Towards an understanding of the relationship between diglossia and literacy[(*)]

John Myhill
University of Haifa

0.0. **Introduction**. One of the most important issues affecting language policy is the connection between the acquisition of literacy and the relationship between the spoken language of the child and the written language which s/he is learning.* It has been recognized for some time that it is problematic for a child to begin to learn to read and write a written form which is understood to be a different language from the language which s/he has grown up speaking. Thus already in the 1950s UNESCO observed that:

> On educational grounds, we recommend that the use of the mother tongue be extended to as late a stage in education as possible. In particular, pupils should begin their schooling through the medium of the mother tongue because they understand it best and because to begin their school life in the mother tongue will make the break between the home and the school as small as possible. (UNESCO 1953:47-8)

This position has been supported by numerous empirical studies (see e.g. Gudschinsky 1977, Okedara and Okedara 1992, Dutcher and Tucker 1997, Mehrotra 1998, etc.).

However, while these studies have left no doubt of the validity of the position which UNESCO has adopted on mother tongue instruction, they have not addressed the related question of the importance to the acquisition of literacy of the difference between the spoken language of the child and the written language s/he first learns to read when they are considered to be forms of the **same** language. In some languages, the written and spoken forms are very similar to each other, in particular because the written language has been developed based closely upon the spoken language, whereas in other languages—which have been referred to as `diglossic' since the seminal paper of Charles Ferguson (Ferguson 1959)—the difference between the written and spoken forms is very great. This difference can result from a number of (not necessarily independent) factors, e.g. (1) although written languages are originally based upon forms which were spoken at some point in time, in later years--because spoken language naturally and inevitably changes more rapidly while written

---

[(*)]The Initiative for Applied Research in Education's *Language and Literacy Domain Committee* commissioned the scientific survey.*

- The findings are in the author's own words and the conclusions reached are his own.
- Any mention or quote from the survey must be referenced in the following manner: Myhill, J. (2009), Towards an Understanding of the Relationship between Diglossia and Literacy, a Survey Commissioned by the Language and Literacy Committee. http://education.academy.ac.il

language tends to be more conservative--a noticeable 'gap' generally appears between the two, the magnitude of which varies from case to case (see e.g. Aitchison 1981, Ravid 1995), (2) written languages are also often elaborated by the conscious use of prescriptive rules and constructions which were made up by linguists rather than being part of the normal spoken language, and (3) written languages may be based upon a particular dialect of a language which for some speakers may be very different from their own native dialect. Given the existence of significant differences between written and spoken language in many cases, the natural assumption given the findings referred to above would be that literacy would be more quickly and easily acquired in cases in which the written version of a language is closer to the spoken language, so that the child learning to read is to a large extent simply learning to associated written symbols with the vernacular s/he already speaks (Verhoeven 1994a:10). However, this question has not been systematically investigated. In fact, there have been relatively few studies of the effect of linguistic difference within the same language upon the acquisition of literacy, and what studies have been done have only addressed this issue within a single language (e.g. Burger and Häcki Buhofer 1994, Häcki Buhofer and Burger 1998, Schmidlin 1999, Schneider 1998, Saiegh-Haddad 2003, Khamis-Dakwar 2005, 2007), so that their findings cannot be incorporated into a general framework for understanding linguistic factors affecting the acquisition of literacy.

The purpose of the present paper is to begin to develop a comparative framework of this type. It will necessarily be based upon what limited comparative data are available, and the conclusions which can be drawn from such data can of course only be tentative. It should be kept in mind throughout the following discussion that efficiency in producing a literate population is only one of the considerations which policy makers weigh in designing a language policy, and that decisions of this type can be decisively influenced by issues related to national representation and symbolism independent of educational efficiency; at the same time, however, it is important that policy makers at least have some sort of concrete idea regarding the cost in efficiency of adopting one language policy as opposed to another, and hopefully the present study can be helpful in this respect.

1.0. **Scope of the study**. The term *diglossia* has generally been used to describe a situation in which the spoken language in a community, known as L (for 'low') differs significantly from its written language, known as H (for 'high'), with the understanding that in some cases L may have some limited written usages (e.g. for folk poetry, songs, children's books, etc.) while conversely in some cases H may have some spoken usages (e.g. in television news, the language in which teachers speak to students, etc.). Different studies have differed with regards to the question of which cases should be considered to constitute diglossia and which should not (see e.g. Ferguson 1959, 1991, ,Wexler 1971, Fellman 1975, Eckert 1980, Scotton 1986, Berger 1990, Daltas 1993, Schiffman 1997. Hudson 2002). According to the preference of the researcher, the reference of the term diglossia may be limited to cases in which H and L are considered to be versions of the same language and H is not the everyday language of anyone in the same country (e.g. Standard Arabic (H)

vs. Colloquial Arabic (L) in e.g. Syria), or it may also be used to refer to cases in which H is spoken as the everyday language of some geographically or ethnically distinct group in the same country (e.g. Italian (H) vs. Sicilian (L) in Italy or Standard English (H) vs. Black English (L) in the United States), or it may even include cases in which H and L are different languages (e.g. Urdu (H) vs. Punjabi (L) in Pakistan).

There are potentially interesting theoretical arguments for defining diglossia in one particular way or another. However, because the present study was prepared for presentation in a symposium which dealt specifically with Arabic and Black English, I have for the purposes of this study defined diglossia so as include these two cases and focus in general upon cases of a similar type. In this understanding, diglossia includes all of those cases in which H and L are considered to be `the same language,' regardless of whether or not H is based upon some group's vernacular usage and, if it is, whether or not this group lives in the same country as the L-speaking group. I therefore include situations of `standard-with-dialects' such as Black English while excluding cases in which H is considered to be a different language.

2.0. **Linguistic distance**. It seems reasonable to suppose that, generally speaking, the closer the written language is to the spoken language, the easier it will be to acquire literacy (see e.g. Saiegh-Haddad 2003). It is not necessarily easy, however, to determine the extent to which this is true, for a number of reasons. Linguists have not developed generally applicable tests of linguistic distance—and indeed as a linguist it is not clear to me how such a test could be developed. There have been studies of linguistic distance between specific sets of spoken vernaculars or between one language and various other languages, based upon comparison of general impressions or relative ease of acquisition for speakers of a given dialect or language (e.g. Kessler 1995, Chiswick and Miller 2004, Gooskens and Heeringa 2004, Heeringa 2004, Nerbonne and Hinrichs 2006), but such a methodology cannot be straightforwardly applied to determine, e.g. the magnitude of the distance between Black English and Standard American English in comparison to the distance between Colloquial Israeli/Palestinian Arabic and Modern Standard Arabic. It would be quite important to determine the extent to which specific attempts in specific languages to change the standard language so as to bring it closer to the spoken language have affected literacy rates, but as far as I have been able to determine there is very little real-time data of this sort (although I will refer to those data of this type which I have been able to find).

3.0. **A comparative study**. The present paper is intended to investigate how--in situations in which the standard language and the spoken vernacular are understood to be versions of the same language--the acquisition of literacy is affected by language policy, specifically the characteristics of the standard language (H), its relationship to the everyday spoken language of the learner (L), and the use of H and L in different functions both inside and outside of school. In this section, I will describe the comparative study which I conducted for the purpose of researching this issue.

Because the detailed studies which have been done to date of the acquisition of

literacy in diglossic situations have been limited to a few features in a few languages, they cannot serve as the basis of the present study, which has the purpose of investigating the **general** effect of diglossia (as it has been defined for the purposes of this study) upon the acquisition of literacy. For this purpose, it seems to me that the only data which give a global perspective on this issue are reported basic adult literacy rates,[1] and I have accordingly taking these rates as the starting point for this study.

I have gathered data regarding language policy in various countries from a number of different sources, including published research, the responses to a questionnaire which I distributed over the internet to specialists in a variety of diglossic languages (given in the Appendix to this paper), and my own personal knowledge. I have then seen what strong correlations existed between different types of language policy and higher or lower literacy rates, and I have suggested what seem to be reasonable explanations for these correlations, based upon the idea that language policy can have an effect upon the acquisition of literacy. Obviously these explanations should be understood to be tentative, given the exploratory nature of this investigation and the limitations of the available data.

Before proceeding, it is necessary to note a number of limitations of this study. Basic literacy rates are measured somewhat differently in different countries, so that simple comparison can in certain cases be misleading. My approach to this problem was to look for and report overwhelmingly powerful general patterns, each supported by data from a fairly large range of countries, rather than basing conclusions on data from a few countries which might be misleading in one way or another. Of course, even these apparently general patterns may be misleading, but the more countries with data supporting these patterns, the less likely this will be. It should be emphasized that, in spite of the shortcomings in the data on comparative basic literacy, they are by far the best data available to get a broad perspective on the questions focused upon in this paper.

The present paper will focus on rates of **basic** literacy, associated with people 'who can with understanding both read and write a short simple statement on his everyday life,' rather than **functional** literacy rates, associated with someone 'who can engage in all those activities in which literacy is required for effective functioning of his group and community and also for enabling him to continue to use reading, writing, and calculation for his own and the community's development' (from UNESCO's *Revised Recommendation concerning the International Standardization of Educational Statistics*; see e.g. Gray 1956, Levine 1994, Verhoeven 1994b, 1997), for the simple reason that to my knowledge there are no lists of comparative functional literacy rates from a large number of countries, in addition to which criteria for measuring functional literacy vary even more greatly than do criteria for measuring basic literacy. A number of high-income countries do not collect data on basic literacy rates but rather assume that basic literacy among their citizens is more or less universal, and the data source for basic literacy rates which I have used conventionally gives the literacy rate for these countries as 99.0%. Obviously a study which was focused on correlating language policy with the acquisition of literacy in

countries with such high basic literacy rates would necessarily have to focus upon **functional** literacy rates and to use such comparative data on functional literacy as are available. The present study, however, focuses particularly on cases in which the written language differs significant from the spoken language, and as we will see, in countries with language policies resulting in such situations, the basic literacy rate is typically far lower than 99.0% and there is good reason to believe that such language policies have a strongly negative effect on the acquisition of even basic literacy. At the same time, it should be kept in mind that even in e.g Holland, which is assumed to have a basic literacy rate of 99%+, functional illiteracy among native speakers may be as high as 18%, depending upon how this is measured (Doets 1994), so that a country which has a basic literacy rate of e.g. 90% is likely to have a much lower functional literacy rate and a real literacy problem in terms of the employability and productivity of the population.

Available comparative literacy data refer specifically to the **attainment** of literacy by **adults**, which is not the same thing as the **acquisition** of literacy by **children**. It would obviously be preferable for the purposes of the present study to rely upon the latter type of data, but unfortunately comparable data of this type from a wide variety of languages do not exist. In such a situation the best which can be done is to assume that the correlations which are found between language policy and adult literacy data reflect the effect which these policies have upon the acquisition of literacy by children, particularly if a plausible account can be given to explain these correlations.

The distinction between data on adults' attainment and data on children's acquisition is particularly problematic in countries in which there are a significant number of immigrants who are not native speakers of the national language. In practice, however, this phenomenon is almost entirely restricted to Western European and Anglophone states in which the basic literacy rate is in any case assumed to be at least 99% (although it is certainly very likely that the relatively large proportion of immigrants in these states affects the **functional** literacy rate). For the purposes of the present study, it is a much more pressing question why so many countries have basic literacy rates which are far lower than 99%, and so I will not be addressing this issue here (aside from one comment regarding Israel, where the relatively low basic literacy rate of 97.1% may be partially due to immigration).

In order to get a good idea of the effect of language policy upon literacy, it is necessary to identify other factors affecting literacy rates and to hold them constant so that they do not confuse the issue. It is clear that, generally speaking, wealthier countries have more money to spend on education and can thus be generally expected to have higher literacy rates, and this is clearly supported by the UNESCO data:[2]

Table 1—National literacy rates according to income of each state

High income	99.0%
Middle income	89.9%
Low income	60.2%

So as to eliminate the confounding effect of this factor, I will present data for different countries ranking them in terms of both literacy rates **and** GDP per capita, and I will base my conclusions upon a **comparison** of these rankings for each country--that is, the **difference** between literacy ranking and GDP per capita ranking--rather than upon literacy rates alone. If it then turns out that a particular type of language policy is consistently associated with a ranking which is **higher** in literacy than in GDP per capita, this will suggest that this language policy is **efficient** with regard to literacy, while on the other hand if it turns out that a particular type of language policy is consistently associated with a ranking which is **lower** in literacy than in GDP per capita, this will suggest that this language policy is **inefficient** with regard to literacy. It is of course the case that some states devote a higher proportion of their resources to education than do others, and it would in principle be preferable to use data on spending on education rather than on GDP per capita, but unfortunately such data are available for much fewer countries; I have therefore restricted myself to checking my findings based upon GDP per capita against such data on education spending as are available, and in fact these data have if anything strengthened the conclusions I have drawn on the basis of data on GDP per capita (see reference and discussion in fns. 5 and 9).

In this section I will report the results of this study. In section 3.1 I will consider cases in which the written language is directly based upon the local spoken language, while in section 3.2 I will discuss situations in which the written language differs sharply from the spoken vernacular but the two are nevertheless considered to be versions of the same language.

3.1. States in which the written language is based directly on the local spoken language. This linguistic ideology has most overtly been realized in two situations, that is, (1) the Soviet Union and the ex-Soviet states, and (2) the Slavic languages in general (these two categories overlap with Russia, Ukraine, and Belarus). In the former case, Bolshevik ideology supported developing literary languages based directly upon the spoken language for every single group in the Soviet territory (Ornstein 1968, Lewis 1972, Pool 1978, Azrael 1978, Simon 1991). Furthermore, the main criterion for drawing political borders inside the Soviet Union was creating political units which were homogeneous in terms of spoken language, so that the written language which was established for each political unit corresponded closely to the spoken language for the **entire** population of that political unit, to the extent that this was possible; this was the basis for the distinctions between (1) Russia, Ukraine, and Belarus, (2) Latvia and Lithuania, and (3) Kazakhstan, Kyrgyzstan, Uzbekistan, and Turkmenistan, with the spoken languages of each of these sets of neighboring republics being related to but distinct from each other. This policy was put into practice in the course of the 1920s, in terms of introducing policies focusing on mother tongue literacy and, where necessary, developing literary languages for languages which had not previously been written, and at least for the languages of the 15 Union Republics this continued to be the policy until the end of the Soviet regime. Data for the ex-Soviet Union Republics in 2007 are given in table 2:[3]

Table 2—Literacy and GDP per capita for ex-Soviet Republics (out of 177 countries for literacy and 179 countries for GDP per capita)

	Literacy rate	Literacy rank	GDP/capita	GDP/ capita rank	GDP-lit
Tajikistan	99.5%	9	$522	155	146
Uzbekistan [4]	99.4%	--	$753	142	127
Moldova	99.1%	14	$1,187	126	112
Georgia	100.0%	1	$2,186	108	107
Armenia	99.4%	11	$2,248	107	96
Kyrgyzstan	98.7%	55	$633	149	94
Ukraine	99.4%	11	$2,830	100	89
Belarus	99.6%	7	$4,013	81	74
Kazakhstan	99.5%	9	$6,314	65	56
Azerbaijan	98.8%	36	$3,633	88	52
Lithuania	99.6%	7	$10,472	51	44
Russia	99.4%	11	$8,612	54	43
Latvia	99.7%	5	$11,826	47	42
Estonia	99.8%	3	$15,310	41	38
Turkmenistan	98.8%	53	$5,055	76	23
Average		16		93	77

(NB—'GDP-lit' refers to the GDP ranking minus the literacy ranking, e.g. 155-9=146 for Tajikistan)

As can be seen from the data in table 2, there is a general pattern which holds for all 15 ex-Soviet Republics of having a higher international ranking in literacy than in GDP per capita.[5] Thus the data from all of these groups, when considered together, suggest that the language policy of these states is generally very efficient in terms of resulting in high literacy rates relative to the amount of money which is available to be spent on education.

We see a similar pattern if we look at Slavic languages in general. The Slavic-speaking peoples have adopted a general ideology of dividing up their territory into a large number of different linguistic units (11 in all) so that each area is relatively homogeneous in terms of spoken language and each written language is therefore very close to the spoken language of essentially the entire population; this stands in radical contrast to the policy of Western European states such as France, Germany, Great Britain, and Italy, where the general pattern is that on the one hand there is a single standard language which was originally based upon the spoken dialect of a particular region and on the other hand there are a variety of other spoken dialects which do not have corresponding written forms in active use, whose speakers read and write in the single national standard written language, so that even though the written language is based upon a particular spoken dialect, there are many people for whom this is quite different from their own spoken language (see e.g. Myhill 2006). Literacy and GDP per capita data for the Slavic-speaking countries are presented in table 3:

Table 3—Literacy and GDP per capita rankings for Slavic-speaking countries

	Literacy rate	Literacy rank	GDP/capita	GDP/ capita rank	GDP-lit
Ukraine	99.4%	11	$2,830	100	89
Belarus	99.6%	7	$4,013	81	74
Russia	99.4%	11	$8,612	54	43
Poland	99.0%	15	$10,858	49	34
Slovakia	99.0%	15	$13,227	44	29
Slovenia	99.7%	5	$22,079	30	25
Bosnia	96.7%	69	$3,400	93	24
Czech Republic	99.0%	15	$16,372	35	20
Macedonia	96.1%	71	$3,574	91	20
Bulgaria	98.2%	60	$5,116	74	14
Croatia	98.1%	61	$11,271	48	-13
Average		31		64	33

As can be seen, there is a general pattern of these countries having a higher ranking in terms of literacy than in terms of GDP per capita; this holds for all of the Slavic-speaking countries but Croatia, and the average difference in ranking is very large (33 places).

Table 4 shows the remarkable success of Soviet language policy in increasing literacy in Central Asian republics whose languages had no previous literary tradition:

Table 4--Literacy percentages for Soviet republics whose languages have no previous literary tradition

	1897	1926	1939	1959
Azerbaijan S.S.R.	9.2%	28.2%	82.8%	97.3%
Kazakh S.S.R.	8.1%	25.2%	83.6%	96.9%
Kyrgyz S.S.R.	3.1%	16.5%	79.8%	98.0%
Moldavian S.S.R.	22.2%	(no data)	45.9%	97.8%
Tajik S.S.R.	2.3%	3.8%	82.8%	96.2%
Uzbek S.S.R.	3.6%	11.6%	78.7%	98.1%
Turkmen S.S.R.	7.8%	14.0%	77.7%	95.4%

(from Lewis 1972:175; data from 1897 are from the areas of the republics, which did not yet exist as political entities, and they are for languages other than those of the republics, which had not yet been written)

As can be seen, literacy rates had only increased moderately by 1926, as the policy emphasizing mother-tongue literacy had only begun to be implemented and had not yet substantially affected the older population who had grown up before the Revolution, but already by 1939 dramatic increases had taken place and this trend was even stronger in 1959. Consider by comparison the statistics on literacy rates for ex-British colonies and ex-French colonies shown in tables 5 and 6:

Table 5--Literacy rates of non-English-speaking ex-British colonies

Israel	97.1%
Cyprus	96.8%

Singapore	92.5%
Palestine	92.4%
Jordan	91.1%
Sri Lanka	90.7%
Myanmar	89.9%
Zimbabwe	89.4%
Malaysia	88.7%
Malta	87.9%
Namibia	85.0%
South Africa	82.4%
Lesotho	82.2%
Botswana	81.2%
Swaziland	79.6%
Kenya	73.6%
Egypt	71.4%
Tanzania	69.4%
Nigeria	69.1%
Zambia	68.0%
Uganda	66.8%
Rwanda	64.9%
Malawi	64.1%
India	61.0%
Sudan	60.9% [6]
Burundi	59.3%
Ghana	57.9%
Pakistan	49.9%
Nepal	48.6%
Bangladesh	47.5%
Sierra Leone	34.8%
Average	74.0%

Table 6--Literacy rates of non-French-speaking ex-French colonies

Gabon	84.0%
Syria	80.8%
Tunisia	74.3%
Cambodia	73.6%
Madagascar	70.7%
Algeria	69.9%
Laos	68.7%
Morocco	52.3%
Mauritania	51.2%
Ivory Coast	48.7%
CAR	48.6%
Senegal	39.3%
Benin	34.7%
Guinea	29.5%

Niger	28.7%
Chad	25.7%
Mali	24.0%
Burkina Faso	23.6%
Average	51.6%

There is considerable variation in literacy rates between the different ex-British/ French colonies, but the general pattern, supported by data from a large number of countries, is that the literacy rates are much lower than those in territories which the Russians colonized and incorporated into the Soviet Union; in fact, the lowest literacy rate in the ex-Soviet republics is still higher than highest literacy rate in the ex-British/French colonies. This difference can be traced to the language policies of the British and the French, which differed radically from those of the Soviets. In many cases, the British and the French tried to teach even basic literacy in English and French in preference to written versions of local languages, which was essentially never the case for Russian in the Soviet Union. When the British and the French did take the trouble to help to develop written versions of local languages, they generally focused on particular languages--e.g. Swahili, Hausa, Urdu--which were intended to serve as the basic written medium for speakers of a wide range of dialects and distinct languages, rather than making a systematic effort to develop written languages corresponding to all of the distinct languages spoken in the territories they controlled. When archaic written versions of local languages existed which differed radically from the spoken language (as with e.g. Arabic, Tamil, Telugu, etc.), the British and the French preferred using these to creating new languages based upon spoken languages (as opposed to, for example, the Soviet decision to create modern literary Uzbek in preference to using the older written language Chagatai). And the British and the French did not generally draw either international or internal political borders so as to create linguistically homogeneous political units. The result was that in many cases children were initially educated in languages which were very different from their own spoken languages, and this can be directly related to the generally low literacy rates in the countries which Britain and France ruled (see discussion in e.g. Calvet 1974, Bokamba 1984, Phillipson 1992, Dumont and Maurer 1995, Alidou 1996, Fishman, Conrad, and Rubal-Lopez 1996, Bokamba and Tlou 1997, Gill 1999, Powell 2002, Salhi 2002).

Thus there is clear evidence that basing a written language on a local spoken language, and drawing political borders so as to general match linguistic reality, results in very high literacy rates even in countries which have very little money to spend on education. In the following section, we will see that the situation is very different when the written version of the language is markedly different from the spoken version.

3.2. Situations in which children are educated in a version of their language which is very different from their spoken language. For the purposes of the present paper, I will refer to such a situation as `diglossia', while recognizing that this is

only one of the ways in which this term can be used.[7]

I will distinguish three different types of cases. In one type, the diglossic situation is limited to a particular geographic region or ethnic group, with the L language being the regular everyday spoken language of people in this region or ethnic group and the H language being based upon the everyday spoken language of some other geographic region and/or some other ethnic group **within the same country** (an 'internal standard'); examples of this include Sicilian (L) vs. Italian (H), Black English (L) vs. Standard American English (H), etc. For the second type, the H language is based upon the everyday spoken language of a group **in another country** (an 'external standard'); examples of this type are found e.g. on Caribbean islands such as Antigua and Jamaica, where an English-based creole is the spoken language but the standard language is based upon British or American English. For the third type, the H language is not based upon **any** group's contemporary usage but rather upon a set of prescriptive rules characteristically derived—in theory, at least—from an archaic text or set of texts, as in e.g. Arabic or Tamil. My purpose in making this distinction is to give some sort of objective criterion for categorizing cases in which the standard and the spoken language are distinct forms of the same language according to what may be understood as the 'accessibility' of the distinctive standard to the child who speaks a nonstandard dialect. In the first two cases, the standard language is unproblematically **cognitively** accessible in the sense that it is at least based upon **someone's** natural everyday usage--even if it has undergone a degree of elaboration--so that a large portion of the task of the child acquiring literacy is simply to learn this other groups' dialect; among standards of this type, an 'internal' standard will presumably be more **experientially** accessible than an `external' standard in the sense that, generally speaking, a child speaking a non-standard dialect will simply have more exposure to an `internal' standard, although this is not the case in every single instance and it may be affected in particular cases by local factors (e.g. de facto racial segregation in the United States). On the other hand, in the third case the standard is less **cognitively** accessible to the child acquiring literacy because it is not based upon anyone's everyday usage so that (1) there is no 'model' to be imitated and (2) its 'rules' are more likely to be the invention of grammarians rather than the product of natural language acquisition. Notwithstanding the fact that there may be individual cases which do not fit this categorization scheme unambiguously, in practice this is rarely a real problem, and as we will see, there is a clear and strong correlation between how languages in particular countries are categorized according to this scheme and literacy rates in these countries, so that there is good reason to believe that this scheme generally reflects what I am calling `accessibility.' I will discuss these three types in 3.2.1, 3.2.2, and 3.2.3 respectively.

3.2.1. Standard-with-dialects 'internal' diglossia. The countries with situations of this type which I was able to investigate have basic literacy rates of around 99%, and I was not able to find studies suggesting that **basic** literacy is lower in such countries for people who speak nonstandard dialects than for people who speak the standard dialect. Based upon the data which I could glean from studies and observations by linguist-informants, there seems to be a general feeling that, for this type, literacy is

somewhat more problematic for speakers of nonstandard dialects but that this is to a large extent restricted to older speakers and/or functional rather than basic literacy. The picture of the effect of linguistic difference for this type is further complicated by the fact that speakers of nonstandard dialects tend to be generally poorer than speakers of standard dialects, which is likely to have an effect upon their quality of education and thus their literacy. This means that even if literacy rates are lower for speakers of nonstandard dialects, it is unclear how much of this should be attributed to linguistic differences per se.

In such cases, it is possible to make increased use of the local/ethnic language in education, e.g. teachers may speak in it in some or all classes, a standardized written form may be used to teach reading in the lower grades, and it may even be used in all written functions, partially or entirely replacing the traditional H written language. The question for language policymakers which is of central concern to the present paper is whether such an approach helps literacy—again with the understanding that this is likely to be reflected in functional literacy rather than basic literacy. Unfortunately, however, I have not been able to find any studies bearing on this important question, because educational policy in the Western European and North American countries which have situations of this type has suppressed all educational use of nonstandard dialects—spoken or written, in lower levels or in higher levels—so there are almost no cases in which researchers may compare the literacy skills of speakers of a nonstandard dialect who have been educated only in the standard written language with the literacy skills of speakers of the same nonstandard dialect who have been educated at least partially in a written version of the nonstandard dialect.

In situations of this type, even in areas in which nonstandard dialects are normally used in most day-to-day interaction, speakers of these dialects get a relatively large amount of exposure to the spoken version of H, because it is after all the everyday vernacular of a subset of the population of the country. They hear it frequently on television and radio and (particularly in the cities) in day-to-day interaction, and this means that they are generally able to understand spoken H quite well. But this does not necessarily mean that they can **speak** it very well. Additionally, young children may in some cases have had relatively limited exposure to H when they begin school, which can result in problems in the acquisition of literacy in H. It should be emphasized, however, that as far as can be determined, the difficulties posed by such a situation to the acquisition of literacy seem to be generally limited to the higher levels of literacy, so that demonstrating the effect of these difficulties and comparing the results of different language policies will be dependent upon comparative studies of higher levels of literacy.

3.2.2. Situations in which the H language is based upon the spoken language of a group of people **in another country** (an 'external standard'). This is the case, for example, in Jamaica, where the local language, known to linguists as Jamaican Creole, is the spoken language, but the standard written language is based upon the spoken English of England or the United States. Data for countries with a language policy of this type are given in table 7:

Table 7—Literacy rates of countries in which the H language is based upon the everyday usage of people living in another country

	Literacy rate	Literacy rank	GDP/capita	GDP/ capita rank	GDP-lit
Antigua	85.8%	112	$12,968	49	-63
Jamaica	79.9%	129	$3,998	82	-47
Cyprus	96.8%	67	$26,386	28	-39
Trinidad	98.4%	58	$15,908	38	-20
Switzerland	99.0%	15	$56,711	7	-8
Saint Lucia	94.8%	76	$5,747	69	-7
Grenada	96.0%	72	$5,162	73	1
Average		76		49	-27

(all of these countries have British or American English as their official language, except Switzerland, which has German, French, and Italian, and Cyprus, which has Greek, based upon a Peloponnesian dialect).

As can be seen, there is a general tendency in such countries for the literacy rate to be lower than would be expected based upon GDP per capita, but this is not really consistent, being only strong in four of the 7 countries of this type for which there are data (although it should be noted that there are no countries of this type which strongly show the opposite trend). It is reasonable to suppose that to the extent that these countries have lower literacy rates than would be expected, this can be attributed to the distance between the H and the L in combination with the fact that, because in these cases the H language is based upon a form of the language in everyday usage **in a different country**, the amount of exposure to colloquial usage of H is considerably less than it is in cases in which H is spoken in the **same** country (as above). Again, however, because there has been essentially no usage of L in education in any of these countries, there have been no studies which have given specific evidence regarding the potential effect which such a policy would have.

3.2.3. Situations in which the H language is not based upon anyone's everyday usage anywhere. This is the situation which is most directly relevant to Arabic language policy, as Arabic is a language of this type. In this case, the H language is based upon an earlier version of the language which is no longer spoken as an everyday language by anyone and which has been elaborated by a set of formal prescriptive rules developed by grammarians, so that it is in effect fossilized. Aside from Arabic, this is also the situation in Persian and a number of languages of the Indian subcontinent. Countries, or states in India, with languages of this type are listed in table 8, together with their literacy rates:

Table 8—Literacy rates for languages in which the H language is not spoken as the regular everyday language of anyone anywhere but is rather based upon an older version of the language

Sri Lanka (Sinhala)	90.7%
Iran (Persian)	82.4%
Tamil Nadu (Tamil)	73.5%

Arab states (Arabic)	70.3%
Karnataka (Kannada)	67.0%
Andhra Pradesh (Telugu)	61.1%
Bangladesh (Bengali)	47.5%

(all of these except Arabic and Persian are spoken on the Indian subcontinent)[8]

As can be seen, the literacy rates for these states are quite low. Furthermore, they are generally considerably lower than would be expected given the wealth of these states. Table 9 shows data on literacy and GDP per capita ranking for Arab states:

Table 9—Literacy ranking and GDP per capita ranking for Arabic-speaking countries (out of 177/179 countries)

	Literacy rate	Literacy rank	GDP/capita	GDP/ capita rank	GDP-lit
Qatar	89.0%	99	$70,754	3	-96
UAE	88.7%	100	$42,275	16	-84
Oman	81.4%	123	$15,412	40	-83
Bahrain	86.5%	111	$22, 109	29	-82
Saudi Arabia	82.9%	119	$15,416	39	-80
Libya	84.2%	117	$10,840	50	-67
Kuwait	93.3%	79	$32,259	24	-55
Morocco	52.3%	159	$2,368	105	-54
Algeria	69.9%	140	$3,702	87	-53
Tunisia	74.3%	134	$3,313	94	-40
Sudan	60.9%	153	$1,257	125	-28
Yemen	54.1%	157	$1,020	130	-27
Egypt	71.4%	138	$1,739	115	-23
Syria	80.8%	126	$1,928	112	-14
Jordan	91.1%	90	$2,741	102	12
Average		123		71	-52

(comparable data were not available for the Palestinian territories and Iraq)
(NB—The first five countries on this list have the lowest GDP-lit in the world)

As can be seen, literacy rankings are lower than would be expected for every Arab state except Jordan, and they are generally **much** lower, **52** places on the average.
It is particularly striking to see the case of Qatar, ranked third in the world with a per capita GDP of over **$70,000,** but ranking only 99th in the world in literacy, with a rate of only 89%.[9] Particularly striking in this respect is the study reported in Wagner, Spratt, and Ezzaki 1989, who found that among 5th grade children in Morocco, Berber speakers read Arabic just as well (or as badly) as Arabic speakers; standard Arabic is thus from a cognitive perspective as much of a foreign language for Arabic speakers as it is for speakers of Berber, which is not even a Semitic language.

States with this type of diglossia generally show a low level of literacy in comparison with GDP per capita. Iran is ranked 83th in the world in terms of GDP per capita but only 120th in terms of literacy rate, and within India, Karnataka ranks 11th out of 32 states in terms of per capita income, but 20th in terms of literacy, while the state of

Andhra Pradesh is 16th, or about average in per capita income, but 27th, near the bottom, in terms of literacy.[10]

This generalization is also supported by comparison of such cases with directly comparable situations in which standards based upon spoken language have been used. Data can be found for three cases like this, comparing Maltese with Arabic, Tajik with Persian, and Demotiki (vernacular-based Greek) with Katharevousa (the Greek H).

Maltese is the national language of Malta. Linguistically it can be considered a dialect of Arabic, but Standard Maltese is based upon spoken Maltese. It is written with the Latin alphabet with a number of diacritics (for example ħ is used to represent the voiceless pharyngeal fricative). (1) gives a typical example of written Maltese:

> (1) Il-bnedmin kollha jitwieldu ħielsa u ugwali fid-dinjità u d-drittijiet. Huma mogħnija bir-raġuni u bil-kuxjenza u għandhom igibu ruħhom ma' xulxin bi spirtu ta' aħwa.
>
> `All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.' *(Article 1 of the Universal Declaration of Human Rights)*

The literacy rate for Maltese is 87.9%; on the other hand, the literacy rate for the Arab states, where the language policy is diglossic with an H based on the fossilized classical language, is only 70.3%.

Until the Soviet period, Tajik was understood to be a dialect of Persian and, to the extent that speakers of what is known today as Tajik were literate at all, they used Persian as their literary language. As we have seen, the Soviet government developed Tajik as a distinctive written language, based upon the spoken language of Tajikistan, and today the literacy rate of Tajikistan is 99.5%; on the other hand, in Iran, the language of literacy is Standard Persian, which uses a fossilized H, and the literacy rate is only 82.4%, in spite of the fact that the GDP per capita of Iran is more than five times as that of Tajikistan.

Greek gives another example of this. Until 1976, Greek had a diglossic language situation, but since then the H language, Katharevousa, which was based upon the Byzantine language rather than any group's spoken usage, has been replaced as the standard language by a written language based upon the spoken language called Demotiki (meaning `the language of the people'; see Browning 1982, Frangoudaki 1992). After this was done, literacy rates increased drastically, as is shown in table 10:

Table 10—Literacy rates in Greece with and without diglossia

1971 (with diglossia)	86%
2006 (without diglossia)	96%

The question then arises as to why this type of diglossia is particularly problematic in terms of literacy, as we have seen to consistently be the case (the case of Sinhala, which is an exception to this pattern, will be discussed below). My best guess at present is that this is because in these cases—and **only** in these cases—the standard (H) language is not based upon any group's contemporary usage but rather upon older texts and grammatical rules which grammarians have constructed, in principle upon the basis of these texts. It may have been the case that these texts were based upon an **earlier** spoken version of the language, although it is not clear that this is the case. But independent of this question, such standard languages at present are conceptualized as a collection of inflexible grammatical rules which have not evolved to meet the natural needs of language users. On the other hand, in diglossic situations in which the H is based upon some group's everyday usage, the standard language does evolve—albeit more quickly in some cases than in others—as the spoken usage of the model group evolves, and this ongoing input from spoken language keeps the standard language relatively natural, even for speakers of other dialects.

This supposition is further supported by data from Hebrew, where the language which is understood to be prescriptively 'correct' is based upon Biblical and Mishnaic texts. Although this situation has not been traditionally described as diglossic, particularly because the actual distance between the spoken and written language is not so great, owing to the fact that Hebrew was only revived as a spoken language, based upon these texts a little over a century ago, nevertheless the same general pattern appears as with diglossic languages with fossilized Hs—Israel is ranked 31st in terms of GDP per capita but only 66th in terms of literacy. While this difference is considerably less extreme than Arab countries such as Oman, Bahrain, and Saudi Arabia, which have similar GDPs per capita, it is still striking, particularly given the unusually high rate of spending on education by the Israeli government (7.5% of GDP, ranked 17 out of 132 countries (see reference in fn. 5)). The lesser differential between GDP ranking and literacy rate in Israel as opposed to Arab countries might be attributed to this higher rate of spending on education or to the smaller linguistic distance between the spoken language and the written language in Hebrew as opposed to Arabic (or to a combination of these factors). On the other hand, it is worth noting that Israel does have a relatively large immigrant population and this in itself may account for its relatively low literacy rate independent of the nature of the standard language.

It should be emphasized that this account of why languages with fossilized Hs have significantly lower literacy rates than would be expected is only a guess. The general **pattern**, however, is quite clear, whatever the explanation for it may be, and the explanation for this pattern is something which should be investigated in future studies and considered in future policy decisions.

It is important to consider the case of Sri Lanka in more depth, because this is the one case of a diglossic language with a fossilized H which clearly goes **against** the general trend. Sri Lanka ranks 118th in terms of GDP per capita but 93rd in terms of literacy rate, which is radically different from other diglossic languages with fossilized Hs. This literacy rate is even more impressive when it is considered that

about 20% of the population of Sri Lanka speak Tamil as their native language, and although separate literacy figures are not available, it is safe to assume that the Tamil speakers pull down the overall literacy rate, as the literacy rate in Tamil Nadu in India is only 73.5%. Consider for example the data in table 11, which compares literacy rates in Sri Lanka and Arabic-speaking countries with a GDP close to that of Sri Lanka:

Table 11—Literacy rates and GDP per capita for Sri Lanka and those Arabic-speaking states with comparable GDP per capita

	Literacy rate	Literacy ranking	GDP per capita	GDP ranking
Sri Lanka	90.7%	93	$1,558	118
Egypt	71.4%	138	$1,739	115
Morocco	52.3%	159	$2,368	105
Sudan	60.9%	153	$1,257	125
Syria	80.8%	126	$1,928	112
Yemen	54.1%	157	$1,020	130

Sri Lanka shares with these Arab states a diglossic language policy with a fossilized H, and the GDP per capita is generally in the same range, but as can be seen, literacy rates in Sri Lanka are far higher than in these Arab states. It should also be noted that in both Sri Lanka and the Arab states it is normal for teachers to speak L to the students, which was not the case for **any** of the other diglossic situations which I investigated.

The question then arises as to why the literacy rate in Sri Lanka is so much higher, in comparison with GDP per capita, than it is in any of the other states which have fossilized Hs. This is a question which needs to be investigated in more detail, but my research has suggested a possible explanation for this difference. The most obvious distinguishing feature in Sri Lanka is that **reading in Sinhala is taught in L for the first four years of school, with students only beginning to learn to read in H in the 5th grade**. This is not the case for Arabic or for that matter for any of the other diglossic languages for which I was able to find data; in these languages, L is not used as a written language in school at all.

This finding would be generally consistent with research, referred to above, showing that literacy is most effectively acquired through the medium of the native language in the early years, with the switch being made to learning to read and write another language only after a few years of schooling, which has formed the basis for UNESCO policy. In the case of Sinhala, this switch is not to `another language' but rather to the H form of what is understood to be the `same language,' but the general pattern is nevertheless the same, because this is significantly different from the spoken language of the child.

In fact, it might be possible to carry this argument a step further, because in fact literacy rates in Sri Lanka are **better** than would be expected given GDP per capita—that is, not only is the effect of a fossilized H cancelled out but the situation is actually **reversed**. While I am venturing into speculation here, it may be that the conscious use of a markedly different H language motivates educational authorities to use--in the early years old schooling-- a written form of L which is **extremely** close to colloquial speech, closer than it would be in a language such as English or French in which the difference between the spoken language and the formal written language is not as great; that is, early readers in Sinhala may more freely be modeled upon colloquial usage because there is no pretense that this is `real' written language, as there is in

e.g. English or French, and this makes acquisition relatively easier (notwithstanding the relatively low economics status of the country). In such a system, `real' written Sinhala is only introduced at a later stage when children are cognitively able to handle it. This account is of course speculative, but it does seem reasonable based upon available data.

3.3. Summary. The comparative study reported in this section has shown that even in cases in which the standard written language is assumed to be 'the same language' as the local spoken language, there are radical differences in terms of literacy rates depending upon the relationship between the spoken and standard language. In cases in which the standard language has been consciously based upon the contemporary spoken language so that there is minimal difference between the two, it is possible to have essentially universal literacy even in very poor states which spend proportionally little of the money they have on education (see tables 2 and 3 and fn. 5). At the other extreme, in cases in which the standard language is not based upon **any** group's everyday usage but is rather to a certain extent the creation of grammarians and thus not 'natural', in which there is no target group whose everyday usage can be learned and imitated, basic literacy is much lower than would be expected based upon national wealth and the relative resources devoted to education (see table 9 and following discussion and fn. 9). Between these extremes, it seems that in diglossic situation, standard languages based upon dialects spoken in the same country are easier to learn than standard languages based upon dialects spoken in a different country, even when we factor in the consideration that the latter situation is generally associated with poorer countries (see table 7).

These findings are consistent with the hypothesis that not only is it easier to learn to read and write one's own native language as opposed to a foreign language--as has been confirmed by the studies referred to at the beginning of this study--but it is also easier to learn to read and write a more `accessible' version of one's own native language. The findings suggest that **the most `accessible' version is one based upon one's own specific native dialect and that it is possible to attain essentially universal literacy in a standard language based upon the local spoken dialect even with extremely limited financial resources**; in such a case, children learn to read and write by simply learning a written version of the spoken dialect which they already know. Progressively less `accessible' standard languages are those based upon other dialects spoken in the same country, followed by those based upon dialects spoken in other countries, and finally those which are not based upon contemporary spoken dialects at all. It should be noted that aside from the observation that there is good reason to believe that it is easiest to learn to read and write a standard language which is literally based upon one's own native dialect, this study has **not** addressed the question of the relative ease of the acquisition of literacy based upon **linguistic** distance--that is, given two possible standard languages which are both based upon dialects other than one's own native dialect, it is **not** clear that it will be easier to learn to read and write the one which is linguistically closer to one's one native dialect; there is, however, reason to believe that it will be easier to acquire literacy in the standard language to which one is more **exposed** in one way or another. To put this in

concrete terms for the sake of exemplification, there **is** reason to believe that, e.g., people living around the border of France and Italy will have an easier time acquiring Standard French or Standard Italian depending upon their relative exposure to the dialects (Parisian and Tuscan respectively) upon which these standards are based, while there is **no** particular evidence that I am aware of suggesting that they will have an easier time acquiring one or the other of these standards depending upon the linguistic proximity of the dialects upon which these standards are based to the local dialect of the border region (linguistic proximity **may** have an effect in such cases, but this has not been suggested by the data in the present study or any other study which I am aware of).

My discussion here has generally assumed that in cases in which the standard language has been understood to be 'the same language' as the local spoken language, education will be in **the same version of the same language** from beginning to end. I have made this assumption simply because, aside from Sinhala in Sri Lanka, I do not know of a single case in which there is a change from one version of the language to another at a certain point in the educational process, and there is therefore very little information on the effect of such a policy. The case of Sri Lanka suggests, however, that the effect of using a standard language which is very different from the local spoken language may be considerably mitigated (or even reversed) by using a written form based upon the local spoken language in the early years of schooling and only switching to the real standard language for later education (in Sri Lanka this starts in the 5th grade). Such a compromise seems like the most efficient choice for a state which for one reason or another insists upon using a standard language which is not based upon the local spoken language but which wants to minimize the negative effects of such a policy upon general literacy. But it is important to note that even in the case of Sri Lanka the resulting rate of literacy of 90.7% is not as high as many states would like and a good deal lower than states such as Tajikistan, Uzbekistan, Moldova, and Kyrgyzstan, which have a lower GDP than Sri Lanka but base their standard languages on the local spoken language.

4.0. **Diglossia with fossilized Hs and technological change: A historical perspective**. Prior to the invention of the printing press in the mid-15th century, Western Europe was to a large extent in a situation like the one I have discussed in section 3.2.3. The only language which was recognized as fully legitimate was Latin, which had remained essentially frozen as a written language even while its spoken version developed and diversified radically. Over the course of time, various written versions of the different spoken dialects of Latin developed, and these have since come to be known as the Romance languages—French, Italian, Spanish, Portuguese, Romanian, Moldovan, Catalan, etc. Written texts presently understood to be early examples of these languages began to appear from the 12th century, e.g. *La Chanson de Roland* in French (dating from the mid-12th century), *El poema del Cid* in Spanish (dating from the early 13th century), Llull's *Llibre de Meravelles* in Catalan (dating from the late 13th century), the *Cancioneiro da Ajuda* in Portuguese (dating from the late 13th century), and Dante's *La Divina Commedia* in Italian (dating from the early 14th century). But notwithstanding the significance which such texts have to modern

literature and linguistics, their production did not have a significant effect **at the time** upon the general dominance of Latin, which was associated with entrenched conservative interests, in particular the Catholic Church, who did not support the idea of a fully literate and intellectually active public (see discussion in Anderson 1983).

There was at the time a generally parallel situation in Slavic-speaking Eastern Europe (where the H language was Church Slavonic), Greek-speaking areas (where the H language was the ancestor of Katharevousa), and Arabic-speaking areas (where the H language was based upon the language of the Koran). But it is safe to assume that the distance between the H language and the spoken language was less in these cases than it was in Western Europe, for the simple reason that the H language there, Latin, had been essentially frozen for far longer—over a thousand years—than in any of these other cases. Although it is not possible to determine relative literacy rates in the 15th century, the very antiquity of the H language in Western Europe was doubtless a considerable obstacle to literacy, and it is quite likely that for this reason literacy rates there were relatively low at the time and this contributed to the relatively backward state of the area in comparison with the East.

But over the next few hundred years a remarkable change was to take place. Western Europeans put the printing press to widespread use, on a scale far beyond anything which was done to the east, where the Ottoman Empire banned its usage fairly soon after its introduction, because its leaders recognized that this technology had the capacity to undermine the existing political order (see Anderson 1983). The new technology came to be associated particularly with the new written languages which were developing in Western Europe, based upon local spoken languages, which were incomparably easier for the general public to read; Latin continued to dominate in handwritten manuscripts, but the output of this medium was soon overwhelmed by that of the printing press. By the end of the 16th century, Latin was in retreat and the new written languages were advancing everywhere. The linguistic unity of Western Europe was shattered, but Western Europeans became at the same time incomparably more literate and more intellectually productive and creative, as they developed and used written languages based upon local spoken languages in which more of them could write and read more easily and more effectively. It was specifically during this period of time—and, it may be argued, to a large extent for this reason—that Western Europe, though linguistically divided, began to be the most advanced civilization in the world. The Ottoman Empire, on the other hand, remained politically united but fell farther and farther behind Western Europe; having rejected the printing press and therefore not developed written versions of vernacular languages, the Ottomans continued to rely on fossilized written languages—the written Ottoman language and Classical Arabic--which were inaccessible to the overwhelming majority of their subjects.

We are currently at the very beginning of another technological revolution in written language, and if history is any indicator the effects of this will be no less dramatic. Just as the printed word began to replace the handwritten word in the second half of the 15th century, so in the last 25 years the electronic word has begun to replace the

printed word (see e.g. Danet and Herring 2003). Already many people read and write more through electronic media—email, SMS messages, blogs, electronic journals, etc.—than through printed media, and in many of these contexts the tendency is to write in a manner which directly reflects vernacular speech. Thus just as with the printing press, a new technological medium is beginning to bring with it new ways of writing.

The recency of this development is reflected in markedly different conventions for writing electronically in different age groups. People older than 30 today were introduced to electronic writing at a later stage in their lives, after they had already developed literacy based upon the printed word, and it was therefore natural for them to simply transfer their established writing conventions to electronic media when these became available. On the other hand, people still below the age of 30 today were generally introduced to electronic writing at a much younger age, before they had fully developed literacy skills based upon the established norms associated with the printed word, and this means that in certain contexts in which there is a feeling that the language should be more colloquial—particularly email messages, SMS messages, and often blogs—they feel relatively free to ignore normal writing conventions to achieve this end. In cases in which the spoken language is in any case not too different from the written language, as in English and Hebrew, such young speakers can basically use the established norm with a few modifications to make it feel more colloquial (e.g. u for *you*, 4 for *for*, etc.). But in cases in which there is a radical difference between the spoken language and the established written language, such as those situations which I have referred to as being diglossic, this is not possible, and writing colloquially essentially involves dropping the established standard language entirely and making up a way to write the colloquial language. And this is what is happening: Every person who responded to my internet survey—about Sicilian, Persian, Tamil, Swiss German, etc.—reported that among young people in particular L is commonly used in electronic media, even in languages which previously had no tradition of writing L, and that it is the normal usage for young people in contexts such as emails, SMS messages, and in most cases blogs. In practice, this means that in those cases in which there is no established way to write the spoken language, young people have no choice but to devise an ad-hoc orthography, not necessarily even in the same alphabet as that which is used to write their H language.

This is also happening in Arabic, and this phenomenon has begun to receive the attention of academic researchers (see e.g. Warschauer, El Said, and Zohry 2002, Wheeler 2003, Palfreyman and al-Khalil 2003). Garra 2007 has shown that various orthographic systems for representing the different varieties of spoken Arabic are beginning to take shape, parallel to what started to happen in Western European languages in the generation or two after the introduction of the printing press, as people are representing their spoken language in a manner which is less ad-hoc and increasingly based upon imitation of general usage for their own spoken dialect. Even more strikingly, there are cases in which local written standards are beginning to develop: For example, Arabic speakers from the central region in Israel today generally write electronically using a written version of the spoken language of the

northern region rather than their own dialect (Garra pp. 91-4); when writing in Latin script, Lebanese use <ch> and <ou> for the sounds which Israelis/Palestinians would write as <sh> and <u> respectively, because of the influence of the French writing system in Lebanon; the letters saad and taa' are typically represented by <9> and <6> in the Gulf States but <s> and <t> elsewhere (Garra pp. 67-8). There is still a good deal of variation in some usage even within a single spoken dialect area; Garra found that for her Israeli/Palestinian data, ghayn could be represented (in descending order of relative frequency) as <g>, <3'>, <gh>, <8>, or <g'>, with none of these usages particularly dominating (pg. 66), and there is ongoing inconsistency with regards to writing vowels in Latin letters, particularly regarding the question of whether to represent them phonetically or transliterate from Arabic orthography (pp. 72-7)—but there is a general development towards regularization in this case, with transliteration being the norm except for the definite article (pp. 74, 95; it should be noted that this concession to spoken language particularly in the case of the definite article is parallel to orthographic conventions in e.g. French and Italian). A typical example of how Arab Israelis write their spoken language using Latin letters is given in (2), from the Panet forum (pp. 89-90):

> (2) Kolhen be2refo, bs Haifa elle 3anjad btestahal la2ano btjanen ow jamalha tabe3e.. ama elba2aya kolhen 3amaleyat tajmeel.. matalan dina hayek ma heye bte2ref shu 7elo feha ya3ne?? wala elissa mahe tomha a3waj ow mesh 7elwe shelleama zoo2 3aleko ya nas..lesh najwa karam 7elwe?? araaaaaaaaf!! wala amal 7ejazy mhye zai el amwat manzarha belzat bel look eljded!!! welko ya nas shu sayebko?? hadol!! 7elwat???? shelle la2
>
> They[celebrities who participated in a beauty contest]'re disgusting, Haifa [a famous Arab singer] deserves [to win the beauty contest] because she's gorgeous and her beauty is natural. But all the others [singers] have had cosmetic surgery. Like Dina Hayek, what's beautiful about her? She's ugly! And Elissa, her mouth is twisted! She isn't beautiful at all! What kind of taste have you people got? And Najwa Karam is beautiful? Get real! Ugh! And Amal Hijazi!! She looks like a corpse, especially her new look. What's wrong with you people? These women are beautiful? Absolutely not!

The same sort of thing is happening when the colloquial language is written with Arabic letters, and in fact in this case the trends towards regularization is stronger because there is less confusion about how to represent vowels (Garra pp. 81-5). Thus speakers from the Gulf will commonly write yaa' for standard jeem and urban speakers from Israel/Palestine will write hamza for qaaf, in order to reflect their own dialectal pronunciation; Egyptians will write jeem but understand that they are pronouncing it [g]; and speakers who pronounce kaaf as a voiceless palato-alveolar affricate in some situations may write this with jeem, with the understanding that in this case it should be pronounced as voiceless rather than voiced.

What this means is that the next generation of Arabic schoolteachers will be in a radically different situation from the present generation. At the moment, as in other

diglossic situations (excluding exceptional cases like Sinhala), there is a general understanding in the school and the society in general that the spoken language is simply not written, and there has even been some sort of underlying feeling that it **cannot** be written, or at the very least it is not clear how to write it in a systematic way. But this is no longer the case for young Arabic speakers, and as these speakers grow up and get jobs it will soon also not be the case for Arabic-speaking **teachers** either. As electronic writing becomes more and more dominant, the idea that colloquial Arabic cannot be written, or cannot be written systematically, will disappear and the inconvenience and inefficiency of writing in a fossilized standard language will become more and more obvious and undeniable.

This will mean that it will be much easier in Arabic-speaking societies to instantiate a program such as is being used with Sinhala, where a written version of the colloquial language is used to teach reading in the early grades. Previously this appeared to be problematic in Arabic because there was no established way to write the spoken language, but now because of the advent of electronic writing, such a system has begun to develop spontaneously, and the next generation of teachers will know it quite well.

In considering the advisability of taking such a step, it should be kept in mind not only that the evidence suggests that it would increase literacy, but also that it would represent an innovative use of a new media for writing which is clearly the wave of the future throughout the world. History has shown that Western European civilization advanced particularly when Western Europeans enthusiastically adopted a new medium for producing and reproducing written language and took the lead in terms of using this new technology to structure their use of written language--**even though this resulted in political fragmentation**. It is reasonable to suppose in this context that the most successful societies of the future will similarly be those which make the most productive use of electronic writing and structure their own use of written language around this, and the groups which are best placed to do this are those whose own speakers are engaged in actively developing new means of writing specifically for electronic purposes today, that is, speakers of languages which have until now been diglossic.

5.0. **Conclusion**. This study has been limited by the availability of comparable data regarding the acquisition of literacy in a wide variety of languages. In the absence of such data, I have gathered such data as exist, superficial and potentially problematic though they may be, and restricted myself to conclusions which were so overwhelmingly clear and supported by so many cases as to obviate the dangers associated with shortcomings of the data, particularly when supplemented with the results of the questionnaire study which I conducted. This appears to be the best which can be done under the circumstances; obviously, further research will be necessary before conclusions can be drawn with a significantly greater degree of confidence. The best course of action to expedite such a process would be to establish consistent and regular contacts, through email networks, journals, conferences, research teams, etc., between linguists and reading experts who are interested in

researching this question from a cross-linguistic point of view. In conducting the research for this paper, it became apparent to me to what extent such contacts are still missing and needed.

Footnotes

*I would like to thank Eliezer Ben-Rafael, Ruth Berman, Avital Darmon, Tami Katzir, Dorit Ravid, Elinor Saeigh-Haddad, Liliana Tolchinsky, and Yossi Zelgov, for their helpful comments on earlier drafts of this paper. I also thank the informants in my questionnaire survey. All remaining errors are my own.

[1] Unless otherwise indicated, basic literacy data for individual countries which I will refer to in this study are taken from http://en.wikipedia.org/wiki/List_of_countries_by_literacy_rate (based upon UNESCO data).

[2] GDP per capita data which I will refer to in this study are taken from http://en.wikipedia.org/wiki/List_of_countries_by_GDP_%28nominal%29_per_capita.

[3] It should be noted that in the overwhelming majority of cases, the numbers in table 2 indicate literacy in the language of the state rather than in Russian. There is no reason to suspect that Russians are any more literate than are non-Russians in ex-Soviet states; indeed, the three republics with the highest proportion of ethnic Russians—Kazakhstan, Latvia, and Estonia—all have **higher** literacy rates than Russia itself.

[4] The Uzbekistan data are from 2003, because more up-to-date data were not available; I have therefore not listed the literacy ranking for Uzbekistan, but a rate of 99.4% would be tied for 11th.

[5] It is possible that the government of the Soviet Union invested a relatively high proportion of their resources in basic education and that this would result in a relatively high rate of literacy compared to GDP per capita. While this hypothesis is certainly worth investigating, it should be pointed out that the data in tables 2 and 3 are from 2007, 16 years after the dissolution of the communist government of the Soviet Union, that I do not know of evidence that the Soviet Union spent a high proportion of its resources on education, and that in fact at present the countries listed in table 2 spent are if anything spending a disproportionately **low** percentage of their GDP per capita on education (see http://www.nationmaster.com/graph/edu_edu_spe-education-spending-of-gdp, which has data for all of these countries other than Uzbekistan and Turkmenistan), averaging only 4.2% and a ranking of 76 out of 132 countries, making their literacy rates even more impressive.

[6] The data from Sudan are only from northern Sudan; southern Sudan is entirely non-Arabic speaking and has been in a state of almost constant war against the north for the last 50 years, so presumably the literacy rates are lower there.

[7] It should be noted that there is a complication in this understanding in that it may not be clear whether the spoken language and written language are understood to be `the same language.' For example, all three Swiss informants in my internet study told me that Swiss people do not consider Swiss German to be `the same language' as Standard German (which was historically based upon an Upper Saxon dialect). In theory, this would mean that the situation in the Germanic-speaking areas of Switzerland does not represent an example of the situation focused upon in this paper. However, as far as I could determine, German speakers outside of Switzerland **do** think that Swiss German is `a version of German' (and furthermore Ferguson included this case as one of his prototypical examples of diglossia in his seminal 1959 paper), and so I did include this case in my survey.

[8] Literacy data from Indian states are from http://cyberjournalist.org.in/census/cenlit0.html. Data from Sri Lanka, Iran, the Arab states, and Bangladesh are from the website referred to in fn. 1.

[9] This is not because of traditional restrictions on female literacy, as the male and female literacy rates in Qatar are basically the same. It is also worth noting that for the six Arab states for which there are data available regarding spending on education as a proportion of GDP in http://www.nationmaster.com/graph/edu_edu_spe-education-spending-of-gdp (Yemen, Morocco, Tunisia, Oman, Lebanon, and UAE), the average amount of money spent on education is 5.2% of GNP, considerably higher than the rate for ex-Soviet states (see fn. 5), while the average rank is 60 out of 132 countries listed, evidence suggesting that the problem is **not** that insufficient money is being spent on education in Arab states.

[10] Per capita income data from Indian states are taken from http://sampark.chd.nic.in/images/statistics/SDP2005R6.pdf. Tamil Nadu is about the same in terms of literacy and per capita income, while Bangladesh is only slightly lower in terms of literacy but is in any case almost at the bottom of the world rankings in both categories.

References

Aitchison, Jean. (1981). *Language change: Progress or decay?* Cambridge: Cambridge University Press.

Alidou, Ousseina. (1996). "Francophonie, World Bank, and the collapse of the Francophone Africa educational system". *CAFA Newsletter* 11.

Anderson, Benedict. (1983). *Imagined communities: Reflections on the origin and spread of nationalism*. London: Verso.

Azrael, J.R. (ed.) (1978). *Soviet nationality policies and practices*. New York: Praeger Publishers.

Berger, M. (1990). "Diglossia within a general theoretical perspective: Charles Ferguson's concept 30 years later". Multilingua 9.285–295.

Bokamba, Eyamba G. 1984. French Colonial Language Policy and its Legacy. Studies in Linguistic Sciences 14

Bokamba Eyamba and Tlou, J. (1997). "The consequences of the language policies of African states vis-à-vis education". *Proceedings of the VII Conference on African Linguistics*. Columbia, OH: Hornbeam.

Browning, Robert. (1982). Greek Diglossia Yesterday and Today. *International Journal of the Sociology of Language* 35. 49-68.

Burger, Harald, and Häcki Buhofer, Annelies. (1994). Spracherwerb im Spannungsfeld von Dialekt und Hochsprache. Zurcher Germanistische Studien , 38. Bern: Peter Lang

Calvet, L.-J. (1974). *Linguistique et Colonialisme: Petit Traité de Glottophagie*. Paris: Payot.

Chiswick, Barry R. and Miller, Paul W. (2004). "Linguistic Distance: A Quantitative Measure of the Distance Between English and Other Languages." *IZA Discussion Paper No. 1246*.

Daltas, P. (1993). "The concept of diglossia from Ferguson to Fishman to Fasold". In I. Philippaki-Warburton, K. Nicolaidis, and M. Sifianou (eds.), *Themes in Greek Linguistics: Papers from the First International Conference on Greek Linguistics*,. Amsterdam: Benjamins. 341-348

Danet, Brenda and Susan Herring. (2003). The Multilingual Internet: Language, Culture, and Communication in Instant Messaging, Email and Chat. Special issue of the Journal of Computer Mediated Communication 9.1.

Doets, Cees. (1994). "Assessment of adult literacy levels: The Dutch case." In L.Verhoeven, (ed.). *Functional Literacy*. Amsterdam: John Benjamin Publishing Company pp. 321-32.

Dumont, P. and Maurer, B. (1995). *Sociolinguistique du Français en Afrique Francophone*. Vanves: EDICEF.

Dutcher, N. and G.R. Tucker. (1997). *The Use of First and Second Languages in Education: A Review of Educational Experience*. Washington D.C.: World Bank, Country Department III

Eckert, Penelope. (1980). "Diglossia: separate and unequal". *Linguistics* 18.1053–1064.

Fellman, Jack. (1975). "On diglossia." *Language Sciences* 34. 38–39.

Ferguson, Charles. (1959). Diglossia. *Word* 15, 325–340.

Ferguson, Charles. (1991). Diglossia revisited. *Southwest Journal of Linguistics* 10 (1).214 –234.

Frangoudaki, Anna. (1992). "Diglossia and the present language situation in Greece: A sociological approach to the interpretation of diglossia and some hypotheses on today's linguistic reality". *Language in Society* 21 (3).365-81.

Fishman, Joshua A., A.W. Conrad and A .Rubal-Lopez (eds.). (1996). *Post-Imperial English*: *Status Change in Former British and American Colonies, 1940–1990*. Berlin: Mouton de Gruyter.

Garra, Eman. (2007). *From a dialect into a language: The cases of English and Arabic*. MA Thesis. Haifa: University of Haifa English Department.

Gill, H. (1999). "Language choice, language policy and the tradition–modernity debate in culturally mixed postcolonial communities: France and the francophone Maghreb as a Case study". In Y. Suleiman (ed.) *Language and Society in the Middle East and North Africa.* Richmond: Curzon Press.

Gooskens, C. and W. Heeringa. (2004)."Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data". *Language Variation and Change* 16.189-207.

Gray. W.S. (1956) *The teaching of reading and writing*. Chicago: Scott Foresman.

Gudschinsky, S.C. (1977). "Techniques for functional literacy in indigenous languages and the national language". In T.P. Gorman (ed.), *Language and literacy: Current issues and research*. Teheran: International Methods.

Häcki Buhofer, Annelies, and Burger, Harald. (1998). *Wie Deutschschweizer Kinder Hochdeutsch lernen.* Stuttgart: Franz Steiner Verlag.

Heeringa, W. (2004). *Measuring dialect pronunciation differences using Levenshtein distance*,

Dissertation. Groningen: University of Groningen.

Hudson, Alan. (2002). "Outline of a theory of diglossia". *International Journal of the Sociology of Language* 157.1-48.

Kessler, B. (1995). "Computational dialectology in Irish Gaelic". *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics*, *EACL*, Dublin, pp. 60-67.

Khamis-Dakwar, Reem. (2005). "Children's attitudes towards the diglossia situation in Arabic and its impact on learning". *Language, Communities and Education* 1. 75-86.

Khamis-Dakwar, Reem. (2007). The Development of Diglossic Morphosyntax in Palestinian Arabic-speaking Children. Unpublished dissertation, Columbia University, New York.

Levine, Kenneth. (1994). Functional literacy in a changing world. In L. Verhoeven,.(ed.) *Functional Literacy*, pp. 113-31.

Lewis, E. Glyn. (1972). *Multilingualism in the Soviet Union: Aspects of language policy and its implementation*. The Hague: Mouton.

Mehrotra, S. (1998). *Education for All: Policy Lessons From High-Achieving Countries*. New York: UNICEF Staff Working Papers.

Myhill , John. (2006). *Language, religion, and national identity in Europe and the Middle East: A historical study*. Amsterdam: John Benjamins.

Nerbonne, John, and Erhard Hinrichs. (2006). *Linguistic distances. Proceedings of the Workshop on Linguistic Distances*, pp. 1-6. Sydney: Association for Computational Linguistics

Okedara, J.T. and C.A. Okedara. (1992)."Mother tongue literacy in Nigeria". *Annals AAPSS* 520 (1). 91-102.

Ornstein, Jacob. (1968). "Soviet language policy: Continuity and change" . In E. Goldhagen (ed.), *Ethnic minorities in the Soviet Union*, pp. 121-46. New York: Frederick A. Praeger.

Palfreyman, David, and Muhamed el-Khalil. (2003)."'A funky language for teenzz to use': Representing Gulf Arabic in instant messaging". *Journal of Computer-Mediated Communication* 9 (1).

Phillipson, Robert. (1992). *Linguistic imperialism*. Oxford: Oxford University Press.

Pool, Jonathan. (1978). "Soviet language planning: Goals, results, options". In J.P. Azrael (ed.), *Soviet nationality policies and practices*, pp. 223-49. New York: Praeger Publishers.

Powell, Richard. (2002). "Language Planning and the British Empire: Comparing Pakistan, Malaysia

and Kenya". *Current Issues in Language Planning* 3 (3).205-78.

Ravid, Dorit. (1995). *Language change in child and adult Hebrew: A psycholinguistic perspective.* Oxford: Oxford University Press.

Saiegh-Haddad, Elinor. (2003). "Linguistic distance and initial reading acquisition: The case of Arabic diglossia". *Applied Psycholinguistics* 24. 431-51.

Salhi, Kamal. (2002). "Critical Imperatives of the French Language in the Francophone World: Colonial Legacy – Postcolonial Policy". *Current Issues in Language Planning* 3 (3).317-45.

Schiffman, Harold. (1997). "Diglossia as a sociolinguistic situation". In F. Coulmas (ed.), *The Handbook of Sociolinguistics*. Oxford: Blackwell. pp. 205–216

Schmidlin, Regula. (1999). *Wie Deutschschweizer Kinder schreiben und erzählen lernen.* Tübingen: Francke.

Schneider, Hansjakob. (1998). *Hochdeutsch, das kann ich auch: Der Erwerb des Hochdeutschen in der deutschen Schweiz. Eine Einzelfallstudie zur frühen mündlichen Sprachproduktion*. Bern: Lang.

Scotton, C. (1986). "Diglossia and code switching". In J. Fishman et al (eds.), *The Fergusonian Impact: In Honor of Charles A. Ferguson on the Occasion of his 65th birthday*, vol. 2,: Sociolinguistics and the Sociology of Language. Berlin: Mouton de Gruyter. pp. 403 – 415

Simon, Gerhard. (1991). *Nationalism and policy toward the nationalities in the Soviet Union: From totalitarian dictatorship to post-Stalinist society*, trans. By Karen Forster and Oswald Forster. Boulder: Westview Press.

UNESCO. (1953). *The use of vernacular languages in education*. Paris: UNESCO.

Verhoeven, Ludo. (1994)a "Modeling and promoting functional literacy". In L. Verhoeven (ed.), pp. 3-34.

Verhoeven, Ludo, (ed.). 1994b. *Functional literacy: Theoretical issues and educational implications*. Amsterdam: John Benjamins.

Verhoeven, Ludo. (1997). "Acquisition of literacy by immigrant children." In *C. Pontecorvo (ed.), Writing development: An interdisciplinary view*. Amsterdam: John Benjamins. pp. 219-40

Wagner. D.A., J.E. Spratt, and A. Ezzaki. (1989). "Does learning to read in a second language always put a child at a disadvantage? Some counter-evidence from Morocco." *Applied Psycholinguistics*. 10. 31-48.

Wexler, Paul. (1971). "Diglossia, language standardization, and purism: parameters for a typology of

literary languages". *Lingua* 27. 330 –354.

Wheeler , D.L. (2003). "The Internet and youth subculture in Kuwait." *Journal of Computer-Mediated Communication* 8 (2).

Warschauer, M., G. El Said, and A. Zohry. (2002). "Language choice online: Globalization and identity in Egypt". *Journal of Computer-Mediated Communication* 7 (4).

Appendix

Please answer the following questions about diglossia in your language:

1. What is the language of school instruction (that is, the language the teacher speaks in)? Does this vary between subjects and/or levels and if so, how?
2. Which language are the students expected to speak in class?
3. Which language is used in teaching reading? Is it only H from the beginning, or is L used at an early stage? If so, until which stage?
4. What is the language of textbooks, H or L? Please specify if there is a difference between subjects or between beginning and advanced learners.
5. How much exposure is there to the H language before formal school begins(e.g. television, children's readers, etc.)?
6. Have there been studies of the effects of the diglossic situation on reading comprehension, mathematical literacy, scientific literacy, etc.? It would be particularly helpful if any of these studies compared reading in H as opposed to reading in L. Specific references (in any language) would also be appreciated.
7. What is the literacy rate for H among the L-speaking community (approximately)?
8. Is L being used these days (particularly by young people) for electronic usages, like blogs, SMS messages, email messages, etc.?