Orthography development

Frank Seifart

Introduction

Written records, such as transcriptions of video-recorded speech events, are essential components of language documentations. Much of the success of a language documentation depends on casting these records in an orthography that appeals to the speech community. As a matter of fact, if it is accepted that the documentation has to be accessible to the speech community, the development and implementation of a practical orthography in the speech community is an absolutely necessary task in an early phase of a documentation project. Nevertheless, orthography development is usually not given much attention by linguists. The idea persists that a good orthography is simply one that represents all phonological contrasts. However, orthography development is in fact a highly complex issue, which involves not only phonological, prosodic, grammatical, and semantic aspects of the language to be written, but also a wide variety of non-linguistic issues, among them pedagogical and psycholinguistic aspects of reading and writing and the sociolinguistic situation.

Given the variety of language structures and sociopolitical situations found throughout the world, it is neither feasible nor desirable to propose a step-by-step model, which would lead to an optimal orthography. The aim of this chapter is rather to give an outline of the most important general issues involved in orthography development. It does so primarily by identifying a number of "factors" that are relevant when making decisions about orthographic design and by discussing the application of these factors to examples of various languages with special reference to situations of language endangerment. The focus is exclusively on the practical decisions that have to be made in the process of developing an orthography or in reforming an existing one. Wider issues of the impact of introducing literacy to oral cultures (see, e.g., Fishman 1991; Mühlhäusler 1996) or the differences between written and spoken communication (see, e.g., Ong 1982) are not considered here. The scope of this chapter is further limited in that its main focus is on alphabetical writing systems.

The basic procedure for developing an orthography outlined in this chapter begins with the analysis of the structure of a given language which will typically reveal a number of options for its orthographic representation. E.g., word-final devoicing could be represented in an orthography or not. These options are then evaluated with respect to factors that are independent of the linguistic structure, e.g. the learnability of certain types of orthographies for beginners. These non-linguistic factors will be decisive in choosing one option over the other. However, these factors are often conflicting. For instance, an orthography that represents word-final devoicing may be easier to learn for beginners, since the written form corresponds more closely to the pronunciation. However, an advanced reader may benefit from an orthography that maintains a constant written form of a morpheme, regardless of whether or not its final consonant is devoiced in some context. Thus, an essential task in developing orthographies is balancing the advantages and disadvantages of the different options and making compromises. It should be noted that these basic principles apply not only in situations where new orthographies are developed from scratch, but also in the reform of existing orthographies.

This chapter is organized as follows: Section 1 introduces the basic concepts in orthography development. Building on these concepts, Section 2 identifies a number of non-linguistic factors for making decisions in orthography development, among them psycholinguistic, sociopolitical, and technical issues. How these factors apply in specific instances is illustrated with a number of case studies in Section 3. Throughout the chapter, the following well-established conventions are used for the different kinds of representation of linguistic data: [ ] – phonetic representation; / / – phonological representation; ⟨ ⟩ – orthographic representation.

1. Basic concepts

In this section, the term *orthography* is defined and a brief overview of the typology of writing systems is given. Then the terms *orthographic depth*, *functional load*, and *underrepresentation* are introduced. These basic concepts will be further elaborated and exemplified in the sections further below.

Writing systems are systems that allow readers to reconstruct a linguistic message on the basis of written signs. Orthographies are writing systems that are standardized with respect to

a. a set of graphic symbols (*graphemes*), such as signs, characters, letters, as well as diacritics, punctuation marks, etc.; and
b. a set of rules/conventions, such as orthographic rules and pronunciation rules, rules for writing word boundaries, punctuation rules, capitalization rules, etc. (Coulmas 2003: 35; see also Coulmas 1996: 1380; Rogers 2005: 2ff.).

Importantly, then, an orthography is defined as the *conjunction* of a set of graphemes, such as an alphabet, and a set of accompanying rules regulating their use. The third defining feature is that both the symbols and their usage are standardized and codified. The actual visual shape of the graphemes that a writing system uses, e.g. the Latin or the Arabic letters, is called its script.

As a starting point to the following discussion, it is useful to take a brief look at the typology of writing systems. Most typologies of writing systems are based on the smallest unit of a system, i.e. its basic graphemes (Coulmas 1996: 1381; Rogers 2005: 269ff.). Different types are distinguished according to what kind of linguistic unit the basic graphemes correspond to.[1] Following this principle, a first type recognized in the typology of writing systems are morphographic writing systems. The basic set of graphemes of morphographic systems correspond to morphemes, i.e. linguistic elements that have a meaning.[2] A prototypical example of a morphographic writing system is Chinese. Each grapheme (i.e. character) of Chinese stands for a morpheme of the language.

The second main type of writing systems are phonographic writing systems. The basic units of these systems refer to elements of the sound structure of a language. Phonographic writing systems in turn fall into two main subtypes: syllabic writing systems and alphabetical writing systems. A prototypical example of a syllabic writing system is the Japanese Kana writing system. The graphemes of this system each refer to a syllable of the language. In alphabetical systems, the basic set of graphemes are letters that correspond (more or less directly) to the phonemes of the language. Well-known examples are the Greek and Latin writing systems.

It is important to note that within alphabetical writing systems "the range of correspondences between phonemes and graphemes varies both in consistency and in completeness" (Katz and Frost 1992: 67): A single phoneme may be represented by combinations of graphemes, such as di- or trigraphs (e.g. German ⟨sch⟩ – /ʃ/) or by combining letters with diacritics

(e.g. French ⟨a⟩ – /a/ vs. ⟨â⟩ – /ɑː/). There may also be phonemic distinctions that are not represented by letters (e.g. vowel length in Latin). Finally, a single phoneme may be represented by a number of graphemes (e.g. English /f/ – ⟨**f**un⟩, ⟨**ph**oto⟩, ⟨lau**gh**⟩) and a single grapheme may represent a number of phonemes (e.g. English ⟨b**u**ll⟩ – /bʊl/ vs. ⟨b**u**lk⟩ – /bʌlk/).

The two main types of writing systems, morphographic and phonographic writing systems, hardly ever occur in a pure form. Rather, most if not all writing systems combine phonographic and morphographic aspects. For instance, the English writing system is basically phonographic, i.e. the letters of the English alphabet represent phonemes (even though the correspondences of letters and phonemes are quite complex, as just mentioned). However, it can also be observed in English that the same morphemes are written with the same sequences of letters, even though they may be pronounced differently in different contexts. The constant written form vs. the variable pronunciation of the stems *wild* and *reduc-* and of the plural suffix *-s* are illustrated in the following examples.

(1) a. [ˈ**w**a͡ɪ**l**d] ⟨**wild**⟩
b. [bɪˈ**wɪl**ə**d**əmənt] ⟨be**wild**erment⟩

(2) a. [ɹɪˈ**djuːs**] ⟨**reduc**e⟩
b. [ɹɪˈ**dʌk**ʃən] ⟨**reduc**tion⟩

(3) a. [hæt**s**] ⟨hat**s**⟩
b. [hɛd**z**] ⟨head**s**⟩

Despite the pronunciation differences, the graphic representation of the morphemes highlighted in boldface in examples (1)–(3) is preserved. This is an example of a morphographic principle that is at work within a basically phonographic writing system.[3] Likewise, phonographic features are usually observable in primarily morphographic systems. For instance, each sign of the Chinese writing system corresponds not only to a (meaning-bearing) morpheme, but also to a syllable of the spoken language. Thus, these signs also have a phonetic value each. As such, they can be used to write words of foreign languages, such as *Frankfurt* (example (4)).[4] In this use, their correspondence to a meaning-bearing morpheme (represented in the first line of example (4)) becomes irrelevant.

(4)

‘law’	‘flower’	‘gram’	‘luck’
fǎ	*lán*	*ke*	*fú*
法	兰	克	福

Thus the terms “morphographic” and “phonographic” can be viewed as principles that are at work within one and the same writing system, rather than describing writing systems as a whole. Understood as such, the distinction between phonographic and morphographic writing systems is closely related to a first basic distinction that is of central importance for orthography development, namely that between “deep” and “shallow” orthographies (Katz and Frost 1992; Bird 1999b; Ellis et al. 2004). The metaphor of the “depth” of an orthography refers to the level of linguistic structure at which forms are orthographically represented. Shallow orthographies approximate a correspondence between an orthographic representation and the surface realization of linguistic forms to the extent that they may specify the phonetic realization of these forms as they are pronounced in a given context. Examples of such orthographies are Serbian and Croatian, which use the same writing system, but different scripts, Cyrillic and Roman (Feldman and Barac-Cikoja 1996). In these orthographies, allomorphy and even regional pronunciations are represented (see Katz and Frost 1992: 69f.), and a close relation between the written form and its pronunciation is thus maintained. A deep orthography, on the other hand, approximates a correspondence between orthographic representation and underlying forms. Deep orthographies thus typically represent each morpheme of the language with one, invariable written form, and do not specify the morphophonological changes that these morphemes undergo in context. Deep orthographies are thus typically less specific with respect to the phonetic realization of a given form. A tendency towards such an orthography can be observed in the English examples (1)–(3), above.

Deep orthographies are widely in use for languages with many morphophonological changes, i.e. languages where the morphophonological representation is quite distinct to the phonetic representation, such as in English (Liberman et al. 1980; Frost and Katz 1992: 69ff.). A deep orthography for such languages can be understood as a technique for preserving the visual image of morphemes, which would be blurred in a shallow orthography. Shallow orthographies, on the other hand, tend to be used for languages with relatively few morphophonological changes, e.g. Serbian and Croatian. In these languages, the morphophonological representation is close to the phonetic representation. Consequently, a shallow orthography of such a

language may preserve the graphic identity of morpheme to the same degree as a deep orthography of a language with many morphophonological changes.

The term *orthographic depth* thus refers – broadly speaking – to the level of linguistic structure at which the features represented in the orthography are located. Another important question is which of the manifold features present in a spoken message should be represented in an orthography at all. Linguistic analysis is crucial here since it reveals the distinctive features of the language, e.g. phonological contrasts. From a strictly structural point of view, a single minimal pair is enough for a given feature to count as distinctive. However, some features are clearly more important than others in the sense of "the extent to which users of the orthography rely on that feature in reading and writing the language" (Bird 1999b: 14). This is referred to as the *functional load* of a linguistic feature. For the development of an orthography it is important to evaluate the functional load of a linguistic feature in order to decide whether or not it should be represented in the orthography.

Functional load can be approximated by assessing how many words or utterances a given feature differentiates. For instance, in English some words are distinguished by stress, e.g. *cónvert* vs. *convért*, *prótest* vs. *protést*. These words are homographs in English, and in a list of isolated words there would indeed be ambiguity (and these words could count as minimal pairs in such a context). However, these words are not many, which is already indicative of the relatively low functional load of stress in English, at least with regard to distinguishing basic lexical items. In addition, the members of these pairs belong to different parts of speech (nouns vs. verbs) and thus they are easily disambiguated in context. Hence, it is clear that the functional load of stress in English is in fact very low in the sense that readers do not rely on it for disambiguating lexical items in a written message. Thus, while for the phonologist one minimal pair in a list of isolated words may be sufficient to identify a certain feature as contrastive, for the purpose of developing a practical orthography it is crucial additionally to evaluate the functional load of a potentially contrastive feature in connected texts. And if there are no, or only very few, instances where a given feature (e.g. stress) in fact disambiguates utterances in a sufficiently large text corpus, then the need to represent the distinction is highly diminished. This is particularly important if it would be cumbersome consistently to represent the feature in the orthography, as the writing of stress, e.g. by accent marks, in English would be.

This leads to a final concept to be introduced here, that of *underrepresentation*. While it is true that orthographies should reduce potential ambiguity of a written message, they should also be simple. And in order to achieve this simplicity, it may be justified not to represent features that do not have a high functional load, even if they are contrastive from a strictly structural point of view. Underrepresentation in an orthography leads to homographs, i.e. more than one word is orthographically represented in the same way, and may thus lead to ambiguity. However, readers can in fact tolerate a considerable amount of ambiguity caused by homographs because they can make use of many cues other than those representing phonological distinctions when decoding a written message. Among these are syntactic cues, such as word classes (as in the case of English *prótest* vs. *protést*, mentioned above),[5] semantic cues (e.g. selectional restrictions), and contextual cues from the surrounding discourse. All this is to say that an orthographic representation may differ substantially from a phonological transcription in that a practical orthography may systematically underrepresent distinctive features for the sake of simplicity.

2. Non-linguistic factors in orthography development

This section identifies a number of factors that may be decisive in choosing one option for orthographic representation over another. These options are determined by the linguistic structure of the language to be written. The factors covered in this section, on the other hand, are independent of this structure and may therefore be called non-linguistic factors in orthography development. The basis of these factors is that different orthographic options have particular advantages and disadvantages for different potential users of the orthographies. These advantages and disadvantages are related to a wide range of issues, including pedagogical, sociopolitical, and mechanical or technical aspects of orthographies. Non-linguistic factors of orthography development are discussed in four sections: psycholinguistic and pedagogical issues (Section 2.1), existing orthographies (Section 2.2), dialect varieties (Section 2.3), and technical issues (Section 2.4).

2.1. Psycholinguistic and pedagogical issues

Psycholinguistic research has shown that different kinds of orthographies favor different kinds of users (Venezky 1970). Different user groups from

the point of view of psycholinguistics are readers vs. writers, beginning readers/writers vs. advanced readers/writers, and mother-tongue speakers vs. non-fluent speakers. To make definitive statements about the learnability and usability of a given orthography for a given language, it is necessary to do extensive testing. However, drawing on results reported in the literature, some general statements can be made here.

A first, probably obvious point is that orthographies that reflect the particular structure of the language to be written facilitate the acquisition of the orthography. They do so because they build on speakers' implicit knowledge of the language, which is explicit in its grammatical description. The importance of this point is that conventions used in existing orthographies of surrounding languages, e.g. a dominant language, may be inappropriate to represent the particular structure of the language to be written, and reproducing them in a newly developed orthography may thus lead to problems (see Section 3.1 below for a case study).

The requirement of adhering to language-specific structures is particularly important for the orthographic representation of word boundaries, because words are the basic units for language processing in reading (Reicher 1969). It is well known that languages vary drastically with respect to word boundaries and that the definition of words can be a highly complex issue because there may be conflicting criteria. Careful examination of a wide variety of issues, including prosodic, morphosyntactic, and semantic factors, is thus a precondition for proposing orthographic rules concerning word boundaries (for discussion of some factors, see Dyken and Kutsch Lojenga 1993; see also Chapter 10).[6]

A second, more substantial point to be made here is that from the perspective of psycholinguistics, "the optimal orthography for a beginning reader is not the same as for a fluent reader" (Dawson 1989: 1). This general statement derives from the finding that advanced readers heavily rely on what is called a "sight vocabulary", i.e. written words are recognized as entire units and processed as such, without breaking them down into units of the sound structure. For that reason, advanced readers benefit from orthographies that preserve the graphic identity of morphemes. A sight vocabulary allows readers to quickly recognize words in written messages without much specification of phonetic details. A high reading competence also allows to make full use of contextual cues, which may require some going back and forth in a written message to disambiguate homographs. Because of the relative importance of a sight vocabulary and the relative unimpor-

tance of phonetic detail, advanced readers benefit from deep orthographies rather than shallow ones.

For beginning readers, however, things are different. The acquisition of a deep orthography at first exposure is relatively difficult because the written form may differ significantly from the actual pronunciation and may have to be memorized in a first phase. Compared to these, shallow orthographies, i.e. orthographies that represent linguistic forms in a way that is close to their actual pronunciation in each context, are considerably easier to learn for a beginning reader (and writer), including second language learners. Wherever languages display heavy morphophonological processes, orthography developers face the problem of either choosing a shallow orthography for the beginning reader or a deep orthography for the advanced one.

A further issue is that the process of reading is different from the process of writing. Again, the difference is between shallow and deep orthographies. A sight vocabulary is most helpful in the process of reading in that it allows quickly to retrieve a morpheme from the mental lexicon independent of its phonetic realization. In the process of writing, the advantages of a sight vocabulary are not as clear. In writing, it may be as easy to spell a form according to its pronunciation as to retrieve the underlying form. When making a compromise between an orthography that suits readers vs. writers, it should be taken into account that reading is far more frequent than writing (ideally, a text is written only once but read many times), so the needs of readers are somewhat more important.

A final point on pedagogical and psycholinguistic issues of orthographies concerns the particularities of endangered languages at an advanced stage of language shift. In such a situation, younger members of the speech community, who have not learned the endangered language themselves (at least not as a first language), may make up an important proportion of the potential users of the orthography. This group may be interested in writing their ancestral languages in the context of "third generation pursuit" (Dorian 1993), i.e. in an effort to revalue or revitalize the language that their parents had abandoned. They are thus in the situation of a second language learner, and they may benefit from a relatively shallow orthography that does not make heavy use of underrepresentation. Such an orthography allows them to correctly write a word from its pronunciation and to correctly pronounce a word from its written form without knowing the word. This is particularly important if the orthography is likely to be used primarily for documenting

ancestral knowledge (e.g. narratives, ethno-biological terminology), rather than for everyday written communication.

2.2. Existing orthographies

Already existing orthographies – be they of the language for which the orthography is being developed or of surrounding languages – tend to be an extremely influential factor in orthography development or reform. Dealing with existing orthographies can be a highly delicate sociopolitical matter, since the emblematic function of an orthography emerges most clearly in its visual contrast to surrounding orthographies.

With respect to orthography reform, it cannot be stressed enough that reforming an established orthography may have an enormous sociopolitical impact, in particular if a substantial number of speakers are already acquainted with that orthography and if printed materials that use this orthography already exist. Thus, it may be better to live with an inconsistent orthography – even if inappropriate from a linguistic or psycholinguistic perspective – unless the speech community is really determined to change it.

How a newly developed orthography relates to existing orthographies of neighboring languages depends primarily on the sociopolitical relation of the speech community to the speakers of those languages. In a typical situation of language endangerment, an increasing number of members of the speech community acquire a dominant language to an increasing degree of proficiency. Often, they acquire literacy for the first time in that language or they are keen to do so in order to gain access to institutions of the national society, e.g. higher education. In these cases, an orthography that resembles the orthography of the dominant language may be advantageous in order to facilitate acquisition of the orthography of the endangered language for those who are already acquainted with the one of the dominant language, and to facilitate the acquisition of the orthography of the dominant language for those who acquire the one of the endangered language first.

On the other hand, it is a recurrent phenomenon that speech communities want their newly developed orthography to have a visual appearance that is decidedly different from that of dominant or other neighboring, possibly closely related languages. However, the wish for an emblematic orthography is often satisfied by choosing graphemes with a particular visual shape. These choices do not affect the overall functionality of the orthography, and

this issue is thus often relatively easily resolved when compared to the difficult choices that may be necessary when choosing between a deep or shallow orthography, or whether to represent a given feature at all.

If literacy in the dominant language is already on the way or desired in the future, and if it is accepted that a newly developed orthography is to borrow elements from the orthography of the dominant language, then the question arises how to deal with internal inconsistencies of this orthography. These are difficult to acquire in the dominant language, and would also be difficult to acquire in the endangered language. Thus, idiosyncratic spelling conventions that have come about for purely historical reasons, such as Spanish /k/ – ⟨k, c, qu⟩, should in general not be replicated in newly developed orthographies.

2.3. Dialect varieties

Dialect varieties exist in every speech community. A characteristic often found in speech communities without a written standard is that there is no widely accepted standard variety among the different dialects. This obviously poses a problem for developing an orthography since an orthography by definition involves standardization. There are limited possibilities to represent various dialects using a single orthography, as further discussed in Section 3.4 below. Multidialectal orthographies are more feasible in case of relatively deep orthographies, which may not represent the features that distinguish the dialects, e.g. vowel distinctions that are contrastive in one dialect but not in another. In any case, it is likely that a standardized, new orthography will have to disregard at least some features of one or more of the dialect varieties. Which ones these will be depends again largely on non-linguistic factors, namely the sociopolitical relations among the dialect groups.

2.4. Technical production issues

At a time when typewriters were the main tools for producing written texts (other than handwriting, of course), the limited set of symbols available on a typewriter keyboard as well as the ease with which they could be produced were of major practical import in designing practical orthographies. Creating graphemes that required the use of two or more diacritics on one

base letter resulted in an extremely cumbersome typing process and thus were very rarely adopted. While modern word processors in principle allow for much greater variety and comfortable shortcuts in producing unusual graphemes, technical production and reproducibility remains a major issue.

The main point here concerns the electronic representation of characters other than those used in the Latin alphabet. This issue has unfortunately still not been satisfactorily resolved in our highly computerized age. Special fonts that contain non-Latin characters often have certain software requirements (e.g. they can only be used under a particular version of a particular system) and are thus not safe options in the long run. The newly developed Unicode character encoding standard comprises thousands of graphemes (including those of the Latin alphabet), independent of special fonts (see Chapter 14). However, Unicode is still not yet fully established (e.g. most commonly available fonts only support a small subset of these characters). Furthermore, even if computers are available, the access to special fonts and the technical know-how to install and run them may not be available to the speech community. Thus, the safest option – to ensure usability of the orthography without access to sophisticated software and computer know-how, as well as for safe long-term storage of digital files containing text written in that orthography – is still to use only characters that can be found on the keyboard of a mechanical typewriter or combinations of these (e.g. digraphs or combinations of letters with diacritics).

2.5. Summary

Most of the factors discussed in the preceding sections relate to decisions about orthographies that vary according to two parameters: orthographic depth and the similarity of a given orthography to the orthography of dominant or other neighboring languages, which is particularly important in the case of endangered languages. The advantages and disadvantages of choosing orthographies towards one end or the other of these two parameters are summarized in Table 1.

Table 1. Advantages and disadvantages relating to non-linguistic factors for orthography development

Parameter	Advantages	Disadvantages
shallow orthography (close to pronunciation)	– easier to learn for beginning readers/ writers – easier to learn for non-(fluent) speakers	– may blur graphic identity of morphemes – more difficult to encompass various dialects in one written form
deep orthography (preserves graphic identity of meaningful elements)	– easier for reading in general – easier to handle for fluent readers – easier to encompass various dialects	– harder to learn for beginners – harder to learn for non-(fluent) speakers
using conventions of the orthography of the dominant language	– easier to learn for speakers that are literate in dominant language – facilitates subsequent literacy in dominant language – facilitates technical text (re)production	– may have to live with inconsistencies in the orthography of dominant language – potentially less emblematic
using conventions different from those of the orthography of the dominant language	– potential problems with technical text (re)production	– highly emblematic

3. Case studies: Options and choices

The following sections (3.1–3.5) discuss selected aspects of a number of linguistic systems and the options that these offer for orthographic representation as well as the choices that have been made based on non-linguistic factors.

3.1. Morphemic nasality in Eastern Tucanoan languages

The reform of the orthographies of the Eastern Tucanoan languages is a good example of the need for a thorough linguistic analysis as a basis for orthography development and for the advantage of an orthography that respects the particular structure of the languages as opposed to one that uses conventions of orthographies of surrounding languages with established orthographies.

The Eastern Tucanoan languages are a group of closely related languages spoken in the Vaupés, a region on both sides of the Colombian-Brazilian border in the North West Amazon. Nasality is a pervasive feature in these languages. All oral, voiced phonemes, i.e. the six consonants *b, d, y, g, w, r* and the six vowels, have a nasal counterpart. Nasality used to be spelled out on each segment in the orthographies, as in the following examples (5a)–(5g).[7]

(5)	a.	⟨ãmũmã⟩	[ãmũmã]	'neck'
	b.	⟨gnãmõrõ⟩	[ŋãmõrõ]	'ear'
	c.	⟨jĩnõ⟩	[hĩnõ]	'anaconda'
	d.	⟨gudamĩsĩ⟩	[gudamĩsĩ]	'stomach'
	e.	⟨ojoño⟩	[ohoŋõ]	'banana plant'
	f.	⟨baamĩ⟩	[baːmĩ]	'(s)he eats'
	g.	⟨ĩãbeco⟩	[ĩãbeko]	'he who does not look'

Recent research on Eastern Tucanoan languages has shown that nasality in these languages is a feature of morphemes, in particular lexical roots, rather than a feature of phonological segments. Thus, all simple (i.e. not compound) verbs and nouns are entirely oral or entirely nasal, i.e. the voiced phonemes of these forms are either all oral or all nasal. This characteristic is represented in the new orthographies of the Eastern Tucanoan languages (Gomez-Imbert and Buchillet 1986; Gomez-Imbert 1998), in which nasal morphemes are preceded by "~" (compare examples (6a)–(6g) with (5a)–(5g)). In case of polymorphemic words that begin with an oral morpheme and end with a nasal one, "~" is inserted before the nasal morpheme (examples (6e)–(6f)). In case of polymorphemic words that begin with a nasal morpheme and end with an oral one, "~" is marked at the beginning of the word and "-" is inserted before the oral morpheme (example (6g)).[8]

(6)	a.	⟨~abuba⟩	[ãmũmã]	'neck'
	b.	⟨~gaboro⟩	[ŋãmõrõ]	'ear'
	c.	⟨~hido⟩	[hĩnõ]	'anaconda'
	d.	⟨guda~bisi⟩	[gudamĩsĩ]	'stomach'
	e.	⟨oho~yo⟩	[ohoŋõ]	'banana plant'
	f.	⟨baa~bi⟩	[baːmĩ]	'(s)he eats'
	g.	⟨~ia-beko⟩	[ĩãbeko]	'he who does not look'

According to Gomez–Imbert (1998) speakers are considerably more comfortable with the new orthographies than with the old ones, presumably because the new orthography builds on their implicit knowledge of the structure of the languages. This example thus shows the importance of modeling orthographies as closely to the linguistic structure as possible, rather than taking over conventions of orthographies of better-known languages.

3.2. Palatalization in Miraña

The examples in this section come from Miraña, an endangered Amazonian language spoken in Colombia, to the South of the Vaupés region (Seifart 2002, 2005). Miraña has a linguistically very close variant called Bora, which is spoken mainly in Peru (Thiesen 1996; Thiesen and Thiesen 1998). Today, Miraña has only about 50 speakers out of ca. 400 ethnic Mirañas. All speakers are bilingual in Spanish and most are also literate in Spanish. Palatalization in Miraña will serve as an example for non-linguistic factors being responsible for the choice of a shallow orthography over a deep one.

Miraña has a set of six palatal consonants. Most of their occurrences are easily recognizable as phonetic realizations of their alveolar counterparts in the context of a preceding /i/, such as [n, ɲ] (examples (7a)–(7b)). However, palatal consonants occur not only after /i/, but also after /a/ (example (7c)).

(7) a. [**n**àʔbɛ̀]
'brother'

b. [íɲàʔbɛ̀]
'his/her/their brother'

c. [táɲàʔbɛ̀]
'my brother'

Further analysis of Miraña revealed that what is causing palatalization of alveolar consonants after /a/ is the underlying phoneme /a^{j}/, whose palatal component is realized as [j] before vowels (example (8a)), spreads to alveolar consonants, which are palatalized (example (8b), see also example (7c)), and is suppressed before bilabial consonants, where the distinction /a^{j}/ vs. /a/ is neutralized (example (8c)) (note that tone alternation has no effect on palatalization in Miraña).

(8) a. [à**j**ɯ́hɯ̀]
/à^{j}ɯ́hɯ̀/
'okay'

b. [tát^{j}á?dì]
/tá^{j}-tá?dì/
1ST_PERSON_POSSESSOR-grandfather
'my grandfather'

c. [tàmámɨbà]
/tà^{j}-mámɨbà/
1ST_PERSON_POSSESSOR-trunk
'my trunk'

This phonological analysis results in a fairly simple, symmetric, and parsimonious inventory of consonant phonemes, while the vowel inventory has to be augmented by the complex unit /a^{j}/ (Seifart 2002: 23–30).

The phonological system allows for the options of representing palatalization in a deep orthography, i.e. phonemically, or in a shallow orthography, i.e. phonetically. A deep orthography has the advantage of preserving the graphic identity of morphemes that begin with alveolar consonants, be they lexical roots (see examples (7b)–(7c) and (8b), above), or suffixes such as the inanimate marker (examples (9a)–(9b)) and the restrictive marker (examples (9c)–(9d)). This advantage is particularly important because a large proportion of roots begin with alveolar consonants and the most frequent suffixes also begin with these consonants, including the markers in examples (9a)–(9d) as well as the plural marker.

(9) a. [tsà**nɛ̀**]
/tsà-**nɛ̀**/
one-INANIMATE
'one (inan.)'

b. [tsìːɲɛ̀]
/tsiˌ-ɲɛ̀ʔ
other-INANIMATE
'another (inan)'

c. [ɯ́hɨ́ʔòɾɛ̀]
/ɯ́hɨ́ʔò-ɾɛ̀/
banana-RESTRICTIVE
'just a banana'

d. [ɯ́βíːbàɾʲɛ̀]
/ɯ́βíːbàʲ-ɾɛ̀/
basket-RESTRICTIVE
'just a basket'

However, from the point of view of orthography development there are two major disadvantages of writing palatalization phonologically. Firstly, it differs significantly from actual pronunciation in some instances, e.g. when palatalization spreads across glottal consonants in coda position and is realized in the onset of the following syllable, as in examples (10a)–(10b). Secondly, features that are neutralized have to be written, e.g. when /aʲ/ is followed by a bilabial consonant (see example (8c), above) or when it occurs word-finally (compare example (10c) with (9d)).

(10) a. [tsàhtʲɛ̀]
/tsàʲhtɛ̀/
'Take!'

b. [tɯ́hpaʔjɛ̀]
/tɯ́hpaʲʔɛ̀/
proper name

c. [ɯ́βìːbà]
/ɯ́βìːbàʲ/
'basket'

A delicate choice thus has to be made between orthographically representing palatalization in Miraña phonemically, i.e. as a complex vowel, or phonetically, i.e. in six additional consonants. The phonological writing ensures an invariant graphic image of a large proportion of morphemes and may thus help to build a sight vocabulary, from which advanced readers may benefit.

However, the palatalization process as a whole is rather complex in that palatalization may be neutralized or it may spread across various segments. The phonetic writing, on the other hand, requires no knowledge of the palatalization process. Its disadvantages are that it requires six additional units (the palatal consonants) and that it introduces a lot of redundancy, in particular by writing palatal consonants after /i/, where they are easily recognizable as palatalized realizations (see examples (7b) and (9b)).

A shallow orthography with respect to palatalization was nevertheless proposed (and adopted) for Miraña. An important reason for this decision is that nowadays, many of the younger Mirañas, who are the main users of the orthography, did not learn Miraña as their first language, and many of them hardly speak it at all. Thus, they do not have an implicit knowledge of the structure of the language to the same degree as, e.g., most users of the orthographies of Eastern Tucanoan languages have about their languages. The main use of the Miraña orthography is to document myths, songs, and ethno-biological terminology, which younger speakers elicit from older ones. The proposed orthography serves these purposes well in that it provides an intuitive system for spelling and pronouncing Miraña words unknown to non-fluent speakers.

3.3. Writing tone

All languages make use of pitch in some way. However, while pitch is used in some languages, e.g. Chinese, to differentiate a vast amount of lexical items, its function in other languages is mostly limited to conveying intonational distinctions. From the point of view of orthography development, pitch is thus a feature that varies drastically from language to language with respect to its functional load in distinguishing lexical items. In languages where it is either very high or very low, the question whether or not to represent it in an orthography does not arise, but there are many intermediate cases that require careful analysis and possibly creative solutions. These issues are discussed in Bird (1999b), from which the examples presented in this section are taken.

Typical characteristics of such "intermediate" systems, which are found in many African, Papuan, and Amazonian languages, are that pitch is widely used to mark grammatical functions and that pitch patterns can only be described in terms of sometimes quite complicated sets of spreading and truncation rules. The processes that underlie the resulting surface tones may thus

be extremely complex and their marking may be very difficult to handle even for experienced writers. This is the case in Dschang, a Grassfield Bantu language spoken in Cameroon. Bird (1999b: 7) reports that in this language, experienced writers attain an accuracy score of only 83.5% correct tone marks and inexperienced ones, only 53% correct tone marks when marking surface tones. Tone writing thus poses a serious problem for the orthography of this language and the question arises to what extent tone actually carries a functional load, i.e. whether it is necessary to write tones at all.

An interesting solution to a similar problem was found in Komo, another Bantu language, which is spoken in the Democratic Republic of Congo (formally Zaïre). In this language, tone is used to distinguish lexical items as well as for marking grammatical functions. With respect to lexical tones, it was found that about 28 minimal pairs are distinguished by their tonal pattern in a representative list of over 3000 words. More than half of these, however, can be easily distinguished in context, either because they belong to different word classes or because of their meaning. Thus, it was decided not to mark lexical tone in this language, even though this creates ambiguity through homographs in a few cases. On the other hand, a considerable amount of inflected and derived word forms in Komo are distinguished by grammatical tones on their first syllable and these can often not be disambiguated by context. Thus, it was decided to mark only grammatical tones on the first syllable in the Komo orthography. Example (11) (data from Paul Thomas, as reported in Bird 1999: 23) illustrates how this tone marking disambiguates inflected or derived forms (examples (11a) vs. (11c), (11b) vs. (11d), etc.), but fails to disambiguate lexical items in some cases (examples (11a) vs. (11b), (11c) vs. (11d), etc.).

(11)	a.	⟨bebhomi⟩	[bèbhòmí]	‘we insulted him’
	b.	⟨bebhomi⟩	[bèbhómí]	‘we did surgery on it’
	c.	⟨bĕbhomi⟩	[bĕbhòmí]	‘we insulted them’
	d.	⟨bĕbhomi⟩	[bĕbhómí]	‘we did surgery on them’
	e.	⟨babhomigi⟩	[bàbhòmìgì]	‘insulters’
	f.	⟨babhomigi⟩	[bàbhómìgì]	‘surgeons’
	g.	⟨bábhomigi⟩	[bábhòmìgì]	‘they insulted habitually’
	h.	⟨bábhomigi⟩	[bábhómìgì]	‘they did surgery habitually’

The Komo orthography illustrates the importance of carefully assessing the functional load of a given feature in order to decide whether it should be

represented orthographically or not, in particular if writing this feature creates major difficulties for the users of the orthography. The solution found in Komo also shows that a given feature – in this case tone – may not have the same functional load in all of its contexts, and, consequently, the possibility of representing a feature such as tone only in those contexts where it effectively helps readers to disambiguate a given form, without overburdening the orthography with tone marking on (almost) every syllable.

3.4. Multidialectal orthographies

The two examples of multidialectal orthographies discussed in this section provide further illustrations of the interaction between linguistic systems and non-linguistic factors in orthography development, in particular, the concept of underrepresentation and the different needs of readers vs. writers.

Sasak is an Austronesian language spoken on the island of Lombok in Nusa Tenggara Barat, Indonesia (Austin 2000). Across the five dialects of Sasak, there are eight phonological vowels, which contrast with each other in different ways in the different dialects. The practical orthography that was established for all Sasak dialects represents only those vowels that are contrastive in all of the dialects and it conflates those that are conflated in the phonological systems of one or more of them (Table 2).[9] The disadvantage of this orthography is that it creates ambiguity through homographs in individual dialects, but it has the great advantage of offering a unified orthography for all dialect groups, and this has apparently been the overriding reason for adopting it.

Table 2. Vowels in the Sasak orthography (Peter Austin, p.c. 2004)

Phonemes	Orthography
a	**a**
e	**e**
ə	
ɛ	
i	**i**
o	**o**
ɔ	
u	**u**

A different solution for representing a number of dialects in one orthography was chosen for Biliau, another Austronesian language, which is spoken in Papua New Guinea (Simons 1994). The phonemic inventories of the dialects of Biliau differ in that in the western dialect /d/ and /z/ are separate phonemes, while in the eastern dialect only /d/ occurs. In this case, an orthography was established (and presumably accepted by all speakers, including those of the eastern dialect) that represents the more complex phonology of the western dialect (example (12)) (Simons 1994: 12).

(12)

		western dialect	eastern dialect	
a.	⟨damom⟩	/damom/	/damom/	'my forehead'
b.	⟨zamom⟩	/zamom/	/damom/	'rotten'
c.	⟨der⟩	/der/	/der/	'a cold wind'
d.	⟨zer⟩	/zer/	/der/	'grass skirt'
e.	⟨badi⟩	/badi/	/badi/	'get up'
f.	⟨bazi⟩	/bazi/	/badi/	'feather'

In the multidialectal orthography of Biliau, speakers of the eastern dialect have to write a distinction that is not present in their phonemic system and have to memorize spellings of these words (examples (12b), (12d), (12f)). Reading will not be difficult, however. Every time speakers of the eastern dialect see a ⟨z⟩, they are taught to pronounce a /d/. Thus, the overall advantage of this orthographic solution is toward the reader. This is a good reason for giving preference to the western dialect in the multidialectal orthography, but this would probably not have been possible if it did not also have "true ascendancy in terms of prestige" (Simons 1994: 20).

A comparison of Sasak and Biliau nicely illustrates that non-linguistic factors are decisive in making choices in the development of an orthography. The phonemic systems of the dialects of Biliau and Sasak offer in principle the same two options for multidialectal orthographies: either representing distinctions that are *not* contrastive in some dialects or neglecting distinctions that *are* contrastive in some dialects. Which of these two options was actually chosen depends crucially on the sociolinguistic situation. In Biliau, one of the dialect groups has a sufficiently high status such that the other dialect group accepts its variant as the basis for a common orthography. This is apparently not the case in Sasak.

3.5. Choosing graphemes

This section briefly discusses the issue of choosing graphemes, using again the example of the Miraña orthography, some aspects of which were discussed in Section 3.2 above. In Miraña, these choices were determined by the Mirañas' sociopolitical relations to two other speech communities which have established orthographies: The Colombian national society, whose language is Spanish, and the Boras, who speak a linguistically very close variant of Miraña (Thiesen 1996: 11, 20; Seifart 2005: 22f.). A first noteworthy characteristic of Miraña orthography is that all of its graphemes are based on Spanish letters. Some of the Miraña graphemes are modified versions of Spanish graphemes, either in their visual graphic form or their phonetic value, as can be observed in Table 3. Miraña speakers also decided to modify the visual appearance of some (Spanish-based) graphemes used in Bora. This can be understood when taking into account that the Mirañas have long struggled to be recognized as a separate ethnic group with respect to the more numerous Boras. Table 3 gives a good impression of the two main conflicting factors that are at work when choosing graphemes: that of adhering to conventions of already known and established orthographies of surrounding languages, and that of giving an orthography a decidedly different appearance in order to fulfill an emblematic function for the speech community.

4. Conclusion

The previous sections have shown that orthography development involves a rich interaction of the characteristics of linguistic systems and a variety of non-linguistic factors. Structural properties of languages often allow for a number of alternative options of orthographic representation of a given feature. These options may correspond to a phonemic representation, but they may as well correspond to a more abstract representation (morphophonemic) or to a more superficial representation (phonetic). These alternative options may favor different potential users of the orthography. The task of the orthography developer is to balance the advantages and disadvantages of these options and find a workable compromise.

Table 3. Some graphemes of the Miraña orthography

Spanish orthography	**Miraña ortho-graphy**	Bora ortho-graphy	IPA	Motivation
(u)	**ʉ**	**u**	**ɯ**	making a difference to Spanish and Bora, local conventions
(i)	**ɨ**	**ɨ**	**ɨ**	new grapheme based on Spanish
qu (before e and i) **c** (other contexts) **k** (in loanwords)	**k**	**k** (before e and i) **c** (other contexts)	**k**	avoiding inconsistencies of Spanish and Bora
v, b (intervocalic pronunciation)	**v**	**v**	**β**	two Spanish graphemes that stand for the same phoneme in Spanish are used for two phonemes in Miraña
v, b (word-initial pronunciation)	**b**	**b**	**b**	
ll, y	**ll**	**ll**	**d͡ʒ**	two Spanish graphemes that stand for the same phoneme in Spanish are used for two phonemes in Miraña
	y	**y**	**j**	
j	**j**	**j**	**h**	local Spanish pronunciation
	'	**h**	**ʔ**	making a difference to Bora
(g, w)	**gw**	**w**	**g^{w}**	making a difference to Bora
(t, d), **(y)**	**ty, dy**	**ty, dy**	**t^{j}, d^{j}**	digraphs based on Spanish graphemes (only two examples included here)

Acknowledgements

I am grateful for comments from Mandana Seyfeddinipur, Ulrike Mosel, Nikolaus Himmelmann, Julia Borchert, Jost Gippert, and the audiences at the *DoBeS* summer school in Frankfurt in 2004 and at the *Instituto Caro y Cuervo* in Bogotá in 2005, where much of the contents of this chapter was presented in seminars. I am also grateful to Peter Austin, Natalia Eraso, Doris Fagua, Elsa Gomez-Imbert, Camilo Robayo, and Maria Trillos for providing examples (along with discussion), not all of which are represented in this chapter. Thanks also to Falk Grollmus for providing the Chinese example.

Notes

1. Further theoretical possibilities to typologize writing systems, such as direction (left, right), axis (horizontal, perpendicular), or lining (top to bottom, bottom to top), are usually disregarded since they yield no insightful classifications.
2. Morphographic systems are sometimes also called "logographic" or "ideographic". Both terms are inappropriate because the units represented in these writing systems are always morphemes, and not words in the sense of units that could be modified by inflection, as the term "logographic" suggests. As a matter of fact, there are no writing systems that represent words in this sense, even though in case of highly isolating languages, such as Chinese, words tend to be monomorphemic. Furthermore, graphemes always refer to linguistic units and never directly to extra-linguistic concepts, as the term "ideographic" suggests.
3. In many cases, the spelling of morphemes is constant in different contexts despite pronunciation differences because the spelling represents an older stage of the language, when these forms were in fact pronounced in the same way. Because such spelling conventions make explicit the etymology of words, phenomena such as the English examples 1–3 can be called "etymological writing". The French orthography – which displays very complex correspondences to pronunciation – also contains many examples of etymological writing.
4. Additionally, for many Chinese signs it may be claimed that they include components with an exclusively phonetic value (Coulmas 2003: 56ff.). This is a further phonographic aspect of this writing system.
5. Note that information about word classes can also be directly represented in an orthography, for instance by capitalization of nouns, as in German.
6. Similar issues apply to the orthographic representation of syntactic units, such as phrases and sentences, which are often orthographically represented with punctuation.

7. I adapted examples 5 and 6 from a report on a workshop on the reform of the orthographies of Eastern Tucanoan languages (Eraso 2003). Examples 5a–f and 6a–f are from Makuna, examples 5g and 6g are from Barasano.
8. Note that in the new orthographies, the representation of /k/ was changed from ⟨c, qu⟩ to ⟨k⟩ and /h/ from ⟨j⟩ to ⟨h⟩ in order to facilitate the use of a single orthography in Spanish-dominated Colombia and Portuguese-dominated Brazil, where these graphemes have different phonetic values.
9. It may be of interest to note here that the writing system of Indonesian, the main contact language of Sasak, conflates the two phonemes /ə/ and /e/ in the grapheme ⟨e⟩ in most publications (although the distinction is maintained in many dictionaries by representing /e/ with ⟨é⟩).