Chapter 6

Language contact

6.1 Introduction

In this chapter, we shall look at a third reason why languages may come to resemble each other: because they are in contact with each other and therefore influence each other.

6.2 Borrowing

In languages, **borrowing** implies that one language takes something from another language and makes it into a permanent part of its own system. In contrast to the borrowing of an object from another person, the borrowing of a linguistic form or concept from another language is not implied to be temporary. This is one of the main differences between borrowing and **code-switching**. Borrowed forms or concepts are **integrated** into the borrowing language, while code-switching, as we shall see, implies the use of two or more different language codes within the same stretch of speech.

    More than two thousand years have passed since the Chinese borrowed the word now pronounced *pútao* 'grape' from an Old Iranian word probably pronounced *bādāwa*, but the word is still there. Nobody except perhaps a few linguists would ever think of it as a foreign element. It belongs to the Chinese language just as much as any indigenous word. It is, in other words, completely integrated in the Chinese language.

    This sense of belonging or integration is a matter of degree. It depends on the extent to which the borrowed items are **adapted** to the phonology, writing system and grammar of the language that borrows them.

    Most Japanese, for instance, will know that *bejitarian* 'vegetarian' and *baiorin* 'violin' are English loanwords. Some will even pronounce the words with an English /v/, which is not originally found in the Japanese sound system: *vejitarian* and

vaiorin. The foreign origin of these words is also obvious from their written form, since, like other loanwords, they are written with the katakana script rather than the otherwise more common hiragana script. Even the Portuguese loanword *pan* 'bread', which has been a part of the Japanese language for several centuries, is written with the katakana script to mark it as foreign. In most respects, however, both *bejitarian/vejitarian* and *baiorin/vaiorin* function as integrated parts of the Japanese language, not to speak of *pan*. Nobody feels that he or she is switching to English or Portuguese when using these words.

An example of varying degrees of adaptation to native grammar is the plural of the English word *cactus,* which may keep its original Latin form *cacti,* or may use its regular English form *cactuses*. Another example is the borrowing into Swahili of the Arabic word *kitab* 'book' and the English word *club*. The first syllable, which is pronounced *ki-* in both words, is reinterpreted as a singular prefix, giving the Swahili forms *ki-tabu* 'book' vs. *vi-tabu* 'books' and *ki-labu* 'club' vs. *vi-labu* 'clubs'.

6.2.1 Donor and recipient languages

Borrowing happens between a **donor** language and a **recipient** language. For instance, when the word *orangutan* is borrowed into English from a Malay word literally meaning 'forest man' (*orang* 'man', *utan* 'forest'), Malay is the donor language and English the recipient language.

In principle, any language can be a donor language and a recipient language, and probably all languages play both roles. Very often, however, the donor language enjoys higher social, cultural and/or political prestige than the recipient language. In our days, English is probably the most common donor language of the world, and most, maybe all, languages of Asia and Africa have borrowed extensively from English. When one country is colonized by another, the local language often borrows substantially from the colonial language, such as Vietnamese from French, Wolof from French (in Senegal) or English (in Gambia), Hong Kong Cantonese from English, Tibetan from Chinese etc.

Other common donor languages are the ones that play a central role within great civilizations. Chinese borrowings are particularly numerous in Japanese, Korean, Vietnamese and other East and South-East Asian languages. Arabic borrowings are equally numerous in Urdu, Persian, Turkish, Swahili, Fula, and a large number of other languages of Muslim cultures in Asia and Africa. Languages like Urdu and Turkish have in addition borrowed heavily from Persian.

In such contexts, even dead languages may function as donor languages. Some dead languages, like Classical Chinese and Sanskrit, are only dead in the sense that they no longer function as anybody's mother tongue. They are far from dead in a cultural sense, and they provide a large repository of linguistic elements that modern languages can pick from, in a similar way that Greek and Latin do in Western languages. Modern Chinese (and, to a lesser extent, Japanese) borrows heavily from Classical Chinese, and Modern Hindi (and other languages in the Hindu cultural sphere) borrows heavily from Sanskrit. In both cases, it is not just a case of borrowing of already existant material, but further innovations on the basis of such material. One of the main differences between the two languages Hindi and Urdu is that Hindi borrows from Sanskrit, while Urdu borrows from Arabic and Persian. For instance, one of the Hindi words for 'thank you' is *dhanyavād,* which is a recent borrowing from Sanskrit, while the common Urdu word is *shukriyā*, which is derived from Arabic.

6.2.2   Cultural vs. core borrowings

Borrowed elements that fill a lexical gap in the recipient language are called **cultural borrowings**. They are often introduced along with a new thing and/or a new concept, such as the Japanese word *sushi,* the Chinese concept *paper tiger* and the Swahili word *safari* in many Western languages.

On the other hand, borrowed elements that roughly correspond to elements already existing in the recipient language are called **core borrowings**. Examples are

> **Cultural borrowings are more common than core borrowings.**

bāi-bāi instead of *zài-jiàn* 'bye-bye' in Chinese and *thainkyū* instead of *dhyanavad* 'thank you' in Hindi and *shukriya* 'thank you' in Urdu. Core borrowings often start their lives as foreign elements in code-switching, but are gradually felt to be parts of the indigenous language.

Cultural borrowings are more common than core borrowings.

6.2.3   Phonetic form vs. semantic content

> **Meaning is more easily borrowed than sound.**

As a general rule, semantic content is more easily borrowed than phonetic form, and in some languages, there are strong norms against borrowing phonetic form. Chinese, for instance, used to have a clear preference for giving loanwords a Chinese form, even when the content

was borrowed, cf. *zìxíngchē* (literally 'self-going vehicle') 'bicycle', *qìchē* (literally 'gasoline vehicle') 'car'. The popularity of English, however, has now made phonetic loans like *ōu-kēi* 'OK' and *bāi-bāi* 'bye-bye' much more common than they used to be.

In most of the cases we examined above, both phonetic form and semantic content have been borrowed (though with varying degrees of accuracy) from the donor language. They are what we might call **direct loans**. In other cases, only semantic content is borrowed, while the phonetic form builds on indigenous material. This may be done in several ways:

1. An indigenous form acquires a new meaning in order to translate a foreign concept. For instance, the Chinese word *dìng*, which originally meant 'fixed; stable', was (and is) used by Buddhists to translate the Sanskrit term *samādhi,* referring to a state of meditative absorption. Such examples are called **loanshifts** (also known as **semantic loans**).

2. A foreign composite form is translated directly, element by element. For instance, the Chinese word *lán-méi* 'blueberry' is borrowed from English; it is a compound of the words *lán* 'blue' and *méi* 'berry'. Such examples are called **loan translations** (also known as **calques**).

3. A new composite form based on indigenous elements is created to translate a foreign concept. For instance, the Chinese word for computer is *diàn-nǎo*, which literally means 'electric brain'. Such examples are called **loan creations**.[1]

Sometimes a foreign composite form is rendered with one element being translated into the recipient language and one element retaining the form of the donor language. For instance, Hindi and Urdu both render the English term *double room* as *ḍabal kamrā,* where *ḍabal* is a direct loand from English *double,* while *kamrā* is the Hindi/Urdu word for 'room'. Such examples are called **loanblends** (also known as **hybrids**).

Table 1. What is borrowed in different types of loans?

	form	content	example
direct loan	yes	yes	*sushi* < Jap. *sushi*
loanshift	no	yes	*write* (orig. 'draw') < Lat. *scribere*
loan translation	no	yes	*paper tiger* < Ch. *zhǐ lǎohǔ*
loan creation	no	yes	Ch. *diàn-nǎo,* lit. 'electric brain' < *computer*
loanblend	partly	yes	Hindi/Urdu *ḍabal kamrā* < *double room*

In addition to the general mechanisms of borrowing examined above, languages may do other things to their borrowed items, e.g. **clipping**, as in Japanese *terebi* and *anime* from English *television* and *animation*.

6.2.4   What can be borrowed?

Lexical items are more easily borrowed than grammar and phonology.

Linguists used to think that languages only borrowed *words* from each other. In fact, it turns out that virtually anything can be borrowed: sounds, meanings, word order, derivational affixes, inflections and grammatical categories. It remains true, however, that words and other lexical items are much more easily borrowed than either grammar or phonology.

Open word classes are more easily borrowed than closed word classes.

Among words, open word classes (nouns, verbs, adjectives, adverbs etc.) are more easily borrowed than closed word classes (pronouns, pre- and postpositions, conjunctions etc.). This follows from the fact that closed word classes constitute fixed sets of words that are not easily changed, while open word classes are groups of words that may be supplemented at any time.

Nouns are more easily borrowed than verbs.

Among open word classes, nouns are more easily borrowed than any other word class. In Japanese, for instance, an enormous amount of nouns have been borrowed from Chinese, while almost no verbs are directly borrowed. Instead, if one wants to express a borrowed verbal meaning, one uses a corresponding noun as

---

[1] The Japanese term *denwa* (literally 'electric speech') for 'telephone' is also an example of a loan creation, while the corresponding term in Chinese, *diàn-huà* is a loan translation from Japanese.

an object for the indigenous verb *suru* 'to do': *hatten suru* 'do development; i.e. to develop'. Other languages have fewer restrictions against borrowing verbs than Japanese, but nouns seem universally to be most easily borrowed.

> **Peripheral vocabulary is more easily borrowed than basic vocabulary.**

As we have seen in the previous chapter, one of our main methods for identifying language families builds on the hypothesis that so-called **basic vocabulary** (such as, for instance, terms for body parts) is not easily borrowed. While this may be true for many, perhaps most, languages, it does not hold for all. Languages in Australia, Amazonia and New Guinea are said to borrow even basic vocabulary quite easily. Even English borrowed from Old Norse a number of words usually considered to belong to basic vocabulary: *they, them, their, are, till, skin, leg, sister, get, take, give, die*. And then there is the question of what exactly constitutes basic vocabulary. Based on experience with Indo-European languages, early linguists thought numerals belonged to the basic vocabulary. But both Korean, Japanese, Vietnamese and Thai, none of which are thought to be related to Chinese, use Chinese numerals (though often in addition to indigenous forms). A speaker of Chinese visiting Bangkok airport can often understand which flights are being announced, because the numbers, although read in Thai, still resemble the Chinese numerals so much.

Fixed expressions consisting of many words, sometimes constituting full sentences, may also be borrowed, though usually as loan translations. For instance, English *it goes without saying* has been borrowed from French *ça va sans dire*.

> **Derivational affixes are easily borrowed.**

Derivational affixes (prefixes and suffixes used to derive one word from another, like the English *pre-* of *prepaid* and *-ness* of *weakness*), are often borrowed—the Latin prefix *pre-* being one example of a borrowed derivational prefix in English. When Western thought was introduced on a large scale to East Asia at the end of the 19th and the beginning of the 20th century, a need was felt to create derivational affixes corresponding to English *-ism* and *-ize*, and languages as different as Japanese, Chinese and Vietnamese often adopted the same strategies, usually invented in Japan, but involving affixes of Chinese origin:

	'socialism'	'europeanize'
Japanese	shakai-shugi	ou-ka
Chinese	shèhuì-zhǔyì	ōu-huà
Vietnamese	chu -nghĩa xã-hội	âu-hoá

One fascinating Vietnamese example of borrowed derivational affixes is the word *in-chê-able* 'impeccable', with the French negative prefix *in-*, the Vietnamese verb *chê* 'denigrate' and the French suffix *-able*.

> **Grammar with a wide scope is more easily borrowed than grammar with a narrow scope.**

Grammar may also be borrowed, albeit less easily than words. In general, grammatical strategies with a wide scope (covering sequences of sentences or clauses) are more easily borrowed than grammatical strategies with a narrow scope (within the clause). For this reason, conjunctions (such as words for

'and', 'or', 'but') are quite easily borrowed, since they often have a bearing on two or more clauses, and sometimes over longer stretches of discourse. Throughout the history of Turkish, for instance, words for 'and', 'or' or 'but' have been repeatedly borrowed from Arabic or Persian. Case particles and markers of tense, aspect or modality, are much less easily borrowed, because they usually only have a bearing on the relation between elements within one clause.

This is the reason why word order is more easily borrowed than inflectional categories, so that, for instance, the Karen language of Burma has borrowed SVO structure from the large area of South-East Asian SVO languages surrounding it, although it belongs to the Tibeto-Burman family, which usually has SOV structure.

The sound system may be changed as a result of lexical borrowing.

Finally, the **sound system** of one language may be changed through borrowing from another language. This usually happens through the borrowing of words that carry the new sounds with them, as when Japanese introduces the voiced fricative [v] in English loanwords like *vejitarian* and *vaiorin*. Tagalog, the main language of the Philippines, used to have only three vowel phonemes /i/, /u/ and /a/, but through Spanish loanwords like *rehiyon* 'region' (from Spanish *región*), /e/ and /o/ have been added to the inventory of phonemes. In many languages, such new sounds are in constant danger of disappearing again. In Japanese, for instance, only those proficient in English are likely to say *vejitarian* and *vaiorin* rather than *bejitarian* and *baiorin*. The South Indian language Telugu, which has borrowed aspirated stops (/b^h/, /p^h/, /d^h/, /t^h/, and so on) from the Indo-Aryan languages of North India, only uses them in higher style; in the colloquial language they are merged with the indigenous unaspirated stops (/b/, /p/, /d/, /t/, and so on).

Suprasegmental features like tones and nasalization seem to be more easily borrowed than segmental phonemes. In large parts of South-East Asia, for instance, the tone system has been developed in much the same way, probably as a result of influence from Chinese.

6.3 Code-switching

In **code-switching**, the speaker (or speakers) switches between two (or more) different language codes during a single stretch of discourse. Take the following example of code-switching between Swahili and English:

(1) Code-switching between Swahili and English (bold-face)

Lakini	ni-ko	**sure**	u-ki-end-a	**after two days**	u-ta-i-pat-a
but	1SG-COP	sure	2SG-COND-go-FV	after two days	2SG-FUT-DO-get-FV

Uchimi Supermarket kwa wingi
Uchimi Supermarket in   abundance
'But I am sure if you go after two days you will find it [=Omo detergent] [at] Uchimi Supermarket in abundance.'

If we look away from the proper noun *Uchimi Supermarket*, the speaker switches from Swahili to English and back again twice during this single sentence.

Note that neither *sure* nor *after two days* are borrowings from English into Swahili. They do not belong to the Swahili language code at all. When we discussed

borrowing above, we were talking of the integration of elements from a donor language into a recipient language. In code switching, there is no such integration. Both languages are "on" at the same time. Code switching, therefore, presupposes a much higher degree of bilingualism than borrowing does. A speaker of Swahili can use English loanwords like *soka* 'soccer; football' without knowing any English himself, because these words have already been integrated into Swahili. To participate in code-switching between Swahili and English, however, he needs to have at least a basic knowledge of both languages.

Although code-switching implies the use of two languages within the same stretch of discourse, the two languages do not participate on equal terms. At any given time, only one of the languages is conceived of as "the language we are speaking", the **primary language**, while the other language is the **secondary language**. In the example above, Swahili is clearly the primary language, with two instances of code-switching into English, which is the secondary language. There is more Swahili than English in the sentence, and the sentence as a whole is dominated by Swahili grammar. This may change during one stretch of discourse, but such changes are not common. Very often, the speaker is more proficient in the primary language than in the secondary language, but this does not have to be so, particularly since the choice of primary language and degree of code-switching is often decided by communal norms rather than just individual choice.

In most cases of code-switching, all grammatical elements belong to the primary language, except when entire phrases or clauses in the secondary language are inserted. In phrases or clauses that contain elements from both the primary language and the secondary language usually do not include grammatical elements from the secondary language. As in the case of borrowing (like in the Japanese example *hatten suru* 'do development' in the previous section), an auxiliary verb of the primary language often carries all inflectional markers when an secondary language verb is used. Consider, for instance, the following case of code-switching between Turkish and Norwegian:

(2) Code-switching between Turkish and Norwegian (bold-face)

Adam-lar	yer-i	**vask-e**	yap-ıyor ('The men are washing the floor.')
man-PL	floor-ACC	wash-INF	do-PROG/3SG

A Norwegian verb in the infinitive (which is the closest one gets to an uninflected verb in Norwegian) is accompanied by the progressive 3rd person singular form of the Turkish auxiliary verb meaning 'to do'. As a result, all necessary grammatical markers may be carried by the primary language.

In **intra-sentential** code-switching, one sentence may contain elements from two or more languages. In most cases - 84 percent by one count - only single words from the secondary language are inserted, such as *sure* in example (1) and *vaske* in example (2). In 10 percent, phrases from the secondary language are inserted, such as *after two days* in example (1). And in 6 percent, entire clauses are inserted, such as the following example of code-switching between German and English:

(3) Code-switching between English and German (bold-face)

Papa, wenn du das Licht ausmachst, then I'll be so lonely. ('Daddy, if you put out
    the light, then I'll be so lonely.')

In addition comes **intersentential** code-switching, examples of which we shall see in example (4) below.

What is the motivation for code-switching? Why would speakers of one language, such as Swahili in example (1), include in their discourse stretches spoken in another language, such as English? There are at least five possible reasons:

First, they might choose to use an English word or expression because it has no exact counterpart in Swahili. The use of English helps to fill a **lexical gap**. None of the two cases in example (1) seems to be of this type. Instead of *sure,* for instance, the speaker might have chosen to use the Swahili word *hakika*.

Second, they might feel that even if Swahili has a corresponding word or expression, it does not express the **pragmatic force** or the "feeling" intended as well as the English word or expression. In example (1), for instance, the speaker might feel that *hakika* would sound less convincing than *sure*.

Third, the use of language is always **emblematic**. The use of both Swahili and English in the same stretch of discourse is an emblem of dual membership of language communities. The speaker signals a mixed social, cultural and/or political identity and solidarity, one represented by Swahili and the other by English. Here as in many other cases, the primary language signals **in-group** membership and expresses solidarity. On the other hand, the secondary language signals an auxiliary **out-group** membership in another language community, which might have, as is often the case with English, greater social, cultural and/or political prestige. It is as if the speaker is conveying a certain pride in being a Swahili-speaker and at the same time an even greater pride in being more than just a common Swahili-speaker.

Fourth, a change of language code may also signal changes of **attitude** during the conversation. The conversation below takes place in Western Kenya and involves a local farmer who speaks Lwidakho and maybe a little Swahili and a local person now employed in an urban centre outside the area in a rural bar. Until this point, the conversation has been in Lwidakho, but then the worker begins to use Swahili (marked with italics) and English (marked with bold-face) as well:

(4) Code-switching between Lwidakho, Swahili (italics) and English (bold-face)

Farmer: Khu inzi khuli menyi hanu inzala- ('As I live here, I have hunger-')
Worker: (interrupting) *Njaa gani?* ('What kind of hunger?')
Farmer: Yenya khunzirila hanu- ('It wants to kill me here-')
Worker: (interrupting again, with more force) *Njaa gani?* ('What kind of hunger?')
Farmer: Vana veru- ('Our children-' [said as appeal to others as brothers])
Worker: *Nakuuliza, njaa gani?* ('I ask you, what kind of hunger?')
Farmer: Inzala ya mapesa, kambuli. ('Hunger for money; I don't have any.')
Worker: **You have got a land.**
　　　　Una shamba. ('You have land [farm].')
　　　　Uli nu mulimi. ('You have land [farm].')
Farmer: ... mwana mweru- ('... my brother-')
Worker: ... mbula tsisendi. ('I don't have money.')
　　　　Can't you see how I am heavily loaded?

By switching from the in-group language Lwidakho to the out-group languages Swahili and English, the worker conveys his negative attitude to the farmer's appeal for solidarity. He also makes a move in the opposite direction, accomodating the farmer by repeating the English-language statement "You have got land" first in Swahili and then in the farmer's preferred language Lwidakho. In the end, however,

he maintains his refusal and closes the matter with an English sentence that the farmer is not likely to understand at all.

Fifth, switching into the secondary language is often used as a **discourse marker**, as a means to mark significant boundaries in the conversation. It may mark emphasis, as is possibly true of the use of *sure* instead of *hakika* in example (1), and especially of the three different versions, one in each language, of the sentence "You have got land" in example (4). It may mark a change of topic, as in the following case of code-switching between English (primary language) and Hebrew (secondary language):

(5) Code-switching between English and Hebrew (bold-face)

Grace: .... no but you know ... see I think, ... that the inability to compliment, ... not all men have that inability, ... but many .. men .. have that inability, ... to compliment, -

Shira: (interrupting) **at rotsa smartut?** ('You want a rag?')

Grace: ... right there, ... I'll get it, ... I'll get it,

Shira: no ... here you go.

Code-switching may also mark a switch from the ordinary informational level of language to meta-level elements like the English expressions *because* and *you know* and the Wolof expression *xam nga* 'my friend' in the following example from Gambia, in which Mandinka is the primary language:

(6) Code-switching between Mandinka and Wolof (italics) and English (bold-face)

a ka itolu fanaŋ nafaa wo le ñaama- ('it is also to their advantage')

because a si a tara - **you know** - **certain things**-oo-lu be jee itolu fanaŋ maŋ meŋ-lu je ì yaa- ('because - you know - there are certain things there which they haven't got in their countries')

xam nga - niŋ ì naa-ta jaŋ *rek*- ('you know - now if they only come here')

niŋ ì taa-ta ǹ la ñiŋ wuloo-lu kono- ('if they go into our woods')

ì si a tara *suma waay*- ('they will my friend')

you know - ì si wo-lu doo-lu je- ('you know - they will see many of them')

you know - a te ke-la ì bulu baa - **you know** ('you know - it doesn't exist at their place - you know')

Such meta-level elements include *look, listen, believe me, you know, (you) see, say, tell me, let's put it this way, I mean, yeah, right, sure, okay, no, well, now, since, because, so, but, and, also, first of all, by the way, oh, uh, uhm* and many others.

6.4 Language shift and linguistic interference

In 1644, the last emperor of the Ming dynasty in China hanged himself after Manchu invaders had entered Beijing, and the Manchus subsequently ruled China until the fall of the empire in 1911. By then, however, virtually all Manchus spoke Chinese. Far from protecting their linguistic heritage, political power had made it all the more important for the Manchus to learn Chinese, and their loss of power eventually also led to the loss of their own language. Today the Manchu language is all but extinct.

This is an example of **language shift**, the process by which the language of one community is replaced by the language of other communities. As long as the

community is still bilingual in the old language and the new language, as was still the case with the Manchus in 1911, the shift is partial. By the time the new language has fully replaced the old language, the shift is complete. This is probably the case with the shift from Manchu to Chinese. Manchu has become, or is soon becoming, an extinct language. Language shift often leads to **language death**.

Both language shift and language death have been common phenomena throughout history. For instance, for the past several thousand years the linguistic descendants of the tribes that spoke the Indo-European protolanguage have spread to most of Europe, large parts of Central and South Asia and, more recently, to virtually all parts of the Americas, Australia and New Zealand, as well as some parts of Africa. The original languages of these areas have either retreated or disappeared entirely.

6.4.1 Interference

The languages that retreat or disappear, however, often leave behind a very noticeable trace in the languages that take over. The languages that take over, the **superstrate** languages, are subject to **interference** from the languages that retreat or disappear, the **substrate** languages.

This is probably the explanation why the Indo-Aryan languages of North India have so many features in common with the unrelated Dravidian languages of South India. Dravidian languages used to be spoken across North India as well (and there are still a few Dravidian-speaking pockets in the north), and when Indo-Aryan languages took over, they picked up many Dravidian traits, such as the use of retroflex stops, which is so characteristic of virtually all Indian languages today.

Interference from substrate languages probably also explains why northern dialects of Chinese have more features in common with Altaic languages, while southern dialects have more features in common with Austroasiatic and Tai-Kadai languages. For instance, words tend to be longer in the north than in the south, tonemes tend to be more numerous and complex in the south than in the north, where stress and accent play a more significant role than in the south. Chinese has spread from its place of origin along the Yellow River, and has replaced other languages originally spoken both in the north and in the south. In the process, however, the Chinese language itself was changed.

The spread of Arabic across a large area in Central Asia, West Asia and North Africa has also resulted in a large number of widely different modern dialects. Some of the dialectal differences are the results of interference from the original languages spoken in the areas. Maghrebi Arabic has incorporated elements from Berber, Syrian and Iraqi Arabic from Arameic, and Sudanese Arabic from Nubian.

In modern days, some of the most significant cases of language shift involve the spread of English. In places with a history of British colonization, such as India and Singapore, more and more children grow up with English as their first language. The shift is only partial, and most of these children are bilingual (or multilingual) in English and one or more indigenous language(s). In these countries, substrate interference has created new and characteristic variants of English, known as Indian English and Singapore English. In India, there are actually a number of different variants of English, depending both on the nature of the substrate language and on the degree of interference. In areas where Hindi is the substrate language, it is common to distinguish between Indian English proper and "Hinglish", the latter having a much higher degree of interference from Hindi lexicon and grammar. A similar distinction is sometimes made between Singapore English proper and "Singlish", the latter

10

having a much higher degree of interference from various Chinese dialects and Malay.

At the outset, interference from substrate languages is a result of incomplete acquisition of the superstrate language. When a speech community begins to shift from one language to another, the first generation meets with all the well-known problems of **second language acquisition**. Most notably, their native language will influence the way they speak the acquired language. The next generation typically grows up as bilingual in the substrate language and something close to the first generation's variant of the superstrate language. If, after some generations, the language shift is complete, the substrate language disappears entirely, but the superstrate language retains features that used to belong to the substrate language.

6.4.2   Lexicon

Vocabulary is often not strongly affected by substrate interference. This is not to say that words from a substrate language are not often borrowed into the superstrate language, but such borrowings are hardly more common than borrowings from other influential languages. They also tend to follow the general trends for borrowing: cultural borrowings are more common than core borrowings, and meaning is more easily borrowed than sound.

The satirical poem "The Prophet", written by the Bombay-based Jewish author Nissim Ezekiel, is written in an exaggerated version of Indian English. In spite of all the exaggerations, however, the only direct loanwords from Hindi occurring in the text are *goonda* 'hooligan' and the cultural borrowing *lassi*, referring to a yoghurt drink. Although this poem is humoristic and cannot be treated as a reliable source of Indian English, it does express pretty well the gist of the language.

The Patriot

I am standing for peace and non-violence.
Why world is fighting fighting
Why all people of world
Are not following Mahatma Gandhi,
I am simply not understanding.
Ancient Indian Wisdom is 100% correct,
I should say even 200% correct,
But modern generation is neglecting-
Too much going for fashion and foreign thing.

Other day I'm reading newspaper
(Every day I'm reading Times of India
To improve my English Language)
How one goonda fellow
Threw stone at Indirabehn.
Must be student unrest fellow, I am thinking.
Friends, Romans, Countrymen, I am saying (to myself)
Lend me the ears.
Everything is coming -
Regeneration, Remuneration, Contraception.
Be patiently, brothers and sisters.

You want one glass lassi?
Very good for digestion.
With little salt, lovely drink,
Better than wine;
Not that I am ever tasting the wine.
I'm the total teetotaller, completely total,
But I say
Wine is for the drunkards only.

What you think of prospects of world peace?
Pakistan behaving like this,
China behaving like that,
It is making me really sad, I am telling you.
Really, most harassing me.
All men are brothers, no?
In India also
Gujaratis, Maharashtrians, Hindiwallahs
All brothers -
Though some are having funny habits.
Still, you tolerate me,
I tolerate you,
One day Ram Rajya is surely coming.

You are going?
But you will visit again
Any time, any day,
I am not believing in ceremony
Always I am enjoying your company.

-- Nissim Ezekiel

In addition to the direct loans come a few specifically Indian meaning variants (possibly loanshifts), like *too much* meaning 'a lot of' and *harass* meaning 'bother' (with an impersonal subject). Another writer, Pico Iyer, relates how, in the *Times of India,* he found a section devoted to "matrimonial notices" (the Indian variant of contact ads), "in which prospective brides were glowingly described as 'homely' and 'artful' and 'wheat-coloured' (which, in the logic of Indian English, means domestically minded, culturally inclined and fair-skinned)". Iyer concludes that "the [British] empire never left, it just settled down in a back street in Madras, and started to tell its story from the other side". [2]

Examples of borrowings in Singapore English include *makan* 'to eat; a meal' (from Malay), *ang mo* 'westerner' (from the Hokkien Chinese, literally 'redhair'), and the loanshift *blur* 'confused; ignorant'. Few such loanwords enter into written Singapore English, one exception being *kiasu* 'afraid of losing out; mildly exploitative' (from Hokkien). This word is even used in the loanblend *kiasuism*, which is sometimes mentioned humorously as a national characteristic of Singaporeans.

Above we noted that code-switching often involves the use of discourse markers. The same is true of substrate interference. Singapore English has picked up from Hokkien and Cantonese a number of (usually sentence-final) particles used as discourse markers, such as *ah* (often implying an expectation of agreement), *lah* (expressing strong assertion) and *what* (implying a correction of some earlier statement):

(7) Sentence-final particles in Singapore English

a) There's something here for everyone *lah*.
b) Otherwise, how can be considered Singaporean *ah*?
c) No parking lots here, *what*.

Such discourse markers are exceedingly difficult to translate, but are felt by the native speaker (of Hokkien or Cantonese) to be indispensable elements of their language and are therefore carried over into their version of English.

6.4.3   Grammar

Substrate influence on grammar is quite common. It is often frowned upon as "wrong", but eventually becomes a regular part of the new language variant.

"The Patriot" contains a number of Indian English examples: the reduplication of verbs in "fighting fighting", the lack of indefinite and definite articles in "threw stone at Indirabehn" and "all people of world", the use of *one* instead of the indefinite article *a* in "one glass lassi", the excessive use of the definite article *the* in "not that I am ever tasting the wine", the excessive use of the continuous tense (the *-ing* form) in "I am standing for peace and non-violence", the omission of an obligatory object pronoun in "modern generation is neglecting" (instead of "neglecting this") etc. Again, this is a satirical poem rather than a genuine example of Indian English, but the phenomena exemplified are genuine enough.

In Singapore English, marking of number and tense is optional, reflecting the lack of inflection in all substrate languages (various Chinese dialects and Malay):

---

[2] *Prospect* October 2003 p. 48.

(8) Lack of inflection in Singapore English

a) Got so many car!

b) What happen yesterday?

c) You know what happen lah? Fine! ('You know what happened? I got fined.')

As can be seen from the very short sentence "Fine!", subject pronouns are more easily omitted in Singapore English than in standard English, reflecting the same tendency in Chinese. As in all the substrate languages, adjectives are usually not preceded by the verb *to be*, as in the film title "I Not Stupid".

Substrate influence on grammar often takes the form of **relexification**, the use of a lexical item from the superstrate language to express a meaning associated with the substrate language. An example from Singapore English is the use of the verb *give* to express passive:

(9) Chinese-style passive marking in Singapore English

John always give his boss scold.

'John is always scolded by his boss.'

Verbs meaning 'to give' are regularly used as passive markers in many Chinese dialects.

6.4.4   Sounds

Substrate influence on phonology is also quite common. Like the influence on grammar, it is sometimes frowned upon as "wrong", but also becomes a regular part of the new language variant.

When the quality of sound segments is changed due to substrate influence, this most often reduces the number of distinctions in the sound system. For instance, Indian English almost consistently pronounces both [v] and [w] as [ʋ]. The two phonemes /v/ and /w/ of standard English are reduced to one phoneme /ʋ/ in Indian English, and the words *vet* and *wet* are no longer distinguished phonologically. Both the actual pronunciation and the underlying system are affected.

Singapore English has gone through a number of substrate-induced sound changes with wide-ranging consequences for the underlying sound system, producing a large amount of new homophones:

1. As in Malay and all Chinese dialects except Cantonese, Singapore English tends to make no distinction between long and short vowels, so that the pairs *pull* and *pool, come* and *calm, sit* and *seat* are often pronounced the same.

2. All final fricatives tend to be unvoiced, so that the pairs *edge* and *etch, rise* and *rice, leave* and *leaf* are pronounced the same. (Chinese allows no final fricatives at all, while Malay allows voiceless fricatives.)

3. There is also a certain tendency for final plosives to be unvoiced, so that the pairs *hob* and *hop*, *bid* and *bit, bag* and *back* are pronounced the same.[3] (Both Malay and the relevant dialects of Chinese allow only unvoiced plosives in final position.)

---

[3] Final plosives in Singapore English also differ from standard English in being unreleased, just like the final plosives of Malay and Chinese, but this does not affect the underlying sound system. Furthermore,

14

4. The vowels in *bad* and *bed* both tend to be pronounced as [ɛ].

The overall result is a tremendous reduction in the number of phonological distinctions in the sound system.

There are, however, also substrate-induced sound changes that do not affect the underlying system, only the actual pronunciation of each sound segment. Consider, for instance, the following "conversion table" from standard British English to Indian English:

[θ] → [t̪ʰ] (as in *think*)
[ð] → [d̪] (as in *thus*)
[t] → [ʈ] (as in *tin*)
[d] → [ɖ] (as in *do*)

Neither the dental plosives [t̪] and [d̪] nor the retroflex plosives [ʈ] and [ɖ] are found in standard English. In this case, however, the number of phonological distinctions made in Indian English and standard British English is the same.

Substrate interference affects not only sound segments, but also the rules for combining these segments into syllables and words (phonotactic rules). For instance, both Indian English and especially Singapore English tend to simplify consonant clusters, especially in final position. Thus, Singapore English pronounces *tact* as [tɛk] or [tɛʔ] and *lift* as [lif].

Substrate interference also affects suprasegmental phenomena like stress and accent. Both Indian English and Singapore English tend to make less use of the distinction between stressed and unstressed syllables than British English. Each syllable tends to be allotted approximately the same amount of time in the speech flow, resulting in what has been described as "machine-gun style".

6.5   Language generation

Language contact may generate new languages, combining elements of the two or more languages in contact as well as adding elements that do not originate in either language. There are two main types of such new languages: **pidgins** and **creoles**.

6.5.1   Pidgins

Pidgins are not full languages. They are used for limited communication between speakers of two or more languages who have repeated or extended contacts with each other, for instance, through trade, enslavement, or migration. Pidgins usually combine elements of the native languages of their users and are typically simpler than those native languages insofar as they have fewer words, less inflection, and a more restricted range of phonological and grammatical options.

As in the case of language shift, many pidgins are based on an asymmetrical relation between their source languages. The socially dominant language is the

---

final *t* and sometimes *d* may be pronounced as a glottal stop [ʔ], reflecting the frequent occurrence of final [ʔ] in all the Chinese dialects of relevance.

superstrate language, which usually provides most of the words and other lexical elements, while the socially subordinate language is the **substrate language**, which tends to have a greater impact on sounds, grammar and meaning. Consider the following example from the Melanesian Pidgin English called Tok Pisin spoken in Papua New Guinea:

(10) Tok Pisin

Nau	wan-fela	master	em	i-kisim mi	...	nau	ol	master	i-kik,	i-kikim	em
then	one-CL	white-man	he	PM-get me		then	PL	white-man	PM-kick,	PM-kick	him

'Then a white man took me ... then the white men were kicking, they were kicking it.'

The words come from English, even the classifier *fela* (from English *fellow*), the plural marker *ol* (from English *all*) and the transitive marker *-im* (from English *him*). The grammatical structure, however, is that of an Eastern Oceanic language, with the use of classifiers, predicate markers, transitive markers, and a preposed plural marker. The meaning of the words has also been largely changed to accomodate local needs, as has the pronunciation.

For a pidgin, Tok Pisin is a fairly complex language. How simple a pidgin can be may be illustrated by Russenorsk, a pidgin used by Norwegians and Russians for trade purposes in the 19th century. This pidgin had only 150-200 words, no inflection, no verb meaning 'to be', no explicit subordination of clauses, and one preposition *på* (originally Norwegian for 'on; at') covers a wide range of meanings: *på moja stova* 'at my house', *sprek på moja* 'say to me', *på Arkangel* 'to Archangel'.

Lack of inflection is not restricted to pidgins, but is also found in many West African and East and South-East Asian languages. The kind of simplicity displayed by pidgins, however, goes far beyond anything found in full languages, especially evident in its limited number of words and lack of semantic and grammatical distinctions.

On the other hand, a pidgin may be gradually expanded until its linguistic resources resemble those of a full language, thus becoming an **expanded pidgin**. Such linguistic expansion is usually the result of extension in the social and/or geographical role of the pidgin. For instance, Chinese Pidgin English in Singapore only acquired a degree of fluency after it began to be used between speakers of mutually incomprehensible dialects of Chinese, rather than between the Chinese and the British. If such social and/or geographical extension does not occur, the pidgin will eventually die when its *raison d'être* ceases to exist. This is indeed what happened to Russenorsk.

6.5.2   Creoles

Like pidgins, **creoles** are the results of intensive contact between two or more languages, and combine elements of these languages. Unlike pidgins, however, creoles are full languages, which children grow up learning as their first language. Creoles typically have a larger vocabulary and more complex grammatical resources than pidgins.

However, the distinction between pidgin and creole is gradual rather than absolute. Some expanded pidgins are structurally indistinguishable from creoles, and even before they have become the native language of any of its speakers, they may serve as their primary language. This is what happened with Tok Pisin in New Guinea and Sango in the Central African Republic.

Creoles sometimes come into being as a result of the **creolization** of a pidgin, which is what happens when a pidgin acquires native speakers. When this happens to expanded pidgins, it does not necessarily imply fundamental changes in the language structure, since expanded pidgins are already close to being full languages. Less commonly, when creolization happens to simpler pidgins with a more rudimentary grammar and lexicon, both structure and vocabulary must be expanded in order for the new creole to function as a full language. Such **abrupt creolization** typically involves a number of new features:

1. Movement rules, such as the marking of topics by placing them sentence-initially.
2. A system of articles, as in *wan buk* 'a book' and *di buk* 'the book'.
3. Tense, aspect and mood particles preceding the verb.
4. Relative clauses and other embedded subordinate clauses.

On the other hand, creoles do not always grow out of established pidgins. Sometimes, the language contact situation simply produces a large variety of individual **interlanguages**, with no communal norms. Creoles emerge when one of the languages in contact, usually the one with the higher prestige, gradually includes elements from such interlanguages.

It is often asserted that creoles tend to have a number of features in common regardless of the structure of the source languages from which they are built. The four new features resulting from abrupt creolization are examples of such alleged common traits. If this is true, how can we account for these similarities?

Some linguists have proposed a hypothesis that all the world's pidgins and creoles are ultimately derived from the same source, say, a Portuguese pidign that came into being in the 15th and 16th centuries. According to this **monogenetic** theory, the rise of pidgin happened once! Unfortunately, this does not seem to correspond to the facts. A **polygenetic** theory of the origin of pidgin seems much more realistic. Another suggestion is that pidgins and then creoles emerge as a kind of "baby talk" or "foreigner talk". We know that both adults talking to children and people talking to foreigners tend to simplify their language in very similar ways in most parts of the world. Some of these simplification strategies resemble those of pidgins. Could it be that pidgins and creoles are the results of similar strategies of simplification? Do they represent attempts by the speakers of superstrate languages to facilitate the understanding and acquisition by substrate speakers?

The problem with this answer, however, is that it is in fact most often substrate speakers who create and use pidgin and creole, while superstrate speakers often even have problems understanding them. It makes more sense, therefore, to look upon pidgins and creoles as a kind of limited attempt to acquire a second language by the substrate speaker. Again, many of the features found in pidgins and creoles are also found in the speech of those who try to acquire a second language. In contrast to the second language learner, however, the substrate speaker may not really attempt to acquire a target language, just to find a medium for communication with other language groups.

A more radical explanation of the similarities between different creoles is that creoles are the inventions of the first generations of children who use them as native languages, and that these children are biologically "programmed" to expand the original pidgins in specific ways. This would explain why features like the ones mentioned above appear even in creoles based on pidgins and source languages that

lack these traits. It would also explain why child language and creole have certain features in common.

However, it is still a matter of controversy to what extent creoles actually do share features that cannot be explained on the basis of their source languages. The phenomenon of language interference discussed above may go a long way in explaining the nature of creoles. Some linguists have suggested that Indian English and Singapore English are creoles, while Hinglish and Singlish are pidgins.

-

-          -

1   Introduction

The term *diglossia* is a fundamental sociolinguistic concept that is used recurrently in the literature with, at times, varying definitions. In most cases, diglossia describes a situation where two linguistic systems coexist in a functional distribution within the same speech community. One system is assigned the status of high variety (H), while the other receives the status of low variety L). The H variety is used in more formal domains while the L variety is typically limited to oral informal communication. Since the mid-twentieth century, the term, nevertheless, has been applied, at times with major conceptual modification, to a wide array of situations in different parts of the world. Some of these situations included contact between varieties of the same language, languages belonging to the same family, or distant languages. The aim of this book is to refocus the concept of diglossia and situate it within current theories of language contact. My central argument is that diglossia, as a theoretical framework, has not been appropriately discussed in the field of contact linguistics, partly because it has been extended to describe cases that in reality bear very few similarities to classical diglossic cases where genetic relatedness is a key element.

In this introductory chapter, I outline the development stages of the concept of diglossia and lay the theoretical ground for the analyses that follow in the subsequent chapters. I argue that classical diglossia, understood as a situation that involves closely related language varieties, is a more useful concept for studies on language contact than the notion of extended diglossia that describes any pair of languages, or more, which are in complementary distribution regardless of their genetic relatedness.

1.1   Defining diglossia

Taken in its original literal meaning, diglossia describes the coexistence of two words, i.e., vocabulary doublets, in a given language to refer to the same concept or entity. It was first used to refer to the Greek language situation, one of the defining cases mentioned by Ferguson (1959a) in his seminal article discussed in detail below. In Greece, and up to 1976, two varieties of Greek were in use. Demotic was the spoken vernacular that evolved over time from classical Greek.

1

It was naturally acquired and widely used by all native speakers of Greek. Katharevousa, on the other hand, was artificially developed in the nineteenth century from classical Greek, without consideration of the vernacular features then present in Demotic. Katharevousa was declared the official language in 1834 and became the language of education and administration until Demotic, which had gone through a process of standardization, was recognized as the new national language, officially ending the diglossic situation in 1976 (Alexiou 1982). Although nowadays, the situation of Arabic is often recognized as the embodiment of diglossia, in fact it was the Greek situation that prompted the adoption of the concept in modern linguistics. Toufexis (2008: 207) indicates that Karl Krumbacher (1902) had already used the term diglossia to refer to this situation around the turn of the twentieth century.

One of the major voices in favor of Demotic Greek was Jean Psichari, who, in his article "Un pays qui ne veut pas de sa langue" ("A country that does not want its language"), also used the term diglossia to describe the debate surrounding the Greek language question at the time:

La diglossie – le fait pour la Grèce d'avoir deux langues – ne consiste pas seulement dans l'usage d'un double vocabulaire, qui veut qu'on appelle le *pain* de deux noms différents: *artos*, quand on est un homme instruit, *psomi*, quand on est peuple; la diglossie porte sur le système grammatical tout entier. Il y a deux façons de décliner, deux façons de conjuguer, deux façons de prononcer; en un mot, il y a deux langues, la langue parlée et la langue écrite, comme qui dirait l'arabe vulgaire et l'arabe littéral (Psichari 1928: 66)[1]

In this early definition, Psichari separates the term from its literal meaning of vocabulary duality and extends it to describe the existence of two linguistic systems that diverge considerably, although they still belong to the same historical language. His inclusion of all levels of structure as part of a diglossic situation and the claim that we are dealing with two nearly separate languages would remain at the center of subsequent studies on diglossia. This first extension of the concept of diglossia would be paralleled later by another extension to practically any situation where two languages are in a functional distribution, regardless of the degree of their genetic connectedness (Fishman 1967, 1980). Initially, however, Psichari, with the Demotic/ Katharevousa situation in the foreground, perceived diglossia as the end result of the creation and maintenance of an artificial system, Katharevousa in this case, imposed by the literate elite on the rest of the population. His line of reasoning is that there

---

[1] "Diglossia – the fact that Greece has two languages – doesn't consist in the mere use of vocabulary doublets, which means that *bread* is called by two different names: *artos*, when you are an educated man, *psomi* when one belongs in lower class; diglossia concerns the grammatical system as a whole. There are two ways to decline, two ways to conjugate, two ways to pronounce; in a word, there are two languages, the spoken language and the written language, say vulgar Arabic and written Arabic" (my translation).

was a vernacular form in use by all members of the community but the formulation of a norm, which is heavily inspired in the classical language and written sources and which is not the native language of any member of the speech community, created the diglossic situation between the standardized variety and the vernacular. This process is guided by what he calls "contempt for the spoken language," on the one hand, and "fetishism of the classical language" on the other (Psichari 1928: 72 and 76). The issue, then, is with writing the vernacular, not necessarily with its oral use in ordinary communication (see Chapter 7 for the role of writing in ending diglossia in the case of the Romance languages). After all, the vernacular is the native language of the entire society and it would be difficult to forcefully prevent its use (Psichari 1928: 118).

This is a key observation since, as I will argue throughout this book, a paradox is observed in the attitude and behavior of speakers in a diglossic situation: they degrade the very same language they faithfully transmit to their children, thus perpetuating the situation of which they consciously disapprove. Both the illiterate speakers, often the ones who are deprived of any possible gains associated with knowing the standard form, and the educated speakers share a negative perception of the vernacular. This is what I propose to call here the *diglossia paradox* and what I believe lies beneath the emergence and perpetuation of diglossic communities in cases where both varieties belong to the same historical language.[2] This paradox also explains the rapid changes in the vernacular and the static, and at times utterly cumbersome, structure of the standardized form. The resistance to include loanwords, naturally incorporated into the vernacular, is one feature that results from this paradox and may lead, for example, to forced morphological creations in the standard variety that are opaque and, at times, do not follow the morphological rules of the classical or the vernacular forms of the language. As an example, Psichari discussed the case of the Italian word *bomba*, which was adapted as *vomva* in Katharevousa Greek, given that the sound /b/ does not exist in classical Greek, leading to an impossible combination in this language of /mv/.

The application of the concept of diglossia to the Arabic situation, specifically in the context of North Africa, was first attempted by William Marçais (1930) in his article "La diglossie arabe" ("Arabic diglossia"). Marçais, who spent several years in Algeria and Tunisia around the turn of the twentieth century, established that the main characteristic of the language situation in the region was the existence of two varieties of Arabic playing against each other, the written form and the spoken vernacular. While the written form had not changed much since it was codified early on in the Islamic period, the dialects had gone through a considerable number of changes and, at the time of Marçais,

---

[2] The situation is different in cases of nongenetic diglossia, given that the indigenous language and the superposed one both have native speakers, even if they do not share the immediate same space.

they were still not satisfactorily described. As a result, he argued that it was challenging for non-native speakers to learn the dialects, given the great degree of divergence existing between the different North African cities where he had lived. Yet, he knew that the written form was not of much use to the largest sector of the population, who, under the colonial rule, were deprived of access to education altogether.

Although he himself did not express a preference regarding which variety should be taught, Marçais presented diglossia as an irreconcilable situation, comparable to a two-headed monster. He argued that this situation was behind the difficulties faced by the colonial schooling programs with which he was very familiar in his positions as Inspecteur général de l'enseignement des indigènes and Directeur de l'école supérieure de la langue et littérature arabes de Tunis. He fervently summarized the situation and his position as follows:

Disons deux états d'une même langue, assez différents pour que la connaissance de l'un n'implique pas, absolument pas, la connaissance de l'autre; assez semblable pour que la connaissance de l'un facilite considérablement l'acquisition de l'autre. En tout état, un instrument pour l'expression de la pensée qui choque étrangement les habitudes d'esprit occidentales; une sorte d'animal a deux têtes et quelles têtes! Que les programmes scolaires ne savent trop comment traiter, car ils ne sont pas faits pour héberger les monstres. On conclura peut-être que le premier responsable de la crise de l'arabe, c'est l'arabe (Marçais 1930: 409).[3]

The anguish that Marçais seems to experience over this situation stems largely from the belief, in his "Western mind," of what constitutes a favorable situation for education: a monolingual monodialectal population. This is reminiscent of the opinions by early observers of attempts to educate bilingual and bidialectal children in other societies, including native American, African American, and Hispanic students in the United States in the first half of the twentieth century. Whether or not the opinion of directly involved parties, such as Marçais, who for a period oversaw *l'enseignement des indigènes* and was fluent in Arabic, added to the marginalization of Arabic instruction under the French rule, the result was that the illiteracy engulfed French North Africa during the colonial rule.

The position that a diglossic situation is unnatural or is not real continued to be at the center of the debate between advocates of a stronger Arabization and advocates of the standardization of the vernacular, although with much less graphic nature. As the French-educated elite accessed the positions of power

---

[3] "Let's say there are two states of the same language, different enough that knowing one does not, absolutely not, imply knowledge of the other; yet so similar that knowledge of one greatly facilitates the acquisition of the other. In any state, an instrument for the expression of thought that strangely shocks the habits of the Western mind; a kind of a two-headed animal – and what heads! – which school programs do not quite know how to treat because they are not made to accommodate monsters. One might conclude that perhaps the one primarily responsible for the crisis of Arabic is Arabic" (my translation).

immediately after independence of the Maghreb, they felt French was a more appropriate language for education and avoided dealing with the diglossic situation altogether, at least initially. A further discussion of the situation of Arabic and diglossia in the Maghreb is presented in Chapter 3.

A full-fledged theory of diglossia was spelled out thirty years later in the seminal work of Charles Ferguson. In his article, "Diglossia," Ferguson (1959a: 336), who was particularly familiar with the situation of Arabic and had carried out extensive research on Arabic dialects, formulated the following definition:

> DIGLOSSIA *is a relatively stable language situation in which, in addition to the primary dialects of the language (which may include a standard or regional standards), there is a very divergent, highly codified (often grammatically more complex) superposed variety, the vehicle of a large and respected body of written literature, either of an earlier period or in another speech community, which is learned largely by formal education and is used for most written and formal spoken purposes but is not used by any sector of the community for ordinary conversation.*[4]

To develop this theory, Ferguson based his observations on four separate situations that became known in the literature as the defining cases of diglossia: Arabic, Modern Greek, Swiss German, and Haitian Creole. After introducing the labels of H variety (the superposed variety) and L variety (the vernacular form), he presented a list of nine features that he argued identify a diglossic situation as such.

First, and foremost, each of the two varieties has a set of functions distributed across the different domains. The H variety is the one used in educational, administrative, and religious discourse, while the L variety is often restricted to personal circles and transactional interactions. Second, the H variety is highly valued by the members of the community, whether they are fully competent in it or not. As a matter of fact, it is considered to be the "real" language they speak or should be speaking while the vernacular is considered a transitory and corrupt version of it and one that came about primarily because of lack of education. Third, the entirety of the literary body that a community records in the written form is in the H variety. Oral poetry and narratives may be produced in the vernacular as long as they are not recorded or published in written form. Fourth, the L variety is the variety that is transmitted naturally and children start to formally learn the H variety only when they start school. The L variety is the native language of all members of the community while nobody has native competence in the H variety. The majority of the speakers never attain higher levels of competence in H variety, especially in societies with high rates of illiteracy. Fifth, the H variety is fully described and dictionaries and grammar books are available for its teaching and learning while the L variety remains under-described. Speakers often have the impression that the grammar of the L

---

[4] Italics used in the original text.

variety is the same as that of the H variety but obscured and corrupted by the lack of knowledge of the "correct" forms. Sixth, diglossia is a stable phenomenon that does not change over a few generations. Both varieties can exist alongside each other for long periods of time. The H variety sees little change from when it was codified based on the written sources while the L variety continues to evolve like any other language that is acquired and transmitted naturally. Seventh, the grammar of the L variety is considerably less complex than the H variety's grammar, which tends to follow a set of conservative rules that reflect the high attention paid to it through the centuries by generations of prescriptive grammarians. Few people have native-like competence in the H variety and profound familiarity with its grammar is interpreted as an index of high levels of education. Eighth, Ferguson states that the largest part of the vocabulary is shared between both varieties, although the H variety has a more extensive technical and specialized lexicon. In addition, Ferguson mentions the existence of "lexical doublets," where the two varieties have settled on different terms for the same concept. Finally, at the phonological level, both varieties share "a single inventory," although the L inventory may show simplification on the one side and interference from other languages on the other. These features will be discussed again as different case studies are introduced in the following chapters.