Creating an Orthography Description

M. Hosken

For those working with minority languages, one of the first needs is the ability to work with the orthography of the language on a computer. The prerequisite for this is an adequate description of the orthography. But how does one go about that description? What is needed to form such a description? This paper provides a detailed outline for such a description with practical suggestions for a variety of script families and addressing particular domain specific issues.

1 Introduction

"Can you help me type my language?" A simple enough question, but it has only one real answer: "How do you write your language?" The purpose of an orthography description is to answer that question in sufficient detail to enable people to create a system in which people can type and process text in the language.

This paper provides an outline for writing orthography descriptions. These descriptions[1] form an essential piece of documentation about a language and are foundational to many other language development activities including computing and literacy.

Unfortunately, not all orthographies are simple to describe and there may be some aspects of a particular orthography that do not lend themselves to easy description. For example, the line breaking of Thai text is an ongoing research topic. Just because some aspect is not easy to describe does not mean that no attempt should be made at an orthography description. Even a partial description is better than no description at all. But readers should be made aware of the limitations of a particular description and be ready to do further research if that is called for.

This paper describes what needs to be in a full description of an orthography. The easiest way to do this, is to make the description relative to another script, e.g. a national script. This is the most common approach used in describing minority language orthographies, presuming there is an orthography description available for the national script. This approach usually results in a listing of the differences between the orthography being described and a reference orthography, such as that for a national language.

We use the term *orthography* to mean one of the ways that people write text in a particular language. While traditionally, the definition of an orthography includes spelling, our concern here is to provide sufficient information about an orthography to be able to create an implementation of a *writing system*. A writing system is often used in the computing domain to as the implementation of a particular orthography (or in some cases, orthographies, for mixed orthography writing systms, like Japanese. This differs from languages which have multiple writing systems).

We use the term *script* or *script family* to describe a common set of characters and behaviours used by related orthographies in different languages. More formally, the script is actually the superset of characters and behaviour from the related orthographies. There is often a dominant orthography which is used to provide a foundational description of a script. For example, the Cyrillic script's foundational orthography is Russian, even though the script is used for many more orthographies and has many more characters than are used in Russian. The term *script family* is used to accentuate this superset characteristic of a script.

1.1 Document or Formalism?

The orthography description is not a database but a prose document. It does not represent a formalism or a form to be filled out, but an outline for a document. As such, this paper

[1] Some languages have more than one orthography.

concentrates on the what of description over the how. Having said all this, a typical orthography description will have many tables in it.

Attempts have been made to formalise orthography description, to the level of producing an XML schema[2]. The problem is that conforming the description to the formal model itself is, in effect, a form of encoding of the orthography. And it is even harder to conform some descriptions to a formalism than it is to express them in prose. This does not nullify those attempts to formalise, but the presupposition taken here is that informal information is far more valuable than no information at all and the world needs this kind of information in any form it can get it!

The process of trying to create an orthographic description on paper is usually done using a computer which means that the orthography needs to be encoded in order to describe it. But whether a font exists for the orthography or photographs are used for the various glyphs, or the whole description is hand written, our focus is on the orthography being described and not the process of description.

The breadth of information and issues that an orthography description may have to address cannot simply be represented using any one formalism. It should also be noted that an orthography description alone may well be insufficient to enable implementation of some of the more complex scripts. For example, it would be unreasonable for a linguist writing such a document to give full details of what constitutes fine typography in the orthography being described. The orthography description is designed to provide a basic description sufficient to facilitate the creation and interaction with legible text. Typographic finesse is beyond the scope of such a document. Such a description would be considered follow up research, which may only be necessary for a small minority of orthographies. Likewise it is also not the purpose of this document to describe a full phonology of the language for the purposes of text to speech conversion.

1.2 Orthography Description or Statement?

There is currently an existing document in use by literacy specialists and linguists, called an *orthography statement*. The purpose of this document is to justify the decisions that went into creating a particular orthography. As such it is written in terms of the orthography's linguistic and sociolinguistic basis and contains the details required to justify the decisions made.

An orthography description is a different, but related, document. It's purpose is to describe an already established orthography.[3] No justification need be given as to why the orthography is the way it is (although explanation of this kind is always welcomed). This allows the description to start with the orthography rather than the linguistics. The link between the orthography and the phonology of the language is given, but in less detail than would be given in an orthography statement.

The orthography description is intended for a wider readership, although it can be assumed that readers will have the necessary background to understand IPA (if they are concerned with the phonological aspects) or sorting principles (for the sorting section).

The outline for an orthography description given here presumes that the orthography in question is an established orthography that is currently widely used. This outline should be modified for use with experimental orthographies that are not yet established and are still changing. Careful note of this should be made if the orthography being described is still undergoing revision.

Section 2 describes the basic outline of an orthography description along with details of the kinds of questions that should be answered in each section of the document. Due to the relative complexities of different script families. Section 3 discusses some of the particular areas of extra description that may be needed for different script families. For example, a description for an Arabic orthography description will address different issues to those of a Roman based

---

2 Albright, Eric S. "Design of an Electronic Method for Describing Writing Systems" (Dallas: GIAL MA Thesis, 2001)

3 An *established orthography* is one that is in widespread use within the language community. It is usually identified with the orthography being actively used and owned by a community wider than the initial developers of the orthography, e.g. by being taught in schools, or newspaper and other ongoing materials being produced in the orthography.

orthography. The final section examines specialist information that may be needed by different users of the document. For example, computer script implementors would like to know if there are existing implementations that they can refer to. Linguists would like to understand the relationship between the orthography and the phonology.

2 Basic Outline

The basic outline lists the major sections in an orthography description and addresses the typical purposes to which the section is used.

2.1 Sociolinguistics

The first section describes the sociolinguistic context of the orthography: how many speakers of the language, readers of the orthography etc. This provides a general context for the orthography in question and gives some idea of potential sizes of user communities. Such information is important for project planning. Does this orthography have a potential user base sufficient to justify the work of creating a solution for sale? To what degree is this orthography established among a user community? Is the use of the orthography dying out or being revived? Who are these people and where do they fit into the general sociolinguistic mileau of the region? The section is best structured as answers to a series of questions:

2.1.1 How many speakers of the language are there?

This should list those who are functional speakers of the language and therefore could be potential readers of the orthography.

2.1.2 What percentage of the speakers are literate in the orthography?

Notice that language communities may be split geographically and/or socially and so speakers of a language may be literate in different orthographies. This question is not looking for a high level of literacy in the orthography. Its primary concern is that the speaker be some kind of user of the orthography.

2.1.3 Is the orthography currently being used? By whom? To what extent?

Is the orthography taught to everyone or just to members of one of the language communities? Perhaps the orthography is not yet established. To what extent is the orthography an established orthography as opposed to a developing or archaic orthography?

2.1.4 What percentage of the speakers could reasonably be expected to become literate in the orthography?

Given that some language groups span national and script boundaries, it will be unlikely that a reader of an orthography based on their national script will also become literate in an orthography based on a different national script. This may differ if there is only one, or one dominant script for the whole language community.

2.1.5 What is the attitude of each community towards the orthography?

Given that a language group may be split into multiple communities, it is important to know the attitudes of each community towards the orthography.

2.2 Rendering

Rendering, in this context, is the process of displaying or drawing text on a screen or printer. The primary mechanism for this is to use a font of glyphs (or drawings) and for a computer to position these glyphs on the screen or printer page and then to display them.

The purpose of this section is to enable the location or creation of a font for this orthography. Note that there is insufficient detail here to ensure that the font might look nice, or conform to

any typographic tradition. But it should be readable. The primary concern is that all characters are listed, along with their basic functions, that all ligatures and contextual variants are listed and that any particular glyph shape requirements are listed. By the end of the rendering section, examples of all the glyphs needed in a font should have been given.

2.2.1 Character Inventory

The character inventory gives a basic list of all the characters used in the orthography. It also provides a mechanism for describing lists of characters that are used elsewhere in the orthography description. This section provides a quick overview of what characters are needed for this orthography.

- List all characters in the alphabet[4], in what is considered the alphabetical order, if there is one. Many scripts do not include all the characters in the alphabet. Diacritics are rarely included, but they are characters just like base characters. It is, therefore, important to list all the consonants, vowels, tone marks, etc.
- List all case relationships. If an orthography has the concept of case (upper and lower case), then the relationships between characters in the various cases should be listed.
- List all the punctuation characters and their functions. Very often such information is borrowed from another orthography. If so, then this should be stated, along with any differences in this orthography. Are all the punctuation characters borrowed from another orthography, or are some not used?
- List all the consonant sequences and vowel sequences. The consonant sequences correspond to consonant clusters, i.e. we are only concerned with consonant sequences within a syllable. It is not necessary to list all syllable final consonants followed by syllable initial consonants.[5]

If any characters have names in the language, they should be listed as well, both in script and in some kind of phonetic form, if possible.

2.2.2 Behaviour

In addition to the characters that are needed to represent the orthography, it is necessary to give behavioural information about those characters. Behaviour describes the visual interactions between characters. Thus two characters next to each other may result in one, or both, changing shape or position. The purpose of this section is to give as complete a description of the behaviours of the characters in the orthography as possible. It addresses such questions as: what do diacritics attach to? Are there any shape variations? Are there any orthography wide stylistic requirements? If there is a separate acute accent used as a tone mark, what characters can it occur over? Only vowels? How about nasals?

The following is a list of questions that guide in the creation of this section.

- For each diacritic, what can it attach to? Include other diacritics, if diacritics can stack.
- Are there any required ligatures? How about optional ligatures? This is particularly pertinent to Indic based orthographies where ligatures are used for conjuncts.
- Are there any particular character shapes this orthography uses? For example there are three types of capital Eng (Ŋ ᴅ ɴ). Which one does this orthography use? Or the orthography may call for a script '*a*' versus a print 'a'.
- Describe any contextual shape variations. For example consonants in Arabic based orthographies have different forms dependent upon the position of the character in a word.

Some orthographies are based on scripts that have very complex behaviour and this section can be quite long. It is important to also describe shape variations that are in free variation, i.e.

[4] The alphabet is the set of characters needed to represent the orthography. Theoretically, it forms a subset of the script on which the orthography is based. But since many scripts are defined purely in terms of a dominant orthography, it is common for minority alphabets to include characters not considered part of the script.

[5] Unless there is some special interaction between them, as in syllable chaining in Myanmar type scripts. See Section 3.4.

those characters that can take different shapes within a single type face (as opposed to shape variations across different type faces).

2.3 Line breaking

For those orthographies based on scripts that have inter word spaces, line breaking is often not considered at all. The assumption is that all scripts break at a space. But there are many scripts that do not have interword spaces. This can mean that the description of where lines can break in a text in such an orthography may be a simple description or a 10 year research project.

Even for those orthographies for which line breaking is not considered an issue, it may be desirable to support hyphenation, especially if words can get long. Hyphenation may be simple to describe or it may be difficult (for example, English).

There is no expectation that an orthography description should solve 10 year research projects, but it should do its best. This may only be to the level of describing some of the problems that make such a description so hard, or the description may make reference to the current state of research in this area.

- Describe where line breaks may occur in a run of text in the orthography.
- If hyphenation is supported by the orthography, give the hyphenation rules that describe where hyphens may go in a word.

Hyphenation is often syllable based, and therefore is often linked to the section on syllable structure found later in the description.

2.4 Sorting

This section describes the sort order for the orthography. Given two strings of text in the orthography is there a standard relative ordering of the two words? Some orthographies have multiple sort orders. They should all be described here.

Sorting is usually described in terms of primary, secondary and even tertiary levels. Where two different strings sort equally at the primary level, then they are compared at the secondary level and so on. The alphabetical ordering is usually used for the primary level. In the case of a Roman based orthography, accents are compared at the secondary level and case is compared at a tertiary level.

Some orthographies may have special sorting requirements. For example, French based sorting sorts accents from the end of the word towards the beginning. Lao compares the final consonant in a syllable before the vowel.

2.5 Sample Text

A sample page of text should be included along with a free translation in the national language and in the language in which the orthography description is written. This text is very useful for testing implementations, and if at all possible, should not be encumbered by a restrictive copyright that does not allow it to be used for testing.

2.6 References

In addition to standard bibliographic references used in the rest of the statement, this section should also include as many of the following references as is possible or reasonable.

- Dictionaries or word lists using the orthography.
- Texts in the orthography, particularly those which are considered to have been typeset well, by the language community.
- A description of the phonology of the language sufficient to relate to the linguistic description of the orthography.

3 Script Specific Details

This section considers some of the particular issues and non-issues that various script families raise when describing orthographies based upon these scripts. Not all script families are included, but only those which are known to be used as the basis for new orthographies.

3.1 Roman scripts

Roman script is used as the basis for most orthographies. It has the advantage of having relatively simple behaviour: letters just follow each other in a sequence with no contextual variation and no changes in relative positioning. Words are easily identified, being separated by whitespace and line breaking is relatively straightforward. Hyphenation can be relatively easy or diabolically difficult.

But as we concentrate on the details of Roman based orthographies, various script specific details do stand out. A common extension of the basic Latin letters is to add accents over various letters, or other diacritics in other positions. An important question to consider is whether such diacritics are letters in their own right (signifying a change in tense, for example), or are simply ways of creating new letters (and so cannot be considered apart from their base characters).

Some diacritic base combinations may be represented in a different way for a particular orthography. If this is the case, a discussion of this should be included in the orthography description. There are also different ways in which diacritics may interact. Are diacritics stacked outwards from the base character or do they position horizontally as in Vietnamese?

Typically Roman based orthographies are based on some other orthography such as English, French, Spanish, etc. Knowing this relationship can help answer the miscellany of unanswered questions that can arise when creating a script solution. For example, how are questions marked? How do quotation marks work? An orthography may have unique approaches to punctuation, but more normally an orthography bases its use of punctuation on some other orthography.

One orthographic characteristic which is particularly important to Roman based scripts is the concept of case. There are generally considered to be three cases: lower case, upper case and title case. Title case is where only the first letter of a word is capitalised. An orthography description should describe when the various cases are used. For example, in English, a word is title cased following a full stop and for proper nouns.

Roman script based orthographies sometimes extend their alphabets by adding new character shapes. These are usually borrowed from related script families like IPA and Greek. If such borrowing occurs, it is helpful to describe the borrowing relationship rather than trying to describe the new character and its shape and possible variations.

Cyrillic and Greek scripts share many of the characteristics of Roman script in regards to orthography description.

3.2 Arabic scripts

Arabic based scripts, including closely related scripts such as Mongolian, Syriac, etc. have the special characteristic of linking. Characters change shape according to how they link to other characters in the word. Thus characters may be described as being right linking (also known as having two forms only), dual linking (link to the left and the right, or have four forms), or non-linking (isolate form only). An Arabic based orthography should describe the linking behaviour of each character and give the various contextual forms of each character.

There can be occasions when a particular character combination results in the linking behaviour being suppressed. For example the Persian plural suffix is separated from the rest of the word by stopping the suffix from linking to the main word. If the orthography in question has this sort of characteristic, it needs to be described.

Vowels are typically represented in an Arabic based orthography using combining marks. In addition combining marks are sometimes used to create new characters as a base + diacritic combination. Typically combinations of combining marks can interact and this interaction

should be described. For example, if kasra (or kasratan) and shadda can co-occur, is the kasra/kasratan rendered below the base character or below the shadda? For rendering purposes, what should be done about diacritics colliding with each other or with base characters? If two diacritics would normally coincide, which one moves and where?

Arabic letters often ligate. Thus when two characters occur in sequence, they are merged to form a ligature. The ligature may also have linking forms that need describing. It is important to know which ligatures are required (the characters in sequence *always* ligate), or discretionary (the characters in sequence *may* ligate, but not necessarily).

There are different types of number forms that are used in different Arabic based orthographies. For the orthography being described are the digits used Arabic, Persian, Urdu or some other?

Typographic style is important within the traditions of Arabic typesetting and orthographies will typically be expressed using a particular style. Appropriate styles for a particular orthography should be listed.

3.3 Indic scripts

Indic based scripts are typified by Devanagari and have two important characteristics: conjuncts and an orthographic syllable structure.

Conjuncts are ligatures representing consonant sequences, used either between syllables or for consonant clusters. Each Indic based orthography has a set of common conjuncts that are used, along with a possible further set of rarely used conjuncts. In some cases it may be that a particular conjunct is never used. Any orthography description for an Indic based script should include a discussion of which conjuncts are used when.

There are different ways of describing the syllable structure of an orthography. From a linguistic standpoint, the phonemic syllable structure is generally considered to be primary. A phonemic syllable consists of a vowel (or vowel glide) optionally preceded by a seqence of one or more consonants, called the onset, and also optionally followed by a coda of one or more consonants. The orthographic syllable is structured differently with a core of a consonant followed by a vowel. The vowel itself may be inherent (not shown) or even killed (not sounded), using a halant.

The orthographic syllable description is particularly important for Indic based scripts because vowels are sometimes rendered before their onset (as in ikar), or a final consonant may be rendered as part of a following syllable (as in candrabindu).

Indic based scripts often have different forms of vowels when they occur in the middle of a syllable (dependent vowels) or isolated (independent vowels). An orthography description should include a discussion of these. For example, if there are no independent vowel forms, the description should say that an independent vowel is simply a dependent vowel in conjunction with a special consonant.

3.3.1 Southeast Asian scripts

Brahmic based scripts are closely related to Indic based scripts, and attempts are sometimes made to describe them using an Indic model. But Brahmic scripts such as Thai, Lao, Myanmar, Khmer, etc. have very different behaviours and are best described using different models. The concept that the orthographic syllable may differ from the phonemic syllable structure along with the issue of independent vowels, remains.

One important aspect of these scripts is that there is no inter word space. This makes the issue of line breaking and the description of where line breaks may occur, a significant aspect of the orthography description.

"Line breaking is obvious! You can break a line between words." The problem is that without any way of marking word boundaries it is very difficult for a computer to locate those possible line breaks. True word based line breaking is impossible without a complete dictionary of the language and even then ambiguities may still arise. There is no expectation that an orthography description should include such a dictionary, even if subsequent implementations use that approach. Instead the orthography description should provide a fall back answer to the

question of line breaking by giving as full a description of the syllable structure, in terms of the orthography, as it can.

Even this may not be completely possible, as in the case of Thai where consonants may occur both syllable initially and syllable finally. This along with the inherent (unwritten) vowel, results in the syllable not necessarily being identifiable. Another particular problem is the idea of the pre-syllable, where a consonant is the prefix to a main syllable and there is a very short inherent vowel between the prefix and the main syllable. But the pre-syllable should not be broken from the main syllable. Pre-syllables also become an issue with sorting. Are words with pre-syllables sorted in terms of the pre-syllable or the main syllable, with the pre-syllable a secondary aspect with regard to sorting?

While Thai based scripts have a particular problem in these areas, other Brahmic scripts do have ways of indicating syllable boundaries and these should be described as part of the orthography description.

Other issues involve the concept of *syllable chaining* whereby the first letter of the next syllable is positioned as a diacritic below (or in relation to) the last letter of the previous syllable. When is syllable chaining used, and when is it not?

One of the identifying features of Brahmic scripts is the sheer number of diacritics. Given that a base character may have more than one diacritic, it may not be possible to see just by looking, which diacritic is written first and which subsequently. The issue of the relative order in which diacritics are written, should be included as part of the orthography description.

Diacritics may also collide, or not collide. Is collision avoidance important for this orthography or do diacritics just overwrite each other?

3.4 Syllabaries

In a syllabic script the consonant and vowel are inflected to form a single glyph (or fidel, in Ethiopic based scripts). Syllabaries are relatively simple to describe, often consisting of one large table listing all the possible character shapes used and their meaning. But there are some issues that need extra description.

Gemination is the process of consonant doubling and is often marked using a diacritic rather than two fidels. If gemination is specially marked, then it should be described.

Loan words often have final consonants. How these are represented in the syllabary is an important part of the orthography description.

3.5 Ideographic Scripts

Ideographic scripts are not well covered by the othography description outline given here. The core of the description of an ideographic script is often a very large dictionary. But there are general issues that do need describing, including the base radicals on which the ideographs are based and the positional and size relationships that are used to build new characters from existing characters.

New ideographic scripts are very rare and the existing ones have all been extensively analysed and described. So the needs in this area are, thankfully, relatively few.

Signed languages share some of the same characteristics of ideographic scripts in that a single character is built up to represent a word. This is an emerging area of orthography development which introduces its own complexities of description.

4 Domain Specific Issues

The main outline, thus far, of an orthography description has been primarily domain neutral. The information does not presume any domain specific expertise. This section discusses sections of the description that presume domain specific knowledge. All the information described here is valuable and should be included if at all possible. It is also of value outside the particular domain that describes it. For example, if there is a standard keyboard layout, then

literacy people would value this information. Likewise the computational linguists would find the linguistic domain information of great, even essential, value.

4.1 Computing

As we enter the computing domain, we head towards implementation details. This is particularly true of the relationship between the orthography and the Unicode standard.

4.1.1 Unicode

The boundary between the encoded form of an orthography and the orthography description itself is often somewhat grey in the computing world. There is always the temptation to encode too early, even while describing the orthography. For derived orthographies whose base orthography is already encoded in Unicode, there is no difficulty in adding a description of how the orthography can be represented in Unicode. If characters exist in the orthography but are not in Unicode, then the orthography description provides an excellent base for a proposal to have the characters added.[6]

One common area of difference between an orthography and Unicode is in the sort order. Sorting is not a normative part of the Unicode standard, so changes are allowed between orthographies. Any difference from the default Unicode sort order should be described.

4.1.2 Other Encodings

It may be that there are other encodings used for this orthography that are in common use. With the rise of Unicode as a dominant standard, these encodings will become less common. But they are still worth noting and if possible, documenting.

4.1.3 Keyboard

Is there a standard keyboard layout used by the language community for this orthography? If so, it should be described here. If there are multiple layouts in use, they should all be described along with a discussion of who uses which keyboard when.

4.1.4 Typesetting

Issues of typesetting can be a major study in themselves. But there are a few easy to address questions that can help to show whether an orthography has a rich typesetting tradition or primarily borrows its style from another orthography. Are there specific issues to address with respect to typesetting?

4.1.4.1 Text size

A common issue is body text sizes. Newly literate societies tend to need larger body text sizes, both because new readers are unused to small text sizes, and also that older people often have poorer eyesight. The complexities of character shapes in a non-Roman script will tend to result in fewer lines fitting on a page, even at small text sizes.

4.1.4.2 Numerals

Sometimes an orthography will use multiple sets of numbers, for example an orthography based set and also the Arabic set of digits. This raises the question which set should be used in which context. Sometimes one set is conventionally used for page numbers with another used in other contexts.

4.1.4.3 Text flow

Each orthography has a normal text flow direction, be it left to right, right to left, top to bottom (with lines proceeding left to right or right to left) and even more complex directions with different lines flowing in different directions. Few computer systems handle vertical text well, and such orthographies need a strategy for being rendered horizontally. Likewise, text that is

[6] With so many characters already in Unicode, great care should be exercised in ascertaining whether something is missing from the Unicode standard, before rushing to propose a new character.

typically rendered horizontally, may need to be rendered vertically, for example in signs, but not necessarily letter by letter.

4.1.5 Font Design

Orthographies tend to fall into two categories when it comes to fonts: those that have an existing typographic tradition (even if borrowed from another language) and those without. The issues raised by each are somewhat different.

4.1.5.1 No existing typographic tradition

The assumption here is that there are no, or very few fonts that can be used for this orthography and that new fonts will need to be designed. In this case a font designer needs as much information and examples as they can get. For orthographies similar, typographically, to another orthography, a difference description is very helpful. Do certain characters have a different shape in this orthography? Are their shape issues that this group feels strongly about?

In general a font designer is looking to see what parts of a letter shape are distinctive and need to be contrasted with other characters. How much variation is appreciated? In what ways can a typographic tradition be developed based on what is, at the moment, a handwritten orthography.

A font designer, therefore, is looking for many different examples of handwriting, with some indication of what readers consider to be good, clear, handwriting and what they consider to be poor handwriting. Obviously good handwriting is of the greater value, but poor handwriting helps to show what is not acceptable. If an artist can be found to draw each character scaled to a height of 2", this is of great help as providing the basis for the creation of a first font.

In some cases, while there is no typographic tradition for the particular orthography, it is related to another orthography for which there is a typographic tradition. This does not necassarily mean that the orthography in question just copies the related orthography in its typographic tradition. There are sometimes sociolinguistic factors that cause the new orthography to want to define its typography in contradistinction to the related orthography in some way. What can then become a problem is deciding whether a glyph is a new glyph or is a variant form of an existing glyph in the other orthography.

4.1.5.2 Existing typographic tradition

Newly emerging orthographies tend to be conservative in their selection of fonts. They tend towards a single body font and then venture towards a heading font. But it is noticeable for well established orthographies where artists have been able to get to work in sign writing, etc. that as soon as the technology is available, there is a flurry of new fonts created in a wide variety of styles.

Some discussion of this variation is of help in deciding where to concentrate further font design efforts and also for typesetters to choose appropriate fonts for the text they are typesetting.

4.1.6 Existing Implementations

A list of existing known implementations, along with some discussion as to the applicability of those implementations, provides an excellent basis for someone wanting to use a script. Clearly, if no such implementations exist, or their applicability is limited, then the information provides a starting point for the creation of other implementations, and as a basis for information regarding data conversion.

4.2 Linguistics

Most orthography statements are written by linguists from a linguistic point of view. Thus they describe the orthography in terms of the underlying phonology. This is helpful, particularly for other linguists, but does not make for an easy to read orthography description for other users. As much of this section should be included as possible. In effect the aim of this section is to indicate how an orthography is read or written. This is particularly useful in the computational domain when trying to ascertain whether a character should be stored as a unit or as a sequence.

It should be noted that all of the sections discussed here are also needed for literacy purposes since their aim is to provide a description of the orthography sufficient to be able to read and write the language in the orthography being described.

4.2.1 Morpho-phonemics

For the most part, orthographies represent the morpho-phonemics of the language. This holds for alphabetic scripts and syllabaries. Ideographic scripts work differently and need a different mechanism for description.

The aim of this section is to describe the correspondance between letters in the orthography and underlying phonemes in the language. In some cases such a description is straightforward, but, particularly with older orthographies (including English), the relationship is very hard to describe. The value of this section is in direct correlation to the simplicity of the description needed. Thus this section carries greater value for simple correspondances than it does for complex ones. In most cases the description can only hope to be an introduction and overview of the topic.

The purpose of this section is not to introduce a complete phonological description of the language. The purpose is to relate the phonemic description (the output of the phonology) to the orthography. This section includes a list of possible consonant clusters and vowel glides.

As part of this section, there should also be a discussion of what constitutes a syllable in the orthography (as opposed to the phonology). This is particularly useful in ascertaining possible line break points and hyphenation points in a text.

Of particular concern when describing the relationship between characters and phonemes are the issues of over differentiation and under differentiation. Over differentiation is where two different characters can be used to represent the same underlying phoneme. This makes spelling difficult since an author has to remember which of the possible letters to use in any one word. Under differentiation is where one letter may represent two different phonemes. This makes reading difficult because faced with an under differentiated letter, the reader does not know how to pronounce the word.

While it is not good practise to design orthographies with these problems, when it comes to a description of an existing orthography, it is too late to make changes. Instead a discussion of the topic should aim to give guidance as to how to deal with the problems. Does one character alternative occur relatively rarely in comparison to another? Are their general contextual rules to help with the differentiation?

Another issue is whether an orthography represents the underlying phonemic structure of the language or represents something nearer the surface phonetic form of the language. The description should relate to the appropriate level of phonology that corresponds to the orthography rather than any deep level phonological analysis.

4.2.2 Loan Words

One of the hardest issues an orthography has to deal with is how to spell words which are imported from another language. Do the words keep their original spelling or are they spelled according to how they sound when spoken in the language we are describing? Loan words also often use sounds that are not in the phonology of the receptor language. Are the words transliterated (each letter is respelled in the receptor language, such that the original spelling may be reconstituted in the source orthography) or are they transcribed (the sound of the word spelled in the receptor orthography with no intention of reconstructing the original spelling of the source orthography)?

Loan words, therefore, are important as test cases for an orthography. How are they spelled? Do they break all the standard rules of spelling, syllable structure, morpheme boundary, etc, or are they forced to conform? A study of such words is an important part of the description of an orthography.

4.2.3 Linguistic phenomena

Different languages express different linguistic phenomena. For example: prosody (tone and register), vowel length, advanced and retracted tongue root, fortis and lenis, vowel glides.

There are numerous such phenomena and it is an important part of a description that there be a discussion of how they are represented, if at all, in the orthography.

5 Conclusion

This conclusion concentrates on the relationship between the practical description approach presented here and the more formal approaches taken elsewhere, and as such is only of indirect relevance to those interested in the question that started our enquiry: “Can you help me type my language?”

This paper has described a practical outline for describing any orthography, with the aim of capturing as much relevant information about that orthography as possible. Students of the theory of writing systems may wonder how this work relates to the existing literature. This paper follows in the illustrious heritage of the work of Daniels and Bright, and Smalley. It aims to go further in its description with an aim of providing sufficient information for at least basic computational implementation to be created purely from such a description. It also aims to provide pointers to sources of extra information for those interested in areas not covered by this description.

The attempt here has been to encourage authors to use a theory neutral approach, and if necessary, to develop their own models of description as needed by the material they are trying to describe. By encouraging a prose description, authors are encouraged to concentrate on describing the orthography before them, rather than the process of conforming that description to a particular model. If authors find that what they need to describe does not fit well into the outline given here, they are encouraged to extend the outline or adjust it, as appropriate, to fit the data they have.

The temptation for those developing script taxonomies is to try to fit an orthography description into a model to aid easy categorisation. Instead of working towards a taxonomy based on script categorisation, the nature of the descriptions outlined here should result in a taxonomy based on relationship between orthographies. Orthographies are developed within a local sociolinguistic mileau, whether by direct inheritance or through interaction. How different orthographies interact is a fascinating subject in its own right. The most common interaction is borrowing. But beyond that human creativity seems to take over. There are occasions of contradistinction: "We want our characters to be like theirs, but look different[7]." There are cases of simplification, where some of the complexities in a related orthography are dropped during the borrowing process. More recently there is the resurrection of previously morabund scripts with the inherent regularising which arises when an orthography is printed.

The danger of this approach is that once we say that an orthography is based on another orthography, we do not give the new orthography the independent respect it deserves. The study of *non-Roman* scripts is a classic example of this problem. It is a falacy to think that there are two major categories of scripts: Roman and non-Roman. Even if we cut the pie into more categories, it is still necessary to study each orthography very carefully within its own sociolinguistic context, giving each orthography the respect it deserves as the primary written form of communication for a language community.

There are many topics for further work in this area. There is a huge number of orthographies that await good description, which is the primary research that is so desperately needed. These form the basis for all further research in the area of script taxonomy or script categorisation. Beyond this, there is work to be done in providing formalisms that allow computers to make use of the descriptions in a semi-automated way. Then, as the rich way in which orthographies inter-relate becomes more evident, the current taxonomies can be updated and the categories refined. But the quality of the resulting formalisms and categorisations will only be as good as the data on which they are based.

[7] This usually arises from reasons of language group identity, even to the extent of the creation of completely new scripts.

6 Bibliography

Albright, Eric S. 2001. "Design of an Electronic Method for Describing Writing Systems" (Dallas, Texas: GIAL MA Thesis)

Daniels, Peter T. and Bright, William, ed. 1995. *The World's Writing Systems* (New York: Oxford University Press)

Smalley, William, A., ed. 1964. *Orthography Studies* (London: United Bible Societies)

Sproat, Richard W. 2000. *A Computational Theory of Writing Systems* (Cambridge, Cambs: Cambridge University Press)

The Unicode Consortium. 2000. The Unicode Standard, Version 3.0 (Reading, Mass: Addison-Wesley)