University Press, Oxford
2. Gaur A (1984) A history of writing. Cross River Press, New York
3. Harris R (1986) The origin of writing. Open Court, La Salle
4. Lewis MP (ed) (2009) Ethnologue: languages of the world, 16th edn. SIL International, Dallas. Online version: www.ethnologue.com
5. Nunberg G (1990) The linguistics of punctuation. CSLI: Center for the Study of Languages and Information, Menlo Park
6. Parkes MB (1993) Pause and effect: an introduction to the history of punctuation in the West. University of California Press, Berkeley
7. Sampson G (1985) Writing systems. Stanford University Press, Stanford
8. Sproat R (2000) A computational theory of writing systems. Cambridge University Press, Cambridge

Further Reading

The voluminous references offered here may suggest interesting further reading.