

The “Languages in Danger” website has been developed within the INNET project at Adam Mickiewicz University, Poznan , Poland. The project has received funding from the 7th Framework Programme under Grant Agreement n° 284415: Innovative Networking in Infrastructure for Endangered Languages (project acronym: INNET). Official INNET project website: [innet-project.eu](http://innet-project.eu)



# LANGUAGES IN DANGER

home

interactive  
map

book  
of  
knowledge

teaching  
materials

what can  
you do

about

## AUTHORS

Nicole Nau

Michael Hornsby

Maciej Karpiński

Katarzyna Klessa

Tomasz Wicherkiewicz

Radosław Wójtowicz

## LANGUAGE EDITOR

Michael Hornsby

# Book of Knowledge

[Home](#) > [Book of Knowledge](#)

The “Book of Knowledge” (BoK) is a collection of texts providing descriptions of languages in danger in terms of the diversity of their structures, sounds, writing systems as well as issues related to multilingualism, language policy, documentation, revitalisation or cultural, social or technical problems related to the languages. The BoK is an integral part of the present website, it offers links to the related [Interactive Map](#) tasks and data included in another sections as well as to the external services and a wide variety of references for further reading. A downloadable version in PDF is available [HERE](#).

## ■ TABLE OF CONTENTS:

1. Languages of the World
2. Exploring Linguistic Diversity
3. Language Structures
4. The Sounds of Language
5. Writing
6. Language and Culture
7. Multilingualism and Language Contact
8. Language Endangerment
9. Endangered Languages, Ethnicity, Identity and Politics
10. Language Documentation

[Glossary, subject index](#)

[List of Languages](#)

[Let's Revise! – Interactive Revision Exercises](#)

[Phonetic exercises](#)

\* [Show detailed Table of contents](#)

## BOOK OF KNOWLEDGE

Chapters: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#)

[List of all languages](#) referred to in the Book of Knowledge and other sections of the website.

[Glossary](#)

**DOWNLOAD** and print out the Book of Knowledge.

## LET'S REVISE!

Go to the [Let's Revise section](#) to see what you can learn from the Book of Knowledge or to test how much you have already learnt!

## PHONETIC EXERCISES

Do you wish for some phonetic practice? Take a look at the exercises in the [Phonetic Exercises section](#).

# Languages of the World

Home > Book of Knowledge > Languages of the World

## ■ CHAPTER AUTHOR: RADOSŁAW WÓJTOWICZ

### Chapter contents:

How many languages?  
Where is the greatest number of languages found?  
What is language?  
Is human language different from animal communication?  
Where does language come from?  
How are languages classified?  
Is it good to have so many languages?  
Notes  
References & further reading

## ■ HOW MANY LANGUAGES?

It is estimated that there are around 7,000 languages worldwide. Contemporary linguistic studies provide different numbers. The Ethnologue – an encyclopedic study of languages of the world – contains descriptions of 7,106 languages as of 2014. In 2009, the number was 6,909, while in 1996 – 6,703.

In his book *Języki świata i ich klasyfikowanie*, the Polish linguist Alfred F. Majewicz (1989) mentions the existence of approximately 20,000 so-called **linguonyms**, that is the names of language varieties. Some languages have more than one name: for example, the Finno-Ugric **Udmurt** language, which is used in the Udmurt Republic (Udmurtia) in Russia, is also called *Votiac*. The first name is used by the community speaking the language (the Udmurts), while the second is preferred by non-Udmurts living in the same area and having daily contact with that people. It could seem that having access to modern technologies and scientific studies, we should be able to give a precise number of languages spoken in the world nowadays. Why are we unable to do this? There are several answers to this question.

First of all, some regions of the world are still inaccessible and unexplored. In Papua – the island with the greatest number of languages in the world – there are tribes which deliberately avoid contact with the outside world (find out more about them [here](#)). We can only speculate about their languages. It is highly probable that there are still some undiscovered languages in the world. We know very little or simply nothing about 80% of the world's languages.

### ACTIVITY PROPOSAL

Take a recording device, e.g. use the recording option in your mobile phone. With the recording mode on, name as many languages as you can in 30 seconds. Save your recording when you're done – we'll get back to it soon!

The inability to clearly determine what a **dialect** of a language is and what constitutes a separate language is another reason why estimations about the number of the world's languages are best preceded by the word "approximately". The main criterion for distinguishing between varieties of language is whether users of two ethnolects understand each other. If they do, we are dealing with dialects of one language. If not, then these are separate languages. The question remains: what does it mean that users understand each other? To what extent do they have common vocabulary? Are grammatical structures overlapping in whole or in part, but the differences do not hinder communication? What if users of different varieties have difficulty in understanding each other, but they feel representatives of the same community?

Nowadays, a large role in determining the boundaries between a language and a dialect is assigned to the so-called **political criterion**. Two varieties can be linguistically or dialectally regarded as dialects of the same language, but for some reasons their users need to recognize them as separate languages. This happens, for example, when new countries are formed. In India, the official language is **Hindi**. The users of this language have no trouble understanding people who speak the **Urdu** language of neighbouring Pakistan. However, due to political and religious differences, we officially talk about two different languages. In addition, these languages use two different systems of writing, which is also connected with religion: for Urdu, the Arabic script is used while Hindi is written in a Devanagari script, reflecting Muslim and Hindu influences respectively. The Saami people – indigenous peoples of northern Europe – see themselves as members of one nation. Only recently did people start to talk about **the Saami languages** in the plural, and not about dialects of the Saami language as before. Slightly oversimplifying, when we take the criterion of mutual understanding into account, it appears that the Saami people speak nine different languages.

### BOOK OF KNOWLEDGE

Chapters: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#)

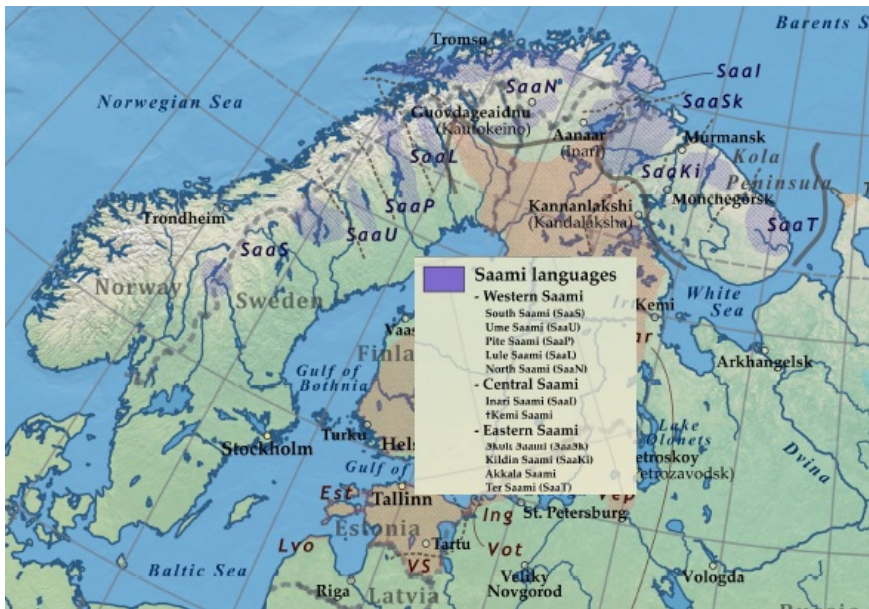
[List of all languages](#) referred to in the Book of Knowledge and other sections of the website.

### [Glossary](#)

**DOWNLOAD** and print out the Book of Knowledge.

### LET'S REVISE! – CHAPTER 1

Go to the [Let's Revise section](#) to see what you can learn from this chapter or test how much you have already learnt!



The violet fields on the map show areas where the Saami languages are spoken (Creative Commons ShareAlike license 4.0)

Nowadays, the political criterion is increasingly important in recognizing a language variety as an actual language. People supportive of the recognition of the Silesian ethnolect as an independent language say that because some Silesians are convinced their national identity and their dialect is an important part of the Silesian culture, it is necessary to speak of a **Silesian** language. On the other hand, opponents point to the fact that the Silesian dialect resembles the Polish language, and therefore it still has to be regarded as a variety of Polish. It is hard not to admit that they both have a point. Distinguishing languages from dialects is a continuous compromise between linguistic and political criteria. More discussion on these issues can be found in **chapter 9**.

Providing a precise number of languages spoken on our planet and distinguishing between separate languages and dialects of one language are thorny issues for linguists – for several reasons. Peter Mühlhäusler, a linguist working on the endangered languages of the Pacific region, wrote the following in his book of 1996: “The very view that languages can be counted and named may be part of the disease that has affected the linguistic ecology of the Pacific and (...) an obstacle to attempts to reconstruct the linguistic past.”

**STUDY QUESTIONS**

- (1) What do you think is the ‘disease’ mentioned by Peter Mühlhäusler?
- (2) Do we need languages to be named and counted? Think of at least two arguments for and against.

You’ll find proposed answers to these questions at the end of this chapter.

**■ WHERE IS THE LARGEST NUMBER OF LANGUAGES FOUND?**

The world’s ten biggest languages are used by 43% of the population of our planet. Of the approximately 7,000 languages, more than two thousand are used in Africa, and about one thousand in Papua alone. The country with the world’s largest concentration of languages is a small state of Vanuatu in the Pacific Ocean – it has 106 languages, while in Europe 285 are used.

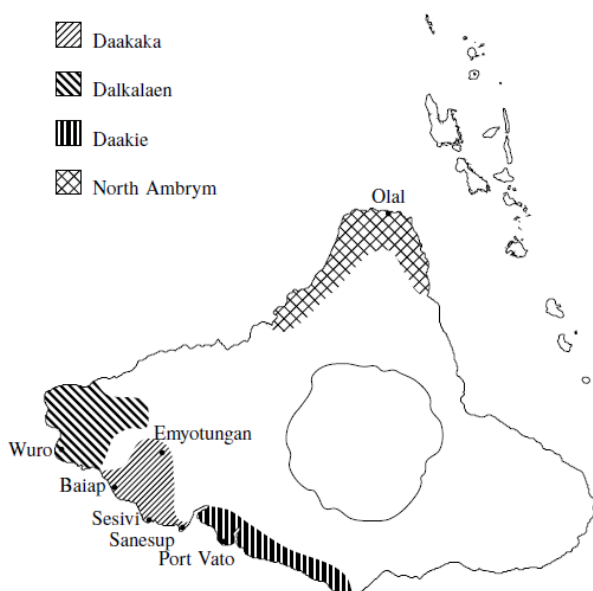
continent	languages	
	number	percentage
Africa	2,146	30,2
North and South America	1,060	14,9
Pacific	1,312	18,5
Asia	2,303	32,4
Europe	285	4,0
together	7,106	100



Table 1. Language distribution among continents (source: Lewis, M. Paul et al. (eds.) 2014)

One reason for the different geographical distribution of languages is the varying conditions shaped by landforms. Countries on plains are often monolingual: this means that a single language is used in them. Mountainous areas are in turn very diverse in terms of language. Witold Maciejewski (1999) speaks of the mountains as of areas of refuge – specific places where various ethnic groups find a kind of asylum, each living in a relatively small area. In countries where there are mountainous areas, the language of the central government has much less influence in the mountains than on the plains, where it has its centre.

On the small island of Ambrym, which is part of Vanuatu, six languages are used: **Daakaka**, **Dalkalaen**, **Daakie**, **North Ambrym**, **Lonwolwol**, and **Port-Vato** language. The picture below shows the island.



The Island of Ambrym (Vanuatu). © Kilu von Prince

As you can see, the island has roughly the shape of a triangle. Each of the three “corners” forms a distinct socio-cultural area. The languages of their inhabitants are more similar to the languages spoken on the islands located near each of the corners of Ambrym than to each other. The reason is that in the middle of the island there is a volcanic desert that is hard to cross – there is no road. Residents of Ambrym corners are in an easier position to contact the inhabitants of the neighbouring islands than the communities on their own island. In this case, the language differences reflect the intensity of contacts between users of different language varieties.



Volcanic desert on Ambrym. © Kilu von Prince

Nevertheless, the type of geographical conditions does not necessarily automatically translate into a particular language being used on a wide or limited area. Some communities just require a much larger area to live on than others. For example, a necessary condition of existence for the Nenets, a people of the Siberian tundra, is to have large areas of grassland because they engage in reindeer grazing and are forced to wander along with these animals in search of food. While the **Nenets** language is spoken today by 31,000 people, the

range of the Nenets language is far larger than the territory of Poland. On the other hand, in South-East Asia we very often encounter the situation in which just a small piece of land ensures food for the household and is enough for a family to live. The population density in these areas is very high, but residents do not feel “crowded” in any way because of that. In areas of the world such as South-East Asia one can find many languages in a relatively small area.

High linguistic density is also characteristic of northern parts of Australia, home to many native Aboriginal languages. For the indigenous people of Northern Australia, it is common to speak several local languages, and each of these languages belongs to a certain place. Nicholas Evans, a linguist working on these languages, reports that in the northwestern part of Arnhem Land (Northern Territory, Australia) every person has their father language, i.e. a language they have special rights and duties in due to the fact that they belong to a certain clan. Visitors to areas which have not been traveled to for a certain time are best accompanied by somebody for whom the language of the area is the father language – this is because the person is able to communicate with the local spirits and thus can ensure spiritual security to the visitors. But on a territory of another clan it is necessary to speak the language which belongs to the place – one cannot simply step foot on a territory without knowing its language nor can one speak one’s own father language there, even if the locals understand it. In his book ‘Dying words: Endangered languages and what they have to tell us’, Evans cites one of his **Kayardild** informants who tells a story showing how serious a breach of that rule is to Kayardild people [14]:

“A more extreme illustration of this principle comes from a story Pluto Bentinck, another old Kayardild man (...). When asked if traditional law included sanctions to be taken against trespassers, he cited an incident during World War II, when a hapless white airman swam ashore on Bentinck Island after his plane crashed in the sea. Pluto told me the man had said *danda ngijinda dulk, ngada warngiida kangka kamburij* (“this is my country, I just speak this one language”), as he struggled ashore without his Berlitz Kayardild phrasebook. When I asked him how he knew what the man had said, when he himself knew no English, Pluto replied: *Marralwarri dangkaa, ngumbanji kangki kamburij!* (“He was an ear-less (crazy) man, he spoke your language!”). Speaking English on Bentinck Island, in Pluto’s view, was tantamount to claiming it for English speakers. *Nyingka kabatha birdiya kangki! Ngada yulkaanda mirraya kangki kabath!* he had replied to the man (“You found the wrong words! I’ve found the right words, since forever”). *Ngada bunjiya balath, karwanguni*, Pluto continued: “And I clubbed him in the back of the neck”. (Evans 2011:8–9).

#### STUDY QUESTION

(3) Is your language linked to the place where you were born in the same way that **Kayardild** is to Bentinck Island? What are the differences and what is similar?



Native Australian languages also feature on the [Interactive Map](#). Find them and solve the exercises!

Compared to native Aboriginal cultures of Australia, some other cultures display much greater expansiveness than others, which under certain conditions results in the range of their languages expanding. During the Age of Discovery, Europeans conquered territories unknown until that point. The Portuguese arrived in South America and on the islands comprising present-day Indonesia, and the British conquered India and much of Africa. Territorial expansion often entails economic and social dominance. The fact that the largest language family in terms of the number of users is now the family of Indo-European languages is a direct result of years of European domination in trade in the above-mentioned areas.

#### ■ WHAT IS LANGUAGE?

So far, we have been talking about the different languages of the world and the border between language and dialect (if you are interested in dialects, go to [chapter 6](#) – is there really so much difference between ‘language’ and ‘dialect’?). There is no doubt that the concept of language appears in everyday life quite often: we are talking about the language of the media or the language of the writer, and in school we learn foreign languages. But what exactly is language? And how does human language differ from other languages, such as computer programming language? Intuitively, we feel that such phenomena we also call languages, such as body language, differ significantly from, say, English or **Sirionó** [1]. Let us look at the factors that make the difference.

Human language is, first of all, **universal** – it is used for many different purposes: naming concrete and abstract objects, expressing emotions, creating poetry. In a programming language, we will not buy ice cream at the market or chat with a shop assistant about the weather. Moreover, in languages such as the mathematical language of logic we express one thing in exactly one way. Human language is more **flexible** in this respect. We can communicate that we want someone to go away in many different ways: *Go away already!* or *Could you leave, please?*, or, for example, leaving a space for the recipient’s to guess: *I’m sorry, but I’m very tired*. Finally, the criterion for distinguishing human language from other above-mentioned languages is its so-called **productivity**, i.e. the ability of language to create an unlimited number of expressions, construct new sentences on the basis of elements that the language users have seen or heard somewhere.

There are as many definitions of language as there are different approaches to it. Some focus on the functions that language serves: for example, it determines whether one belongs to an ethnic group, it is a means of expressing emotions and a communication tool. Those for

whom language is primarily its grammar will emphasize the formal aspects of language as a multi-level system of meanings. The most general definition of language says that it is a system of signs used for communication. The science which is concerned with the study of human language is called **linguistics**.

The main function of language is **communication**, and therefore the language above all enables its users to transmit and receive information. Language is the code by which the sender sends the recipient a **message** via a communication channel. An element inherent to the communication situation is the **context** – the sender and the recipient of the message must have, at least to some extent, a common knowledge of what is communicated. If a plumber called by a university professor to replace the sanitary installation his home starts to explain her actions by using a large number of **jargon** terms specific to professional plumbers, the message will not be received, even if the professor is an outstanding linguist.

A classic of structural linguistics, Ferdinand de Saussure, was one of the first to draw a distinction between language as a set of rules which is located in the minds of the community of its users (*langue*) and a specific, one-time use of language – an act of speech (*parole*). Every time we talk on the phone, tell a joke or catch someone's attention by shouting *Hey!*, we use the words, grammar, intonation rules that are stored in the brains of our language community. On the other hand, language is not only in the community, between people, but also in the individual's mind – after all, everyone uses some form of language. Each of us, consciously or not, uses his or her "favourite" words or sayings: for greeting one person might always say *sup!* and another *hi!*. This is the kind of individual language characteristic of a particular person and is called an **idiolect**.

For most people in the world, multilingualism – the use of more than one language – is everyday reality (You'll learn more about this important phenomenon from [chapter 7](#) – the whole of it is dedicated to multilingualism). In school or on a language course a person learns a foreign language, and at least to some extent gets to grips with it. Language is, therefore, both a social as well as an individual fact. It also has a psychological dimension: it is stored in the brain. We can look upon language from different angles, depending on which aspect is of interest to us. The study of the psychological dimension of language, the complex mental processes of combining signs with their meanings is **psycholinguistics**. A psycholinguist will be interested in, for instance, how the person learns a language or what types of associations he uses to learn the vocabulary of a foreign language. The place of language in society and the social significance of its varieties – dialects, sociolects, etc. – is of interest to **sociolinguists**. Sociolinguistic research focuses on, for example, on variation in language and which variety enjoys linguistic prestige within a particular speech community.

In all the languages of the world we find sounds. Every time we use **speech**, we use a set of sounds of a language. Speech is inextricably linked with language: we cannot speak without using language, but language can be used without using speech. Another way of projecting the language stored in the brain is by **writing** – although the vast majority of languages do not exist in written form, but are only spoken varieties, it cannot be denied that they exist. And just like English or Spanish, they are complex systems of interpersonal communication. Each particular language – whether it is [Dyirbal \[2\]](#) or Portuguese – is an example of a human language.

Signing is another way of using language. Contrary to popular belief, sign language (of the deaf) is by no means a set of gestures that refer to objects in external reality. Signs are not iconic in this way and but are the result of a convention. Signs differ from, say, gestures in that they are used consciously to convey meaning that could be otherwise expressed by e.g. speech. Everybody can gesture but not everybody can sign. It is worth mentioning that due to the differences between the systems of signs we should rather speak of **sign languages** than of a sign language. The sign which in British Sign Language means 'mystery' stands for 'father' in Chinese Sign Language. Sign language can be learned via the Internet: for example you can check [here](#) for a material on American Sign Language or [here](#) for British Sign Language.

Sign languages, as well as other human languages are systems of signs. Sign is a symbol that refers us to a situation, a particular object or concept. Language signs are, for example, words and sentences. For instance, the word cat is a sign, it has its meaning: namely, it refers to domesticated animals moving on four paws, having whiskers and issuing specific meowing. Language signs are conventional, their meaning is based on agreement. A group of speakers agree that the word has a specific meaning. For example, the aforementioned language sign – the word written as cat – refers to the above-described animal. Any other language also applies a convention: for the determination of a meowing animal with a whiskers, users of [Friulian](#) language use a sign written as gjat and [Bengali \[3\]](#) – ■■■■■■■■. Whereas the word for a cat in [Russian](#) sound like [this](#). Even onomatopoeias – words which mimic sounds – are conventional. It might seem that such words are somewhat determined in advance, after all, they are like the sounds for example made by animals. However, even onomatopoeias are different in different languages: for example, for the referring to the sound produced by insects Polish uses the word *brzęczenie* and [Finnish](#) uses *pöriä*, *ininä*, and many others, depending on what insect is mentioned.

We encounter various other systems of signs in our everyday lives. Take road signs for example: they are a set of symbols used to communicate a variety of information to road users. Similarly to signs of a language, road signs are conventional: 'Give Way' is expressed with an equilateral triangle with the apex pointing down, having a yellow field and red border. It refers to any situation in which while on an intersection, we have to give way to a vehicle traveling on the main road. There is no doubt, however, that the languages spoken by people use different systems other than road signs. A characteristic feature of human language is called **double articulation**. Road signs cannot be combined with each other in such a way that they would form an entirely new sign. It is different in the case of human language: we create words from morphemes (see [chapter 3](#) what is meant by this) and phrases and sentences from words, and so on. Each word is a separate sign, it has a meaning that more than just the sum of the meanings of individual morphemes. The system consists of both language characters belonging to different subsystems, as well as the rules governing the way to connect these characters.

## ■ IS HUMAN LANGUAGE DIFFERENT FROM ANIMAL COMMUNICATION?



Human languages, as opposed to, for example, artificial logic or programming languages are **natural languages**. This raises the question – what about the languages of animals? Birds as well as humans communicate by using sounds. Are ultrasonic signals through which dolphins communicate, or the so-called waggle dance of the bees, a set of movements used by bees to communicate where the food can be found, to be considered equivalent to the communication systems of human languages?

Both human language and the systems of communication characteristic of animals occur naturally and are the result of the evolution of species. Gatherer bees use a system of movements to inform their fellow bees where the food is, but they are unable to communicate that the food was in the same place a week earlier. In addition, people, unlike bees, have the opportunity to talk about general phenomena, abstract concepts such as love or life. Human language, as opposed to dancing bees, is characterized by **abstraction** and consequently, by the independence from the stimulus: people do not need to have direct access to the fact about which they talk with other users of the language.

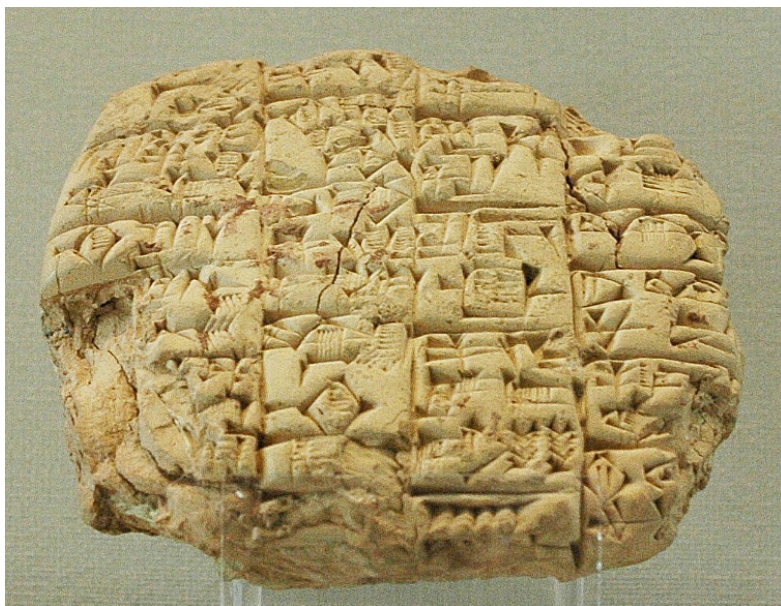
On the other hand, we know that, for example, some chimpanzees are able to arrange meaningful messages made up of more than a hundred simple characters. It has already been scientifically proven that among acoustic signals used by dolphins are those that allow the identification of individual animals in the herd – the dolphin names. However, there is no evidence that dolphins have evolved **metalinguistic** abilities, that is, abilities which would allow them to communicate anything about the language itself. In addition, the diversity of human languages – around 7,000 – is unique: unlike humans, for example, all dogs in the world are probably able to communicate and do not need to learn another variety of dog language.

In sum, in several respects the line between animal communication systems and human language is not always completely clear and a radical view that human language is so fundamentally different should be adopted with caution. There are, however, some very important differences between human language and the means of communication used by other species which make human language unlike other natural languages.

### ■ WHERE DOES LANGUAGE COME FROM?

With the appearance of the theory of the evolution of species in the nineteenth century, people boldly began to move away from religiously-based theories explaining the origin of man and his creations, and turned to scientific explanations. The same trend dominated the search for answers to the question about the origins of human language. Finding it turned out to be extremely difficult, and a precise placement of the beginnings of human language in time is virtually impossible.

The world's oldest cave paintings depicting some form of pictographic writings are about 17,000 years of age. The oldest [Sumerian](#) [4] tablets with cuneiform inscriptions date back to the fourth century BC. Human languages had certainly existed in the minds of users much earlier, but unfortunately we know very little about their development. The problem of determining the period of emergence of human language lies in the fact that to this day nothing that would directly indicate the emergence of human language endured. Muscles, such as the tongue, the oral cavity and soft tissues of the mouth very quickly undergo biodegradation upon death. Only by the shape of the skull and the shape of other bone structures of our Homo Sapiens ancestors can we point out when, more or less, the development of language might have started.



An example of an ancient cuneiform letter from around 2400 BC (licensed under Public domain via [Wikimedia Commons](#). User: Jastrow).

There are theories that our ancestors could use language even 3.5 million years ago. According to researchers, the decisive factor for communication with the sound systems is our oral cavity system. The pharyngeal cavity characteristic only to humans in the shape of the letter L appeared with the adoption of the bipedal position by Australopithecus – the ancestors of Homo Sapiens. Unfortunately, there is no evidence preserved that would prove the phases of development of the language, but it is assumed that it developed gradually.

With the evidence provided by genetics, we know that about 40,000 years ago Homo Sapiens had already learned a complex communication system based on sounds. Some researchers claim that it was language which was the reason why the Homo Sapiens, and not another human species dominated the world. All human languages are likely derived from a common ancestor. This means that language evolved only once, and at some point in history it began to branch out, which gave rise to language families. Languages developed separately with man's leaving the cradle of humanity – Africa – no later than 60,000 years ago, and the spread of species on other continents.

### ■ HOW ARE LANGUAGES CLASSIFIED?

Languages can be classified in many ways depending on the criteria employed. The basic assumption of the genetic classification of languages is that they derive from a common ancestor. For example, most of the languages spoken in Europe are Indo-European languages and originated from a Proto-Indo-European language. The Indo-European language family can be divided into several different groups as, for example, the **Slavic** language group which includes such languages as **Lower Sorbian**, Polish and **Macedonian**. Each was born out of Proto-Slavic.

Within language families, language subfamilies can be distinguished, further divided into groups, then subgroups, then even smaller units. Often scholars are not in complete agreement about how exactly a language family should be divided. There are many instances when a language family or group has several different proposed divisions which function at the same time, such as the case of **Bantu languages**. You can read about language families for example in the Britannica Online Encyclopedia. Follow this [link](#) for information about the Dravidian languages.



Is it possible to establish the degree of genetic relationship between languages not knowing them? Find out by solving exercises from **the Khanty language** on our [Interactive Map](#)!

The **typological classification** is another way to classify the languages of the world. This method can be applied through comparing similarities of structures which exist within these languages. For example, the feature shared by tonal languages (see the Chapter 4: **The sounds of language**) is the existence of tones. Tones are variations in the pitch of one's voice which carry meaning. In tonal languages words can be identical regarding the sequence of sounds they comprise of, so it is the pitch of one's voice that helps differentiate word meaning. **Yoruba** [5], **Wānsōhōt** [6] (Puinave), **Chinese** and **Burmese** [7] are examples of such languages. If word order is taken into consideration, then languages such as **Estonian** [8], **Totoli** [9] and English can be ascribed to the same group in which the basic word order is SVO (subject – verb – object). This means that declarative sentences takes the form of, for example, Anne is eating an apple. Another group encompasses languages such as **Welsh** [10], **Agta** [11], **Mixtec** [12] and **Malagasy** [13], the basic word order of which is **VSO** (verb – subject – object). Declarative sentences, in this case, take the form similar to the expression Likes Anne apples. From a typological perspective, then, languages can be included into the same group although they differ from each other genetically. Linguistic typology is dealt with more elaborately in [chapter 3](#).

### ■ IS IT GOOD TO HAVE SO MANY LANGUAGES?

In the biblical story of the Tower of Babel, God confused human languages as a punishment. For their pride and desire to be on a par with God people were sentenced to mutual incomprehension. Interestingly, a similar legend of divine anger caused by the construction of the great pyramid of Cholula was also known in the tradition of the Toltecs from central Mexico. It is widely accepted that the multiplicity of languages in the world causes a problem for humanity. But maybe the fact that we speak so many different languages is a blessing, not a curse?



No one is sure how many languages were spoken in Teotihuacán – the holy city of the ancestors of the Toltec and Aztec people. (© Katarzyna Miszczyszyn)

It could seem that it would be more practical and easier if we all spoke one language. We would not have problems with communication

with people living in the farthest corners of our planet nor would we have to spend time learning foreign languages or employing interpreters to translate books and documents. Such a view, however, reflects a certain way of understanding language: as an external tool that speakers use consciously to conduct activities. But language is not only something we employ for purposes connected with leisure, work or education – it is also part of the identity of its speakers and of the local environment. People who wish only one language was spoken all over the world are probably not aware of an important fact: namely, that languages tend to naturally develop in different directions as they are used by different communities living in different places. Let us take English for example – it is the lingua franca of today and the mother tongue for millions of people from all around the globe. Since the colonial era, the language that the British people brought to the places they conquered has evolved into what we now know as American English, Indian English, South African English, etc. These varieties differ from one another with respect to vocabulary, grammar, prosody and other phonetic features. Also the English of the British Isles presents a great deal of variety (as it did back during the Age of Discovery and in 1788, when captain James Cook landed in Australia): for example, Scottish English is famous for its rolled [r]-sound; it has certain grammatical characteristics that make it different from British English (e.g. varying usage of the present continuous tense) and features Scotticisms such as bairn ‘child’ or muckle ‘big’ which do not exist in the language people speak in London or Cornwall. These differences often present difficulty for those who learn English as a foreign language: even those who achieve high proficiency in British English find it hard to understand native English speakers from New Zealand or India. Therefore, it is sometimes more accurate to speak of different Englishes, i.e. geographical varieties of English. You’ll learn more about different language varieties, as well as language being part of one’s identity, from [chapter 6](#).

The world’s linguistic richness is one of the aspects of the **cultural diversity** of humanity, which is widely recognized as a value in itself. We feel that cultural diversity enriches us, lets us feel the way of how a different community functions, learn its system of values and customs. International organizations devote enormous resources to support local cultures. This also applies to languages, particularly **endangered languages**. We often hear about the effort to save endangered species. We assume that biodiversity is important for the proper functioning of ecosystems. The disappearance of a species has often disastrous consequences for the environment, for example, if in a given ecosystem there were no species that would deal with vermin. According to some researchers, the rapid decrease of the number of languages in the world entails a danger comparable to the extinction of species. Although in the case of a dying language we are dealing with a slightly different situation than when a species dies in nature, these two phenomena have much in common. The first important similarity is that the two different types of diversity overlap: hotbeds of linguistic diversity are places enormously rich in flora and fauna ([compare them here](#)). Furthermore, in places where languages die out, also the local wilderness is at risk. As British linguists Daniel Nettle and Suzanne Romaine show in their famous book about linguistic endangerment, the passing of languages is symptomatic of dangers to local ecosystems. For example, fishermen from the Pacific region are famous for their detailed knowledge about favourable fishing conditions in the oceanic waters, the behaviour and habitats of fish species and about the management of marine resources. The knowledge is reflected in their native languages: e.g. in the **Tobian** language of Palau, there exists a classification of fish species which tells how different types of fish behave [\[15\]](#). The Tobian term *hari merong* ‘always bites, takes any bait’ refers to a type of groupers which are easily caught. *Moghu* is a name for a fish species and for a disease which is cured by grounding up the fish and eating it. In other words, rather than on formal characteristics of species known from e.g. Latin descriptions, the Tobian taxonomy relies on functional criteria. This knowledge is very practical as it makes the life on Palau easier. With slightly more than a dozen of speakers, Tobian is now critically endangered, as are many other languages which preserve invaluable knowledge of nature and ways to protect it. The once balanced and renewable ecology of the Pacific water now experiences overfishing due to the impact of Western technology, education, and economy. Nowadays, anyone can purchase a handheld spear gun and go fishing without asking permissions from elders who have traditionally regulated and guarded fishing rights, relying on the knowledge of the environment they had acquired from their ancestors via the local languages.

Importantly, speakers of endangered languages work out this precious knowledge before Western science does. In Nettle’s & Romaine’s book, we find numerous examples: legends telling the precise time when Tasmania separated from the Australian mainland have been passed orally through the native Aboriginal languages for thousands of years, but it was not until the 20th century that Western scientists found their evidence convincing enough to finally agree when exactly the separation happened. The Europeans discovered that quinine treats malarial fevers in the 17th century, but the healing properties of the chinchona tree bark which contains quinine had been known to the indigenous inhabitants of the Andes much earlier, and so on. We will probably never learn how much knowledge which is still to be discovered was locked in the languages that have already vanished.

When a language dies, the way in which its users had understood the world dies with it. In this way we lose to some extent what makes us, as humans, so unique: a tool for understanding other people and the ability to gain a different way of looking at reality. But there are also more mundane dangers which come with the extinction of languages: lost languages are lost wisdom – wisdom which might be already gone by the time we need it. In [Chapter 8](#) you are going to learn how many of the 7,106 languages currently spoken on our planet are not going to make it to the next century. But before you move on to the chapters which are concerned with language endangerment, do see what the upcoming ones offer – there are so many fascinating aspects of linguistic diversity to discover!

#### FOOD FOR THOUGHT

Listen to your recording. For each of the languages you have named, check if it appears in the [UNESCO Atlas of the World’s Languages in Danger](#) – to do so, just type the name of the language into the ‘Language name’ box. Languages that are not there are most probably ‘safe’. How many of the languages you have on your recording are endangered? What do the proportions between safe and endangered languages on your recording tell you?



## LET'S REVISE! – CHAPTER 1

Go to the [Let's Revise section](#) to see what you can learn from this chapter or test how much you have already learnt!

### Notes

- [1] The Siriono (also known as: Siriono and Mbia Chee) language of the Tupi language family is spoken by roughly 400 people in the village of Ibiato (Bolivia). Siriono is critically endangered.
- [2] Dyirbal is a now extinct Australian language. It is famous for its mother-in-law variety – find out more about mother-in-law languages in [chapter 6!](#)
- [3] An Indo-European language spoken natively in Bangladesh and India. One of the ten largest languages of the world (first and second language users taken together).
- [4] The ancient Sumerian language was a language of Mesopotamia spoken up to 3rd millennium B.C.
- [5] The Yoruba language belongs to the Niger-Congo language family. It is spoken by around 19 million people in Nigeria and Benin.
- [6] Wansöht, otherwise known as Puinave (or: Waipuniavi) is a language isolate spoken along the river Orinoco in Colombia and Venezuela.
- [7] Burmese is a Sino-Tibetan language with official status in Myanmar.
- [8] Estonian is one of the largest Uralic languages. It is spoken by over a million people, mainly in Estonia where it has official status.
- [9] Totoli is an Austronesian language spoken by around 25,000 inhabitants of Central Sulawesi (Indonesia).
- [10] The oldest language spoken in Britain, around 19 per cent of the population of Wales (some 568,500 people) claim to speak it. It belongs to the Celtic branch of the Indo-European family.
- [11] Agta (or Aytá) are several languages of the semi-nomadic Negrito people of Luzon, Philippines.
- [12] The Mixtec language continuum comprises several indigenous languages spoken by around 550,000 people in southern Mexico. Mixtec belongs to the Oto-Manguean language family.
- [13] The Austronesian Malagasy language is the national language of Madagascar. It is also spoken on Mayotta and the Comoros.
- [14] Kayardild is a critically endangered Aboriginal language spoken on the South Wellesley Islands, Queensland, Australia.
- [15] Tobian is a critically endangered Austronesian language of Palau.

### References and further reading

- Bauer, Laurie. 2007. The Linguistics Student's Handbook. Oxford: Oxford University Press.
- Crystal, David. 2010. The Cambridge Encyclopedia of Language. Cambridge: Cambridge University Press.
- Evans, Nicholas. 2010. Dying words: Endangered languages and what they have to tell us. Chichester: Wiley-Blackwell.
- Lewis, M. Paul (ed.) 2009. Ethnologue. Languages of the World. 16th edition. Dallas: SIL International.
- Lewis, M. Paul, Gary F. Simons & Charles D. Fenng. (eds.) 2014. Ethnologue. Languages of the World. 17th edition. Dallas: SIL International: <http://www.ethnologue.com>
- Lyons, John. 1981. Language and Linguistics. An Introduction. Cambridge: Cambridge University Press.
- Maciejewski, Witold. 1999. Wielka Encyklopedia Geografii Świata. Tom XIV: Świat Języków. Poznań: Wydawnictwo Kurpisz.
- Majewicz, Alfred F. 1989. Języki świata i ich klasyfikowanie. Warszawa: PWN.
- Mühlhäusler, Peter. 1996. Linguistic ecology: language change and linguistic imperialism in the Pacific region. London: Routledge.
- Nettle, Daniel & Suzanne Romaine. 2000. Vanishing voices: The extinction of the world's languages. Oxford: Oxford University Press.
- Vajda, Edward. Linguistics 201 <http://pandora.cii.wvu.edu/vajda/ling201/ling201home.htm>
- von Prince, Kilu. 2012. A grammar of Daakaka. Doctoral dissertation, Humboldt Universität zu Berlin.

**Chapter translation from Polish :** Radosław Klimczak. **Translation update:** Nicole Nau, Michael Hornsby, Radosław Wójtowicz.

To access solutions to study questions click [here](#).

[back to top](#)

# Exploring Linguistic Diversity

Home > Book of Knowledge > Exploring Linguistic Diversity

## ■ CHAPTER AUTHOR: NICOLE NAU

### Chapter contents:

Exploring Linguistic Diversity: What is different? What is common?

- Languages and LANGUAGE
- The quest for universals
- How can different languages be described and compared?

Words and their meaning

- Basic vocabulary
- Word meaning and categorization
- Words for the world we live in

Notes

References & further reading

## ■ EXPLORING LINGUISTIC DIVERSITY: WHAT IS DIFFERENT? WHAT IS COMMON?

Some people find it hard to believe that there are so many different languages in the world – 6,500 or even 7,000? Are these all real languages? Aren't most of them just dialects? People used to the situation in Europe often think that a “real language” has a written form and a written tradition, it has been standardized and described in dictionaries and grammars, it is taught in schools, and it has official status within a state. Languages that lack these characteristics are thought to be “only dialects”, which often implies that they are incomplete and somehow inferior to languages such as English, French or Dutch. In the 19th century the European feeling of superiority was overtly expressed by labelling smaller languages of non-European cultures “primitive languages”. Today we know that there is no such thing as a primitive language. In fact, languages spoken by communities that lack western technology often have highly complicated systems and are therefore of great value to linguists.

On the other hand, when people say that something is “only a dialect” they may have in mind its similarity to another language, usually one that is better known and has a standard and/or official status. For example, they may think of [Kashubian](#) as a dialect of Polish, because these languages are similar and Polish is the dominant language in Poland. In this view, the question “language or dialect” can be answered by determining the grade of difference between two ethnolects. Unfortunately, this is not an easy task if the two ethnolects in question are genetically closely related (belonging to the same branch or subgroup of a language family, as [Kashubian](#) and Polish do; note that today [Kashubian](#) is recognized as a separate language). Most linguists today agree that the question which “lects” belong to one language and which to another cannot be decided by one or two simple criteria. Many different factors have to be taken into account, and mutual intelligibility, or grade of difference, is only one (see also [Languages of the World](#) and [Endangered languages, ethnicity, identity and politics](#) for discussion). This is one of the reasons why we are not able to give the exact number of languages spoken today. Anyhow, we are sure that there are several thousand languages in the world that differ from each other at least as much as English and German do.

## BOOK OF KNOWLEDGE

Chapters: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#)

[List of all languages](#) referred to in the Book of Knowledge and other sections of the website.

[Glossary](#)

**DOWNLOAD** and print out the Book of Knowledge.

## LET'S REVISE! – CHAPTER 2

Go to the [Let's Revise section](#) to see what you can learn from this chapter or test how much you have already learnt!



## Languages and LANGUAGE

The question “how different are the languages of the world?” can also be approached in a more principled way. It has intrigued linguists and philosophers for a long time and still causes some controversy.

Some hold that all languages are essentially the same, as they are all products of the human mind – they are manifestations of LANGUAGE, a human capacity. Children all over the world acquire their mother tongue in about the same time, which shows that from the point of view of the native speaker there are no “complicated” or “easy” languages. It has been claimed that children could not acquire a language from what they hear (the “input”) alone, but that they must possess an innate set of general linguistic rules and principles that allows them to build up the system of the language or languages of their surroundings. This innate mechanism is called Universal Grammar; it determines which structures are possible in individual languages. Some linguists claim that the aim of linguistics is to discover this abstract Universal Grammar, not to describe the different languages that are spoken in real time and space. In its most radical form, this approach may lead to the conclusion that it is enough to study one language thoroughly, but language comparison is necessary only to test one’s hypotheses. From this perspective the fact that many small languages are dying out is not a big issue, for Universal Grammar can as well be discovered from the study of English, Japanese and Arabic alone. This line of thought was very influential in linguistics, especially in the USA during the 1960s – 1980s, and its most prominent proponent was [Noam Chomsky](#), probably among the most famous linguists of the 20th century.

On the other hand, there are linguists who hold that each language is a unique system that has to be studied and described in its own right. One cannot make generalizations or draw conclusions about language in general from the regularities found in one or two, or even a hundred, individual languages. To explore LANGUAGE, the human capacity, we need in principle all the languages that ever have been spoken by humans. As this is not possible, we cannot determine which structures are impossible in human languages, and linguists have to set themselves more humble tasks. The idea that each language has its own system that has to be studied as such, without drawing inferences from one’s knowledge of other languages was generally accepted in the research tradition called structuralism which dominated linguistic research in Europe and the USA during the 1930s – 1950s.

Most linguists today probably would place themselves somewhere between the extreme positions “all languages are essentially the same (so it is enough to study one)” and “all languages are essentially different (and one cannot make any generalizations)”. Languages vary widely in their structures, but there is also some common ground and that is the reason why we find similar structures and categories in different languages all over the world. The study of languages that have not been described before may reveal something completely new, but it will add also evidence for regularities that have already been discovered in other languages.

### The quest for universals and the establishment of linguistic types

Features that are common to all languages are called **linguistic universals**. There are different approaches to their study and different understandings of universals. In the approach that uses the concept of “Universal Grammar” mentioned above, universals are abstract principles that underlie concrete structures. For example, there is assumed to be an abstract feature [tense] with a certain place in an abstract structure [VP] (the name VP comes from verb phrase). Overt markers of tense that correspond to this abstract feature are, for example, the English morphemes *-ed* and *will* in *We walked in the park* and *She will come tomorrow*. However, claiming that the abstract feature is part of Universal Grammar does not mean that in all languages of the world there must be an overt marker of tense.

**Absolute universals** (statements that are true for all languages) are usually fairly general, for example: “in all languages there are means to negate a statement”, or: “in all spoken languages there are vowels and consonants”. There are not many such general statements that really hold for all languages of the world. Some are disputed, for example, the question whether all languages have nouns and verbs as separate word classes. Furthermore, what we call a verb in one language may be very different from a verb in another language. The linguist Wolfgang Klein once wrote that using the term “verb” for such different phenomena as the verb in Latin and the verb in Chinese may be like using the same term with reference to rice and potato and calling rice the potato of the Chinese (Klein 1995: 81). We cannot even be sure that what is a verb in one language will also be translated by a verb in another language.



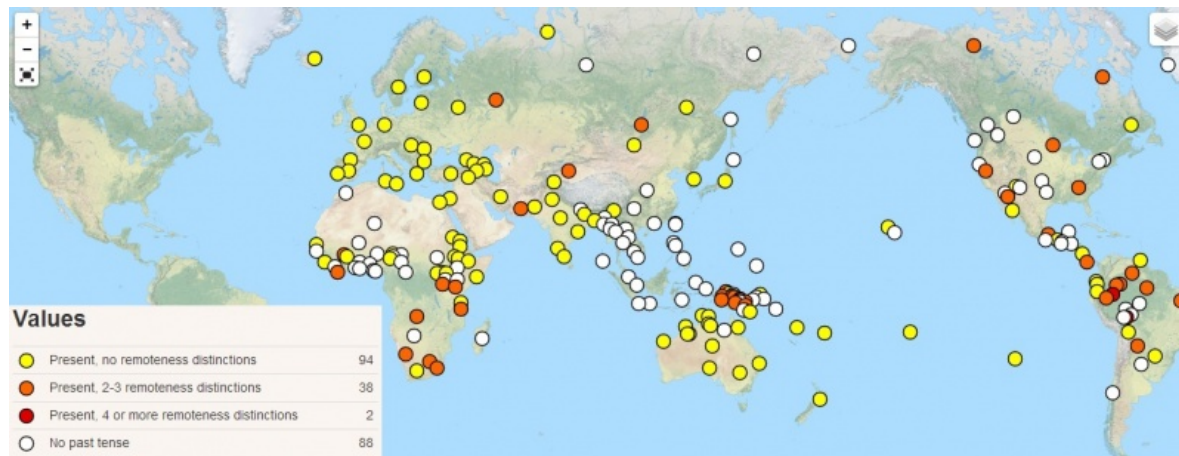
Find examples from languages where parts-of-speech differ from European languages on our [Map](#): exercises for Iwaidja (ex. 3) or Hoocak (ex. 2).

Another type of universals are **implicational universals**. They are stated in the form “if a language has feature A, **then** it also has feature B”. For example: “If a language has a dual then it also has a plural” – that means there is no language that marks nouns or pronouns for dual (meaning ‘two’) and not for plural (meaning ‘more than one’), but there are languages that have both a dual and a plural and languages that only have a plural. This kind of statement does not claim that all languages have feature B (in our example – that all languages mark nouns for plural). The search for implicational universals has brought to light many interesting regularities that characterize human languages.

In addition to universals in the strict sense (= what is true for all languages without exception), linguists are also interested in features that are found if not in all, then in the great majority of languages. Sometimes these features are called “statistical universals”, or, better, universal tendencies. For example, in most languages that have a preferred word order, the subject comes earlier in the clause than the object (see [Chapter 3 Language Structures](#) for details). Many implicational universals are **universal tendencies** rather than absolute

universals.

The search for features that are common to all or most languages is complemented by cataloguing and systematizing the differences. This is what **linguistic typology** is concerned with (see also [Chapter 1](#)). To continue the example given above, a research question for linguistic typology is which number systems are found in the languages of the world, and the result of this research is the distinction of the following types: 1) languages that don't mark number, 2) languages that distinguish singular and plural, 3) languages that distinguish singular, plural, and dual, and so on. Some typological investigations start with a simple yes/no-question, for example: Does a language mark past tense (that is, distinguish formally between the present and the past)? Such questions divide all the languages of the world into two groups. In the next step, the languages of one of the groups are considered further, for example by investigating further differences that are made in languages that mark past tense. Some of these languages distinguish between a remote past and a past further away from the moment of speech. The Amazonian language **Yagua** stands out by distinguishing five grades of remoteness: 'a few hours ago', 'a day ago', 'about a week to a month ago', 'more than a month ago, up to two years', and 'a long time ago' are all marked by different suffixes on the verb (find the details in [Dahl & Velupillai 2013](#)).



Past tense distinctions (Dahl & Velupillai 2013)

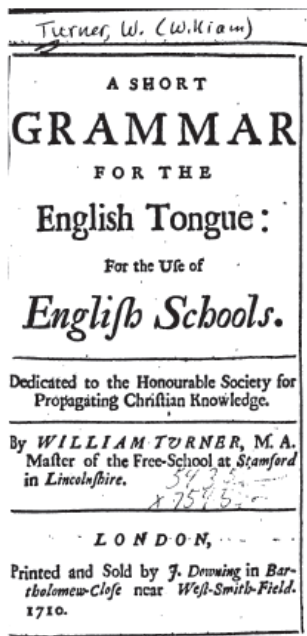
Certain linguistic features and types occur more often in some part of the world than in others. Interestingly, the geographic distribution of linguistic features often cuts through the classification of languages into families.

**The World Atlas of Language Structures (WALS)** shows and discusses the distribution of various features that have been studied by typologists ([Dryer & Haspelmath 2013](#)). The range of features that are presented is very large: it includes sounds (for example: [Which languages have nasal vowels?](#); [Where do we find tone languages?](#)), words (for example: [Which languages use the same word for 'hand' and 'arm'?](#); [Where does the word for 'tea' come from?](#)), and grammar, and there is also a chapter about [writing systems](#).

### How can different languages be described and compared?

There are many possible ways of describing linguistic structures, words, sentences, or sounds. Different traditions have been established in different parts of the world, for example, there is a European tradition, an Indian tradition, an Arabic tradition. These traditions have sprung from the analysis and description of one or two particular languages, most often focusing on a literary variant – the language of sacred texts, or that of renown writers. The European tradition was founded by Greek grammarians and later developed further for the description of (written) Latin. Since the early Middle Ages this “Latin grammar” was then used as a model to describe first other European languages, and later on also non-European languages, for example, the native languages of the American territories conquered by Spain. The more the languages under description differed from Latin, the less suitable this method was for their description. The categories and structures known from Latin were supposed to be found in all languages, but other categories were often ignored – a good example for the wisdom that one finds only what one is looking for. For example, English was described as having an ablative (“of the book”) and a vocative case (“Oh book!”), while the use of the definite and the indefinite article was not treated at all (Latin didn't have articles). Over the centuries the Latin tradition was slowly modified and adapted to better suit modern European languages. European national languages developed their own tradition, which nevertheless was (and is to this date) based on the Latin model.





**§ In other Languages a Noun is declined with six Cases; The Nominative, Vocative, Accusative, Genitive, Dative, and Ablative. But the English Noun being the same in all these six Cases, it will be sufficient to decline it in each Number, as above. Nevertheless the Teacher may, if he pleases, let the Scholar decline it with Cases as follows,**

	Singular.	Plural.
Nom.	A Man	Men
Voc.	O Man	O Men
Accus.	A Man	Men
Gen.	Of a Man	Of Men
Dat.	To a Man	To Men
Abl.	With a Man	With Men

**¶ Defective Substantives.**  
Some want the Plural Number; as *bread, beer, ale, honey, silver, gold, hay, poverty, honesty, righteousness, &c.*  
Some want the Singular Number; as *goods, riches, villas, &c.* *ajour, cloaths, pains, (i.e. Labour.)*

**Declining of Nouns Adjectives.**  
Note. Adjectives stand most commonly before their Substantives, and are the same in both Numbers,

Singular.	Plural.
A good Boy	Good Boys.
A great Place	Great Places
A small Fish	Small Fishes
A wife Man	Wife Men

Except.

A 18th century grammar of English – with a full declension of the word “man”

A serious challenge to the European tradition arose in the 20th century, when anthropologists and linguists in the USA became interested in native American languages and found that traditional European grammar was not suited for their description. The study of these languages was an important impetus towards the development of new methods of language description, and at the same time it triggered a new understanding of the motive and the aim of language descriptions, a new approach to “grammar”. In this new view, a grammar has to state objectively which forms and constructions are used by speakers of a language and must not judge which of these forms are “right” or “wrong”, “good” or “bad”. Every structure that native speakers of a language regularly use and accept is “right”. This view is in opposition to the traditional understanding that associates grammar with the literary standard of a language and the acceptance of structures by certain authorities, for example teachers.

Modern linguistic research (such as that behind the [WALS](#)) uses methods that are suitable for the description of very different linguistic structures and thus can be used with any language.

This can be seen most clearly in the description of the sound system of languages. The International Phonetic Alphabet ([IPA](#)), which was developed at the end of the 19th century, is a set of symbols that can be used to describe the sounds of all human languages. Its values have been fixed: each symbol stands for a sound that is produced in a certain manner of articulation (among others, by bringing one’s tongue and lips in a certain position). Letters of an alphabet, in contrast, are associated to sounds in words of particular languages. For example, the letter < r > stands for different sounds in English, Spanish, or French, and even within one language a letter may symbolize different sounds, for example the letter < a > in the English words *has*, *was*, *ask*. The IPA symbols [r] and [a] are unambiguous. Note that in this paragraph you can observe a convention used by linguists: to distinguish letters from sounds, different sorts of brackets are used.

#### STUDY QUESTION:

Find words in your native language where the same letter stands for different sounds.

#### PRACTICAL TIP

Many languages of the world have already been described using the symbols of the [IPA](#). Knowing this system is very useful to people interested in the languages of the world. Learn the IPA [here](#).

For a small sample of the sounds of a language that could then be compared with others, the International Phonetic Association created a simple version of a traditional tale – The North Wind and the Sun. The translation of this tale is read by a native speaker and the recording transcribed using the IPA. Find examples of The North Wind and the Sun (recording and transcription) [here](#).

The common basis for describing the different sounds of human languages is the “device” by which they are produced that is common to

all humans (the voice box, mouth, tongue, lips). Read more on this and on different sound systems in [Chapter 4](#). It is much more difficult to find a common base for the description of words and grammatical structures. The latter will be discussed in [Chapter 3](#), while the remaining part of the current chapter considers what is common and what is different in the vocabulary of the languages of the world.

## ■ WORDS AND THEIR MEANING

In any fully functioning language one can express everything that the speech community needs to express. Obviously, different speech communities, but also groups within one speech community, may have different needs, and this is reflected in differences in the vocabulary. For example, Germans who like mushroom picking know many different names of mushrooms while the average German may know no more than three or four. Many interesting questions arise when comparing words across languages, for example:

- Which words are common to all languages? Are there such words? Do all languages have a word for ...?
- Do words in different languages mean the same? How do they differ?
- Which differences are related to differences in culture? What can we learn about differences in culture by studying vocabulary?
- What may be the consequences of differences in the meaning of words? Are differences in vocabulary related to differences in thinking?

These questions will be addressed in this section.

### Basic vocabulary

Several linguists have tried to devise a list of words that most probably have equivalents in all or almost all languages. These lists are often used to compare languages and to find similarities between two languages that are genetically related (belong to the same family). The most famous of these lists was started by the American linguist [Morris Swadesh](#) in the 1950s. There are several versions, ranging from about 100 to 200 words. Swadesh's motivation was to provide a list of words that could be used to establish the degree of relationship between languages. Languages that are closely related (belong to the same branch or subgroup of one language family) share a large part of their vocabulary. Words that two or more languages have inherited from the same ancestor language are called **cognates**, for example English *blood*, Dutch *bloed*, German *Blut*, or Polish *głowa*, Croatian *glava*, Russian *голова* (*golova*) 'head'. Cognates are found most easily in basic vocabulary: words for everyday concepts that are less dependent on cultural differences and not likely to be borrowed from other languages. The Swadesh list contains among others words for body parts (*head, eye, blood*), everyday human activities (*eat, drink, see, walk*), natural phenomena (*sun, water, fire*), small numbers (*one, two, three*), and personal pronouns (*I, you, he*).

#### SWADESH LIST

There follows a fragment of a [Swadesh list](#). Which words may be cognates? Fragment from a Swadesh list for three languages of the [Hokan or Yuman family](#) (Native American languages spoken in California and Mexico)

Nº	English	Ipai (Kumeyaay)	Kiliwa	Cocopa
22	One	'ehink, 'uun	msíg	Shitt
23	Two	hewuk	juwak	Xwak
24	Three	hemuk	jmik	Xmuk
25	Four	chepup	mnak	Spap
26	five	sawrrup, saarap	salchipam	Ssrap
36	woman	siny	kekóo	s'ak
37	man (adult male)	'iikwich	kumeey	'apá
64	blood	'ehwatt	kujat	(ny)xwatt
65	bone	aq	ják	(ny)yak
72	head	hellytaa	iy	Mkur
74	eye	eyiiv	yuuv	'iyú
83	hand	esally	sal	'isháálly
84	wing	wirewir	wálu	'irwir, 'isháálly
92	to drink	wesii	chee	Ssi
93	to eat	wemaa, wesaaw	tmaa	Ma
107	to sleep	hema	smaa	Shma
147	Sun	'enyaa	eniaay	Nya
148	moon	hellyaa	ja'la'	xlly'a
150	water	'ehaa	ja'	Xa
154	sea (as in ocean)	'ehaasilly	ja' tay ("big water")	xakwss'illy
159	earth (as in soil)	'emat	maat	Matt



167 fire	'aaw	a'aw	'a'á
172 red	'ehwatt	kwál	Xwatt
175 white	nemeshap	msaap	Xmaally
176 black	nyilly	nleeg	Nyilly

Using a method called **lexicostatistics**, Swadesh tried to calculate the degree of relationship between languages on the basis of the number of cognates found in his list. This approach is disputed in linguistics. On the one hand it offers a feasible method for historic comparison of languages of which we don't have historic records. On the other hand it is doubtful how reliable this method is. For example, though it is true that basic vocabulary is less often borrowed from other languages, such borrowing can be found everywhere in the world. Without historic documents it is often not possible to determine whether similar words in two languages are the result of a common heritage or of borrowing.

Apart from language comparison, Swadesh lists and other collections of basic vocabulary are also used in **language documentation**. They are a useful tool and starting point for collecting vocabulary of a language.

#### PRACTICAL TIP:

Examples of these lists in various languages can easily be found on the Internet by searching for "Swadesh list". Good places to start are Wikipedia (entry "[Swadesh list](#)" in several languages), Wiktionary, and the [Rosetta](#) project, which hosts a growing collection of such lists in lesser used languages (search for "Swadesh").

A weak point of these lists is that the point of departure is usually a collection of words in a certain language (most often English) that are then translated into other languages. This method works on the assumption that these words indeed can be translated and that the translation equivalents are more or less exact correspondences. However, this cannot be taken for granted. Even seemingly basic concepts such as 'hand' or 'brother' can be covered by very different words. While English distinguishes *hand*, *arm*, and *finger*, other languages use the same word for 'hand' and 'arm', or for 'hand' and 'finger' (see Brown 2013a and 2013b in WALS, <http://wals.info/chapter/129> and <http://wals.info/chapter/130>). And while European languages usually have a word 'brother' designating a male sibling, other languages may make other distinctions. For example, in [Samoan](#) [3] a man or boy refers to a sister with the word *afafine* and to a brother with the word *uso*. But this last word (*uso*) is used also by girls and women to refer to a sister, while they call a brother *tuagane*. Thus the English words *brother* and *sister* cannot be translated into Samoan (or only in a clumsy way, saying for example that a *brother* is either *tuagane* or a male *uso*).

A more neutral way of collecting vocabulary is by defining topics, for example 'body parts', 'kinship terms', 'natural phenomena', 'colour terms', etc. and then collect the means of expression a given language uses for these topics and try to give their precise meaning without assuming that they have an exact equivalent in English (or whichever language is used for the description). This method is basically used in the project *The Intercontinental Dictionary Series* (IDS), where vocabulary belonging to 22 different topics is collected from a great variety of languages. The collection can be browsed at the site <http://lingweb.eva.mpg.de/ids/>.

#### Word meaning and categorization

When we learn another language we often find that the meaning of words is not exactly the same as in corresponding words in our mother tongue. Sometimes a word in one language combines the meaning of two words in another language. For example, English *wood* may refer to a place with many trees or to a material. German has two different words: *Wald* and *Holz*, respectively. In Latvian, on the other hand, the word *koks* refers to both the material and a single tree. Compare:

	a place with trees	material from trees	a tree
English	wood		tree
German	Wald	Holz	Baum
Latvian	mežs	koks	

Such examples show that the vocabulary of a language is not an inventory of labels for things and phenomena that exist independently of the human mind. The categories that stand behind individual words are created and maintained by these words, by the way speakers use these words and contrast them with other words in their language.

Each individual language includes an invitation to look at the world in a certain way. What is lumped together in one language may be highly differentiated in another. For example, European languages usually use one verb for the action of transporting an object while holding it on some part of one's body – in English this is the verb *carry*. Now look at some of the verbs that are used in [ǂAkhoe Hai||om](#) [4] to express this meaning:

## ‡AKHOE HAI||OM WORDS FOR 'CARRY'

ton	'carry on one's shoulder'
!guri	'carry on one's head'
gobe	'carry on one's back'
aba	'carry (a baby) on one's back'
‡khore	'carry a load'

We may say that ‡Akhoe Hai||om (and many other languages of the world) invites us to see the different ways of carrying, depending on what is transported and how it is held, while languages such as English invite us to see the similarity of these actions. Knowing that there are different ways to look at something makes us wiser. None of these different ways is superior to others, all are reasonable and useful.

The categories defined by the meaning range of individual words are first and foremost lexical categories (related to words). A much disputed question is in which way they are related to conceptual categories, to categories of the mind, and how strong this relation is. Does the vocabulary of a given language determine the way speakers of this language perceive the world? Does language determine thought? Or is it merely a reflection of the way our mind works – and as our mind may work in different ways, there are different possibilities from which to choose when associating a word with a meaning?

The strong hypothesis that language determines perception has been proven untenable. The different lexical categories for the expression of 'tree', 'place where trees grow', and 'material of trees' do not indicate that speakers of English or Latvian see no difference between trees growing and the substance they consist of – after all, there are further words in these languages that single out smaller categories, such as English *forest*. Europeans are equally able to perceive the difference between carrying a sack on one's shoulder and carrying a baby in one's arms, or carrying a suitcase by its handle, while speakers of a language that has different verbs for each of these actions will not deny that they have something in common. These and similar examples from well-known languages should make us cautious when we hear about some exotic language whose speakers allegedly perceive the world completely different from Europeans because their words have different meanings.

A good case to study the relation between lexical categories and perception are words for colour. The languages of the world differ in the range of words denoting colours. For example, many languages use the same word for 'blue' and 'green' [5], other languages on the other hand have different words for colours that English lumps together as blue. However, experiments have shown that speakers of such different languages nevertheless perceive colours in the same way. Speakers of a language with one word for 'blue' and 'green' distinguish the colour of the sky as clearly from the colour of grass as speakers of English. Furthermore, the word they use for both (which linguists sometimes call "grue") is not defined as something in-between green and blue (say, a shade of turquoise) or the changing colour of the sea, but rather as both the colour of the sky and the colour of grass. Even more striking for Europeans are languages with only two colour terms, where the colour of blood is named by a word that also means 'white' (or the one that also means 'black').

There is much more we can learn from the study of colour terms [6]. For example, there seems to be a universal tendency which colours are named most often by a simple, non-derived word (such as English *red*, but not *golden*, which is derived from *gold*). This tendency has first been described by Brent Berlin and Paul Kay in their book *Basic color terms: Their universality and evolution* (1969). The authors propose the following hierarchy (simplified here):

first	black and white (both words are also used with reference to other colours)
then	red
then	green (or "grue") and yellow
then	blue
then	brown
then	others (pink, orange, purple, grey)

This means, for example, that if a language has only three basic colour terms, one of them will mean 'red' (but not 'blue' or 'yellow'), or that if a language has a word for 'blue', it also has words for 'red', 'green', 'yellow', but not necessarily for 'brown' or 'pink'.

## EXERCISE

The web-site [sorosoro.org](http://www.sorosoro.org) includes short video clips presenting selected words in several endangered languages. Go to <http://www.sorosoro.org/en/colors> and watch the clips about colours in six languages of Central Africa (Akele, Baynuk, Benga, Menik, Mpongwe and Punu). How many and which colours are named in these clips? Do the inventories in these languages correspond to the hierarchy proposed by Berlin and Kay?

## Words for the world we live in

Differences in vocabulary may reflect differences in culture and society and are therefore an interesting subject not only for linguistic investigations, but also for anthropology and cultural studies. Furthermore, words can be the carrier of traditional wisdom, of knowledge accumulated by a people through experience over a long time. In recent time also researchers from disciplines such as biology, medicine, or even astronomy have started to pay attention to lexical categories found in languages in different parts of the world. As Nettle and Romaine (2000: 60) put it: “The vocabulary of a language is an inventory of the items a culture talks about and has categorized in order to make sense of the world and to survive in a local ecosystem.” The preservation of endangered languages is part of the preservation of knowledge about local ecosystems, which in turn can be vital for the preservation of this ecosystem.

Probably all languages have a differentiated vocabulary for certain animals that are important for a community. In European languages we may find many words for cattle, with different terms for males and females (bull, cow), young ones (calf), and several distinctions regarding fertility and the use the animal has for humans (bullock, ox, heifer...). In [Baka](#), the language of a community of hunter-gatherers in Central Africa, there are many different words referring to elephants [\[7\]](#):



Baka word	Explanation
ijà	elephant in general
ndzàbò	very big male elephant, king of elephants
sèmē	old big male elephant
kàmbà	big male elephant (but not as strong as the above)
mòsèmbi	male elephant (between kàmbà and mòbòngò)
mòbòngò	smaller male elephant
èkwāmbē	young male elephant living alone; male or female elephant that has lost its mother and become solitary
likòmbà	female adult elephant
bèndùm	elephant calf

To the Baka, elephants are an important source of food, and hunting elephants is an important activity (note that the Baka traditionally didn't hunt elephants for ivory and that their need of elephant meat is not the cause that elephants have become an endangered species). The different terms for elephants of different strength reflect knowledge that is vital for the hunters.

The two examples above are concerned with differentiations within one species that are of direct importance to a given community. Another aspect is the amount of names for different species, which reflects a community's awareness of the biological diversity surrounding them. Some of the most biodiverse places in the world are found in the Pacific. Thousands of different species of fish and other marine creatures live in the reefs of the Coral Triangle [\[8\]](#). Languages spoken in that area usually distinguish between 300 and 500 different fish names (Pawley 2011: 269). Many species had long been known and named by local people before western scientists discovered and catalogued them. Recently the importance of that local knowledge has been recognized by biologists and conservationists. Lists of names of species in local languages complete lists of the Latin terms, and sometimes they are the better inventories.

### MILNE BAYE CONSERVATION INITIATIVE

Watch [a short clip](#) about a conservation initiative on a small island of Milne Baye, Papua New Guinea, shot by James Morgan for [USAID](#). At the beginning of the clip you see how a list of fish names is used in making an inventory of the species that have to be protected.

DIYADIYANA	Striated surgeonfish	<i>Acanthurus lineatus</i>
WULOALAOALAU	Orangespine unicornfish	<i>Chelodactylus</i>
OSAOSA	Bullethead parrotfish	<i>Naso lituratus</i>
OSAALALAWA	Yellowbarred parrotfish	<i>Chlorurus sandaka</i>
HINEGAYUYU	Barred rabbitfish	<i>Scarus dimidiatus</i>
DERI	Silver spinefoot	<i>Siganus dalatius</i>
MAMAI	Humphead Maori Wrasse	<i>Siganus argenteus</i>
LAUNAPELO	Coral Trout	<i>Chelinus undulatus</i>
SABUGABUBU	Blackspot snapper	<i>Plectropomus leugnis</i>
WILITA'ATA'AI	Blackspotted grouper	<i>Lutjanus fulviflammus</i>
AU'AU'U	Blacktipped grouper	<i>Cephalopholis cyanostictus</i>
ULUTAPOTAPOI	Bigeye Bream	<i>Epinephelus spilotoceus</i>
SABAWA	Sabre Squirrelfish	<i>Monotaxis grandoculis</i>
LAWATAI	Moray eel	<i>Sargocentron</i>

In [another film](#) by James Morgan about people of the Coral Triangle and the disastrous effects of blast fishing on the environment and on people, the following questions are posed:

- What is the future of marine conservation?
- What are we really trying to conserve?
- And who should we look to for the answer?
- Who really understands the ocean?
- And why aren't we listening to these voices?

<http://jamesmorganphotography.co.uk/film/the-bajau-laut>

Biology and environmental studies are examples of fields where listening to local voices has been recognized as valuable for modern science. A related field is medicine: the knowledge of people about the healing capacities of local plants can be very important for the development of new pharmaceutical products. The key to access this traditional knowledge is language – very often, a language spoken by a small community that is endangered by cultural change and pressure from other, more powerful languages. Recall that the hotspots of biological diversity (and endangerment of species) coincide to a considerable extent with the hotspots of linguistic diversity (and language endangerment; see [Chapter 1 on Languages of the world](#)). Collaboration between members of local communities and scholars from various fields is needed to uncover and preserve this knowledge. Today, linguists, anthropologists and biologists often undertake field trips together and profit from each other's knowledge and skills. Other disciplines that have become interested in local knowledge and local languages are geography (including cartography) and cultural astronomy (collecting names for stars and constellations, knowledge and myths about celestial objects, etc.). By documenting endangered languages and assisting in revitalization projects, linguists can also give something back to the community whose knowledge has been explored.

Also reflected in language are social structures within a community and relationships between its members. A classic field of research in cultural anthropology and in anthropological linguistics is kinship and the names for relations between members of a family. As speakers of a European language we may think that the kinship relations expressed by our terms for 'father', 'brother', 'uncle', 'cousin' etc. are basic and should be found anywhere in the world. In fact, kinship terminology varies greatly across cultures, but there is also some order in this variation: anthropologists have identified six basic types of kinship systems that can be found all over the world. They are called each after one of the languages where the type is found: the Hawaiian system, the Eskimo system, the Omaha system, the Crow system, the Iroquois system, and the Sudanese system. The Eskimo system is the one found in many European languages, including modern English. Old English, on the other hand, supposedly belonged to the Sudanese type, where there are different terms for father's and mother's brother (all "uncles" in the Eskimo type), and different terms for their children (all "cousins" in the Eskimo type). The system with the least terminological differentiation is the [Hawaiian](#) system: cousins are referred to by the same term as brothers and sisters (in Hawaiian *kaikua'hine* for girls and *kaikua'ana* for boys), and in the parent generation there is one term for mothers and aunts (Hawaiian *makuahine*), and one for fathers and uncles (Hawaiian *makuakane*).

## LEARN MORE

Learn more about kin terms at one of these sites (tutorials for students of anthropology):

- [Kinship Terminologies](#) by Brian Schwiemmer, University of Manitoba; includes examples from real languages;
- [Kinship. An introduction to Descent Systems and Family Organization](#), by Dennis O'Neil, Palomar College, San Marcos, California

Despite their diversity, kinship terms are relatively easy to classify and to compare across languages because they are based on some simple parameters: relation by blood or by marriage, generation, age, and gender. Kinship terms not only reflect the kin system of a given community, they also maintain the systems. Children acquire these words at a young age and develop an awareness of the relation named by the terms.



## ACTIVITY

Watch a clip about kinship and the importance of family in an Australian aboriginal culture (the Yolngu people of Arnhem Land) at the site [Twelve Canoes](#) (when the coloured rectangles appear, choose the topic “kinship”).

Families are important in all human societies. But what exactly is a family? The meaning of the word for ‘family’ may be different both with respect to who belongs to a family and how important certain family ties are for individuals. One doesn’t have to go far away to study such differences, they can be observed even in neighbouring European countries such as Poland and Germany. There are many differences between the meaning of the Polish word *rodzina* and the German word *Familie*, although these words are treated as translation equivalents in dictionaries and in texts. In general, to Poles ‘family’ includes more persons and is of more importance for an individual than it is for Germans. An anecdote says that in international language courses where each participant has to speak about the topic “my family”, Polish participants always take the longest time. In German a grown up person may say “Ich habe keine Familie” (‘I don’t have a family’), meaning that he or she is not married and doesn’t have children. Translated into Polish, this sentence would imply that the speaker is completely alone in the world, without parents, sisters, brothers, without a single uncle, aunt, or cousin. Also other social relations such as friendship have different meanings in different cultures, which may be reflected in the vocabulary or in the meaning and use of individual words. In Polish there are two words for ‘friend’, *przyjaciel* (female *przyjaciółka*) and *kolega* (female *koleżanka*), and for someone learning Polish as a second language it is not easy to find out to which category their new Polish friends belong. There are also differences within one language, when it is spoken in different societies: the English word *friend* has different meanings in England, the USA, or Australia.

[Anna Wierzbicka](#), a Polish linguist who has lived and worked in Australia since 1972, has discussed these issues in several of her works. She emphasizes the importance of a careful study of word meaning for comparative sociological and psychological studies:

“Consider, for example, the following question: how do patterns of friendship differ across cultures? One standard approach to this question is to use broad sociological surveys based on questionnaires, in which respondents are asked, for example: How many friends do you have? How many of them are male and how many female? How often, on average, do you see your friends? [...] The procedure seems straightforward – except for one small point: if the question is asked in Russian, or in Japanese, what word will be used for *friend*? The assumption behind such questionnaires, or behind comparative studies based on them, is that, for example, Russian, Japanese, and English words for *friend* can be matched. This assumption is linguistically naive and the results based on it are bound to present a distorted picture of reality.” (Goddard & Wierzbicka 1995: 59)

Another example Wierzbicka gives are emotions. She argues that the labels for emotions (such as English *fear*, *anger*, *happiness*) are language specific and culture specific and as a rule do not have exact matches in other languages. From her own experience as a bilingual she reports that even after having lived for a long time in an English speaking community in Australia, she still categorizes (and experiences) her feelings with Polish expressions such as *żal* and *przykro*, terms that only roughly may be translated by English *sorry*, but include various nuances specific to Polish culture (Wierzbicka 2001).

Some words have a special significance in a given culture, as they label concepts that are important parts of this culture. Such words have been called (by Wierzbicka and others) “keywords of culture”. The detailed study of these keywords, their meaning and use, gives us insights into different cultures – also our own.

## STUDY QUESTION

What could be a “keyword of culture” in your native language? Why? What is special about it – what does it include that is not easily translated into other languages?

In the previous section, when discussing colour terms or different words for ‘wood’, or for ‘carry’, we rejected the thesis that lexical categories (the meaning of words) determine the way people perceive the world and think about it. In this section we argued that the vocabulary of a language reflects the natural and social environment of a speech community. These two approaches may be reconciled, as in the quote from Goddard & Wierzbicka, to which we would do well to subscribe:

“Culture-specific words are conceptual tools which reflect a society’s past experience of doing, and thinking about things in certain ways; and they help to perpetuate these ways. As a society changes, these tools, too, may be gradually modified and discarded. In that sense the outlook of a society is never wholly ‘determined’ by its stock of conceptual tools, but it is clearly influenced by them. Similarly, the outlook of an individual is never fully ‘determined’ by the conceptual tools provided by his or her native language,

because there are always alternative ways of expressing oneself, but one's conceptual perspective on life is clearly influenced by his or her native language." (Goddard & Wierzbicka 1995: 58).

## LET'S REVISE! – CHAPTER 2

Go to the [Let's Revise section](#) to see what you can learn from this chapter or test how much you have already learnt!

### Notes

- [3] Samoan data by courtesy of Ulrike Mosel.
- [4] Source: ꞥAkhoe Hai||om word list (2010). The letters ꞥ, ||, and ! represent clicks (see [Chapter 4: The Sounds of Language](#))
- [5] The Wikipedia entry "Distinction of blue and green in various languages" discusses many examples. The map number 134A in WALS shows the distribution of different ways to express 'green' – by a separate word, by a term for both 'blue' and 'green', and other possibilities.
- [6] Readers interested in this topic are referred to the first chapter in Taylor (1995) and to the [World Color Survey](#) at the University of Berkeley. See also the [WALS Chapter on Basic Colour Categories](#) by Kay & Maffi 2013.
- [7] Baka spoken in Gabon, Source: Paulin 2010: 294.
- [8] "The Coral Triangle is a geographical term so named as it refers to a roughly triangular area of the tropical marine waters of Indonesia, Malaysia, Papua New Guinea, Philippines, Solomon Islands and Timor-Leste that contain at least 500 species of reef-building corals in each ecoregion." Wikipedia, "Coral Triangle", accessed 30.08.2014.

### References & further reading

- Bartmiński, Jerzy. 2006. *Językowe podstawy obrazu świata*. Lublin: Wydawnictwo Uniwersytetu Marii Curie –Skłodowskiej.
- Brown, Cecil H. 2013a. Finger and hand. In: Dryer, Matthew & Haspelmath, Martin (eds.), *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library, chapter 130. Available online at <http://wals.info/chapter/130>.
- Brown, Cecil H. 2013b. Hand and arm. In: Dryer, Matthew & Haspelmath, Martin (eds.), *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library, chapter 129. Available online at <http://wals.info/chapter/129>.
- Dahl, Östen & Velupillai, Viveka. 2013. The Past Tense. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library, chapter 66. Available online at <http://wals.info/chapter/66>.
- Dryer, Matthew & Haspelmath, Martin, eds. 2013. *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library. Available online at <http://wals.info>.
- Goddard, Cliff & Wierzbicka, Anna. 1995. *Key words, culture and cognition*. Philosophica 55: 37-67.
- Kay, Paul & Maffi, Luisa. 2013. Number of non-derived basic colour categories. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library, chapter 132. Available online at <http://wals.info/chapter/132>.
- Klein, Wolfgang. 1995. *Das Vermächtnis der Geschichte, der Müll der Vergangenheit*. Zeitschrift für Literaturwissenschaft und Linguistik 100: 77-101.
- Nettle, Daniel & Suzanne Romaine. 2000. *Vanishing voices. The extinction of the world's languages*. Oxford: Oxford University Press.
- Pawley, Andrew. 2011. What does it take to make an ethnographic dictionary? On the treatment of fish and tree names in dictionaries of Oceanic languages. In: G. Haig et al., eds. *Documenting endangered languages*. Berlin: de Gruyter Mouton, 263-287.
- Paulin, Pascale. 2010. *Les Baka du Gabon dans une dynamique de transformations culturelles. Perspective linguistiques et anthropologiques*. Thèse du doctorat, Université Lumière Lyon 2. [available on-line at: [http://www.ddl.issh-lyon.cnrs.fr/fulltext/Paulin/Paulin\\_2010\\_These.pdf](http://www.ddl.issh-lyon.cnrs.fr/fulltext/Paulin/Paulin_2010_These.pdf)]
- Taylor, John R. 1995. *Linguistic categorization*. Oxford: Oxford University Press.
- Wierzbicka, Anna. 1997. *Understanding cultures through their key words: English, Russian, Polish, German, and Japanese*. New York: Oxford University Press.
- Wierzbicka, Anna. 2001. A culturally salient Polish emotion: *przykro*. In: J. Harkins & A. Wierzbicka, eds. *Emotions in crosslinguistic perspective*. Berlin, New York: Mouton de Gruyter.
- ꞥAkhoe Hai||om word list. Based on Tertu Heikkinen's wordlist. Adapted by Thomas Widlok, Christian Rapold and Gertie Hoymann. Expanded by Gertie Hoymann. Work in Progress. March 2010. [available online at the [DoBeS archive](#), node: ꞥAkhoe Hai||om -> DoBeS -> Unsorted Sessions -> Akhoe Word List.]

[back to top](#)



# Language structures

Home > Book of Knowledge > Language structures

## ■ CHAPTER AUTHOR: NICOLE NAU

### Chapter contents:

Words in texts

Words in spoken language

The inner structure of words

- Word-formation
- Techniques for building words and word-forms

Grammatical categories

- Person
- Gender
- Classifiers

Word order

Polar questions

How to express possession

How to show the structure of words and clauses

Notes

References & further reading

Languages differ in the way meaningful elements are put together: how words are made up and how they are combined into sentences. The branch of linguistics that studies the formal make-up of words is called **morphology**, while the combination of words into phrases and of phrases into clauses is studied in **syntax**. This chapter considers the structural diversity found in the languages of the world and introduces some basic terms and techniques for its description.

## ■ WORDS IN TEXTS

Texts consist of words, but in order to understand or produce a text in any given language one has to know more than the meaning of individual words. Words may appear in different forms according to their function in a clause, and various techniques are used to combine word-forms. We cannot translate a text from one language into another by translating word after word. Languages differ with respect to how much and what kind of information can (or must) be packed into one word. Therefore, the number of words used to express one and the same meaning varies greatly across languages. Compare the headline of the Universal Declaration of Human Rights in Estonian and Tok Pisin! One can see at once that these languages use very different techniques. Differences can also be observed in closely related languages, such as German and Dutch.

Language	'Universal declaration of human rights' [1]	Word count
<b>Estonian</b>	Inimõiguste ülddeklaratsioon	2 words
<b>Tok Pisin</b>	Toksavə long ol raits bilong ol manmeri long olgeta hap bilong dispel giraun	13 words
<b>German</b>	Allgemeine Erklärung der Menschenrechte	4 words
<b>Dutch</b>	Universele verklaring van de rechten van de mens	8 words

Analyzing these examples we find two reasons for the different number of words. On the one hand, what are separate words in one language may be expressed by a **compound** in another language: 'human rights' in **German** is *Menschenrechte*, composed of *Menschen* 'people' and *Rechte* 'rights'. **Estonian** likewise combines the nouns *inimene* (root *inim-*) 'man, human being' and *õigused* 'rights' into one word. The meaning 'universal declaration' is also expressed by a compound in Estonian (*ülddeklaratsioon*). New words can also be made by **derivation** – for example, the **English** adjective *universal* is derived from the noun *universe* (more examples for compounding and derivation will be given below).

On the other hand, languages may use more or less **function words** – small words that are used to link words or phrases together, or to express grammatical meaning, for example plural or definiteness. In the **Dutch** example, the preposition *van* 'of' has such a linking function, and *de* is the definite article: *de rechten van de mens* literally translates "the rights of the people". In **Tok Pisin** the equivalent of Dutch *van* or English *of* is the word *bilong*. The Tok Pisin word *ol* marks the plural, for example: *buk* 'book' – *ol buk* 'books'. This plural marker is also found in the example above: *ol raits bilong ol manmeri* 'rights of people' = 'human rights'. Words that express concepts like 'man', 'rights', 'declare', 'speak', 'universal', on the other hand, are called **content words**.

Broadly speaking, there are two possibilities for expressing grammatical meaning: by separate function words (as Dutch *van*, *de*, English *of*, *the*, Tok Pisin *ol*, *bilong*) or by **inflection** of content words, that is by changing their form, for example by adding or subtracting an ending. Polish and Hungarian are examples of languages that mainly use inflection.

## BOOK OF KNOWLEDGE

Chapters: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#)

[List of all languages](#) referred to in the Book of Knowledge and other sections of the website.

[Glossary](#)

**DOWNLOAD** and print out the Book of Knowledge.

## LET'S REVISE! – CHAPTER 3

Go to the [Let's Revise section](#) to see what you can learn from this chapter or test how much you have already learnt!

	'Universal declaration of human rights'
Polish	Powszechna deklaracja praw człowieka
Hungarian	Az emberi jogok egyetemes nyilatkozata

The Polish example contains four content words ('universal', 'declaration', 'rights', 'man'), each in an inflected form that shows its function in the phrase and its relation to other words. The last word, *człowieka*, is the genitive singular of *człowiek* 'man', formed by adding the suffix – *a* to the basic form: *człowiek-a* 'of (the) man'. The word-form *praw* is the genitive plural of *prawo* 'right' (here, the basic form *praw-o* has an ending, while the genitive plural form does not): *praw* 'of (the) rights', *praw człowieka* = 'of human rights', equivalent to Dutch *van de rechten van de mens*.

In the Hungarian version we find one function word (the definite article *az*) and four content words 'human', 'rights', 'universal', 'declaration'. Plural is marked by a suffix on the noun: *jog-ok* 'rights' (*jog* 'right'). The relation between 'declaration' and 'human rights' is marked on the word *nyilatkozat-a* 'declaration' by the suffix *-a*. (See below: [Possession](#))

Not only content words, but function words may also be inflected. This can be seen in German: in the phrase (*Erklärung*) **der** *Menschenrechte* '(declaration) **of the** human rights' the word-form *der* is the genitive plural form of the definite article; in the nominative it has the form *die*. Compare: **die** *Menschen* 'the humans' / **der** *Menschen* 'of the humans', **die** *Rechte* 'the rights' / **der** *Rechte* 'of the rights', **die** *Menschenrechte* 'the human rights' / **der** *Menschenrechte* 'of the human rights'.

## EXERCISE

Compare translations of Universal Declaration of Human Rights, find function words in different languages! View or download this exercise [here](#).

## ■ WORDS IN SPOKEN LANGUAGE

For texts written with a modern alphabetic script, words can be defined easily: a word is a string of letters separated from other strings by spaces or punctuation marks. More precisely such a string is called a **graphic word**. Some doubts may arise in cases such as English *I'm* (= *I am*), *I've* (= *I have*) or *he's* (= *he is*). In speech, however, the doubtful cases are often more numerous than the clear cases, and it is quite difficult to define the borders of words exactly. Determining what should be written as one graphic word and what should be written separately is one of the first difficult tasks when developing an orthography for a formerly unwritten language (see also: [Writing](#)), and even for languages with a long tradition of writing there often remain points of disagreement. The notion of word and the distinction made above between function words and inflection are idealizations that work best for written language.

## STUDY QUESTION

Find cases in your native language (or another language you know well) where the boundary of words is not clear.

Function words are often short and only weakly stressed and therefore tend to fuse with a neighbouring word. An element that always fuses in that way with another word without really becoming part of it is called a **clitic**. Examples from English are <'s>, <'ve> and <'m>. Clitics differ from suffixes in that they can attach to different words – suffixes are usually specialized for a part of speech (verb, noun, or adjective) and have a fixed place. However, there is no fixed boundary: in the course of time a function word may become a clitic (for example English <am> becomes <'m>) and a clitic may become a suffix. Consider the following example from Polish (see [below](#) for the technique used in these and following examples and the meaning of the abbreviations):

*śpiewa-ł-a*= **by** '(she) would sing'  
sing-pst-f=cond

*gdy*= **by** *śpiewa-ł-a* 'If she sang...'  
if=cond sing-pst-f

The past tense marker *-ł-* and the feminine marker *-a-* have a fixed place in the verb-form, while the conditional marker *-by* can attach to the verb-form or to a conjunction. In earlier times this element was a function word, now it is a clitic on the way to becoming a suffix.

## ■ THE INNER STRUCTURE OF WORDS

Words may contain several components. The smallest units of a word that bear a meaning are called **morphemes**. “Meaning” here includes grammatical meaning. For example, the word-form *rights* contains two morphemes, the root (*right*) and the plural marker (-s). The word *unbelievable* contains three meaningful elements (morphemes): the prefix *un* (negation), the root *believ(e)*, and the suffix *able* (element for building adjectives from verbs). A morpheme can have different forms of expression. For example the forms < *un* > (in *unbelievable*), < *in* > (in *incredible*) and < *im* > (in *impossible*) are different forms of the same morpheme. The concrete spoken or written form of a morpheme is called a **morph**. Thus it would be more correct to say: the word-form < *rights* > contains two morphemes that are expressed by the two morphs < *right* > and < *s* >. If a morpheme has several forms, they are called **allomorphs**: in written English the prefixes < *in* > and < *im* > are allomorphs of one morpheme. On the other hand, the < *in* > in < *incredible* > and the < *in* > in < *income* > are morphs that belong to different morphemes because they express different meanings.

The following types of components of a word are commonly distinguished:


<b>root</b>	a morph(eme) bearing a lexical meaning ( <b>right-s</b> , <b>un-believ-able</b> )
<b>suffix</b>	a morph that follows a root ( <b>right-s</b> , <b>unbeliev-able</b> )
<b>prefix</b>	a morph that precedes a root ( <b>un</b> -believable)
<b>affix</b>	a cover term for suffix, prefix, prefix etc. (see below)
<b>(inflectional) ending</b>	the last suffix that expresses a grammatical meaning ( <b>right-s</b> , <b>untouchable-s</b> )
<b>stem</b>	the part of a word to which an inflectional ending is attached, if there is one; a stem may contain only the root ( <b>right-</b> ), a root plus one or more affixes ( <b>untouchable-</b> ), or more than one root with or without affixes (German <i>Menschenrecht</i> -, Estonian <i>inimõigus</i> -)

## Word-formation

The word *word* is ambiguous: it may refer to a certain form that is part of a spoken or written text (for example, if we count the words of a text), or it may refer to a more abstract unit of meaning and form (for example, when we say that < *book* > and < *books* > are forms of the same word). A technical term for “word” with the second meaning is **lexeme**, but what we see in a text are **word-forms**. The building of word-forms that belong to one lexeme is called **inflection**. The building of words in the sense of lexeme is called **word-formation**.

One way of building words (lexemes) is by **compounding**. A compound is the combination of two (sometimes more) roots in one word. We saw above German *Menschenrechte* and Estonian *inimõigused* ‘human rights’. Some compounds can also be found in the Tok Pisin example: *toksave* ‘declaration’ contains the roots *tok* ‘speak’ and *save* ‘know’; *manmeri* ‘people’ consists of *man* ‘man’ and *meri* ‘woman’. Further examples:

Language	compound	meaning	components
Teop	beiko moon	‘girl’	beiko ‘child’ + moon ‘woman’
Logba	iwónḑú	‘honey’	iwó ‘bee’ + nḑú ‘water’
Sheko	bōw kuṭṣu	‘palm’ (of the hand)	bōw ‘belly’ + kúṭṣú ‘hand’
	yārb suku	‘vein’	yārbm̃ ‘blood’ + súkú ‘rope’
	ṣūbū bambù	‘grave’	ṣūbū ‘death’ + bambù ‘pit’

 Go to the [Interactive Map](#) and try exercises on compounds in Wilamowicean.

In many languages words (lexemes) are built by adding suffixes or prefixes to a root or to a stem. This way of building words is called **derivation**. For example:

Hungarian	<i>ember</i> ‘man’ (noun) -> <i>ember-I</i> ‘human’ (adjective)
Polish	<i>prawo</i> (root <i>praw-</i> ) ‘right; law’ -> <i>praw-nik</i> ‘lawyer’
Dutch	<i>verklar-</i> ‘declare’ (verb stem) -> <i>verklar-ing</i> ‘declaration’ (noun)

More examples and other techniques of derivation will be presented below.

Word-formation can be a “shortcut” to express a meaning which otherwise had to be described using several words. For example, the meaning of the **Chocław** word *tononoli*, derived from *tonoli* ‘to roll’, is expressed in English as *to roll back and forth*. On the other hand, in the Tok Pisin text of the Universal Declaration of Human Rights, the meaning ‘universal’ is expressed as *long olgeta hap bilong dispel giraun*, literally ‘at all places of this world’.

## Techniques for building words and word-forms

Several formal means are used in inflection and derivation. Most widespread is the use of **affixes**, especially **suffixes** (elements that follow a root) and **prefixes** (elements that precede a root), for example:

Language	base	forms with suffix, prefix, or both
Sheko	<i>íík</i> 'be old' (verb)	<i>ííkńs</i> 'old' (adjective)
	<i>sūb</i> 'be red' (verb)	<i>sūbńs</i> 'red' (adjective)
	<i>ʒááz</i> 'be good' (verb)	<i>ʒéénf</i> 'good' (adjective)
Puma	<i>khim</i> 'house'	<i>un̄khim</i> 'my house', <i>kakhim</i> 'your (sg.) house'
Logba	<i>gbla</i> 'teach'	<i>ɔgblawo</i> 'teacher'
	<i>ʒ</i> 'sell'	<i>ɔʒwo</i> 'seller'
		Note: ɔ- is a prefix and -wo is a suffix

Another type of affix is the **infix**, which is inserted into a root, for example:


Language	base	forms with infix
Mlabri	<i>peelh</i> 'to sweep the floor'	<i>prneelh</i> 'broom'
	<i>tɛk</i> 'to hit'	<i>trnɛk</i> 'hammer'
	<i>chrɛɛt</i> 'to comb'	<i>chnrɛɛt</i> 'comb'
Lakhota	<i>máni</i> 'he sings'	<i>mawáni</i> 'I sing'
	<i>aphé</i> 'he hits'	<i>awáphe</i> 'I hit'
	<i>hoxpé</i> 'he coughs'	<i>howáxpe</i> 'I cough'

While an infix splits the base, a **transfix** (also called **confix**) is itself split into parts that are inserted into the root. This kind of morphological process is found in Semitic languages ([Arabic](#), [Hebrew](#)). Various vowels are inserted into a root of consonants, sometimes prefixes or suffixes are added. For example, in Egyptian Arabic the root meaning 'write' is *k-t-b*, and examples of word-forms are *k a t a b* 'he wrote', *k i' t a a b* 'book', *mak 't a b a* 'bookshop', *mak 't u u b* 'written' (Bauer 1988: 25).

**Reduplication** is the repetition of a word or parts of a word. This technique is very widespread in the languages of the world. In Europe it is rare, but it is found, for example, in some Latin verbs that build the perfect stem by reduplication of the first part of the word. Reduplication may also affect a part from the middle of a word, as in the examples from Choctaw below. Data from languages from all over the world are collected in the [Graz Database on Reduplication](#). You can find more information on reduplication in the languages of the world in [Chapter 27 of the WALS](#) (World Atlas of Linguistic Structures).

Language	form without reduplication	form with reduplication
Latin	<i>curr-o</i> 'I run' <i>tend-o</i> 'I span' <i>pung-o</i> 'I sting'	<i>cucurr-i</i> 'I have run' <i>tetend-i</i> 'I have spanned' <i>pupung-i</i> 'I have stung'
Choctaw	<i>tonoli</i> 'to roll' <i>binili</i> 'to sit'	<i>tononoli</i> 'to roll back and forth' <i>bininili</i> 'to rise up and sit down'

Amele	ana 'where' me 'good' ʔela 'long' dahing 'ears' eben 'hands' gasuena 'he searches'	anaana 'wherever' meme 'very good' ʔeʔela 'very long' dadahing 'the ears of everyone' ebeben 'the hands of everyone' gasu-gisu-ena 'he searches repeatedly'
-------	---	--


 Go to the [Interactive Map](#) and try exercises: on reduplication in [Totoli](#) and in [Teop](#).

Another technique used in inflection and derivation is the modification of a stem. This includes the following phenomena:

- **ablaut**, where a vowel changes within the base: English *man* – *men*, *sing* – *sung*, *sing* – *song*;
- **consonant alternation** at the end or the beginning of a stem (English *believe* (verb) – *belief* (noun)). Consonant alternation at the beginning of a stem is typical for [Celtic languages](#), for example [Welsh](#) *cartref* 'home' – *gartref* 'at home';
- change of stress: English *'import* (noun) – *im'port* (verb);
- change of tone in tonal languages. Look at the examples from Logba and Sheko, where tones (in writing indicated by accents) mark grammatical categories such as tense or person:

Language	example	
Logba	Matúkǐ ubón adzísíadzí. Matukǐ ubón adzísíadzí.	'I was going to farm every day.' 'I go to farm every day.'
Sheko	M̩baadúra hadũfũ. M̩baadúra hádũfũ.	'Did you hit my younger brother?' 'Did he hit our younger brother?'

Two or more techniques may be combined, for example suffixation plus consonant alternation plus vowel alternation. In Polish the nominative singular of the word meaning 'wood' is *las*, pronounced [las]. The locative form is built by adding the suffix -e, changing the vowel from [a] to [ɛ] and the final consonant from [s] to [ʃ]: *w lesie* [ɛʃɛ] 'in the wood'.

 Go to the [Interactive Map](#) and try exercises on consonant alternation in [Celtic languages](#).

## ■ GRAMMATICAL CATEGORIES

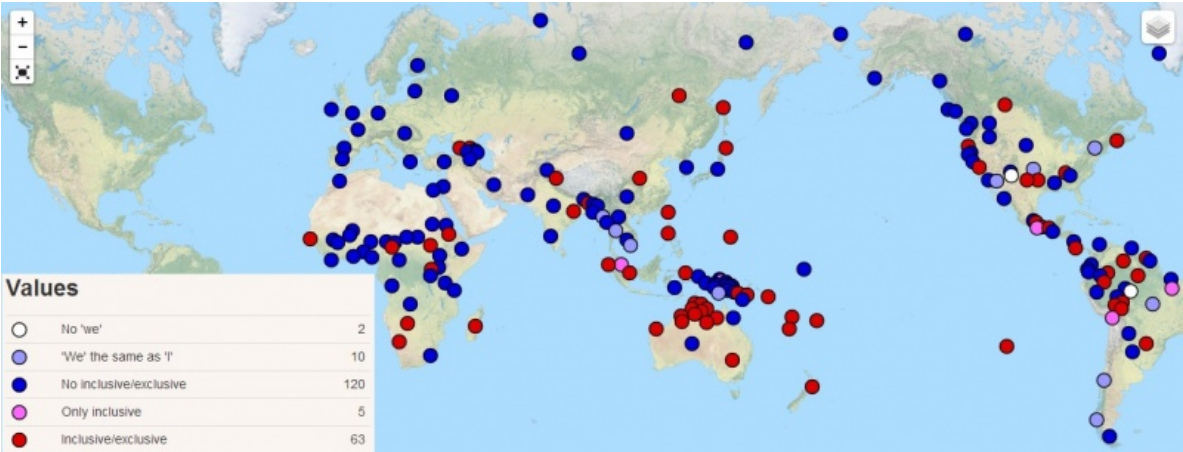
Function words, affixes and the other techniques for building word-forms described above are used to express grammatical categories, such as number (singular, plural), tense (past, present, future), or case (nominative, accusative, dative). One of the characteristics of grammatical categories is their obligatory use in a certain language. For example, in many European languages nouns are always marked for number: we either talk about (a/the) *book* or *books*. In many languages of Australia or North America, on the other hand, nouns are not marked for number, or only some nouns are. For example in [Warrgamay](#) (Australia), *ŋulmburru* may mean 'woman' or 'women' (Corbett 2001: 84, citing Dixon 1980).

Another example for a category that is obligatory in some languages but not in others is **definiteness**. If we want to translate the Polish sentence *Kupiłam książkę* into English, we have to decide between *I bought a book* and *I bought the book*. This shows us that in English definiteness (the difference between *a book* and *the book*) is a grammatical category, while in Polish it is not. We may say in Polish (and in English) things like *Kupiłam tę książkę* 'I bought that book' or *Kupiłam jakąś książkę* 'I bought some book', but this is not the same as the obligatory and regular choice between definite and indefinite article in languages like English or German. On the other hand the Polish sentence tells us that the "I" who bought the book is female – a man would have said *Kupiłem książkę*. This is because **gender** (masculine, feminine, neuter) is a grammatical category in Polish that has to be marked in past tense forms of verbs (as well as on adjectives and pronouns), while in English gender is not a grammatical category. If we want to translate the English sentence *I bought a book* into Polish, we have to know whether the speaker is a man or a woman in order to build the verb form – there is an obligatory choice between -a- and -e- in the frame *kupił\_m* 'I bought'. In English gender is important only in the choice of the third person pronoun: **He** bought books vs. **She** bought books. Hungarian in turn does not make this distinction, both these sentences are translated as (ő) *vett könyveket*, where *ő* may mean 'he' or 'she' (and it is not necessary in this sentence).

There are many different grammatical categories of which individual languages make their choice. Nevertheless, we find the same categories in many languages all over the world. The most widespread categories are: person, number, gender, definiteness, case, tense, aspect (e.g. perfective, imperfective, continuous), mood (e.g. imperative, conditional), voice (e.g. active, passive), and some others that are not known in western European languages. For each category, there is again a limited choice of options. For example regarding **number**, most languages distinguish between singular (one) and plural (more than one), but some languages make more distinctions: singular (one) – dual (two) – plural (more than two), or singular (one) – paucal (few) – plural (many). We will now look in more detail at two of these categories: person and gender.

Person

The category of **person** is concerned with participation in the speech act. It is formally marked in personal pronouns (*I, you, we ...*) and/or in personal forms of verbs, for example in Polish: *kocham* 'I love', *kochasz* 'you (sg.) love', *kocha* 'he/she loves', *kochamy* 'we love', *kochacie* 'you (pl.) love', *kochają* 'they love'. Most often we find a threefold distinction, called **first person** = the speaker (I), **second person** = the addressee (you), and **third person** = other persons or things not participating in the speech act (he, she, it, they). This system is often combined with a twofold number distinction (singular vs. plural) so that, for example, 'we' is defined as first person plural. However, this is not precise – *we* is not the plural of *I* in the way *trees* is the plural of *tree*. It usually does not refer to several speakers but to a combination of the speaker with someone else, either with a second person or a third person. In English the question *Will we meet again?* may mean 'will **you and me** meet again?' (first + second person) or 'will **(s)he and I** meet again?' (first + third person), depending on the context where it is uttered. The first meaning is called **inclusive** (because the addressee is included), the second **exclusive**. Many languages distinguish these meanings by different pronouns. For example in [Chamorro](#) there are two words for 'we': *ta* (inclusive, i.e. 'you and I') and *in* (exclusive, 'I and he/she'). Read more about this distinction in [WALS, Chapter 39](#) and [40](#) (Cysouw 2013a, 2013b).



Inclusive/Exclusive Distinction in Independent Pronouns (Cysouw 2013a)

If the inclusive vs. exclusive distinction is combined with a number distinction of singular vs. dual vs. plural, we get still more possibilities. Compare the following sentences in [Puma](#), which are different ways to say 'we eat rice', depending on whether the addressee is included or not and whether two or more people are referred to:

'We eat rice'	where 'we' =	category label
<i>keci roŋ caci</i>	'you (sg.) and I'	dual inclusive
<i>ke roŋ cee</i>	'you and I and at least one other person'	plural inclusive
<i>kecika roŋ cacika</i>	'(s)he and I'	dual exclusive
<i>keka roŋ ceeka</i>	'they and I'	plural exclusive

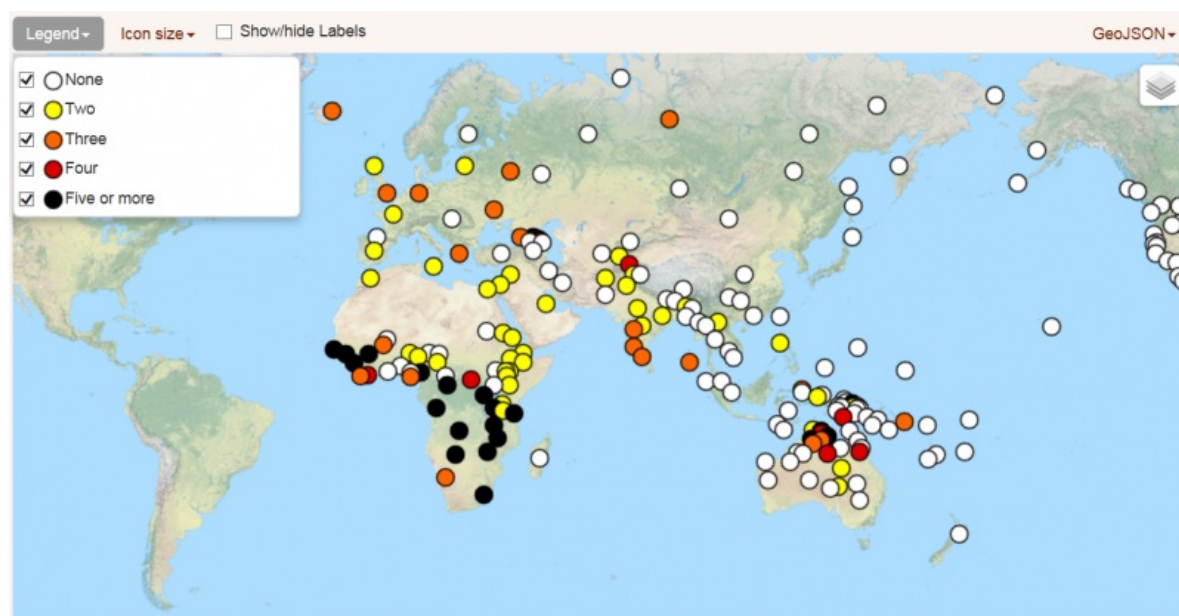
\*Note: the first word is the personal pronoun ('we'), the last word is the verb 'eat' inflected for person.

 Go to the [Interactive Map](#) and try the exercise no. 4 for [Daakaka](#) (Vanuatu).

Gender (noun class)

In European languages gender most often is manifest as a distinction between masculine and feminine, or between masculine, feminine and neuter. In other parts of the world we also find languages with four and five genders. The Nigerian language [Fula](#) even distinguishes twenty classes. On the other hand, many languages of the world do not distinguish noun classes at all. In [The World Atlas of Language Structure](#) (Corbett 2013), we find the following information: In a sample of 257 languages from all over the world, 145 did not have gender as a grammatical category (as Hungarian and English), 50 languages had two genders (in Europe for example French and Latvian), 26 languages showed three genders (as do Polish and German), 12 four and 24 five or more. Languages with more than five classes are found most often in Africa, but also in Papua New Guinea and Australia.





Number of genders in the languages of the world (Corbett 2013)

The “idea” behind gender is that nouns belong to different classes and that this is reflected in **agreement**, that is, in the form of other words in the same phrase or clause. The grouping of nouns into such classes (genders) may have a semantic motivation, for example nouns denoting human beings belong to one class, nouns denoting trees belong to another class and so on. The traditional terms “gender”, “masculine”, “feminine” are based on the fact that in European languages nouns denoting human beings and some animals belong to different classes according to the gender of the referent (the word for ‘man’ is masculine and the word for ‘woman’ is feminine). But for most nouns gender assignment lacks semantic motivation, it is a purely grammatical classification. In Polish, for example, the word *książka* ‘book’ belongs to the feminine class, while *czasopismo* ‘journal’ is neuter and *artykuł* ‘article’ is masculine. The following sentences show how adjectives, pronouns and some forms of verbs agree in gender with the respective noun:


**Ten** dobr-**y** artykuł został wydan-**y** już dawno. ‘This good article was published long ago.’

**Ta** dobr-**a** książka został-**a** wydan-**a** już dawno. ‘This good book was published long ago.’

**To** dobr-**e** czasopismo został-**o** wydan-**e** już dawno. ‘This good journal was published long ago.’

“Noun class” is a more neutral term than “gender”, because it avoids association with natural gender (sex). The individual classes may be simply counted as “class 1”, “class 2” and so on. The following examples are from *Yimas*, a language with 10 classes. The suffixes on the words for ‘my’ and ‘big’ show that the words for ‘foot’, ‘basket’ and ‘voice’ belong to different classes, just as in the Polish example the words for ‘article’, ‘book’ and ‘journal’ did:

<i>namtampara</i>	<i>amana</i>	<i>kpa</i>	‘my big foot’
foot	my:class9.sg	big:class9.sg	
<i>antuk</i>	<i>amana-wŋ</i>	<i>kpa-wŋ</i>	‘my loud voice’
voice	my-class10.sg	big-class10.sg	
<i>impran</i>	<i>amana-m</i>	<i>kpa-m</i>	‘my big basket’
basket	my-class7.sg	big-class7.sg	

 Go to the [Interactive Map](#) and try exercises on Noun classes in [Logba](#) (Ghana)

## Classifiers

Another device where a classification of nouns is reflected in grammar is the use of **classifiers**. Classifiers are function words that are used in certain constructions with nouns. A typical construction where classifiers are found is with numerals, as in the following examples from the Austronesian language *Minangkabau*, spoken in Indonesia:

<i>sar-urang</i> one-clf	<i>padusi</i> woman		'one woman'
<i>duo</i> two	<i>ikue</i> CLF	<i>anjiang</i> dog	'two dogs'
<i>tigo</i> three	<i>batang</i> CLF	<i>pituluik</i> pencil	'three pencils'

We may try to imitate this construction as “one person woman”, “two animal dog”, “three object pencil”, but we have to keep in mind that the words *urang*, *ikue*, *batang* in Minangkabau are function words, not nouns.

Several languages from North America use classifiers that point to the object of a verb. In such languages, the verb meaning ‘give’ has different forms depending on what is given. In the examples below from [Cherokee](#) the classifier is put into the sequence *gà-\_\_\_\_\_nèè’a* which means ‘she is giving him’ (the first word in each line is the noun denoting the object – ‘cat’, ‘water’, ‘shirt’):

<i>Wèésa</i>	<i>gà- káà -nèè’a</i>	'She is giving him a cat' ( <i>káà</i> for living beings)
<i>Àma</i>	<i>à- nèèh -nèè’a</i>	'She is giving him water' ( <i>nèèh</i> for liquids)
<i>Àhnàwo</i>	<i>gà- nǒó -nèè’a</i>	'She is giving him a shirt' ( <i>nǒó</i> for flexible objects)

## ■ WORD ORDER

In most languages the order in which words are combined to phrases and clauses is important: there may be only one possibility, or different orders have different meanings. English is very strict in allowing only the order subject – verb – object in transitive clauses. For example, we can only say *He loves me*, but not *\*Loves he me*, *\*Me loves he*, *\*He me loves* etc. (The asterisk indicates that the construction is not grammatically correct, or possible only in very restricted contexts, for example, in a poem). In German, the most important rule for simple declarative sentences is that the verb be the second element of the clause. Both *Er liebt mich* and *Mich liebt er* are grammatically correct sentences. If we add another word, for example *vielleicht* ‘maybe’, we can form the correct sentences *Vielleicht liebt er mich* ‘maybe he loves me’, *Er liebt mich vielleicht* or *Mich liebt er vielleicht*, but not *\*Vielleicht er liebt mich* with the verb in the third position. The difference between the German sentences *Er liebt mich* and *Mich liebt er* is that in the second variant the object (*mich* ‘me’) is emphasized, while the first variant is neutral with respect to emphasis. Many languages use word order in clauses for emphasis, usually together with a distinctive intonation. If a simple shift of word order is not possible, there may be special constructions which allow to place the object at the beginning of the sentence, for example in English *It's me he loves* or *I am the one he loves*. The order which is (most) neutral with respect to emphasis is called “basic word order”.

Linguists have investigated the basic word order of simple sentences and found out that there are certain regularities. In simple sentences with the three elements subject (S), object (O) and finite verb (V), there are six logical possibilities for their order: SOV, SVO, OSV, OVS, VSO, VOS. If the choice were completely random, we would expect that each pattern were found equally often in the languages of the world, but this is not the case. Instead, the two patterns that start with the subject are by far more frequent than the rest. Here is the result of an investigation of 1377 languages ([Dryer 2013a](#); consult this source for more information and example sentences):

Basic order	Example	Number of languages
<b>SOV</b>	Turkish, Saliba	565
<b>SVO</b>	French, Lelemi	488
<b>VSO</b>	Welsh, Maori	95
<b>VOS</b>	Malagasy, Tsotsil	25
<b>OVS</b>	Hixkaryana	11
<b>OSV</b>	Nadëb	4
<b>no dominant order</b>	Hungarian, Nunggubuyu	189
<b>total</b>		<b>1377</b>

From these data we may infer that there is a strong preference to put the subject before the object – only in 40 of the 1377 investigated

languages the object precedes the subject in basic word order (in the patterns VOS, OVS and OSV). Note that in this investigation only nominal subjects and objects were considered (like in *The cat chased the bird*), not those expressed by a pronoun (like in *He chased it*) or a person marker on the finite verb.

## ■ POLAR QUESTIONS

Another function of word order in German is to distinguish declarative sentences from interrogative sentences, or more precisely sentences that express polar questions (those that can be answered by “yes” or “no”). While in declarative sentences the verb is at the second position, in polar questions it is put at the first place, usually followed by the subject: *Liebt er mich?* ‘Does he love me?’, *Liebt er mich vielleicht?* ‘Does he love me, maybe?’, *Liebt er vielleicht mich?* ‘Is it maybe me he loves?’. This technique of marking questions is found mainly in European languages (German, Dutch, Swedish, Czech, Spanish and others), only occasionally in other parts of the world. The different techniques used in polar questions and their occurrence in the languages of the world are described in [WALS, Chapter 116](#) (Dryer 2013b).


The most popular technique for marking questions in the languages of the world is by a question particle, as Polish *czy*. In Polish and many other languages the question particle is placed at the beginning of the sentence, while in other languages it is placed at the end, or after the first word of the clause.

Language	Example	
<b>Polish</b>	<i>Jan kupił książki.</i> <i>Czy Jan kupił książki?</i>	‘Jan bought books.’ ‘Did Jan buy books?’
<b>Maybrat</b>	<i>ana m-amao Kumurkek a</i> 3pl 3-go Kumurkek <b>q</b>	‘Are they going to Kumurkek?’
<b>Mono</b>	<i>Charley =wā?</i> <i>mia-pi</i> Charley = <b>q</b> go-perf	‘Has Charley left?’ (the question marker is a clitic that attaches to the first word)

Some languages mark questions in the verb-form. In this case, it is more common to have a special inflectional form used in questions, while the verb-form of declarative sentences is unmarked. However, in a few languages the opposite constellation is found: declarative sentences contain an obligatory marker (for example, for mood) which interrogative sentences lack. The following examples are from two languages spoken in Ethiopia:

Language	Example	
<b>Zayse</b>	<i>hamá-tte-ten</i> <i>háma-ten</i>	‘I will go’ ‘Will I go?’
<b>Sheko</b>	<i>ŋ-māāk-ā-m</i> <i>ŋ-māāk-ā únà sókú tuurúk'à tɕ'ádɬ kiákə</i> <i>únà sókú tuurúk'à tɕ'ádɬ kìa</i>	‘I will tell’ ‘Shall I tell?’ ‘In the past, <b>there has been</b> war in Sheko.’ ‘In the past, <b>has there been</b> war in Sheko?’

In most languages questions have an intonation different from declarative sentences, and often this is the only feature that distinguishes interrogative sentences from declarative sentences.

 Go to the [Interactive Map](#) and try exercises on [ɛAkhoe](#), exercise 1

## ■ HOW TO EXPRESS POSSESSION

There are several ways to express the same content, even in one language. A good example to show this is **possession**, that is the meaning ‘someone has something’, ‘something belongs to someone’. For example, in English we can say *She has red hair*, or *Her hair is red*, *She is red-haired*, *She is a redhead*... The possibilities vary with the things possessed. We cannot say \* *She is red-cared* for ‘she has a red car’, and we don’t say \* *She owns red hair*, while *She owns a red car* is fine. In most languages of western Europe (including Polish) the construction with a verb meaning ‘have’ is the most basic; it is used with very different kinds of “possession” (compare: *I have a car* / *I have a brother* / *I have time* / *I have a headache* ...).

### Terminology

**possession** the relation between a possessor and a possessum, meaning ‘have’, ‘own’, ‘belong’

<b>possessor</b>	the “owner” in a broad sense: someone or something that has something; in the English sentences John has red hair, John owns two houses, This car belongs to John, and in the phrase John’s father the name John expresses the possessor
<b>possessum</b>	what belongs to someone or something; in the sentences Suzie has a car, This car belong belongs to Mary, My mother’s car is red, the possessum is (a/this) car

In Greenlandic, the verb ‘have’ and the noun expressing the thing possessed are combined in a compound verb:

Language	example	
<b>Greenlandic</b>	<i>angut taanna qimmi-qar-puq</i> man that dog-have-3sg.ind	‘That man has dogs’, literally: “That man dog-owns”

In constructions of the ‘have’-type the possessor is encoded as the subject of the clause. Many languages of the world prefer other constructions. In a sample of 240 languages, only 63 had an equivalent of English *have*, and many of these are spoken in western or central Europe (Stassen 2013, [chapter 117 of WALS](#)). Languages that don’t have a verb meaning ‘have’ usually use a construction with a verb meaning ‘be’. This is therefore called the ‘be’-type. It is more widespread in the languages of the world than the ‘have’-type. Here, it is the possessum that is encoded as the subject, and the verb expresses existence or location. The possessor is expressed in various ways, for example in a dative form, as in Hungarian and Sheko, or marked by a preposition, as in Irish. In other languages the meaning ‘I have a car’ is expressed in a construction that can be translated literally as “a car is with me”, or “my car exists”, or “speaking of me, there is a car”, etc.

Language	Example	
<b>Hungarian</b>	<i>Istvan-ak új autója van.</i> Istvan-dat new car.poss <b>is</b>	‘Istvan has a new car’ (Literally: ‘To Istvan is his new car.’)
<b>Sheko</b>	<i>dādū t’āāgħ ðf-kħ kiákə</i> child two she-dat exists	‘She has two children.’ (Literally: ‘There’s two children to her.’)
<b>Irish</b>	<i>Tá cat beag agam.</i> is cat small at.me <i>Níl madra agam.</i> is.not dog at.me	‘I have a small cat.’ ‘I don’t have a dog’
<b>Avar</b>	<i>dir mašina bugo</i> 1sg.gen car iii-be.pres	‘I have a car’ (Literally: ‘My car is’)
<b>Tondano</b>	<i>si tuama sie wewean</i> anim.sg man top exist <i>wale rua</i> house two	‘The man has two houses’ (Literally: ‘As far as the man is concerned, there are two houses’)

Possession is expressed not only in clauses, but also in phrases like *my car*, *John’s house*, *the father of my friend*. As you can see in these examples, English uses several techniques in such phrases: a different word-form (*I – my*, *he – his*, *we – our*), a clitic attached to the last word of a noun phrase (*John – John’s*, *[the new teacher] – [the new teacher]’s*), or a preposition (*[my friend] – of [my friend]*). In all these cases the relation “possession” is marked at the word or phrase that denotes the possessor, while the possessum is expressed in the basic form (see table 1).

In Hungarian we also find several construction types, but in contrast to English, it is the possessum that always bears the mark of the relation while the possessor may be unmarked, for example: *István könyv-e* ‘István’s book’, *a diák könyv-e* ‘the student’s book’ (see table 2). Recall the example ‘declaration of human rights’ from the beginning of this chapter: in Hungarian the relation between ‘human rights’ and ‘declaration’ is marked by the suffix -a at the word *nyilatkozat* ‘declaration’.

*emberi jogok nyilatkozat-a*  
human rights declaration-POSS

In the case of first and second person, the possessor is expressed in Hungarian as a suffix on the noun denoting the possessum, for example *könyv-em* ‘my book’, *könyv-ed* ‘your book’.

In another Hungarian construction both the possessor and the possessum are marked: *István-ak könyv-e* ‘István’s book’, *a diák-ak a könyv-e* (see table 3).

Table 1: Possession marked at the possessor (English *my car, his house, John's book, the father of my friend, declaration of human rights*)

possessor		possessum	
basic form	as possessor	basic form	as possessed
I	my	car	=
he	his	house	=
John	John's	book	=
my friend	of my friend	the father	=
human rights	of human rights	declaration	=

Table 2: Possession marked at the possessum (Hungarian *István könyve* ‘István’s book’, *a diák könyve* ‘the student’s book, *emberi jogok nyilatkozata* ‘declaration of human rights’; *könyvem* ‘my book’, *könyved* ‘your book’).

possessor		possessum	
basic form	as possessor	basic form	as possessed
István	=	könyv ‘book’	könyv-e
diák ‘student’	=		
emberi jogok ‘human rights’	=	nyilatkozat ‘declaration’	nyilatkozat-a
én ‘I’	-(e)m	könyv	könyv-em
te ‘you (sg.)’	-(e)d		könyv-ed

Table 3: Possession marked at both the possessor and the possessum (Hungarian *Istvának a könyve* ‘István’s book’, *a diáknak a könyve* ‘the student’s book’)

possessor		possessum	
basic form	as possessor	basic form	as possessed
Istvánadiák ‘the student’	István-aka diak-ak	a könyv ‘the book’	a könyv-e

Further examples for the three strategies:

Language	Example	Strategy: marking of...
Chechen	<i>mashie-an maax</i> car-gen price‘The price of a car’	possessor
Yurakaré	<i>shunñe a-pojore</i> man 3sg.p-canoe‘the man’s canoe’  <i>ti-bba</i> ‘my husband’ 1sg-husband	possessum
Southern Sierra Miwok	<i>cuku-ŋ hu:kiʔ-hy</i> dog-gen tail-3sg‘the dog’s tail’	possessor and possessum



## Puma

*uŋ-bo uŋ-khim*  
1sg-gen 1sg-house 'my house'

possessor and possessum

*kenci-bo kenci-khim*  
2dua-gen 2dua-house

'your house' ("the house of you two")

*khokkuci-bo kaci-khim*  
3pl-gen 3pl-house

'their house'

## Asmat

*Warsé ci* 'Warsé's canoe'  
*Warsé* canoe *no cem* 'my house'  
I house

neither possessor nor possessum (rare)

Many languages use different constructions for different kinds of "possession". For example, one construction is used for kinship ('my sister', 'my father') or body parts ('my hair', 'my nose') and another construction for things that can be owned and sold ('my house', 'my book'). The first type is called **inalienable possession**, the second type **alienable possession**. The following examples are from *Saliba*, an Austronesian language spoken in Papua New Guinea (Mosel 1994):

<i>sinagu</i>	my mother	inalienable possession
<i>sinana</i>	his/her mother	
<i>tamana</i>	his/her father	
<i>nimana</i>	his/her hand	
<i>Maui nimana</i>	Maui's hand	
<i>yogu numa</i>	my house	alienable possession
<i>yona numa</i>	his/her house	
<i>Maui yona numa</i>	Maui's house	



Go to the [Interactive Map](#) and try [Daakaka](#) exercises on possession.

## ■ HOW TO SHOW THE STRUCTURE OF WORDS AND CLAUSES

In this chapter, examples from various languages were presented using a technique that is called "interlinear translation" or "morpheme-by-morpheme glossing"; linguists often simply call it "glossing". This technique helps to understand the structure of examples from languages that we don't know. For example, talking about questions in section 4 above, a sentence from the West Papuan language *Maybrat* was shown in this way:

ana m-amao Kumurkek a  
3PL 3-go Kumurkek Q

The glosses in the second line tell us that the first word is a pronoun for third person plural, the second word starts with a prefix that marks third person, followed by a root meaning 'go', the third word is a proper noun, and the last word is a question particle. With this information, we can construct the meaning of the whole sentence. Note that it is not possible to translate each word of the *Maybrat* example by an English word – there is no question particle in English, nor is there a marker for third person (the marker – *s* in *talk-s*, *smile-s* etc. is more specific, it marks third person singular). Glossing is largely independent of the grammatical structure of the language into which we translate. Only lexical roots are translated into this language, but all grammatical information that a word contains is indicated by a grammatical label such as pl for plural, 3 for third person. Grammatical must of course be explained (commonly by giving a list of abbreviations, as below).

The words of the language you want to describe (the object language) are segmented into components (morphs), separated by hyphens, for example:

*The boy scream-ed and ran quick-ly to his mother.*

Then the meaning of each component is written exactly below the segment. The number of hyphens has to be the same in both lines. If a segment includes more than one meaning, the glosses of this segment are separated by a period in the translation. The lexical roots are

translated by words of the language of the description (the meta-language). Grammatical morphemes, including function words, are translated by grammatical labels; function words may also be translated by a corresponding function word, if there is one. A morpheme-by-morpheme glossing of the above sentence into Polish may thus look like this (note: the function words *and* and *to* could also be translated by a Polish equivalent, *i* and *do*, respectively):

<i>The</i>	<i>boy</i>	<i>scream-ed</i>	<i>and</i>	<i>ran</i>	<i>quick-ly</i>	<i>to</i>	<i>his</i>	<i>mother.</i>
ART	chłopak	krzyczeć-PST	CONJ	biegać.PST	szybko-ADV	PREP	3SG.M.POSS	matka

Glossing is a very effective tool for linguistic description, especially if we want to compare the structures of very different languages. Reading glosses is not difficult, it just needs some training. Here are morpheme-by-morpheme glosses of the examples from the beginning of this chapter – the phrase “universal declaration of human rights” in four European languages:

Estonian

<i>Inim-oigus-te</i>	<i>üld-deklaratsioon</i>
man-right-GEN.PL	general-declaration

German

<i>allgemein-e</i>	<i>Erklär-ung</i>	<i>der</i>	<i>Mensch-en-recht-e</i>
general-PL	declare-NOUN	ART.GEN.PL	man-AFX-right-PL

Polish

<i>powszechn-a</i>	<i>deklaracj-a</i>	<i>praw</i>	<i>człowiek-a</i>
general-NOM.SG.F	declaration-NOM.SG	right.GEN.PL	man-GEN.SG

Hungarian

<i>az</i>	<i>ember-i</i>	<i>jog-ok</i>	<i>egyetemes</i>	<i>nyilatkozat-a</i>
ART	man-ADJ	right-PL	general	declaration-POSS

For details on glossing you may consult the “[Leipzig Glossing Rules](#)”, a standard way of glossing used by many linguists.



Go to the [Interactive Map](#), find [Teop](#) and try reading glosses in exercise 3.

## EXERCISE

Try to make an interlinear translation of a short text fragment (1 paragraph) in your mother tongue, using English as a metalanguage!

## Abbreviations used in the glosses in this chapter

ADJ	adjective (affix for building adjectives)
AFX	affix (unspecified)
ANIM	animate
ART	article
CLF	classifier
COND	conditional
CONJ	conjunction
DAT	dative

DUA	dual
F	feminine
GEN	genitive
IMPF	imperfective (aspect)
IND	indicative (mood)
NOM	nominative
NOUN	affix for building nouns
PERF	perfect
PL	plural
POSS	possession
PREP	preposition
PRES	present tense
PST	past
Q	question particle or affix
SG	singular
TAM	marker of tense, aspect and/or mood
TNS	tense
TOP	topic (what is talked about)

## LET'S REVISE! – CHAPTER 3

Go to the [Let's Revise section](#) to see what you can learn from this chapter or test how much you have already learnt!

### Notes

[1] The text of the Universal Declaration of Human Rights in various languages can be found at <http://www.ohchr.org/EN/UDHR/Pages/Introduction.aspx>

### Further reading

On different language structures

- Dryer, Matthew & Haspelmath, Martin (eds.). 2013. *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library. Available online at <http://wals.info/>
- Dürr, Michael & Schlobinski, Peter. 2006. *Deskriptive Linguistik: Grundlagen und Methoden*. Göttingen: Vandenhoeck und Ruprecht. [Third edition; earlier editions had the title „Einführung in die deskriptive Linguistik“]
- Haspelmath, Martin & Sims, Andrea. 2010. *Understanding morphology*. 2nd edition. London: Hodder Education.
- Payne, Thomas E. 1997. *Describing morphosyntax. A guide for field linguists*. Cambridge: Cambridge University Press.
- Payne, Thomas E. 2006. *Exploring language structure: A student's guide*. New York: Cambridge University Press.

### References

Sources of the language data

**Amele:** [Graz Database on Reduplication](#), citing Roberts 1991; **Asmat:** [Nichols & Bickel 2013](#), citing Voorhoeve 1965; **Avar:** [Stassen 2013](#), citing Kalinina 1993; **Chamorro:** [Cysouw 2013a](#), citing Topping 1973; **Chechen:** [Nichols & Bickel 2013](#); **Cherokee:** Seifart 2010, citing Blankenship 1997; **Choctaw:** [Rubino 2013](#), citing Kimball 1988; **Greenlandic:** [Stassen 2013](#), citing Fortescue 1984; **Irish:** Data provided by Mike Hornsby; **Lakhota:** Albright 2000; **Logba:** Dorvlo 2008; **Maybrat:** [Dryer 2013b](#), citing DoI 1999; **Minangkabau:** [Gil 2013](#); **Mlabri:** Rischel 1999; **Mono:** [Dryer 2013c](#), citing Norris 1986; **Puma:** Sharma et al. (online); **Saliba:** Mosel 1994; **Sheko:** Hellenthal 2010; **Southern Sierra Miwok:** [Nichols & Bickel 2013](#), citing Broadbent 1964; **Teop:** Mosel 2007; **Tondano:** [Stassen 2013](#), citing Sneddon 1975; **Yimas:** Seifart 2010, citing Foley 1991; **Yurakaré:** Van Gijn 2006; **Zayse:** [Dryer 2013b](#), citing Hayward 1990.

- Albright, Adam. 2000. The productivity of infixation in Lakhota. Unpublished paper prepared for publication in *UCLA Working Papers in Linguistics*. Available [online](#).
- Bauer, Laurie. 1988. *Introducing linguistic morphology*. Edinburgh: Edinburgh University Press.
- Blankenship, Barbara. 1997. Classificatory verbs in Cherokee. *Anthropological Linguistics* 39, 92-110.
- Broadbent, Sylvia M. 1964. *The Southern Sierra Miwok language*. University of Chicago Press.
- Corbett, Greville G. 2004. *Number*. Cambridge: Cambridge University Press.
- Corbett, Greville G. 2013. *Number of genders*. In: Dryer, Matthew & Haspelmath, Martin (eds.), *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library, Chapter 30.
- Cysouw, Michael. 2013a. *Inclusive/exclusive distinction in independent pronouns*. In: Dryer, Matthew & Haspelmath, Martin (eds.), *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library, Chapter 39.

- Cysouw, Michael. 2013b. *Inclusive/exclusive distinction in verbal inflection*. In: Dryer, Matthew & Haspelmath, Martin (eds.), *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library, Chapter 40.
- Dixon, Robert M. W. 1980. *The languages of Australia*. Cambridge: Cambridge University Press.
- Dol, Philomena. 1999. *A grammar of Maybrat: A language of the Bird's Head, Irian Jaya, Indonesia*. University of Leiden.
- Dorvlo, Kofi. 2008. *A grammar of Logba (Ikpana)*. Proefschrift, Universiteit Leiden. Available [online](#).
- Dryer, Matthew S. 2013a. *Order of subject, object and verb*. In: Dryer, Matthew & Haspelmath, Martin (eds.), *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library, Chapter 81. /li>
- Dryer, Matthew S. 2013b. *Polar questions*. In: Dryer, Matthew & Haspelmath, Martin (eds.), *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library, Chapter 116.
- Dryer, Matthew S. 2013c. *Position of polar question particles*. In: Dryer, Matthew & Haspelmath, Martin (eds.), *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library, Chapter 92.
- Foley, William A. 1991. *The Yimas language of New Guinea*. Stanford: Stanford University Press.
- Fortescue, Michael. 1984. *West Greenlandic*. Croom Helm.
- Hayward, Richard J. 1990. *Notes on the Zayse Language*. School of Oriental and African Studies, University of London.
- Gil, David. 2013. *Numeral classifiers*. In: Dryer, Matthew & Haspelmath, Martin (eds.), *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library, Chapter 55.
- Graz Database on Reduplication at <http://reduplication.uni-graz.at/redup/> [30.05.2012]
- Hellenthal, Anneke Christine. 2010. *A grammar of Sheko*. Proefschrift, Universiteit Leiden. Available [online](#).
- Kalinina, E. 1993. Sentences with non-verbal predicates in the Sogratl dialect of Avar. In: A. E. Kibrik, ed. *The noun phrase in the Andalal dialect of Avara as spoken at Sogratl*, 90-104. Eurotyp Working Papers.
- Kimball, Geoffrey D. 1988. Koasati reduplication. In: W. Shiplay, ed. *In honor of Mary Haas*, 431-442. Mouton de Gruyter.
- Mosel, Ulrike. 1994. *Saliba*. München: LINCOM.
- Mosel, Ulrike, with Yvonne Thiesen. 2007. *The Teop sketch grammar*. Version 2007. Online publication available at the [DOBES archive](#).
- Nichols, Johanna & Bickel, Balthasar. 2013. *Locus of marking in possessive noun phrases*. In: Dryer, Matthew & Haspelmath, Martin (eds.), *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library, Chapter 24.
- Norris, Evan J. 1986. *A grammar sketch and comparative study of Eastern Mono*. University of California at San Diego.
- Rischel, Jørgen. 1995. *Minor Mlabri. A hunter-gatherer language of Northern Indochina*. Copenhagen: Museum Tusculanum Press.
- Roberts, John R. 1987. *Amele*. Croom Helm.
- Roberts, John R. 1991. Reduplication in Amele. In: T. Dutton, ed. *Papers in Papuan linguistics*, No. 1, 115-146. *Pacific Linguistics*, A-73, 1991.
- Rubino, Carl. 2013. *Reduplication*. In: Dryer, Matthew & Haspelmath, Martin (eds.), *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library, Chapter 27.
- Sharma, Narayan P., Balthasar Bickel, Martin Gaenszle, Arjun Rai, and Vishnu S. Rai. *Personal and possessive pronouns in Puma (Southern Kiranti)*. [Online publication](#).
- Seifart, Frank. 2010. Nominal Classification. *Language and Linguistics Compass* 4/8 (2010): 719-736.
- Sneddon, James N. 1975. *Tondano phonology and grammar*. Australian National University.
- Stassen, Leo. 2013. *Predicative possession*. In: Dryer, Matthew & Haspelmath, Martin (eds.), *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library, Chapter 117.
- Universal Declaration of Human Rights in various languages: <http://www.ohchr.org/EN/UDHR/Pages/Introduction.aspx> [15.05.2012]
- Van Gijn, Erik. 2006. *A grammar of Yurakaré*. Proefschrift (PhD thesis), Radboud Universiteit Nijmegen. Available [online](#).
- Voorhoeve, Clemens L. 1965. *The Flamingo Bay dialect of the Asmat language*. University of Leiden.

[back to top](#)

# The sounds of language

Home > Book of Knowledge > The sounds of language

## ■ CHAPTER AUTHOR: MACIEJ KARPIŃSKI

### Chapter contents:

Different languages, different sounds  
Different sounds, different manner of articulation  
Strange sounds of strange languages  
Sound classes: phonemes  
Tone and intonation  
Accent and rhythm  
Archiving and reconstructing sound systems  
How to transcribe sounds of a language?  
Visible sound  
Less widely used languages and technology

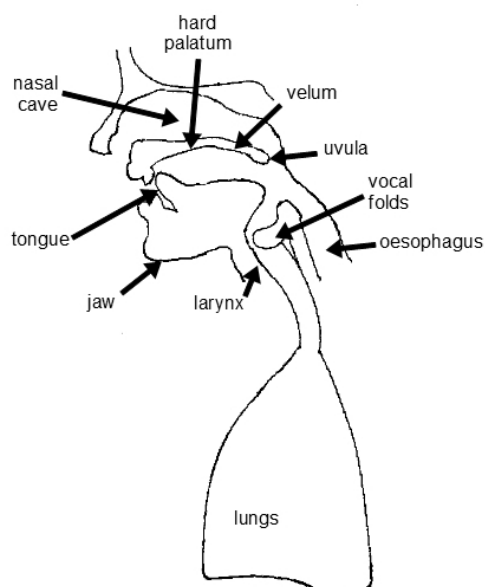
References & further reading

## ■ DIFFERENT LANGUAGES, DIFFERENT SOUNDS

It is often said that languages differ by sound or melody. What does this mean? Only when one begins consciously the process of learning a foreign language, does one notice that the language in question possesses sounds far removed from those in one's own, and not even produced in the same manner. Sometimes, there are also sounds which sound similar, yet prove to be different by a minute, but essential, detail. Those sounds cannot simply be replaced by sounds one knows from their own language. Such a replacement could change the meaning of a word or phrase, or even cause the sentence to become incomprehensible. Correct articulation can prove to be of great difficulty and may require arduous and repetitive practice. Several different sounds may sound the same to a non-native speaker, and at the same time, deceptively similar to a sound from their own mother tongue. Although awareness of such phenomena increases with every new foreign language learnt, only a few realise just how much variety of sound exists in the languages of the world.

## ■ DIFFERENT SOUNDS, DIFFERENT MANNER OF ARTICULATION

The differences in sound are the result of different manners of articulation – the way they are pronounced. Speaking in a native or a very well-known language doesn't require much thought about the positioning of the lips, the tongue or a possible closing of the air flow through the nasal cavity. Fully conscious articulation would be too slow. There are only a few elements whose position can be controlled by conscious will (see the figure below). Nonetheless, it still allows a great variety of sounds to be pronounced. Apart from speech, this can also be observed, perhaps with greater ease, in singing. Each natural language uses but a small part of the great phonetic potential. As such, languages usually consist of only several dozen such sound units, which are then used to build words and utterances.



## BOOK OF KNOWLEDGE

Chapters: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#)

[List of all languages](#) referred to in the Book of Knowledge and other sections of the website.

### [Glossary](#)

**DOWNLOAD** and print out the Book of Knowledge.

## LET'S REVISE! – CHAPTER 4

Go to the [Let's Revise section](#) to see what you can learn from this chapter or test how much you have already learnt!

## PHONETIC EXERCISES

Do you wish for some phonetic practice? Take a look at the exercises in the [Phonetic Exercises section](#).



It is mainly the vocal folds which are responsible for voicing and pitch. The process by which the vocal folds produce certain sounds is called 'phonation'. It is possible to feel them vibrating by placing fingers on one's Adam's apple. Women's vocal folds are of a smaller size than those of men. The rest of the vocal tract, its exact shape differs from person to person, decides on what sound is to be initiated. The changes, be they conscious or automatic, during the articulation process also influence the produced sound. Vowels are the sounds produced with a widely open articulatory tract. If in the process of articulation, an obstruction occurs in the vocal tract (i.e. the tongue touches the palate, the mouth is closed), the produced sound is a consonant. The sounds 'in-between', articulated with a narrowed vocal tract are called approximants (for example the English sounds which are represented in orthography by the letters *and*, and in the phonetic alphabet of the International Phonetic Association by the same symbols /w/ and /j/, (slash brackets mean broad transcription, see below). In a sense, they are a bit like 'incomplete consonants').

### ■ STRANGE SOUNDS OF STRANGE LANGUAGES

'Strange' is of course a term used half-jokingly here and should be understood rather as 'rare'. The readers of this text are probably familiar only with major European languages, thus they might find some of the articulation phenomena or sounds existing in languages of Asia and Africa quite different from their own.

It is often assumed that vowels are always voiced as that is the case in the most commonly spoken languages in Europe. When whispering, however, the vowels of the aforementioned languages can be rendered voiceless, all the while remaining comprehensible and distinguishable. Nevertheless, such cases occur rarely. Languages whose sound systems comprise voiceless vowels include American Indian languages like *Zuni* or *Cheyenne*, or *Japanese*. Laryngealization can be regarded as a particular voice quality. It is realised as a kind of croak, especially at the end of sentences. It appears when the vocal folds vibrate irregularly as a result of low volume-velocity of the air flow. There are, however, languages such as *Kedang* (an Austronesian language spoken in Indonesia) or *Jalapa Mazatec* (spoken in Mexico), where a creaky voice is an important sound quality. Although aspirated consonants are quite popular and exist in many languages of the world, aspirated vowels present in *Jalapa Mazatec* or *Gujarati*, remain quite a rarity.

#### TRY IT YOURSELF

In order to produce an aspirated sound, produce just any consonant breathing out simultaneously. You will find detailed instructions in [this video](#).

And [here](#) is an example of how aspiration works in a specific language.

\*\*\*

In order to produce a laryngalised sound, try to pronounce any vowel continuously for some time, gradually decreasing the speed at which you breath out the air. Finally, you should hear and feel that your vocal folds start to vibrate irregularly and produce this peculiar sound of "vocal fry".

When speaking of rare sounds, clicks are often mentioned. They occur mainly in the languages of Southern Africa, but not exclusively since they are also known to be present in a ceremonial language called Damin, which used to be spoken in Australia.

#### TRY AND CLICK

Watch a [video clip](#) introducing the four click sounds of *Khoekhoegowab*

Try producing these sounds! It may be easy in isolation, but it will prove far more difficult in continued or uninterrupted speech. Undoubtedly, native speakers will not have any difficulty in articulating the clicks, even when singing.

The South African singer Miriam Makeba popularized clicks with a song in *Xhosa*, which originally is called "Qongqothwane" but outside Africa is known as "The Click Song".

#### CLICKS IN XHOSA

Watch a performance of [Miriam Makeba's song](#), with introductory remarks by the singer.

In transliterations based on the Latin alphabet, a click may be transliterated by using an exclamation mark (see examples in the next subchapter).



Go to the [Interactive Map](#) and try exercises for Taa and ǀAkhoe languages.

## ■ SOUND CLASSES: PHONEMES

As is apparent, every language can possess a different set of sounds which, in turn, are differentiated by various features. Traditionally, researching the phonetic inventory of a natural language relies on observing the context where the particular sounds appear and also their influence on the meaning of a given phrase. Thus, the so-called *minimal pairs* are being sought after. *Minimal pairs* are pairs of words which differ from each other only by one single sound, such as Polish *dama* 'lady' and *tama* 'dam'. The same type of differing qualities between phones can be essential in distinguishing sound classes in one language, and completely irrelevant in another. If a change of one phone in a word carries a change in meaning and it changes into another, separate word, it means that this feature is phonologically important. Thus, those two phones belong to different classes – they represent different phonemes. The features that normally make phones sound different are related to the place and manner of articulation or to the features of the sound, if viewed from the acoustic perspective. Voicing is a distinct feature in many languages of the world. In Polish, for example, the words *dama* and *tama* begin with phonemes that differ only by the fact that the former is voiced, and the latter voiceless. It is phonation that decides the meaning of the word. In some languages, the length (duration) of the sound is of high importance. Most often, there are long and short vowels (for example in Hindi). Two words may mean something different because of one vowel's length (such as German *raten* 'to advise' and *Ratten* 'rats', or English *beat* vs. *bit*; however, if you listen to these pairs closely, you may note that there is also a change to the voice quality). There also exist languages in which it is consonant length that is of great importance. In Polish, vowel length is used, amongst other things, to enhance comprehensibility or expressiveness of the sentence (for example, in the question *co?* 'what', the vowel can be lengthened to indicated astonishment: *coooo?*). Longer or shorter segments do not, however, change their lexical meanings simply because of the length (just as *co*, also *coooo* means 'what'; in phonetic alphabets, lengthening is often marked with a colon: *coooo* = *co:*).

Phonological systems are abstract systems, assumed to function in human minds and containing information about the phonemic inventory of a given language. Such a system comprises of the already mentioned phonemes. A phoneme is a unit that can differentiate meaning even though it does not have a meaning on its own. Phonemes do not carry a meaning, unlike words or sentences. The number of phonemes in languages is astonishingly varied, from a dozen or so in Rotokas, Piraha, or Ainu to more than a hundred in Taa (also known as !Xóǀ or !Xuun – the exclamation mark representing one of the clicks), one of the Khoisan languages spoken in South Africa. Irish also possesses a formidable phonetic inventory of 69 phonemes. The usual number, however, is between 20 and 60.

### LEARN MORE

Learn more about phoneme inventories in the languages of the world in the World Atlas of Language Structure at <http://wals.info/chapter/1> and <http://wals.info/chapter/2>

Traditionally, phonemes are grouped into consonant phonemes and vowel phonemes. Nevertheless, there are those that lie somewhere in-between (approximants). Most often, there are many more consonant phonemes than vowel phonemes (such as in Polish or German), though exceptions do exist. One such exception is the Marquesan language (a cluster of East-Central Polynesian dialects) in which the number of consonants and vowels is comparable. Samples of Marquesan are available on the [language documentation project's](#) website. An already extinct language, Ubykh, had eighty two distinct consonants, but only two distinct vowels. Those vowels, however, could be pronounced differently, depending on their phonetic environment. The language that holds the title of the one possessing the largest number of vowels is Taa, whose one dialect distinguishes between thirty one vowels.

Linguists believe in a complicated system, residing in the mind, which stores all the information about units of a language. It is called the 'mental lexicon', but its functions and structure differ markedly from an ordinary dictionary. One may assume that every known word is stored in the mental lexicon as a phonemic transcription – a sequence of phonemes. By pronouncing it, all the particular phonemes are being realised. They are transformed from being abstract ideas into concrete, physical and perceptible sounds; they become phones. Changing a sequence of phonemes into a sequence of phones is not an easy task, however. During this process, many phenomena may occur, depending on the context and other factors, which may result in the actual word being pronounced a bit differently. The phonemes could be realised in another way, for example due to neighbouring sounds, and yet the meaning would stay the same.

A sound system of a language is not simply an abstract set of phonemes and rules of their realisation in different contexts. It is also a collection of phonotactical restrictions on the permissible combinations of phonemes. Many languages are fond of planting vowels in between consonants. Some, however, do allow unusually complex consonant clusters (i.e. Polish *pstryknąć* /pstrɨknɔʋtɕ/ or Czech *čtvrtek* /tʃtvrtɛk/). The presence of such a cluster in English would surely seem suspicious to a native speaker. As a result, it is not possible to guess the possible inventory of syllables in a language by forming arbitrary pairs, triples etc. of its phonemes. The actual number of possible syllables is much lower than this hypothetical number.

## EXERCISE

Try writing down all the sounds of your own language. Do not pay attention to the writing system or the number of characters in it. Do any pairs or triples of those sounds, placed one after another, sound out of place in your language?

## ■ TONE AND INTONATION

Apart from the abovementioned features of the sound system of language, there are also other elements that may influence the meaning. In tonal languages, every syllable is pronounced with a certain melody to it – its tone. Change of a tone on a given syllable may change the meaning of the word, be it a mono- or polysyllabic word. If indeed a syllable's tone can modify the meaning of the word, it is called 'lexical tone'.

The example below from [Mandarin Chinese](#) is often cited to illustrate the idea of tonal languages. The sound sequence /ma/ has a variety of distinct meanings, depending on the tone chosen.

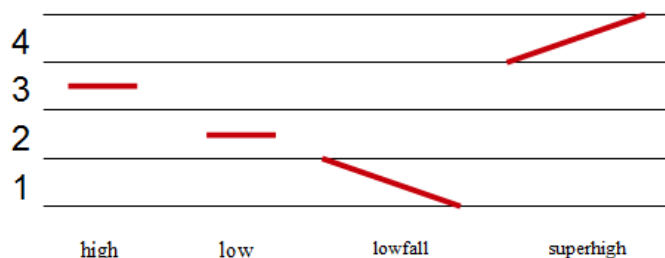
- mā mother (level tone)
- má hemp (high-rising tone)
- mǎ horse (high-falling tone)
- mà scold (low tone)

When talking about tonality, the languages of Southeast Asia, such as Mandarin Chinese, [Vietnamese](#), [Thai](#) etc., are mentioned most often, even though two other continents, North America and Africa (especially in the central west areas) also provide an abundance of such languages. Lexical tone is present, for example, in the [Yaka](#) language (a [Bantu](#) language of central Africa):

- mbókà – village
- mbòká – fields
- mbóká – civet (mammal from the viverrid family)

(data from Kutsch Lojenga 2011)

[Cherokee](#) is an exemplary tonal language from North America. It is generally considered to have four distinct tones:



According to some researches, the tones are gradually disappearing in Cherokee.

Tones (the melody of the syllable) may also have other effects on the language, such as serving grammatical functions. Here is an example of the [Ngiti](#) language spoken in the Democratic Republic of the Congo:

- ma màkpěná 'I whistled' (recent past)
- ma mákpěná 'I whistled' (intermediate past)
- ma makpéna 'I will whistle' (near future)
- ma makpěná 'I used to whistle' (past habitual)

(The tones here have been transcribed by the following diacritics: á – high tone, à – low tone, a – neutral tone, ă – rising tone; Kutsch Lojenga 1994)

Non-standard phonation which does not rely on regular vibration of the vocal folds to produce sounds may cause a problem with choosing an adequate pitch and its variation. This, in turn, might interfere with identifying the particular tone. In those few languages which feature both of these two phenomena (tones and non-modal phonation), contrastive pairs between them are rare indeed.

Many of the European languages are based on intonation. In those languages, variation of pitch does not change the basic, lexical meaning of the word. It can be used to express the speaker's attitude or emotions. Intonation can be used to discern the grammatical category of the utterance (i.e. differentiate between questions and declarative statements, etc.). Intonation exists in tonal languages as well, but it generally needs to be subordinate to the tones (although this claim is still under discussion). A 'hybrid' language, in which intonation co-exists with lexical tonality, such as Cherokee, is a rare oddity.

## A QUESTION

Is your own language tone- or intonation-based? If it is an intonation language, do you think it is possible it has ever been a tonal one? What about the other way around?

## ■ ACCENT AND RHYTHM

Syllabication does not present a problem for most people, even if at times their intuitive syllabification does not match the officially recommended rules for a given language. For a linguist however, a syllable remains a very difficult unit to define. Likewise, a native speaker can easily recognise that some syllables are more *sonorous* or *prominent* than others in their immediate vicinity. Still, a precise definition of this *sonority* remains difficult to pinpoint.

In this context, the concepts of accent, stress and emphasis are worth mentioning. Browsing through entries in dictionaries of English, German, Polish or other languages, one may find specific indications as to which syllables are to be (or may be) accentuated/stressed. It is possible that when dealing with a particularly long word, there may be more than one prominent syllable. One of them will carry the primary (i.e. strongest) stress, while the others represent a secondary or tertiary etc. (i.e. weaker) stress. The particular syllable where the lexical stress falls can be described as a '*potential* position of accentuation'. Consequently, it is possible to pronounce a given word while placing stress on a different syllable, even if at times it may be difficult to do so. Most often it also leads to an erroneous pronunciation or problems with comprehension. The position of stress may be a result of the word's morphological structure, or it may be fixed. For example, the stress falls on the first syllable of a word in Czech, on the final syllable in French, and on the penultimate syllable in Polish. There are also languages, such as Korean, that do not use lexical stress. Instead it only appears within phrases or sentences used to perform certain functions. Some languages are said to have variable stress, such as in the case of Russian. In certain languages a change in position of the stress changes their meaning. In English, *research* with stress on the first syllable is a noun, while on the second it is a verb. How does it work in your own language?

There are several ways to mark a syllable as prominent and make it stand out from the others. It can be pronounced more loudly, it can be said with an altered pitch, or it can be lengthened. Alternatively, any of the aforementioned methods can be applied in any combination.

## EXERCISE

Record a few utterances in your language. Listen to them closely and try to find out which ways of achieving syllabic prominence are possible, and which of them seem to be typical of your language.

Try listening to your own speech and think about which of the techniques are used in your own language. Perhaps record a few spontaneously uttered sentences and then listen to them carefully.

Thanks to the features of particular syllables, such as sonority, length or sequentiality, a certain rhythm arises. This kind of rhythm is slightly different from the well-known musical rhythm which is associated with constant repetition of the same musical patterns. Such repetitiveness is very rare in a language, and it can only be found in poetry, song lyrics and melodeclamation. However, rhythm does exist in language. Listening to English and French and then imitating the speech patterns whilst substituting 'da' and 'dam' for actual words allows one to notice the distinct differences between the rhythmic systems of the two languages. According to one of the hypotheses, rhythm-wise, languages can be grouped into two categories: syllable-timed and stress-timed. In the case of the former, the rhythm of speech is governed by a tendency to maintain a constant syllable length, in the latter, a constant interval between the stressed syllables. Although nowadays this theory is criticised more and more frequently, but it may be useful to get familiar with it. Perhaps you will be able to discern which of the two categories applies to your own languages (see also: Appendix). Some languages may be found problematic in this categorisation framework (e.g. Polish is often mentioned as lying between the two categories).

## ■ ARCHIVING AND RECONSTRUCTING SOUND SYSTEMS

One of the main contemporary methods of documenting a language relies on the recording spoken texts (see [Chapter 10: Language documentation](#) for more information). Sound systems of languages long dead can be replayed and analysed anew owing to such a database of recordings. The first mobile recording devices ([Nagra reel-to-reel recorders](#) were invented in the 1960's, but it was not until the following decade that relatively cheap and light cassette deck recorders became available to the general public. Although there exist recordings of dying languages which date back to the previous century (i.e. [Bronisław Piłsudski's](#) wax cylinders used to record speakers of the [Ainu](#) language), they remain very rare. How, then, is it possible to reproduce the sounds of a language which has ceased to be spoken? If the language in question was written in an alphabetic script, it might be possible, if still very difficult. How can one be sure of the correlation between the signs and sounds? In many cases, such attempts require vast knowledge and extensive research far beyond the matters of the given language only. The questions to ask are: What other, better-known languages or cultures imposed their influence on it? Do any presently-spoken languages originate from it? Are there any similarities in sound due to the same linguistic origin?

It is also important to note that spoken language differs from written language. The differences are far more striking than the utilisation of

prosody and voice quality in speech. Written texts are usually more orderly, neat, thought over and grammatically consistent. Thus, even if written documents in an extinct language exist, recreating how exactly the language was spoken in everyday situations remains demanding, if not impossible. This is just one more argument for keeping linguistic corpora and databases of speech in all languages, but even more so in those less widely used or close to extinction.

## ■ HOW TO TRANSCRIBE SOUNDS OF A LANGUAGE?

Audio recording of spoken texts is no longer a technically demanding task (see [Chapter 10: Language documentation](#)), but its written record is also needed for many purposes. This kind of written text linguists call 'transcription'. It is often very different from an orthographically written text.

There is often no correlation between the graphic signs of a language and its sounds in many languages of the world. Their written symbols frequently do not pertain to the actual sounds of the written texts. They may represent whole words and thus, they do not show the reader of what sound units those words comprise. One could expect the highest level of correspondence between graphic signs and sounds from languages based on an alphabetic script, such as Latin (which, in one form or another, is used by most European languages). Even here, however, one encounters difficulties as one letter may signify different sounds and be read in various ways. In English, a double *o* is pronounced differently in *blood*, *book* and *door*. The fact that sometimes up to four letters are used to represent a single sound (for example, in *though*) may also leave one wondering. Writing systems of natural languages are usually based on tradition and in many cases they prove difficult for linguists to analyse. Moreover, many languages, especially the less widely used or endangered ones, do not even have a writing system. Therefore, there was a need for a universal, well-planned system for writing down the sounds of speech. This system could be used equally well to write utterances in a known as well as an unknown language, as it would be capable of noting down sounds of almost any thinkable manner of articulation. The phonetic transcription system of IPA (*International Phonetics Association*; homepage of the IPA: <http://www.langsci.ucl.ac.uk/ipa/>) fulfils all those requirements. It assigns particular symbols to every possible configuration of the articulators (see: above). An overview of the IPA symbols can be found for example here: <http://www.phonetics.ucla.edu/course/chapter1/chapter1.html>. The system itself is quite complex and its use requires a lot of time and effort spent on mastering it. Even experienced phoneticians may not always agree as to the phonetic transcription of a particular utterance, because the system operates on binary and absolute categories. A given sound can either be transcribed as voiced or voiceless, nasal or non-nasal. In practice, however, a phonetician knows that nasality or voicing are gradable in actual speech and it might be difficult to decide whether the feature has already appeared or not (i.e. if the given segmental is voiced or not).

Knowing the phonemic inventory of a given language (i.e. the set of its 'basic sounds'), it is possible to transcribe speech in a less complicated manner. A broad (or phonemic, phonological) transcription takes into consideration only the phoneme which the particular sound of the utterance belongs to. In this case, the articulatory details of the speaker are irrelevant. Only the key features matter that differentiate the categories of sound. The transcription only contains the number of symbols needed to transcribe all phonemes of a given language. It is, thus, language-related, although at times it can serve a larger number of languages with similar phonological systems. A native speaker can use it without the arduous training required to master narrow phonetic transcription. It can still prove difficult to a person who does not know the language in question as she may not be familiar with what the important features differentiating segments in the given language are. Neither will she know how to assign them to particular phonemes.

In view of some of the unusual symbols of the IPA, some phoneticians decide to use different symbols, based on combinations of letters known from the Latin alphabet. This system is known as [SAMPA](#). It is most often used for broad transcription. SAMPA has been adapted to many languages and thus, there are various 'national' variants of this kind of notation.

When necessary, however, linguists may use transcription systems based on the orthography of the given language. It can be extended with additional symbols in order to transcribe phenomena such as yawning, silence, interjections or to specify the voice quality (i.e. creaky, high, whisper). Some decide not to include punctuation, capital/small letter distinction (if such a distinction exists in the language) or other orthographic rules so as to minimise arbitrariness and subjectivity. Transcription, however, always remains a subjective interpretation of speech.

Here is an example using three different manners of transcription – orthographic, IPA and SAMPA for a phrase “[Pewnego razu Północny wiatr i Słońce sprzeczali się](#)”. To facilitate comparison, the short text was divided into syllables. It is, however, a broad, hypothetical transcription as it can be realised differently (for example by keeping the last segment nasal).

Zapis ortograficzny	pew	ne	go	ra	zu	pół	noc	ny	wiatr	i	słoń	ce	sprze	cza	li	się
IPA	pev	ne	go	ra	zu	puw	noʦ̥	nɨ	vjatr	i	swɔɲ	tse	spʃe	tʃa	li	ɕe
SAMPA	pev	ne	go	ra	zu	puw	not̪s	ny	vjatr	i	swon̪	t̪se	spSe	t̪Sa	li	s̪e

\*) A slightly adapted version of the Polish SAMPA has been used here. It was prepared by J. Kleśta to limit ambiguity caused by the same symbols being used to represent different sounds in different languages

## ■ VISIBLE SOUND

The term may be understood in at least two different ways. Firstly, the simple fact of seeing the speaker's face greatly enhances the ability to identify and decipher their utterances. This has been proved by, amongst other things, the McGurk effect: seeing the form of the mouth of a speaker influences one's perception of the sound they produce.



## TRY IT YOURSELF

Try it yourself, for example using this [video](#) or [this one](#).

If you can edit the soundtrack and you have any camera on your computer, you may try to prepare a McGurk effect movie yourself.

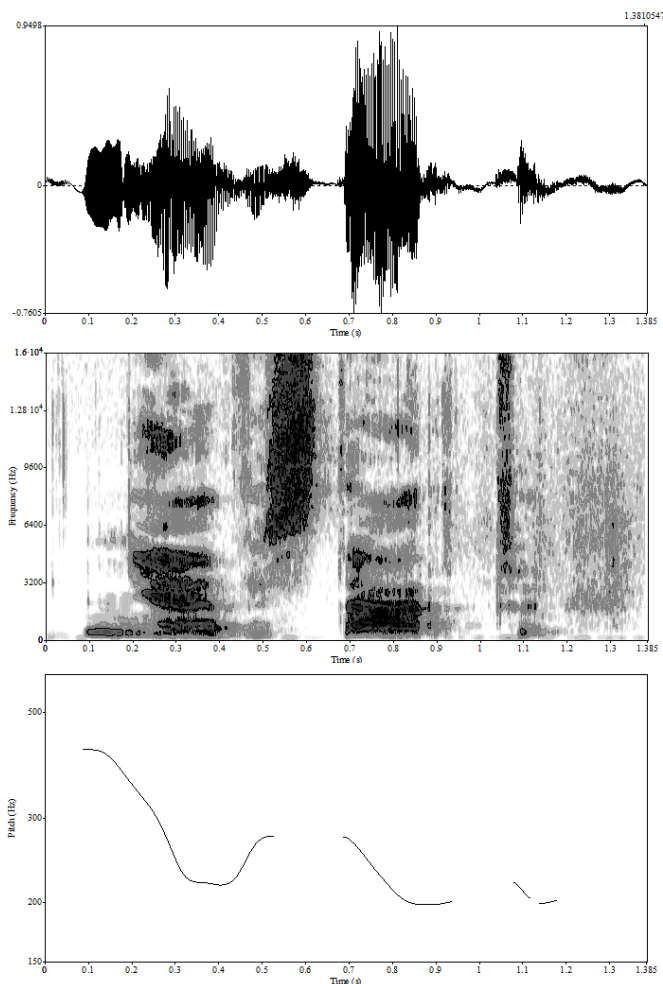
In a conversation over the phone, even if the sound is of high quality, communicational problems and misunderstandings arise more often than in conversations taking place face-to-face. Visualizations of sound signals, produced by appropriate devices or software, can also be called 'visible sound'. They are often used in phonetic research. The least technically complicated visualization is an oscillogram. It displays the sound wave propagation in time and changes of the amplitude. It can also be used to approximately identify the sound signal. If the chart resembles a sine wave, it is probably the sound of a flute or a similar simple instrument. In the case of speech, the oscillogram is far more uneven and edgy because of its many overlapping components. The most informative visualization and also, the one used most often by phoneticians, is a spectrogram. Although shown in 2D, it actually represents three dimensional space. It shows how the energy of the respective frequencies varies with time. The vertical axis is the frequency measured in hertz (Hz), the horizontal axis represents time. The darker (or in the case of colour spectrograms – usually the redder) the given point on the image, the more energy has been recorded in that particular frequency's vicinity within a given time span.

Fundamental frequency ( $f_0$ ) is often measured when studying intonation. ( $f_0$ ) is the main factor responsible for the perceived pitch, and by extension, intonation and tonality. The average frequency is between 100 and 150 Hz for men, and between 180 and 230 Hz for women, and even higher for children.

The described methods of visualizations, however, do not highlight what the viewer wants to concentrate on, neither do they omit what she/he cannot perceive anyway. Thus, they do not take the conditions of perception into consideration.

Shown below are examples of an oscillogram, spectrogram and intonogram for the same utterance: 'nie, wystarczy' /n'evystart^Sy/ which is Polish for 'no, enough' (listen to it: [here](#)). Please pay attention to the fact that the darker areas in the lower part of the spectrogram form horizontal stripes that correspond to so-called formant frequencies. In the upper part of the spectrogram, there are also dark clouds but they do not form stripes. They represent noise that is typical of fricatives. Their energy concentrates in the upper section of frequencies visible. In the final section of the spectrogram (closer to its right edge), there is a dark vertical strip that covers almost the entire range of the frequencies covered in the image. It is preceded by a delicate fog of ambient noise only. This vertical strip represents a plosion which is part of the segment /t^S/. Plosions are anticipated by normally impercivable silence of a few dozens of milliseconds which is required to generate enough air pressure in the mouth and releasing it.

Perhaps, even if you do not know Polish, you can now hypothesize about how to align transcription with the spectrogram.



Oscillogram, spectrogram and intonogram for the same utterance: 'nie, wystarczy' /n'evystart^Sy/ which is Polish for 'no, enough'

## ■ LESS-WIDELY USED LANGUAGES AND TECHNOLOGY

Language technology is beginning to play an ever increasing role in the documentation of lesser-used languages. Thanks to its developments, one can create systems to synthesise, recognise and interpret speech, expert and dialogue systems as well as software enhancing language learning or computer-assisted translation. For less-widely used languages availability of such programmes is limited since investing into such small, and often economically poor, markets does not bring much profit to large companies. As it seems, however, it is possible to create, for example, speech synthesis systems using the [MBROLA](#) system on a shoestring budget and without much effort required. They will not rival the latest developments in the area, but they can be fully functional and find many uses. Are you a user of a small, endangered or perhaps simply a less researched language? You can create a speech synthesis system for it on the base of the MBROLA system. Dafydd Gibbon propagated this method in Africa and India. There have been attempts to design such software for languages such as [Yoruba](#), [Bete](#) (you can listen to a sample [here](#); courtesy of Jolanta Bachan) or [Igbo](#).

## ■ PHONETIC EXERCISES

Do you wish for some phonetic practice? Take a look at the exercises in the [Phonetic Exercises](#) section.

### LET'S REVISE! – CHAPTER 4

Go to the [Let's Revise section](#) to see what you can learn from this chapter or test how much you have already learnt!

## References & further reading

- Phonetic/phonological diversification of languages of the world:
  - Ladefoged, P., Maddieson, I. 1996. The sounds of the world's languages. Oxford: Blackwell.
  - "Vowels and consonants" – a well-known on-line publication by Peter Ladefoged, introducing the basics of phonetics. It includes samples of sounds from many

languages of the world: <http://www.phonetics.ucla.edu/vowels/contents.html>

- Spoken texts and their IPA-authorised transcription:
  - Phonetic transcription and Polish dialects: [http://www.gwarypolskie.uw.edu.pl/index.php?option=com\\_content&task=view&id=72](http://www.gwarypolskie.uw.edu.pl/index.php?option=com_content&task=view&id=72)[http://web.uvic.ca/ling/resources/ipa/handbook\\_downloads.htm](http://web.uvic.ca/ling/resources/ipa/handbook_downloads.htm)
  - "The North Wind and the Sun" – a story used by phoneticians as a standard text for the purpose of reading: [http://en.wikipedia.org/wiki/The\\_North\\_Wind\\_and\\_the\\_Sun](http://en.wikipedia.org/wiki/The_North_Wind_and_the_Sun)
- If you want to read about intonation of various languages of the world
  - Hirst, D., Di Cristo, A. (red.) 1998. Intonation Systems. CUP.
- Speech technology for small languagesSong archives in many languages, recorded on wax cylinders: <http://sounds.bl.uk/World-and-traditional-music/Ethnographic-wax-cylinders>
  - Duruibe, U. V. 2010. A Preliminary Igbo text-to-speech application. BA thesis. Ibadan: University of Ibadan.
  - Gibbon, D., Pandey, P., Kim Haokip, M. & Bachan, J. 2009. Prosodic issues in synthesising Thadou, a Tibeto-Burman tone language. InterSpeech 2009, Brighton: UK.
- Other used or aforementioned texts:
  - Jassem, W. 1973. Podstawy fonetyki akustycznej. Warszawa: PWN.
  - Jassem, W. 2003. Illustrations of the IPA: Polish. Journal of the IPA, 33(6)
  - Kutsch Lojenga, C. 1994. Ngiti: A Central-Sudanic language of Zaire. Volume 9. Nilo-Saharan. Köln: Rüdiger Köppe Verlag.
  - Kutsch Lojenga, C. 2011. Orthography and Tone: Tone system typology and its implications for orthography development. Leiden University / Addis Ababa University / SIL International Linguistic Society of America Annual Meeting – Pittsburg – Jan 6-9, 2011.
  - Ostaszewska, D., Tambor, J. 2010. Fonetyka i fonologia współczesnego języka polskiego. Wydawnictwo Naukowe PWN.
  - Silverman, D., Lankešhip, B., Kirk, P., Ladefoged, P. 1995. Phonetic Structures in Jalapa Mazatec. Anthropological Linguistics, (37), str. 70-88.

**English translation by:** Arkadiusz Lechocki. **Translation update:** Nicole Nau.

[back to top](#)

# Writing

Home > Book of Knowledge > Writing

## ■ CHAPTER AUTHOR: TOMASZ WICHERKIEWICZ

### Chapter contents:

Writing systems of the world / Typology of writing systems

- Semasiography
- Pictographic writing systems
- Ideographic writing systems
- Syllabic writing systems
- Alphabetic writing systems
- Transcription
- Transliteration

Developing a written form of small languages – graphisation

Notes

References & further reading

## ■ WRITING SYSTEMS OF THE WORLD / TYPOLOGY OF WRITING SYSTEMS

The vast majority of languages have existed and functioned predominantly or exclusively in the form of oral transmission. That is why, in the case of both endangered languages and cultures, the heritage of oral tradition is the key to understanding the collective, social and ethnic human memory. However, everything that is to be passed on needs to be remembered, oftentimes in the collective memory of a community. For thousands of years, there have been two means for this goal: oral literature (or more generally oral culture) that was transmitted from generation to generation, and attempts to communicate with language in time (e.g. with the descendants) or in space (e.g. from a distance) by the means of graphic record.

### BOOK OF KNOWLEDGE

Chapters: [1](#) [2](#) [3](#) [4](#) **[5](#)** [6](#) [7](#) [8](#) [9](#) [10](#)

[List of all languages](#) referred to in the Book of Knowledge and other sections of the website.

[Glossary](#)

**DOWNLOAD** and print out the Book of Knowledge.

### LET'S REVISE! – CHAPTER 5

Go to the [Let's Revise section](#) to see what you can learn from this chapter or test how much you have already learnt!

### WRITING AND WRITING SYSTEM

**Writing** is the symbolic representation of language through the use of visual symbols. A writing system is a way of recording a language established with a set of visual symbols and rules of orthography. Writing is the symbolic representation of language through the use of visual symbols. A **writing system** is a way of recording a language established with a set of visual symbols and rules of orthography.

There were a few languages, however, which had been used only in writing (Majewicz 1989: 236), such as kawi (a literary language based on the Old Javanese grammar, with a substantial part of vocabulary borrowed from Old Indian Sanskrit), Classical Tibetan or Classical Chinese wenyan.

Writing, unlike speech, is not based on an innate faculty and not acquired naturally. It has to be learned through conscious and long-lasting work. This may lead to a situation in which some users of a language are not able to use the writing systems, even if these are traditionally established in the language used by a given community (illiteracy).

The history of writing most probably dates back to 5 000 years ago, although the first attempts of narrative graphic expression in the form of cave paintings or clay inscriptions are dated respectively to 20 and 10 thousand years ago. Ideographic cuneiform script carved on clay tablets is the oldest attested form of writing.

The cuneiform script was used to write many languages which became extinct a long time ago. However, linguists managed to decipher some of these languages from the cuneiform writings, e.g. Sumerian, Akkadian, Old Persian and Ugaritic.

The materials used in the process of deciphering the oldest writings were usually the inscriptions carved in stone or imprinted on clay tablets. Other writing materials, such as wood, bark, leather, bones, animal shell, papyrus and paper started to be used only with the later development of writing.

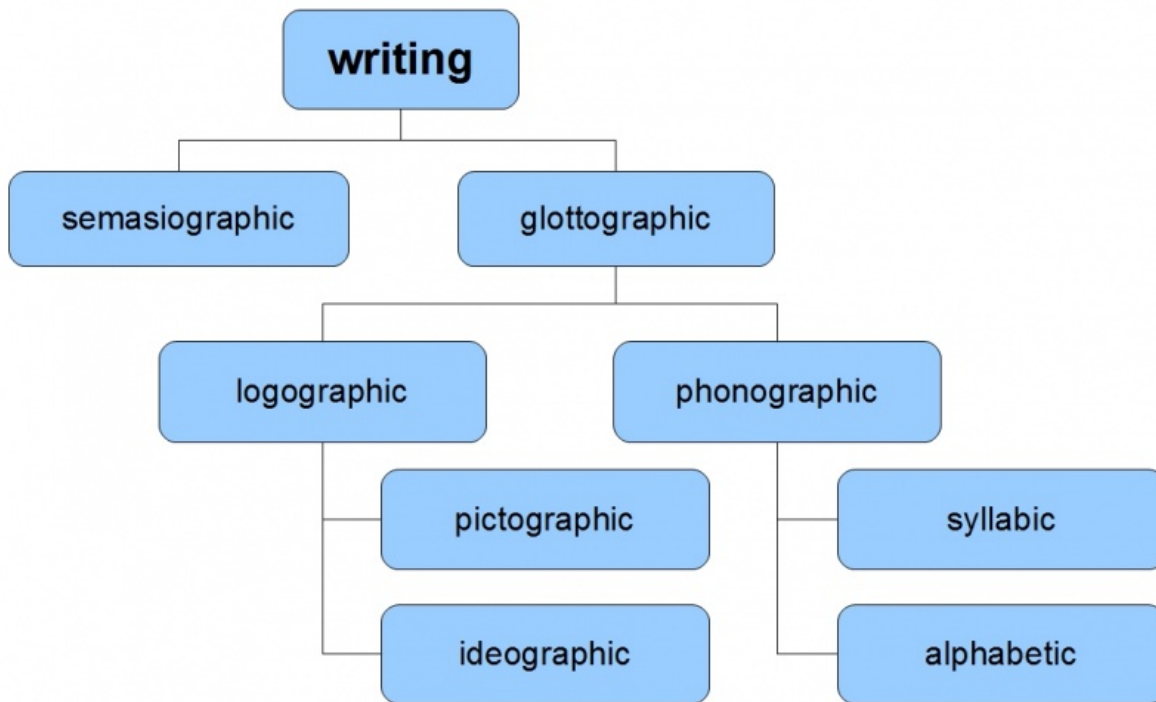
\* \* \*

Writing systems may be divided into several types depending on the relation between the individual visual symbols (called graphemes) and their content:

**semasiographic** systems (from the Greek *σημασία* [semasía] 'meaning, marking') in which graphic symbols represent concepts directly and are not dependent on any linguistic structures (some linguists do not recognise them as proper writing systems);

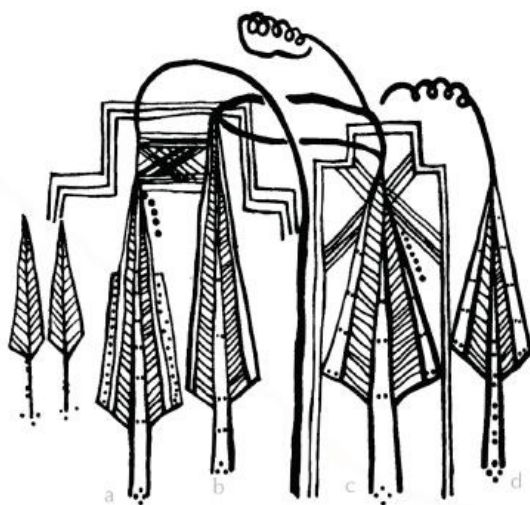
**glottographic** systems (from the Greek *γλῶττα* [glōtta] 'tongue') in which elements of speech are represented in writing by graphemes. Depending on which of the layers of speech they represent, the systems are divided into:

- **logographic** script (from the Greek λόγος [lógos] ‘thought, word, mind’) where graphic symbols indicate particular concepts (simple or complex):
  - **pictographic** script uses graphemes, which constitute visual (pictorial) representations of the objects in the surrounding world and bear much resemblance to the objects they symbolise;
  - whereas, if the graphic symbols are (already) (almost) only conventional representations of concepts (often made of pictograms), this kind of script is called ideographic;
- in **phonographic** writing systems (from the Greek φωνή [phōnē] ‘sound, voice’) each grapheme represents a sound unit of language, e.g. a syllable or a phoneme.



### Semasiography

The illustration below is a letter sent by a Yukaghir girl to a young man. According to the sources, the symbols resembling a conifer are people. Figure c is the author of the letter (the row of dots represents girl's plaited hair). Figure b is the addressee, former lover of the writer, who is now in a relationship with a Russian woman (figure a). The line starting from her and going in between the figures b and c means that she has broken the ties between the Yukaghir couple. Nevertheless, the new relationship is stormy, which is symbolised by crossed lines between the two. Lines around the figure c depict the sadness of the deserted girl living alone in her Yukaghir house (the polygonal structure around figure c). She is still thinking of the man b (shown by the curly line). But there is another young man (figure d), who is thinking positively about the author of the letter (a similar line). If the addressee wants to get back with the Yukaghir girl, he should do this as soon as possible, before there are children in the household of the girl c and man d (two little conifers on the left).





The message transmitted through this picture is relatively detailed. It is also depicted conventionally, which means that in order to understand it, one has to know the conventions behind the symbols and their written equivalents. However, individual symbols do not indicate elements of a spoken language and can be understood (verbalised) in many equally valid ways. Only the reference to the general concept of the message is coherent. Some linguists do not agree that the Yukaghir letter is a piece of writing at all. They consider it a fleeting piece of art produced on the spot on a birch bark, rather than an example of communication through language.

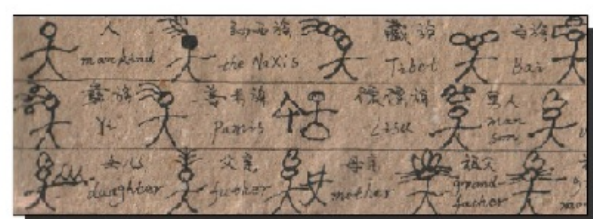
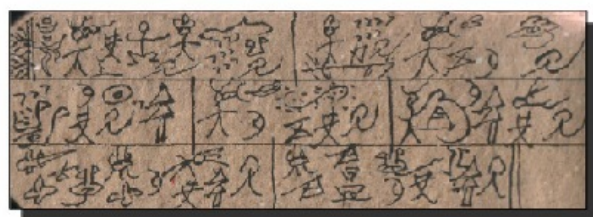
The Yukaghirs live in Sakha Republic and Magadan Oblast in north-eastern Siberia, in enclaves hundreds of kilometres away from each other (see Vakhtin 1991). Northern (Tundra) Yukaghir and Southern (Kolyma or Forest) Yukaghir are classified as an isolated group of languages possibly related to Uralic languages (comprising Samoyedic and Finno-Ugric groups). They are spoken respectively by 90 and 30 people (according to Ethnologue 2009). Since Yukaghir was taught in schools, teaching materials for this language have been published, for example *Букварь для первого класса юкагирских школ* ('A primer for the first grade of Yukaghir schools') by G.N. Kurilov printed in Yakutsk in 1987. It contained the Yukaghir alphabet, which was based on the Russian Cyrillic alphabet and completed with a few additional letters.



### Pictographic writing systems

Pictographic writing systems have been used by the Naxi in China, Native American tribes (e.g. Ojibwa), the Eskimos (e.g. the Inuit from Alaska), the Paleosiberian peoples, ancient Chinese, Egyptians and others. Apart from these, the rongo-rongo inscriptions from Easter Island and the Mayan script may also be classified as pictographic writings. However, the latter may also be considered a mixed ideosyllabic script (Majewicz 1989: 237).

The Naxi (pronunciation [Naʔi]) are a non-Chinese people living predominantly in regions in south western China, the Yunnan Province, adjacent regions of Sichuan and Tibet, and most likely in Burma. These regions were hard to reach until recently. Naxi, of which there are several dialects, is a Tibeto-Burman language. There are 309 000 Naxi speakers, 100 000 of whom are monolingual, meaning that they do not speak any other language apart from their mother tongue. The Naxi writing system, or to be more precise, the Dongba script was used to write sacred texts already by the priests of Bön, the pre-Buddhist religion of Tibetans. It consists mainly of pictographic characters (pictograms), symbols (ideograms) and phonetic syllabic characters. It is one of the last pictographic scripts in the world still in use [1]. Nevertheless, there are few living Dongba priests and it is now rarely used for religious purposes. On the other hand, its popularity among the Naxi themselves is growing. More and more people start to use the script also in everyday situations, e.g. in tourism and services.



Naxi. Photo: Tomasz Wicherkiewicz



Go to the [Interactive Map](#), try the exercises for Naxi.



## RONGO-RONGO

Rongo-rongo is the recently decoded script used in the past to write the **Rapa Nui** language, which is severely endangered.

According to *Ethnologue* 2009, 220 of its 3390 speakers live on Easter Island. Fischer (1997, 2003) writes more on the structure and decoding of the characters.

source: <http://www.rapa-nui.net>

The **Mayan** language was formerly written with a mysterious script which can be classified as either pictographic or a combination of ideographic and syllabic. It now may be easily written with a Latin-derived alphabet based on the Spanish one and completed with several diacritical marks.



Inscription in Late Classical Mayan script from Yucatán,

Photo: Tomasz Wicherkiewicz



Tourist notice board (Uxmal on Yucatán) in Mayan Yucateco language,

Photo: Tomasz Wicherkiewicz

## Ideographic writing systems

The best known ideographic writing system is the Chinese script which is used predominantly to write texts in the Chinese languages (including among others Mandarin~Beijing dialect, Jin, Yue~Cantonese, Hakka, Min with Taiwanese, Xiang, Gan, Wu~Shanghainese, which all differ from each other so much that the communication between users of those languages is possible only through the Chinese script). It is also partially used in Japanese (where the ideograms are completed with the kana syllabaries) and Korean (although a drastic fall in the usage of the Chinese ideograms is noticeable in South Korea, and they were completely abandoned in North Korea already in 1949). The Chinese writing system was also used in Vietnamese, where the ideographic script was substituted by Latin alphabet based on Portuguese orthography and completed with numerous diacritical marks which reflect for example the tonal character of the language.

Another ideographic writing system is the classical script used by the speakers of some of the Yi languages (also known as Nuosu or Lolo). They all constitute the whole of the Tibeto-Burman language group. Six of those languages are recognised by the Chinese administration as separate mutually unintelligible linguistic varieties: Northern Yi (**Nuosu**), Western Yi (**Lalo**), Central Yi (**Lolopo**), Southern Yi (**Nisu**), *Southeastern Yi* (**Wusa Nasu**), Eastern Yi (**Nasu**). Other Yi languages are spoken also in Vietnam, Burma and Thailand. The majority of Yi languages are not classified as endangered yet, although *Ethnologue* 2009 reveals small numbers of speakers of several Yi varieties, such as **Miqie** – 30 thousand users and still dropping. However, intensified cultural exchange and administrative services conducted through the official languages (especially Chinese) have led to great contamination of the Yi languages with borrowed vocabulary and structures. Also, the endangered literature written in Yi needs to be documented and digitised.





Bimo priest holding a Yi manuscript  
 Manuscript written in Yi script  
 Photo: Halina Wasilewska

CECHY TRADYCYJNEGO PISMA NUOSU  
 JEDEN ZNAK - KILKA ZNACZEŃ

„niebiosa”

„ogon”

„robić”

„wysoki”

„dobry”

„w pośpiechu”

„wiatr”

„dom”

„serce”

„zmieniać”

„bestia”

„siedzieć”

„wiosna”

„chmura”

„mózg”

„obszar  
ziemi”

Author: Halina Wasilewska

Characteristics of the traditional Nuosu script. One sign – several meanings			
“heaven”	“wind”	“beast”	“cloud”
“tail”	“home”	“sit”	“brain”
“do”	“heart”	“spring”	“earth territory”
“high”	“change”		
“good”			
“in a hurry”			

Syllabic writing systems

Syllabic scripts may have developed from sets of simplified ideograms or pictograms, which have lost their semantic value. The symbols, nevertheless, have retained their phonetic value and indicate a corresponding syllable. This is how Late Sumerian and Late Egyptian writing systems, and later on the Japanese kana syllabaries (hiragana and katakana), have developed.

Most writing systems of India (abugidas) derived from Brāhmī script are syllabic (or actually alpha-syllabic). They are used in Indo-European languages in North India (e.g. Devanagari, Bengali, Gujarati, Gurmukhi, Oriya and others), but also in Dravidian languages in South India (especially Tamil, Telugu, Kannada, Malayalam).

In North America, there were several attempts to record local American Indian and Eskimo languages with syllabaries. One of the best known of such syllabaries was Tsalagi invented by Sequoyah – a Cherokee from North Carolina. In his script, each grapheme represented a syllable consisting of a consonant and a vowel.

## ᑭᓄᓂ-ᑭᓄᓂᑭᓄᓂ CHEROKEE

webpage heading: [http://www.languagegeek.com/rotononhsonni/tsalagi/tsa\\_syllabarium.html](http://www.languagegeek.com/rotononhsonni/tsalagi/tsa_syllabarium.html)

Another syllabic (or to be more precise, alpha-syllabic) script is **Dené**. It is one of the *Canadian abugidas* developed by Christian missionaries for languages of Aboriginal people in Canada. It has been used mainly in a branch of the Na-Dene language family, the Athabaskan languages, such as **Carrier** (**Dulkw'ahke** – according to *Ethnologue* 2009, 2060 speakers) or **Hare** (**North Slavey** – 1030 speakers).

## ᑭᓄᓂᑭᓄᓂᑭᓄᓂ ᑭᓄᓂ ᑭᓄᓂ ᑭᓄᓂᑭᓄᓂᑭᓄᓂ ᑭᓄᓂ ᑭᓄᓂ NORTH SLAVEY

webpage heading: [http://www.languagegeek.com/dene/kashogotine/north\\_slavey.html](http://www.languagegeek.com/dene/kashogotine/north_slavey.html)



The syllabic script is used on a daily basis even by foreigners residing in Nunavut, autonomous region in Canada inhabited by Inuits (Eskimos) – see above the business card of a Polish pilot working for airlines in the capital city of Iqualit – source: Paweł Torz

### Alphabetic writing systems

Alphabets are writing systems in which each grapheme indicates one speech sound, or to be more precise, one type of sound unit (called a phoneme) in a given language. The earliest alphabetic scripts developed in Semitic languages, such as Phoenician, Hebrew and Arabic. Alphabets in these languages (called abjads) used the basic characters to indicate mainly or even only consonants. Vowels (especially short ones) if represented at all, were indicated by additional diacritic marks. In modern Arabic and Hebrew everyday texts the short vowels are also omitted, and readers must supply them according to their knowledge of the language.

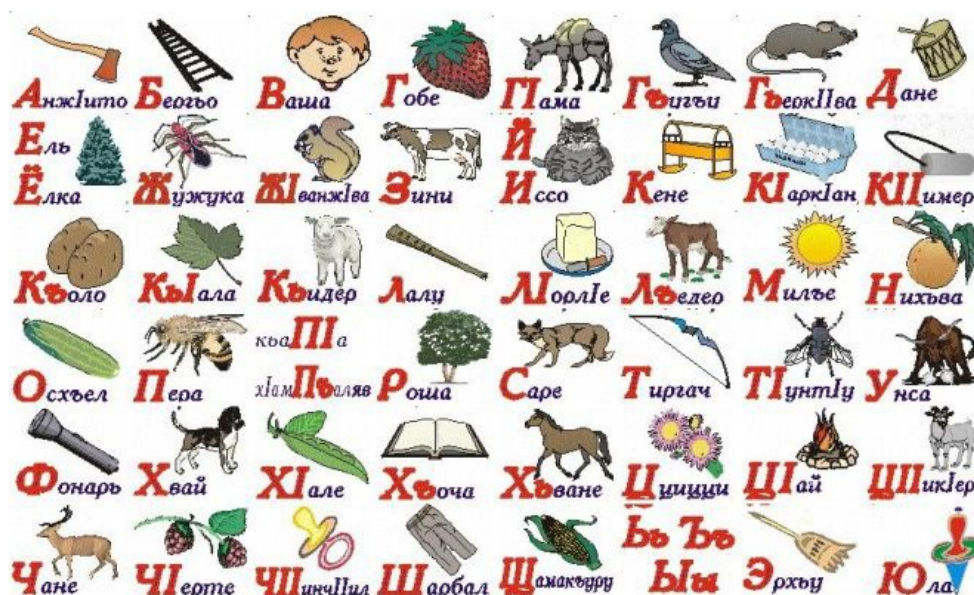
Most alphabetic writing systems known today (such as Greek, Latin, Russian Cyrillic) developed from Phoenician writing. Exceptions are for example Georgian and Armenian, two languages with a highly phonemic orthography (meaning that each consonant or vowel phoneme is represented by a corresponding letter). The two scripts are used almost exclusively to write the Georgian and Armenian languages. Nevertheless, other regional languages of Georgia, such as Svan or Mingrelian, are also written in Georgian script. Moreover, the Armenian alphabet was used in Armeno-Kipchak, a Turkic language spoken in the past by the Armenians in Poland.



Go to the [Interactive Map](#), try the exercises for Svan.

Most small and endangered languages of the world, if they exist in writing at all, use one of the alphabetic writing systems. They are most often based on the Latin or the Russian (Cyrillic) alphabet. Developing writing systems are also based on one of these two alphabets (naturally, the influence of the Russian alphabet reaches mainly countries of the former Soviet Union and Mongolia).

The image below shows the alphabet of the Karata language used in southern Dagestan (a republic of the Russian Federation located in the North Caucasus region). It was invented by the speakers of the language, after the Dagestan authorities and linguists refused repeatedly to do so. According to Ethnologue 2009, there are ca. 5 000-6 000 Karata people, and almost all of them still speak their language. However, the lack of education, resulting also from the prior lack of a written language, makes the intergenerational transmission of the language difficult.



The Mongolian language is an example of a change in the writing system conditioned politically. Traditionally, Mongolian was written in a purely alphabetical script derived from Uyghur. In 1946 Mongolian Communists decided to adopt the Russian Cyrillic script instead, which was supposed to make the classical Mongolian literature inaccessible to the younger generation. Although the classical Mongolian script was reintroduced after the political transformation in Mongolia in 1994, it is used today rather only symbolically and as a means of artistic expression. Nevertheless, the Mongolian script is still used in Inner Mongolia (an autonomous region of the People's Republic of China), where 17% of inhabitants are Mongolian (which amounts to 4 million). The first language of 2.5 million of these people is Mongolian, a number similar to that of inhabitants of the independent state of Mongolia.



МОНГОЛ БИЧИГ

The Mongolian script has been adopted to write other minority languages of the Inner Mongolia area of China, such as Evenki (according to Ethnologue 2009 – 19 000 speakers in China and 1000 in Mongolia), Xibe (30 000) and Oirat (139 000 in Mongolia and 139 000 in Inner Mongolia).

## Transcription

There are writing systems, which are used only by specialists (linguists, ethnographers, ethnomusicologists) to write texts for their own needs, e.g. texts in previously unknown or unwritten languages or languages which do not have their writing systems. **Phonetic transcription** is the most common technique used to record such texts. It is a system of notation in which each phone has its own invariable graphic representation. A set of such symbols is called a phonetic alphabet. The knowledge of orthography of a given language/script is not required to transcribe and read a transcribed text. However, the symbols of a particular system of transcription have to be used consistently throughout the text. Each field of linguistics (e.g. Slavic, Caucasian or Semitic studies and various dictionaries of English language) has developed its own transcription system. Phoneticians, however, have always tried to invent one system that would enable to transcribe and read texts in any language of the world. *The International Phonetic Association* succeeded in establishing such a system in 1888 and has been improving its *International Phonetic Alphabet (IPA)* since then (<http://www.langsci.ucl.ac.uk/ipa>). After numerous modifications and supplementations, IPA consists of 107 symbols representing consonants and vowels (mainly letters derived from the Latin and Greek alphabets), 31 diacritics modifying these symbols and 19 symbols indicating suprasegmental features such as length, tone, stress and intonation. They are all organised into a clear and systematic diagram



[http://www.langsci.ucl.ac.uk/ipa/IPA\\_chart\\_%28C%292005.pdf](http://www.langsci.ucl.ac.uk/ipa/IPA_chart_%28C%292005.pdf)

Here is a link to a website which helps to practice the transcription of texts with the IPA symbols –

[http://www.cambridge.org/resources/0521612357/3744\\_Chapter%201%20additional%20problems.pdf](http://www.cambridge.org/resources/0521612357/3744_Chapter%201%20additional%20problems.pdf).

Sometimes texts written in the International Phonetic Alphabet (or other more or less simplified transcription system) are the only attested texts of small endangered languages, which have not developed their writing systems. Here is a link to a website with materials concerning the documentation and transcription of various languages, including endangered ones: <http://archive.phonetics.ucla.edu>. The website provides examples of audio recordings of these languages and their phonetic transcription.

Writing in IPA requires special symbols, which can now be easily downloaded to any computer. There are also other systems of transcription, which consist of characters available on a standard computer keyboard – one of the most important is SAMPA (*Speech Assessment Methods Phonetic Alphabet* – the rules of transcription in this alphabet can be found here: <http://www.phon.ucl.ac.uk/home/sampa/index.html>). See also Chapter 4: *The Sounds of Language* on phonetic transcription in general and on SAMPA.

## Transliteration

When working with two or more different writing systems, there is often the need to convert one of them into another. Such a process is called transliteration, especially if the source text is to be converted into an alphabet, i.e. a set of letters. The more consistent the transliteration is, the easier it is to work with the transliterated text. In order to avoid ambiguity and misunderstandings, some countries have adopted standardised official transliteration systems to convert their native writing systems into Latin alphabet. For example, the government of the People's Republic of China promotes the *pinyin* system to transliterate the Chinese ideograms [2]. Whereas the Hepburn system of transliteration, called *romaji*, has been used traditionally in Japan.

## ■ DEVELOPING A WRITTEN FORM OF SMALL LANGUAGES – GRAPHISATION

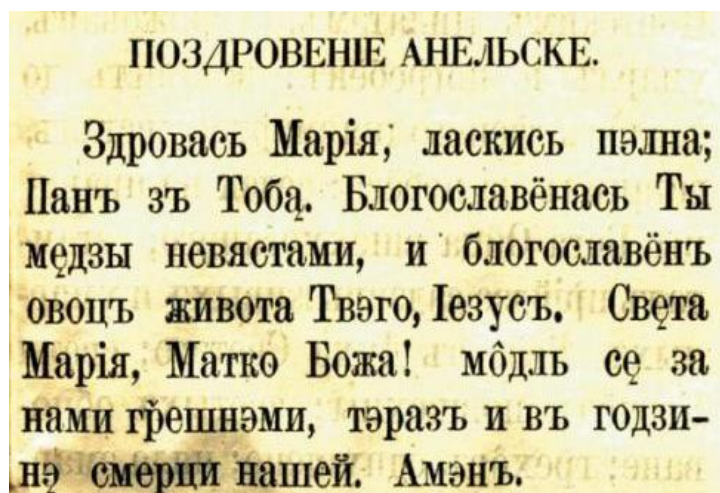
As Mühlhäusler (1990: 205) observed, developing a written form of a language or its variety (**graphisation**) involves not only a simple selection of appropriate orthography but also making decisions concerning cultural, religious, political and historical matters. Writing languages without considering these aspects is usually simply a transcription mentioned above.

Thinking about the graphisation of a language – either as a large-scale group project or an individual's need and initiative – one has to take into consideration the fact whether a given community has ever used any writing system to write their own language or this language has always remained oral.

Another very important aspect is the cultural and religious context behind the functioning of a particular writing system. As Diringer (1968) stated, „alphabet follows religion”. It may be observed that a text written in Cyrillic/Russian script brings associations with the Eastern Orthodox Church, while in Eastern Europe, the Latin alphabet was associated until recently with Western Christianity (especially, the Catholic Church). The Arabic script is associated with Islam, Hebrew with Judaism, scripts of South Asia with Buddhism, Brahmanism and Hinduism, and the Chinese script with Confucianism. Assigning a certain writing system to a language often means connecting it with or strengthening its connection with a particular culture. It may happen that two similar language varieties become two separate languages after they adopt different writing systems, for example Serbian and Croatian, Hindi and Urdu, Chinese and Dungan, and in the past – Romanian and Moldovan.

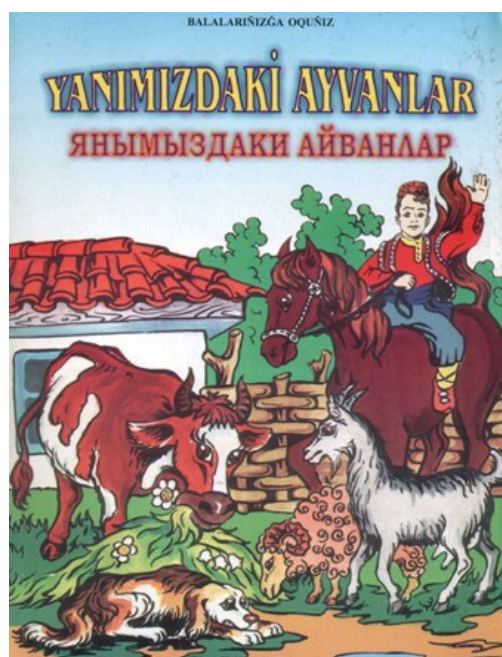
Here is a link to an audio recording in the Dungan language: <http://www.youtube.com/watch?v=X9TMy06J7XU&feature=related>. The Dungans are the descendants of the Chinese-speaking Hui Muslims, who have settled in today's Kyrgyzstan and Kazakhstan. Because they lived in isolation from China and adopted the Russian alphabet, the Dungan language has changed significantly and differs from the original Chinese dialects. It is now considered an ethnolect on its own.

Leaders of nation states have used graphic changes of writing systems as a part of the process of transforming the societies culturally, as for example the Turkish leader M. K. Atatürk, who introduced the Latin alphabet instead of the one based on the Arabic script and at the same time ordered loanwords of Arabic origin to be removed from the Turkish language. The previously mentioned change introduced in the Mongolian language had a similar character. These reforms were supposed to bring the civilisations closer to Western Europe (Turkey) and the Soviet Union (Mongolia). What is worth remembering, after the suppression of the January Uprising, the Russian government intended to change the Polish writing system from the Latin alphabet into the Cyrillic. Naturally, it would have changed the cultural orientation of Polish people as the citizens of the Russian Empire in the long term, and possibly would have led to the endangerment of the Polish language (see: [http://www.niniwa2.cba.pl/cyrylica\\_nad\\_wisla.htm](http://www.niniwa2.cba.pl/cyrylica_nad_wisla.htm) and the Polish text of the *Hail Mary* prayer written in Cyrillic:



The graphisation of minority languages in multinational states may also have a political background. As mentioned in Chapter 7 on *Multilingualism*, the language policy set by Lenin in the 1920's aimed at establishing Latin-derived alphabets for the previously unwritten languages of the Soviet Union. Then, at the end of the 1930's most of these alphabets were modified and based on the Russian script as a part of Stalin's language policy in the USSR.

After these manipulations with graphic systems, some languages of the former Soviet Union have not decided definitively about their writing systems until today. The Crimean Tatar language classified as endangered may serve as an example here. It is used as a native language in Crimea located in the Ukraine and in the Central Asian countries of the former USSR. Below is a Crimean Tatar book with stories for children, which is written with both Latin and Cyrillic letters.



After 1951, the Soviet pattern was followed by the government of the People's Republic of China, when it aimed at the graphisation of many previously only oral languages of various minority groups. What is interesting, none of the non-Chinese languages adopted the ideographic Chinese script – they all remained in their cultural and religious domains, which was expressed by the usage of other traditional writing systems or new scripts invented based on an extended/modified Latin alphabet or phonetic alphabet (e.g. *IPA*). The tradition of numerous writing systems among peoples living in China is reflected for example in the Chinese banknotes. The words featured on the notes are written in the Chinese ideograms, the Latin transliteration of *pinyin*, and at the bottom of the reverse, in Old Mongol, Tibetan, Uyghur based on the Arabic script and severely modified script of the Zhuang language (based on the Latin alphabet completed with the IPA characters).



## ADVANTAGES AND PROFITS

In the opinion of a community speaking a rare language, inventing and introducing a writing system for that language may:

- increase its prestige and make it (more) noticeable (=present) in the public life;
- help to preserve (=record) the important elements of the native culture, such as oral literature;
- facilitate education and thereby hinder the process of abandoning the language;
- make it easier for its speakers to access global and popular culture without the need to use the major (inter)national languages.

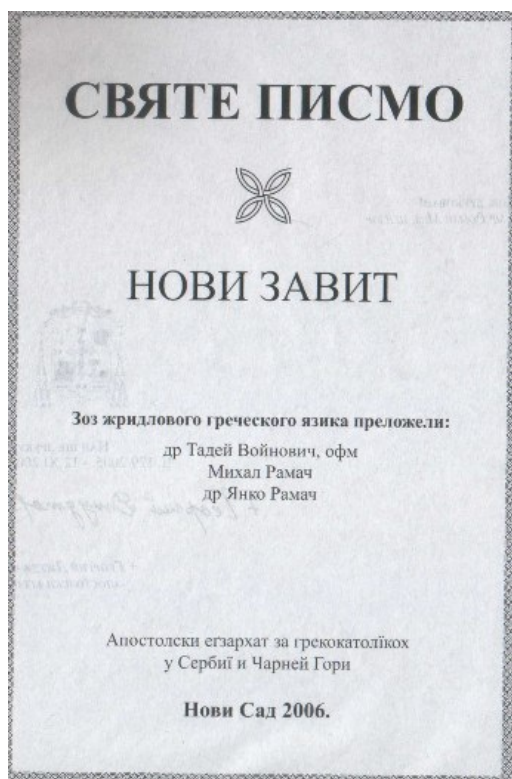
The graphic presence of a minority language in the linguistic landscape may make a given minority a significant element of the local culture visible in real-life domains, e.g. administrative or geographical. Below is a photograph taken in 2007 at the celebrations during a ceremonial uncovering of a bilingual road sign – the first in history – in the Kashubian village of Szymbark/Szimbark in the Pomeranian region. The local prestige of the Kashubian language has increased significantly in the recent years, which is reflected in its more and more frequent presence in the linguistic landscape of the region.



The written variety of an endangered language enables the documentation of important elements of local, folk and previously oral culture. An attempt to create a literary standard for the Wilamowicean language by Florian Biesik in the 1920's may serve as an example here. This endangered, archaic language derived from 13th century Middle High German is spoken by several dozen inhabitants of Wilamowice, a town located in southern Poland. The first page of the manuscript of Biesik's narrative poem describing the material and non-material culture of his native town of Wilamowice is presented below.

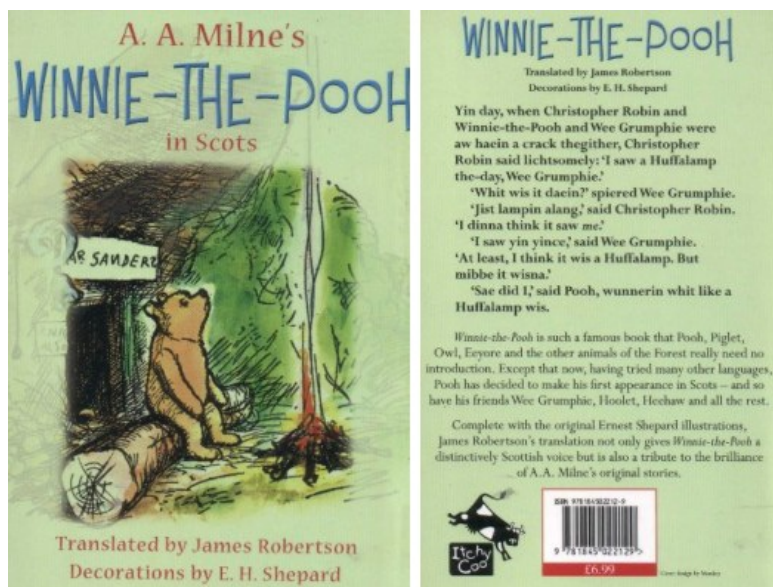






The visible graphic attestation of the language is important for many endangered language communities with insecure intergenerational transmission. Therefore, (young) readers should have access to the written/published versions of the classic works of world literature, more and more frequently including popular literature as well. It would not be possible without the standardisation of the written varieties of those languages. Examples of such initiatives are given below:

- the publication of the classic English book for children, *Winnie-the-Pooh* by A. A. Milne, in standard Scots. The Scots language has very low prestige in Scotland. The intention of the publishers was not only to increase the prestige of the language but also to provide educational reading material which is valued within Scotland.



- the publication of popular children's literature classics, such as *Le Petit Prince* (*The Little Prince*) by A. de Saint-Exupéry and *Asterix* by R. Goscinny and A. Uderzo. The release of these titles in Molise Croatian and Mirandese were the first in history popular literature publications in these languages.





Molise Croatian (on the left) is a South Slavic language spoken by approximately 3 000 people living in three villages in the Italian Molise province. Mirandese (on the right) is the only indigenous minority language of Portugal spoken by ca. 5 000 inhabitants of the city of Miranda de Douro in the northeastern area of the country.

Thus, for more and more endangered languages the existence of a written form/standard of a language has become a subjective or objective condition for the preservation of linguistic identity.

## LET'S REVISE! – CHAPTER 5

Go to the [Let's Revise section](#) to see what you can learn from this chapter or test how much you have already learnt!

### Notes

[1] More about the script: Gaca, Maciej 1997. *Literatura piktograficzna Naxi (Chiny Południowe)*. Stęszew: International Institute of Ethnolinguistic & Oriental Studies.

[2] It differs, however, from the system adopted for Chinese in Taiwan (i.e. the Republic of China).

### References & further reading

- Bouquiaux, Luc & Jacqueline M.C. Thomas 1992. *Studying and Describing Unwritten Languages*. SIL International.
- Coulmas, Florian 2003. *Writing Systems. An introduction to their linguistic analysis*. Cambridge University Press.
- Dinger, David 1968. *The alphabet: A key to the history of mankind*. 3rd revised edition. London: Hutchinson.
- Fischer, Steven Roger 1997. *Rongorongo: The Easter Island Script: History, Traditions, Texts*. Oxford Studies in Anthropological Linguistics 14.
- Fischer, Steven Roger 2003. *A History of Writing*. London: ReaKtion Books.
- Gaca, Maciej 1997. *Literatura piktograficzna Naxi (Chiny Południowe)*. Stęszew: International Institute of Ethnolinguistic & Oriental Studies.
- Lüpke, Friederike 2011. „Orthography development”, w: Peter K. Austin & Julia Sallabank (red.) *The Cambridge Handbook of Endangered Languages*. Cambridge University Press. Ss.312-336.
- Majewicz, Alfred F. 1989 *Języki świata i ich klasyfikowanie*. Warszawa: PWN.
- Mühlhäusler, Peter 1990. “Reducing’ Pacific languages to writing”, w: John Joseph & Taylor Talbot (red.) *Ideologies of Language*. London: Routledge. Ss. 189-205.
- Sampson, Geoffrey 1985. *Writing Systems. A linguistic introduction*. London: Hutchinson.
- Yule, George 2010. *The study of language*. Cambridge University Press.
- Vakhtin, N. B. 1991. *The Yukagir language in sociolinguistic perspective*. Stęszew: International Institute of Ethnolinguistic and Oriental Studies.
- Vaux, Bert & Justin Cooper 1999. *Introduction to Linguistic Field Methods*. Lincom Europa.

Website about Cherokee syllabary: [http://www.languagegeek.com/rotonhsonni/tsalagi/tsa\\_syllabarium.html](http://www.languagegeek.com/rotonhsonni/tsalagi/tsa_syllabarium.html)

**English translation by:** Agnieszka Lewandowska.

[back to top](#)

# Language and culture

Home > Book of Knowledge > Language and culture

## ■ CHAPTER AUTHOR: NICOLE NAU

### Chapter contents:

Culture, cultures, and cultivation

Culture means diversity – so does language!

- Geographical varieties and local identity
- Dialect versus standard
- Dialects don't die!
- Dialect or regional language?
- New dialects and social variation
- Vocabulary for special purposes
- Functions of special vocabulary and a another look at "cultural" language

Language is doing and culture is a verb

- Genres (text types)
- Verbal art: Oral traditions
- Riddles and riddling
- Oral traditions and endangered languages

References and further reading/listening

Most people agree that language and culture are tightly connected. Some people also say "language is culture" or "culture is language". However, such very general statements are not very helpful – what do they mean? If culture and language were simply the same, why would we need two different labels? Not all expressions of culture require language, and not all aspects of language are culture-dependent. It is worth taking a closer look at the relationship between language and culture. In this chapter we will ask what the two concepts have in common and what roles language has in cultural practices. Another aspect, how cultural peculiarities are reflected in language, is dealt with in [Chapter 2 \(Exploring Linguistic Diversity\)](#).

## ■ CULTURE, CULTURES, AND CULTIVATION

The words "language" and "culture" are used both as collective nouns and as countable nouns. In English we may ask "What is language?" or "What is **a** language?".

Answers to the first question will give some characterization of language as a human capacity, or as a means of communication etc. (see Chapter 1), while with the second question we want to know what characterizes and distinguishes individual languages such as [Polish](#), [Irish](#) or [Turkish](#). The same two perspectives can be applied to "culture". **A culture** – with the plural form cultures – is defined in the following way:

- "the ideas, customs, and social behaviour of a particular people or society" ([Oxford Dictionary](#))
- "a particular set of customs, morals, codes and traditions from a specific time and place" ([Yourdictionary.com](#))

One of the ties between language and culture is that ideas, customs and traditions are typically passed on through talking. Some parts of a culture may not rely on words – for example, one may pass on a dance or a traditional craft by showing and imitating – but most customs are related to ideas, beliefs, knowledge that can only be understood when being recounted. Language is especially important for the maintenance of our **intangible cultural heritage**, and at the same time it is part of it.

### Intangible Cultural Heritage

"Cultural heritage does not end at monuments and collections of objects. It also includes traditions or living expressions inherited from our ancestors and passed on to our descendants, such as oral traditions, performing arts, social practices, rituals, festive events, knowledge and practices concerning nature and the universe or the knowledge and skills to produce traditional crafts." ([UNESCO](#))

Read more and find out which cultural practices have already been inscribed into the UNESCO list at [their website](#)!

### BOOK OF KNOWLEDGE

Chapters: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#)

[List of all languages](#) referred to in the Book of Knowledge and other sections of the website.

### [Glossary](#)

**DOWNLOAD** and print out the Book of Knowledge.

### LET'S REVISE! – CHAPTER 6

Go to the [Let's Revise section](#) to see what you can learn from this chapter or test how much you have already learnt!

*Culture* as an uncountable noun (without a plural form) is a more abstract concept that rouses different associations. The Latin word *cultura* was first of all used in the context of agriculture. In order to produce crop, one has to cultivate the ground – just letting plants grow by their nature is not enough. This idea was then extended metaphorically to the development of an individual and of human society. Culture in the broad sense is what humans add to nature in order to achieve something better. In many European languages, the concept is associated to civilization, refinement, education, or arts. For example, in Polish the adjective *kulturalny* (literally: "cultural") means

'educated, refined, sophisticated' when speaking of persons, but also referring to language. Most people agree that there are ways of expressing oneself that are more "cultural", or "cultivated" than others. However, there are different views on what characterizes cultivated speech and its opposite, which may be conceived as "primitive", "vulgar", "uneducated" or simply "careless" speech. At different times and in different parts of European societies, one or more of the following features were (are) held to characterize refined language, or the speech of an educated (cultivated) person:

- having a pleasant voice, speaking with a pleasant rhythm and intonation
- speaking with a particular accent, for example from a region that is thought of as more cultivated than others, or speaking without a local accent (your speech doesn't show from which region you come)
- using sophisticated words, internationalisms of Greek and Latin origin
- speaking in "whole sentences", such as in written language
- speaking politely, showing respect towards the listener

**Study question 1:** What is **your** idea of cultivated ("good") language and its opposite – "bad language"?

**Study question 2:** Which aspect of the concept "culture" is illustrated in this Hungarian poster of the International Mother Language Day? What is its message?

(*anyanyelv* 'mother tongue', *nemzetközi* 'international', *napja* 'day')



Poster by the Hungarian artist Varga Gábor Farkas

#### ■ CULTURE MEANS DIVERSITY – SO DOES LANGUAGE!

Speaking of "a" language and giving it a label such as "Polish" suggests a certain unity. However, within this unity there is also a lot of diversity: the speech of an elder peasant from southern Poland, a young worker from Gdańsk, or a university professor from Poznań is certainly not identical, yet they all speak "Polish". Or think of various texts, such as a poem by the classic poet Adam Mickiewicz, a newspaper report, a discussion in an Internet forum – the language in each of them has different characteristics. The same holds for a culture. According to the definitions given above, "Polish culture" is the set of ideas, customs, traditions of Polish people. Evidently, not all people in Poland share all these ideas and customs, and a particular custom shared by a larger group of people usually shows some variation. At a closer look the set that defines a culture or a language thus consists of several overlapping subsets.

Linguistic varieties – the different ways of using a language – can broadly be divided into three classes:

- geographical varieties – varieties used only in certain parts of the territory where the language is spoken;
- social varieties – varieties used by parts of the society, defined by factors such as age, gender, or occupation;
- functional varieties – variation associated to the situation and the function in which the language is used.

A given variety often does not fit neatly into one of these classes – for example, it may be used within a certain region only by a certain social group, or by a socially defined group only in certain situations and for certain functions. In this section we will mainly be concerned with geographical and social varieties, while typical functional varieties will be discussed in the following section when we will turn to genres.

#### Geographical varieties and local identity

*Dialects remind us of the staggering diversity and beauty of humanity. (dialect blog)*

Geographical varieties may occupy larger or smaller territories. In the case of languages spoken in several states, the language of each state can be considered a geographical variety, for example the **French** spoken in France, Switzerland, Belgium, or Canada. At the other extreme are local rural dialects spoken in one particular village or parish (called *Ortsmundart* in German). In between are dialects of territories such as a county or a cultural region within one state, or sometimes extending across state borders. For example, **Alemannic German** German dialects are spoken in territories across the borders between Germany and France and between Germany and Switzerland. Dialects of a middle range – more than one parish, less than a state –, especially when they are associated to a cultural region, are probably the most important to speakers of a given language.



Go to the [Interactive Map](#), find out about dialects of **Karaim** and try to solve the exercise!

For most speakers of a local dialect, this is the language in which they grew up, the language of home, family and friends. Speaking and hearing this variety gives them a feeling of belonging. It is part of their personal identity, whether they like it (most people do) or not. Those who didn't grow up with a dialect often fail to understand the importance of this kind of variation. They are indifferent or even hostile towards the geographic diversity of their language, and sometimes they make fun of dialects and their speakers. Outside of their speech community, dialects are rarely prestigious varieties of a language, but some are more stigmatized than others. Sometimes there are historical reasons for differences in prestige of dialects. For example, dialects spoken in regions where the peasants were known to be poor may have lower prestige than dialects from wealthier regions.

In many European countries there are wide-spread stereotypes which dialect is “ugly” and in which region people speak “nicely”. How is it in your country? Is the prestige of a dialect connected to the economic success of its speakers, or can you find other historical reasons for differences in prestige?

### Dialect versus standard

In Europe, dialects are usually opposed to a standard language that is common to all speakers regardless of the region they come from. It is important to recognize that the standard language is a variety, too – it is not “the language”, but only part of it. Apart from the geographical spread, several other features tend to distinguish dialect and standard, for example:

- **speakers:** dialects are spoken with people one knows well, with family, friends, or neighbours, while the standard language is used with other people;
- **situation:** dialects are used in informal situations – private conversations, free time activities –, while in formal settings people rather use the standard variety;
- **medium:** dialects are mostly spoken, seldom written, while the standard language exists in spoken and written form – note that in several European languages the concept standard language is expressed as “literary language”;
- **acquisition:** dialects are acquired in a natural way, without any explicit “learning” or studying, while the standard language is additionally taught in school (especially the writing);
- **standardization:** as the term indicates, the standard language is a standardized variety, which means that its form is consciously developed. Dialects, in contrast, are non-standardized varieties of a language – only the actual use, the speakers’ unconscious choices of words and constructions decides about what is right and what is wrong.

These are only typical characteristics, not necessary features. For any given dialect, the situation may be different.

Study question: Think of a local dialect you know well. Which of the given characteristics are true for this dialect, which are not? Are there other differences in the use of this dialect and the standard language? Which differences do you find important, which are less important?

The standard variety is associated with education and schools, with writing and books, with the public sphere of life, and with formal situations that require a conscious and planned use of language. A dialect is associated with the private sphere, informal situations and spontaneous language use. Partly because of these oppositions, dialects sometimes become stigmatized as an “uneducated” variety and only the standard variety is held to be “cultivated” (compare the discussion of culture and cultivation above). Such a view was held by many people all over Europe at various times during the 19th and 20th century. Especially in the decades 1950-1980 many parents didn't

speak the local dialect with their children although it was their own first language, because they thought that raising the children in the standard variety would be the key to a better education and their getting on in life. They probably weren't aware that children are perfectly capable of managing more than one variety of a language and that speaking a dialect at home should not prevent them from learning to speak and write in the standard variety when attending school. Because of this tendency, many dialects of European languages became endangered. For the children of these parents, the dialect wasn't the most natural language any more. Maybe they still picked it up to some degree from their grandparents or from neighbours and friends, but they didn't speak it fluently. In linguistic studies, these people have been described as "semi-speakers". Of course this generation then didn't speak the dialect with their children. This is a typical scenario that quickly leads to severe endangerment of languages and dialects.

### Dialects don't die!

Fortunately for the dialects, attitudes have now widely changed and local varieties have become popular again. People are no longer ashamed of their accent, and words and popular sayings are used as markers of a cultural region to which people are proud to belong. They often turn up in advertisements for local products, or in information for tourists. A recent hit in several European countries are GPS satnavs with dialect speakers. In Germany the first one in the Cologne dialect, launched in December 2009, was met with great enthusiasm. During the first year the voices were downloaded over 25 000 times (<http://www.koelsch-akademie.de/>).

Here is an example for the use of dialect in an advertisement. In Poznan, Poland, *pyra* is the local word for 'potato'; the surrounding region Wielkopolska is famous for potato cultivation.



The local brew and the local dialect (Photo: Nicole Nau)

Dialects, as any language, change over time. The different attitudes described above, ongoing industrialization and urbanisation, individuals' increasing mobility, and the expansion of mass media are factors that heavily influenced the development of European dialects during the past 100 years. Many dialects have become more similar to the standard language, and sometimes all that is left is a couple of different words and a regional accent. A "true" dialect differs from the standard variety also grammatically. A popular misconception in Europe is that a dialect has no grammar. Of course it has, for there is no language without grammar! Only the grammatical system of a dialect is not the same as that of the standard language and in addition it is often not made explicit, not described in grammar books or taught in schools. However, it could be, and in recent time many attempts to write down the grammar of a dialect and to prepare teaching material have appeared in print and especially on the Internet. Here is an example:





Website "bairischlernen.de" (Bairisch lernen = learning Bavarian)

## Dialect or regional language?

When a dialect is used in writing and in public settings, when it is taught in schools and its system is fixed in grammar books and dictionaries, people start to ask "How do you write that?" or "Is this correct?". This means the need for standardization arises. The dialect has lost most of the characteristics of dialects mentioned above, except for its association to a certain place or region. In such a situation it may be more adequate to speak of a **regional language** instead of a dialect (see also [Chapter 9 Endangered Languages, Ethnicity, Identity and Politics](#)). Regional languages are found in many European countries, for example [Low German](#) in Germany, [Kashubian](#) in Poland, [Latgalian](#) in Latvia. Typical for these languages is that they are strongly associated with regional identity and with other parts of the culture of the region. For example, Latgalian is traditionally used in the Catholic church, and Catholicism is an important part of the culture of Latgalia, while other regions in Latvia are predominantly protestant. A regional language is most often used alongside other languages, first of all the respective state language – the speakers are bilingual. Regional languages have much in common with minority languages, but there are also important differences. Speakers of a regional language are not a minority, but part of the majority. For example, speakers of Low German are as much Germans as speakers of High German dialects. It is however not straightforward if we should speak of something as a dialect, a regional language, or a minority language. People usually have different opinions about the status of a particular idiom and use different criteria in their argumentation. Quite often it is a topic of heated public discussion. This shows again how important the issue is.

Study question: Which local varieties in your country have been the subject of public discussion? What were the arguments for calling them a dialect or a (regional/minority) language?

## New dialects and social variation

Traditional dialectology, which emerged as a field of linguistic studies in the 19th century, was most interested in rural dialects of a small area and their relationship to neighbouring and other dialects of the same language. For this kind of research, the ideal speaker was an elderly male person who had limited contact with the standard language and whose speech therefore was more traditional and showed "old-fashioned" features. British dialectologists characterized this ideal with the term NORM = non-mobile, older, rural male (Chambers & Trudgill 1980, cited in Barbour & Stevenson 1998: 110). For many non-linguists, too, the stereotype of a dialect speaker is an elder peasant. However, societies in Europe have changed a lot since the late 19th century and the NORM has become a curiosity. Modern dialectological research takes a broader view at dialects and their speakers. For example, linguists now investigate the use of local varieties by different groups within the community, that is, the correlation of dialect speech with social variables such as age, gender, or class. Such research has shown that the reality is often far from the stereotype "old male peasants in the countryside speak dialect, young female students in cities speak standard". The situation is much more differentiated, and it may be quite different in different parts of Europe. For example, in the southern part of Germany there is often a continuum between "pure dialect" and "pure standard", and the speech of different speakers can be placed at different parts of this continuum. It may also vary according to situation and interlocutor – for example, at home with my grandmother I speak a variety closer to the "pure dialect", at school with my friends my variety is somewhere in the middle between dialect and standard, but in more formal situations I speak standard German with just a slight regional accent.

Study question: What is your stereotype of a speaker of a local dialect? Try to think of five possibly different persons you know who speak a dialect – do they conform to the stereotype? Are there differences in the way they speak the dialect?

While in most parts of Europe the “pure” rural dialects that were documented in the 19th century are coming out of use, new local varieties appear. As more and more people nowadays live in cities, **urban dialects** have gained importance for speakers as well as for linguists. An urban dialect often mixes characteristics of a geographical variety (the rural dialects of the surrounding region) and social varieties (the speech of certain groups of society). For example, the Helsinki urban dialect, called *Helsingin slangi* or *Stadin slangi* in Finnish, was originally created and used by young members of the working class. Later it spread among other parts of the society, and today *slangi* is popular in many different spheres. There is even a *slangi* version of the information platform of [Helsinki City Transport](#). The urban dialect of Paris (*argot parisien* in French) had two roots: the speech of Parisian craftsmen and the secret language of crooks (so called thieves’ argot).

### Example of an urban dialect: Stadin slangi

At the beginning of the 19th century, Helsinki was a small Swedish speaking town, but when it became the capital of Finland and massive industrialization started, many young Finnish speaking people moved to Helsinki to work there. In the 1880s, the population was mixed and the city was multilingual: Swedish, Finnish, Russian and German were in use. Helsinki slang was created by workers whose mother tongue was Finnish. The grammar of this variety was the same as in colloquial Finnish, but the vocabulary was formed mostly from Swedish words, with some Russian and a little German. In the 20th century, when it was used by more and different people, Helsinki slang changed. In its modern form it is more similar to colloquial Finnish. While the Swedish element is still strong, new vocabulary now often comes from English.

There are several terms used to refer to varieties used by certain groups of speakers within a speech community. **Sociolect** or **social dialect** is a broad technical term for such varieties in linguistics. Both linguists and laymen use the term **slang** to refer to varieties of colloquial speech. We have just seen that the urban dialect of Helsinki is called slang. Another example is teenager slang – varieties used by teenagers for chatting among friends, often associated with school. Sometimes teenagers of one school even have their own kind of slang which differs from that used in other schools. An important function of slang is to demonstrate and maintain in-group relationship: you can hear if someone belongs to your group or is an outsider. Sometimes slang is associated with a certain culture (often a so called “subculture”). A good example is hip hop culture which originates in cultural practices of Afro-American and Latino youth in New York suburbs and is associated with their slang. As hip hop culture became popular in other parts of the world, elements of this slang spread along with the customs, especially rap music. Varieties associated with a professional field (for example, medicine) or an activity (such as hunting or weaving) are called **jargons** or **language for special purposes**. A jargon is usually not thought of as non-standard language (while a slang typically is), and it may be used both in speaking and writing. For example, hunters’ jargon is used when people are hunting as well as in professional journals for hunters.

These explanations are only rough guidelines – there is no conformity in the use of such terms. Maybe this is inevitable, because the varieties themselves have many facets and can be classed in different ways. What has been defined as slang above is called dialect by some people, while others use “slang” to refer to a jargon and so on. Another term that is used in different meanings is **argot**. In his book on secret language, Blake defines argot as “a body of non-standard vocabulary used by a group bound by common interest, isolation, or their opposition to authority” (Blake 2011: 211). We may make a distinction between argot, slang, and jargon by considering the purpose of their use: the main function of slang is to show the speakers’ membership of a community while an argot is used in order not to be understood by outsiders. A jargon in turn mainly offers more differentiated means for communication within a certain field or about a topic.

### Vocabulary for special purposes

Slang, argots and jargons differ from the standard variety mainly with respect to vocabulary. How do they build their vocabulary, where do new words come from? There are several techniques that can be found in languages all over the world.

First, the words may come from another language. As mentioned above, the Helsinki urban dialect took its vocabulary mainly from Swedish. Teenager slang nowadays uses many words from English. In medical or academic jargon we find words of Latin and Greek origin. The secret language of British Gypsies is (or was) [Anglo-Romani](#), a language based on English but with many Romani words.

**What is Anglo-Romani?** An explanation by Prof. Yaron Matras from Manchester University:

“It is reported that the older generations used to use many more Romani words in everyday conversation, but that use of the

Romani vocabulary has now declined. Speakers may insert a Romani word occasionally when welcoming Romani guests or when meeting with other family members. Sometimes the use of Romani is for humour, and sometimes British Romanies will use Romani words among themselves in public places in order to prevent Gaujos (non-Gypsies) from understanding what they are saying. Thus, someone might say: 'the moosh is dikkin us!' meaning 'the man is watching us!'. Insertion of the odd Romani word into English conversation is often referred to by scholars as 'Angloromani'." (Yaron Matras: Romani in the UK, at: <http://www.bbc.co.uk/voices/multilingual/romani.shtml>)

Can you understand the following comment to Professor Matras' article? Which words come from Romani, can you guess their meaning?

Bryn Heron, Northampton

*My Puri dai and the Rom spoke the pure, inflected chib. Their grandchildren, me included, have only the pogerdi chib, now. I married away from the kawlo rattee, a gawji whom I love to this day. Apart from my grandparents, I have never heard the pure chib spoken. I agree with Jacqueline, though – if you want the pukkered chib, go to the kawlo ratte, not the Romanes Rai or Rawnee.*

See also the Angloromani dictionary at: <http://romani.humanities.manchester.ac.uk/angloromani/index.html>

Second, new words may be created by giving an existing word another meaning. In German hunters' jargon *Licht* (standard German 'light') refers to the eyes of hoofed game, *Mönch* ('monk') is a stag without antlers, but *Schalen* ('bowls' or 'shells' in standard German) are the claws of ground game (examples from the German Wikipedia entry *Jägersprache*). Examples from British thieves' argot include *pig* for 'policeman', *fork* 'pickpocket', *school* 'prison', and *convent* 'brothel' (Blake 2011: 214). The old (standard) and the new (special) meaning may be linked by **metaphor** – a similarity is seen between the two concepts. For example, a stag without antlers is seen as "bald" like a monk with a tonsure. Another technique is choosing a word with an opposite meaning or opposite associations (as in *convent* for 'brothel'). This technique may be used just for being funny, but also in contexts where the speakers don't want to be understood by outsiders. Saying the opposite of what you mean can also be an indicator of a special situation, something out of the ordinary. The Warlpiri people of Central Australia have a variety used in initiation rites called Jiriwirri or "upside-down language". It consists of reversing the meaning of whole sentences. For example, when the young man says "I am short" it means "you are tall".

#### Jiriwirri (or Jiliwirri): saying the opposite from what you mean

(examples from Riemer 2010: 95, citing Hale 1971)

*kari ka ngurungka karimi*

Ordinary Warlpiri: 'Someone else is standing in the sky'

Jiliwirri: 'I am sitting on the ground'

*ngajurna rdangkarlpa*

Ordinary Warlpiri: 'I am short'

Jiliwirri: 'You are tall'

Third, new words can be formed by **word-formation** (see Chapter 3 Language structure) – especially derivation and compounding. The techniques may be the same as in the standard variety, but in slangs and argots there are often some special means of derivation that mark words as belonging to this slang. Sometimes these involve "playing around" with the material of words. Two widespread techniques found in slang and secret languages, as well as in language games popular with children are (i) to insert additional vowels or syllables into a word, and (ii) to reverse the order of syllables or other parts of a word. These two techniques may also be combined. You can find examples from many languages of the world in the English Wikipedia entry *Language games*. An example for the first technique is Latvian *pupinvaloda* ("bean language"), where a syllable consisting of the consonant p and the previous vowel is inserted after each syllable of the word. Thus, *pu-pin-va-lo-da* becomes *pu-pu-pi-pin-va-pa-lo-po-da-pa* (of course, the fun is in speaking these words quickly). Varieties where the main technique is reversing parts of a word are English Back slang, French Verlans, and Bosnian/Croatian/Serbian Šatrovački.

#### Examples for words created by reversing parts of the original word:

**Back slang** (examples from Blake 2011: 217): *look* > *cool*, *market* > *tekram*, *yes* > *say*, *no good* > *on doog*; *hat* > *tach*, *home* > *eemosh*; *old* > *delo*, *knife* > *eefink*, *back slang* > *kecab genals*

**Verlans** (from the French Wikipedia entry Verlans): *argent* > *genhar* 'money', *cigarette* > *garetteci*, *copine* > *pineco* '(girl) friend', *famille* > *mifa* 'family', *femme* > *meuf* 'woman', *comme ça* > *asmeuk* 'that way'

**Šatrovački** (from the English Wikipedia entry Šatrovački): *pivo* > *vopi* 'beer', *kafa* > *fuka* 'coffee', *smrdi* > *dismr* 'it stinks!', *muž* > *žmu* 'husband', *pazi* > *zipa* 'pay attention!'

Two facts are worth noting here. First, although these are primarily or exclusively spoken varieties of a language, at least English Back slang and French Verlans rely on the spelling of a word, thus, written language. For example, if the English word *knife* [naɪf] were just spoken backwards, we would get *fine* [faɪn]. But the Back slang form of this word is *eefink* [iːfɪnk] – the letter “e”, which is not pronounced in *knife*, is part of the Back slang form, where it is pronounced as it is in isolation. The French word *femme* is pronounced [fam], so if Verlans were based on pronunciation the outcome would be [maf]. Instead, it is [mœf] because that is how the letter “e” is pronounced when stressed. The other interesting fact is that this technique and varieties where it prevails are used by very different groups of speakers – from criminals to children.

### Functions of special vocabulary and another look at “cultural” language

All the techniques for vocabulary formation discussed here can have at least three functions:

- they can serve as a code to conceal the content of a communication from outsiders,
- they can be a marker of identity, of belonging to a group (those who know the technique and are able to understand and create new words are “in”), and
- they can be part of playing with language – something not only children enjoy.

A fourth function was touched upon with the example of Jirivirri –

- special words can be used to mark a situation or a conversation as extraordinary.

This function may be less important in Europe, but it is an important part, for example, of Australian aboriginal cultures. Several Australian languages have special varieties used in conversations between family members where a participant of the conversation is by social convention not allowed to speak in an ordinary way to another person. There are certain taboos, words that must be avoided in the presence of certain persons, and therefore a variety called **avoidance language** must be used. As the taboo often involves in-laws, avoidance languages are also called mother-in-law languages (the variety a man must use when speaking to his mother-in-law). In these avoidance languages we find the same techniques as described above: using words from another language, giving words another meaning, or forming new words by special rules.

“Mother-in-law languages” may strike us as exotic, but the wisdom behind this phenomenon is one shared by European cultures as well: social relations determine the way we use language, and certain situations require special ways of speaking. This may bring us to a new definition of a “cultural” (cultivated) person with respect to language: it is someone who uses different varieties according to the social rules of their culture(s). No variety is in itself “bad language” – it only becomes “bad” when used out of place.

### ■ LANGUAGE IS DOING AND CULTURE IS A VERB

For both culture and language, various scholars have independently noted that these concepts are better understood as activities or processes, not as things – they are something we do or something that happens rather than something that exists or something that we possess. To illustrate this idea, we may try to use the names of these concepts as verbs instead of nouns. For example, we may say “we culture” instead of “we have (a) culture”, or “linguaging is an important activity in human life” instead of “language is an important tool for humans”. The cultural anthropologist [Brian Street](#) used the statement “Culture is a verb” in the title of a paper about the problems of defining culture. Reasoning about the nature of language, Wilhelm von Humboldt argued already in 1836 that language is not a product or result of activities, but the activity itself, and a creative force.

This perspective leads us to new questions regarding the connection between culture and language, for example: How and in which situations do we do culture with language? Which linguistic activities are cultural practices? What forms do they have in different cultures?

We may distinguish everyday cultural practices and those performed only at special occasions. Another distinction is between practices shared by all members of a community and customs which are performed only by special members, because they require more training or talent (such as writing poems) or a special status (such as preaching). Examples for customs performed by ordinary members of a culture are: exchanging greetings, saying grace before a meal, thanking for a gift, writing text messages wishing a happy birthday, singing Christmas carols, sending cards at weddings, reading a newspaper at breakfast, reading bedtime stories to children, writing diaries or blogs.... Some of these customs are universal – greeting and thanking are found everywhere in the world – others are more culture specific. Some are **oral practices** (performed by speaking and listening), others are **literary practices** (using writing and reading).





Language happens everywhere... (seen in Berlin, photo Nicole Nau)

When a practice is widespread among cultures, there are still differences in the way it is performed. We become aware of these differences when we learn another language and visit the place where it is spoken. For example, we recognize that it is not enough to learn the words for saying “hello” and “thank you” – we also have to learn the rules for their use. There may be different rules for when you have to say “thank you” and what the person who is thanked replies; different rules for who greets first when two people meet and which particular greeting is used in which situation (“hello” or “good afternoon, Sir”); different rules for whether you should send a card or rather express your congratulation in person, and so on. The more the culture where we are guests differs from our own, the more we recognize how much of our daily use of language is in fact cultural practices. A certain occasion, for example starting a meal or drinking wine at a party, may require just one or a few words in one culture, while in another culture much more has to be said or written.

People in the Caucasus are famous for performing elaborate toasts instead of just saying “cheers!”. Click [here](#) to watch an example of a toast performed in the endangered language Svan.

## Genres (text types)

Language is a constant companion of our everyday life, and it is used for many more purposes than to convey information. Important social acts such as marrying or welcoming a child to a community are performed by speaking certain formulas. Many celebrations require elaborate speeches. Using language is often an important part of religious practices: saying prayers, performing rites, paying respect to God, speaking to the deceased. Each practice comes in a certain form that can be more or less fixed by tradition. For example, in many cultures the marriage vow has a fixed form. This is the English version of the Roman Catholic marriage vow:

I, \_\_\_\_, take you, \_\_\_\_, to be my (husband/wife). I promise to be true to you in good times and in bad, in sickness and in health. I will love you and honor you all the days of my life. (source: [http://en.wikipedia.org/wiki/Marriage\\_vow](http://en.wikipedia.org/wiki/Marriage_vow))


In this example each word is fixed, the vow has to be spoken in exactly this form. Other ceremonies only determine the structure of a text and require the presence of certain elements, but otherwise allow for variation. A speech given by a student at a graduation ceremony will include elements such as: greeting the guests (in a certain order, for example: director, teachers, parents, fellow students), recalling the past years (maybe including some anecdotes), thanking teachers and parents for the education, expressing wishes for the future.

The different forms of different linguistic practices (whether we think of them as cultural practices or not) are called **genres** or **text types**. The term genre is probably best known from literary studies, where it refers to types of literary works such as the drama, the novel, the poem. In linguistics it is used in a broader sense and may refer to more mundane texts as well, both written and spoken. For example, cooking recipes are a genre, as are greeting cards, forum discussions, or oral exams at school. A genre is characterized by its structure, the choice of words and constructions, the structure and length of sentences, and by certain features of pronunciation. It has been shaped by the situation in which the text is produced and by the function it has. The function of a cooking recipe is to instruct how to do something, therefore we find constructions such as imperatives (“take two eggs”). The function of radio news is to inform, therefore they are read in a neutral voice, while a story read to entertain listeners is delivered in a more vivid mode, and a sermon read during worship requires still



another intonation. If you listen to the radio in a language you don't understand, you often can guess which type of program you are listening to.

In each culture we find very many different genres, and it is probably impossible to make a full inventory of the genres used in one speech community. As cultural practices change, some are given up and some new ones are started, genres also change and new ones may be introduced. Take for example cooking recipes. The typical recipe is a short written text published in a cookbook (or a journal, or a web-site and so on). It is written in the absence of the reader and read in the absence of the writer. Though this text type is known from ancient India and China and is widespread in today's Europe, it is evident that it is not universal. It is more natural to pass on knowledge about cooking by showing and explaining while preparing the meal than by writing a text. We also note that the word for recipe is often borrowed (compare in Europe alone: English *recipe*, French *recette*, Spanish *receta*, German *Rezept*, Russian *recept*, Swedish *recept*, Finnish *resepti*), which shows that the genre itself has spread from culture to culture, alongside the practice of sharing knowledge about meals in this form.

 Find the language **Chipaya** on the [Interactive Map](#) with a recipe for quinoa soup! What is the word for 'recipe' in Chipaya?

### Task

At the [Sorosoro website](#) on endangered languages you find two videos where speakers from Africa explain the preparation of a traditional meal. Compare the video (with the help of the subtitles) with the recipes given at the same site: How do these two forms of explanation differ? What do they have in common?

In many European countries cooking shows on television or videos on the Internet are popular today. How is the preparation of meals explained there – is it more similar to the Sorosoro videos or to the written recipes? (For English, you may try the videos on Jamie Oliver's website at [www.jamieoliver.com](http://www.jamieoliver.com))

## Verbal art: Oral traditions

In each culture there are certain texts or text types that have a special status: they are held in special esteem because they are thought of as representing the culture more than other texts and as showing a more elaborate, artful use of language than texts for everyday functions such as cooking recipes. In European cultures this kind of language use is often associated with the word literature. The word **literature** is historically linked to the word letter (Latin *littera*) and thus to writing. This makes it awkward to use this word when speaking about oral texts and performances. Even if the existence of "oral literature" is acknowledged, written texts are thought of as primary and more important for the concept of literature. For example, the Polish Wikipedia entry on literature begins like this:

**Literatura piękna** – typ piśmiennictwa (także dzieł ustnych) ...  
'Literature – a type of writing (also of oral works) ...'

Writing is much younger and much less widespread among the cultures of the world than oral forms of verbal art, which is a more neutral term. A definition that reflects this relationship might start like this:

**Verbal art** – a type of language use (including written works) ...

At the beginning of this chapter we referred to UNESCO's definition of intangible cultural heritage, of which oral traditions are an essential part. The examples UNESCO gives of oral traditions are also examples of verbal art:

### Oral traditions – verbal art

"The oral traditions and expressions domain encompasses an enormous variety of spoken forms including proverbs, riddles, tales, nursery rhymes, legends, myths, epic songs and poems, charms, prayers, chants, songs, dramatic performances and more. Oral traditions and expressions are used to pass on knowledge, cultural and social values and collective memory. They play a crucial part in keeping cultures alive." [UNESCO](#)

Let's take a closer look at one of these types (genres), the riddle, to understand what is special about oral traditions and verbal art.

## Riddles and riddling

Riddles are one of the two shortest forms of verbal art (the other one is proverbs). They often consist in only one sentence. Riddles are found all over the world, though there are some cultures where they are not known or used rarely. Sometimes riddles of different parts of the world are very similar. Here are some examples (sources of the riddles are given in the reference section at the end of this chapter):

language	riddle	solution
<b>Quechua</b>	<i>Rinki rinki qatisunki.</i> ‘You go and go and it follows you.’	<i>Llantu</i>
<b>Mordvinian</b>	<i>Molʹat, molʹi, latkat, latki.</i> ‘When you move it moves, when you stop it stops.’	<i>Sulʹej</i>



It's your...

language	riddle	solution
<b>Tshanglakha</b> (Bhutan)	<i>Lha khang karp chi nang gomchen serp chi yed mi ga chi mo?</i> ‘Inside a white monastery there is a yellow monk.’	<i>Gong do</i>
<b>Ibanag</b> (Philippines)	<i>Pira y levu na / Vulauan y unag na.</i> ‘What is golden that is surrounded with silver?’	<i>Illuk</i>
<b>Latgalian</b>	<i>Puorsit myuru, atrassi sudabru, puorsit sudabru, atrassi zaltu.</i> ‘Knock down the wall and you’ll find silver, knock down the silver and you’ll find gold.’	<i>Ūla</i>
<b>Turkish</b>	<i>Altun apamaz, gümüş tapamaz, o gırlınca dünya yapamaz.</i> ‘Gold cannot carry it away. Silver cannot find it. Once it is broken, the world cannot repair it.’	<i>Yumurta</i>



Photo Biswarup Ganguly (CC)

But riddles may also be specific to a certain culture so that one needs cultural knowledge to solve them. Look at the following examples that again have a similar answer:

language	riddle	solution
<b>Manx</b>	<i>Myr yeeagh mee harrish boalley chashtal my ayrey honnick mee yn marroo curlesh ny bioee ersooyl.</i> ‘As I looked over my father’s castle wall I saw the dead carrying the living away.’	<i>Lhong</i>
<b>Latgalian</b>	<i>Dzeivs byudams, zaļu krūni nas, numiers, duorga dveseleit.</i> ‘While it is alive it carries a green crown, when it is dead it carries a dear soul.’	<i>Laiva</i>

The answer is “ship” (Manx) or “boat” (Latgalian). What you have to know is that ships and boats are made of wood. If your idea of a ship is a big white ship made of metal, you don’t understand the riddle. In addition, you have to think of trees as something living and consequently of their wood as something dead, and you have to see a tree and wood as being essentially the same thing.

Riddles can have at least three different functions. First, they encode a small piece of knowledge, an observation made about something that seems worth sharing. Giving children riddles is a way of passing on this knowledge. This is the educating function of riddles. For example, the following riddle:

language	riddle	solution
<b>Mordvinian</b>	<i>ʹelʹnʹa ašo, kizna sʹormav.</i> ‘In winter white, in summer checkered.’	<i>Numolo</i>

encodes the information that the animal in question (have you guessed it?) has a different fur in summer and in winter. The educating function of riddles is not so much to teach something new, but to strengthen knowledge, including cultural knowledge, by saying what is known in a new, interesting way.



Photo: Steve Sayles (CC, flickr)

Second, riddles have what may be called a poetic function. As pieces of verbal art, riddles are a way to express an observation or an idea in a poetic way, using techniques such as metaphor, parallelism, word play, rhythm, or rhyme. The latter also help to memorize the riddle. Metaphors are an invitation to look at something familiar in a new way, by comparing it to something else. In traditional riddles typical semantic fields for comparison are natural phenomena, body parts, and objects of everyday use. Here are two riddles where body parts are compared to parts of landscape:

language	riddle	solution
<b>Tshanglakha</b>	<i>Tsho nyig tshing rum la rum la dag pa phu thur gi tok pa hang kharbey?</i> 'Two seas are about to merge but blocked by a mountain. What is it?'	<i>Ming nyig tshing cham ka nawong.</i>
<b>Tshanglakha</b>	<i>Dar phu chig gi nang dren po per gang yed pa sholong?</i> 'A handful of guests present in a cave. What is it?'	<i>So</i>

Sometimes a riddle sounds like a little poem:

language	riddle, solution	translation
<b>English</b>	<i>White bird featherless Flew from Paradise, Perched upon the castle wall; Up came Lord John landless, Took it up handleless, And rode away horseless To the King's white hall.</i>	.
<b>Quechua (Peru)</b>	<i>Tras tras chakicha, Frazada qepicha. (Wiqacha)</i>	'Trot trot little feet carrying blanket.'
<b>Sami</b>	<i>Loddi girdá ja varra goaiku soajáin.</i>	'A bird flies and blood drips from its wingtips'

In the Quechua example we find a play with sounds and rhythm: the words *tras tras* have no meaning, they just imitate the sound of trotting (and in this way give a sound image of the concept that is to be guessed), and the word *frazada* 'blanket' is a loan from Spanish that was chosen because its sound shape fits better to *tras tras* than the corresponding Quechua word.

The Sami example is especially intriguing. [Harald Gaski](#), who published this riddle, writes thus about it: "As with poetry generally you can think of several interpretive possibilities". Here is his suggested solution:

"We must think of an evening hour with the sun setting and a boat being rowed on the water. When we observe the boat from land it looks like a bird flying (typically Sami to see the beautiful and poetic in all motion!) and every time the rower takes a new stroke, water drips from the tips of the oars, which against the light looks like drops of blood." ([Harald Gaski](#))

Third, riddles have an entertaining function: they are used to make fun, to show one's wit, or to tease someone. The second Sami riddle that Harald Gaski mentions at his site is a good example of a funny riddle:

language	riddle	translation
<b>Sami</b>	<i>Olmmoš huiká guovtti vári gaskkas ii ge oaččo veahki.</i>	'A person shouts between two mountains but doesn't get help.'

(Hint: the two mountains are a body part and the shouting a certain sound that sometimes escapes from there.)



Up to now we spoke of riddles as a genre, a type of text. If we recall that culture is doing we easily see that the text is only part of the game. Another part is the rules of playing it. A riddle needs to be performed to be a riddle – or, better, to become an instance of riddling. Performance is essential to oral traditions, to those instances of verbal art that do not depend on writing. Furthermore, riddling is a practice that requires interaction. You may write a poem all alone and keep it in your drawer, but the wittiest riddle is not really a riddle when it is not given to another person to guess. It takes at least two to riddle! This is especially true for the educating and the entertaining function of riddles (riddling). In cultures where the art of riddling is alive, it is often performed in certain settings: there is a time and place for it. For example, in rural European cultures before the industrial revolution riddling was a typical activity when people came together to do evening work. The setting is yet another part of the oral tradition.

As oral traditions are performed in certain settings and with certain functions, they are vulnerable to changes within a community. When the setting does not exist anymore (as in the above example – European peasants do not come together in this way anymore) or the function is taken over by other practices (for example, children receive their wisdom from books instead of being given riddles by adults), the cultural practice gets lost. The genre, the text, may survive – riddles are written down and collected in books, but this is not the same as riddling.

When an oral tradition is lost because a community becomes more and more literate and written literature takes up the greatest and most esteemed part of verbal art, two things often happen. For one, the text of the oral tradition becomes a (written) literary genre. In Europe we find literary riddles already in Latin. A literary riddle exists as a text that is not part of a game, a performance. Another difference to the oral tradition of riddling is that a literary riddle often has a known author. As part of an oral tradition a riddle typically belongs to everyone who can play it, it is not important who was the first to think of it (except for situations where the game includes inventing new riddles). Second, as an oral cultural practice riddling becomes a children's game. Children naturally depend less on the written word, as they are only in the process of getting literate. Many European children like to riddle, while adults rarely do it – they think it is childish. However, in societies where the practice of riddling is fully alive, it is practised by all ages. There may be special riddles for children, while others are exclusively for adults.

### Oral traditions and endangered languages

What was said above about riddles could be said about other oral traditions as well. Regarding the fate of oral traditions in literate societies, another good example is the folk tale. Many people in Europe today think that folk tales (fairy tales) are for children. They are told or, more often, read, to children, not to adults. This is the result of a development that started when the cultural practice of telling tales to different listeners was given up and partly replaced by reading, or later by watching television. Where the practice is still alive, there are often skilled and trained story-tellers to whom adults like to listen. On the other hand, the fairy tale has become a literary genre and as such is cultivated by known authors. The traditional folk tale is embedded in a practice carried out in a certain setting and with certain functions – entertaining, educating, passing on cultural knowledge. This tradition has been lost in many industrialized societies.

**Task:** Watch videos of traditional African stories at the [Sorosoro site](#). Which of the videos were produced trying to show the traditional setting of this practice? What characterizes this setting?

You may say: well, in old times people were telling each other tales, today we have books and films. It is only natural that cultural practices change over time. Is that a bad thing?

It isn't a bad thing, if people are happy with it and don't miss anything, which is often the case when the change is gradual. Those interested in the texts – riddles, fairy tales, songs etc. – may enjoy them in their written form, as films or music recordings, without the former practices. If you speak a "big" language with a long history of documenting cultural practices, such as English, German, or Polish, you won't feel that you lost something because people don't tell riddles and tales anymore (or maybe you do?). However, for smaller and endangered languages the situation is different. Here, the cultural change that comes with industrialization and globalization often is sudden, within a few generations. There is no time and no possibility to record old tales and songs. Customs, genres and texts are lost without a trace. New customs (reading books, watching films, performing pop music) are often adopted together with a new language – English or another "big" language that is important in the region. In such a situation it more than often happens that people feel they are losing (or already have lost) something very important: a part of their identity. Keeping oral traditions alive is a way to maintain a language. And keeping a language alive is a way to maintain cultural diversity. This idea is strongly supported by UNESCO, and we will close this chapter with another quotation from their text about intangible cultural heritage:

"Different languages shape how stories, poems and songs are told, as well as affecting their content. The death of a language inevitably leads to the permanent loss of oral traditions and expressions. However, it is these oral expressions themselves and their performance in public that best help to safeguard a language rather than dictionaries, grammars and

databases. Languages live in songs and stories, riddles and rhymes and so the protection of languages and the transmission of oral traditions and expressions are very closely linked.” (UNESCO)

But the last word in this chapter shall be given to the speaker of an endangered language, **Miyako**, the singer **Isamu Shimoji**:

“Me, I sing songs in the Miyako language. I record them on CDs which are then sold in the whole country. I do this because I want the language to last for the future generations. Actually I don't have the feeling that it's my duty or anything. What I mean is just that there is a world that can only be described with the Miyako language. A world which no Japanese words could express, a world which cannot be translated into Japanese. This is a kind of culture that lives within a language. And so I sing with the intention to convey this world in my songs.” (**Isamu Shimoji**, translated by Aleksandra Jarosz)

Listen to Isamu Shimoji sing in Miyako [here](#)!

## LET'S REVISE! – CHAPTER 6

Go to the [Let's Revise section](#) to see what you can learn from this chapter or test how much you have already learnt!

## ■ REFERENCES AND FURTHER READING/LISTENING

### General references

- Barbour, Stephen & Patrick Stevenson. 1998. Variation im Deutschen. Soziolinguistische Perspektiven. Berlin: Walter de Gruyter. [English original: Barbour, Stephen & Patrick Stevenson. 1990. Variation in German: A critical approach to German sociolinguistics. Cambridge: Cambridge University Press.]
- Biber, Douglas & Susan Conrad. 2009. Register, genre, and style. Cambridge: Cambridge University Press.
- Blake, Barry J. 2011. Secret language: codes, tricks, spies, thieves, and symbols. Oxford: Oxford University Press.
- Hale, K. L. 1971. A note on a Walpiri tradition of antonymy. In: Steinberg, D. and Jakobovits, L. (eds.) Semantics. An interdisciplinary reader in philosophy, linguistics and psychology. Cambridge: Cambridge University Press, 472–482.
- Riemer, Nick. 2010. Introducing semantics. Cambridge: Cambridge University Press.

**For teachers:** Lesson outlines for teaching about oral traditions

- <http://www.museevirtuel-virtualmuseum.ca/sgc-cms/expositions-exhibitions/logan/en/index.php?/md/education/curriculumunits/ImportanceOfOral/L4WhatIsOral>
- <http://www.readwritethink.org/classroom-resources/lesson-plans/exploring-world-cultures-through-91.html?tab=1#tabs>

### Dialects

Read and hear more about local dialects, listen to examples and find links to dialect related sites:

English

- British dialects: <http://www.bl.uk/learning/langlit/sounds/>
- English dialects (American, British, Irish): <http://dialectblog.com>
- English dialects and English spoken by different speakers, including immigrants: IDEA International Dialects of English Archive: <http://www.dialectsarchive.com/england>

German

- Eine Deutschlandreise fürs Ohr (Deutsche Welle): <http://www.dw-world.de/dw/article/0,,4230751,00.html>
- [http://www.planet-wissen.de/alltag\\_gesundheit/lernen/dialekte/index.jsp](http://www.planet-wissen.de/alltag_gesundheit/lernen/dialekte/index.jsp)
- <http://www.dialektkarte.de/>
- <http://regionalsprache.de/>

Polish

- <http://www.gwarypolskie.uw.edu.pl/>



## Sources of the riddles in the section on verbal art

- Ibanag: <http://cagayano.tripod.com/arts/riddles.html>
- Latgalian: Łotysze Inflant Polskich, a w szczególności z gminy Wielońskiej powiatu Rzeżyckiego. Obraz etnograficzny przez Stefanię Ulanowską. Część II. Zbiór Wiadomości do Antropologii Krajowej, Tom XVI, Dz. II, 104-218. Kraków 1892.
- Manx: William Cashen's Manx Folklore. Edited by Stephen Miller. Electronic edition. Onchan, Isle of Man: Chiollagh Books 2005.
- Mordvinian: Mészáros, Edit. 2000. Az erza-mordvin nyelv alapjai. Budapest: Budapesti Finnugor Füzetek 14.
- Quechua: Isbell, Billie Jean & Fredy Amilcar Roncalla Fernandez. 1977. The ontogenesis of metaphor: riddle games among Quechua speakers seen as cognitive discovery procedures. *Journal of Latin American Lore* 3:1, 19-49.
- Sami: Gaski, Harald: Folk wisdom and orally transmitted knowledge – Everyday poetry In adages, rhyme and riddles. <http://www.utexas.edu/courses/sami/diehtu/siida/language/folkevisdom.htm>
- Tshanglakha: Dorji, Tshering. 2007. Khar: The oral tradition of game of riddles in Tshanglakha speaking community of Eastern Bhutan. *Journal of Bhutan Studies* 17, 55-82. Available at: <http://www.bhutanstudies.org.bt/journal-of-bhutan-studies-volume-17-winter-2007/>
- Turkish: Başgöz, İlhan & Andreas Tietze. 1973. *Bilmece: A corpus of Turkish riddles*. Berkeley: University of California Press. Cited after: Jennes, Andrew E. 2011. What is the difference between an undergraduate thesis and a riddle? Parsing the linguistic and cultural structures of folk riddling. Thesis, Swarthmore College.

# Multilingualism and language contact

Home > Book of Knowledge > Multilingualism and language contact

## ■ CHAPTER AUTHOR: MICHAEL HORNSBY

**Chapter contents:** Everyone has the potential for multilingualism

How does multilingualism function? Refining the term 'multilingualism'

- Prestige: Official multilingualism
- Lack of prestige: Situations of unbalanced multilingualism
- Benefits of plurilingualism

Language contact

Results of language contact

- Pidgins
- Creoles
- Code switching and loan words

Translanguaging

References & further reading

## ■ EVERYONE HAS THE POTENTIAL FOR MULTILINGUALISM

Multilingualism is a powerful fact of life around the world, a circumstance arising at the simplest level, from the need to communicate across speech communities' (Edwards 1994:1).

Multilingualism may indeed be a fact of life, as Edwards maintains above, and people use the term freely, but what exactly is meant by it? The definition of multilingualism as used here centres on the practice of using more than one language, to varying degrees of proficiency, among individuals and societies. It includes individuals who use one language at home, and another (or others) outside the home; it means people who have equal ability in two or three languages; it includes people who can function much better in one language but who can still communicate in another (or other) language(s); it refers to societies and nation-states who use more than one language in a variety of situations to varying degrees. Basically, multilingualism is the co-existence of more than one language in any given situation, which, according to Guadelupe Valdés, writing on the Linguistic Society of America website, is actually the norm for most people and not the exception:

Contrary to what is often believed, most of the world's population is bilingual or multilingual. Monolingualism is characteristic only of a minority of the world's peoples. According to figures cited in Stavenhagen (1990) for example, five to eight thousand different ethnic groups reside in approximately 160 nation states. Moreover, scholars estimate that there are over 5000 distinct languages spoken in that same small number of nation states. What is evident from these figures is that few nations are either monolingual or mono-ethnic. Each of the world's nations has groups of individuals living within its borders who use other languages in addition to the national language to function in their everyday lives.

(<http://www.linguisticsociety.org/content/multilingualism>)

## ■ HOW DOES MULTILINGUALISM FUNCTION? EXPLORING THE TERM 'MULTILINGUALISM'

Referring to another definition of multilingualism, that of the European Commission, we come to the interesting notion of how exactly multilingualism works in practice. If multilingualism is 'the ability of societies, institutions, groups and individuals to engage, on a regular basis, with more than one language in their day-to-day lives' (EC 2007:6, see also [PDF](#)), then what does this actually look like? The Council of Europe points out that the mere existence of more than one language in any given territory does not mean that multilingualism affects all individuals there:

Multilingualism refers here exclusively to the presence of several languages in a given space, independently of those who use them: for example, the fact that two languages are present in the same geographical area does not indicate whether inhabitants know both languages, or only one.

(Council of Europe: 2007:17).

For example, here is a sign in Glasgow, Scotland (UK) which reflects local multilingualism:

### BOOK OF KNOWLEDGE

Chapters: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) **[7](#)** [8](#) [9](#) [10](#)

[List of all languages](#) referred to in the Book of Knowledge and other sections of the website.

[Glossary](#)

**DOWNLOAD** and print out the Book of Knowledge.

### LET'S REVISE! – CHAPTER 7

Go to the [Let's Revise section](#) to see what you can learn from this chapter or test how much you have already learnt!



Photo: Nicole Nau, August 2011

As the Council of Europe points out, it is likely that such a poster reflects compartmentalized multilingualism – a Glaswegian would be unlikely to know all of the languages mentioned and certainly no more than one or two, if that. Conversely, it is interesting to note which languages are NOT included. **Gaelic**, for example, which has special status in Scotland and in Glasgow in particular, where thousands of Gaelic speakers live, does not appear. It could be argued that as all Gaelic speakers can also speak English, Gaelic is ‘redundant’. Such an example demonstrates that limits are often imposed, with good reason, on the number of languages to be employed in any situation of ‘official’ or public multilingualism.

Therefore, multilingualism can often be seen to refer more to societies and states rather than individuals. When it comes to individuals’ abilities in more than one language, the term plurilingualism might be more appropriate and this has been defined by the Council of Europe (2007:17) as the use of ‘languages for the purposes of communication ... where a person ... has proficiency, of varying degrees, in several languages and experience of several cultures’. Note that we can talk of different levels of ability in the same individual: a person may speak one of his or her languages more easily than another, but she/he remains ‘plurilingual’.

### **Prestige: Official multilingualism**

Different languages can be granted a high status level if they are recognised by governments as official within a given territory. For example, in Equatorial Guinea, Spanish, **Portuguese** and **French** function as co-official languages within the state. This situation can also exist within smaller units of territory. For instance, the city of Brussels in Belgium is officially bilingual, and (in theory at least) people there can deal with public officials in either **Dutch** or French, according to their preference.

This may not always work out in practice though. Let us take the example of Brussels again. In theory, someone can access public services (for example, when paying your telephone bill or visiting a hospital) in either French or Dutch, but in practice, French tends to dominate, as the following passage shows:

So I went to the outpatient emergency department. A doctor came to see me, a woman who originally spoke a language other than French or Dutch, very nice, but who spoke to me in French. I spoke back in Dutch, but she didn’t understand. In the end, I explained to her what had happened in French. Finally, I was allowed to leave the hospital and I was given a letter for my GP (family doctor). And what happened? The letter was also written in French! ... I am bilingual. I know how to express myself in French. But I just wonder: what if I had been a Dutch monolingual, what would have happened?

Thus we can see that even when a territory has official bilingualism or multilingualism, different languages can occupy different levels of what might be called a linguistic hierarchy, depending on the level of prestige, given to the language. Prestige here means the level of respect accorded to a language or dialect as compared to that of other languages or dialects in a speech community. The concept of prestige in sociolinguistics is closely related to that of prestige (or class) within a society. Generally, there is positive prestige associated with the language or dialect of the upper classes, and negative prestige with the language or dialect of the lower classes. For example, the language variety in the Ukraine known as ‘surzhyk’ (the original term means poor quality bread made of mixed rye and oats) is seen as the language of peasants, and is used on the street or at the bazaar by newly urbanized inhabitants and is widely regarded as ‘a pejorative ... collective label for a wide range of mixed Ukrainian-Russian and Russian-Ukrainian language forms that dissolve and intertwine the structures of the two Eastern Slavonic languages’ (Bernsand 2001: 40 ). The following clip is part of a TV campaign in Ukraine to encourage people *not* to use surzhyk, indicating just how low the variety’s prestige is in Ukrainian society. As the presenter explains, the fact that surzhyk doesn’t ‘officially’ exist doesn’t prevent it from being ‘everywhere’. He says it’s neither Russian nor Ukrainian, but the result of the speech of people who don’t know either Russian or Ukrainian well (note the value judgement in such a statement). He then gives many examples of how surzhyk is not considered prestigious but rather a ‘mixed’ language, one example being the tendency for surzhyk speakers to add the Ukrainian suffix -ти (indicating the infinitive of a verb) to Russian verbs.



A map showing the areas in Ukraine where surzhyk (суржик) is spoken (source *The Guardian*)

### Lack of prestige: Situations of unbalanced multilingualism

The example above of Brussels shows that in official multilingual situations, even official languages can occupy different tiers on a hierarchy and that in the case of Brussels, *French* holds more prestige than does *Dutch*. What happens then in situations where unofficial multilingualism exists? Let us take the case of the Democratic Republic of the Congo: this country has French as its official language, and *Lingala*, *Kongo*, *Swahili* and *Tshiluba* as national languages.

A lingua franca is a working, bridge or vehicular language systematically used to make communication possible between people not sharing a mother tongue.

Conceivably, a citizen of this country might expect to deal with a state official in one of these four languages (and indeed in practice might only be guaranteed service just in French) but what about the other 238 languages spoken in the country? There would be no systematic way of ensuring that a speaker of one of these other languages could communicate with a government official other than in one of the four lingua francas mentioned above. Indeed, the practicalities of making all the languages in the Democratic Republic of the Congo ‘official’ are probably high on impossible, thus making full and equal multilingualism a dream, in this particular case at least.

### Benefits of plurilingualism

For both individuals and societies as a whole, there are considerable benefits to be gained from being plurilingual. These are listed below.

Individual plurilingualism:

1. For the individual, plurilingualism is known to produce cognitive advantage (Bialystok, 2001)
2. It improves performance on a range of tasks related to educational attainment (Ricciardelli, 1992)
3. It facilitates the acquisition of literacy (Kenner, 2004)
4. It makes the learning of additional languages easier (Cenoz & Valencia, 1994)

5. It delays the effects of ageing on the brain (Bialystok et al., 2006).

## Societal plurilingualism

1. There are economic advantages for societies in which adults can use more than one language in commercial contexts (CILT/ InterAct International, 2007)
2. Ensuring that public services are linguistically accessible to all produces a more informed and democratic society (Corsellis, 2005)
3. People who grow up speaking more than one language in their daily lives have the potential to gain personally but also to constitute a valuable resource for wider society.

Some possible drawbacks:

It is generally agreed by linguists and educationalists nowadays that plurilingualism provides more benefits than drawbacks, and views such as those of Jespersen, who wrote in 1922 that bilingualism was bad for a child, are now discredited. The circumstances in which people become plurilingual may be problematic though: some may grow up in families where each parent speaks a different language; some may move from one country to another in the course of their childhood and learn a second language at a later stage than the first; while others may speak one language at home with their family and another at school, among other possibilities. Such experiences typically lead to uneven levels of competence in each of the languages in question (Baker, 2006). In such cases, formal instruction in the language(s) in question may lead to more balanced plurilingualism.

## ■ LANGUAGE CONTACT

Whereas multilingualism can exist in separate enclaves, with speakers of different languages living on the same territory not being able to communicate in each other's language (for example, monolingual speakers of English and French in Canada), in situations of plurilingualism, where individuals are using more than one language in their lives, language contact is likely to occur. By language contact, we mean where groups, or individuals, are using different languages and their use of language is modified as a result. This can occur in several different ways. English, for example, has borrowed a great deal of vocabulary from French, Latin, Greek, and many other languages in the course of its history without speakers of the different languages having actual contact; book learning by teachers causes them to pass on the new vocabulary to other speakers via literature, religious texts, dictionaries, etc. But many other contact situations have led to language transfer of various types, often so extensive that new contact languages are created.

Extensive plurilingualism in a given region can lead to diffusion of both vocabulary and grammar across their languages, examples today being areas of Papua New Guinea, the Amazon basin and the Australian desert.

In some communities, the ability to manipulate two or more languages can lead to very intricate patterns of linguistic swapping. Some communities have highly regular patterns of what is known as 'diglossia', where one language variety is used in informal contexts such as the home, neighbourhood, etc., and another is used in more formal situations, usually due to prestige. Sometimes, specific words or phrases alternate in the same sentence, and not whole languages, and we will look at this phenomenon below.

## ■ RESULTS OF LANGUAGE CONTACT

### Pidgins

A pidgin is a communicative code that allows people of different languages to talk to each other without having to go through the trouble of learning each other's languages. Some English-pidgins are *Liberian English* (Africa) and *Chinese Pidgin English*. A French-based one is *Tay Boi*, spoken in Vietnam. It is characterized by reduced syntax and vocabulary, no fixed order of words and used purely as a language of communication. Here is an example of a pidgin used in Hawaii, which English-speakers and non-English speakers use to communicate: <http://www.youtube.com/watch?v=O7X9AAeDCr4>

### Creoles

A creole is a stable natural language developed from the mixing of parent languages; creoles differ from pidgins in that they have been adopted by children as their primary language, with the result that they have features of natural languages that are normally missing from pidgins. For example, whereas pidgins in their phonology might show cluster reduction (e.g. 'dust' becoming 'dus') and morphologically, pidgins do not mark different verbal forms (e.g. by dropping the s-agreement from verbs: 'he goes' becoming 'he go'), creoles do show such features.

Some examples include *Babalua Creole* (based on Arabic and spoken in Chad), *Negerhollands* (Dutch-based and spoken in the U.S. Virgin Islands) and *Krio* (English-based and spoken in Sierra Leone).

Here is an example of *Guadeloupe Creole*, which derives its grammar and vocabulary from Carib, African languages and French. The sign means: 'Slow down. Children are playing here.'





Sign in Guadeloupe creole on tree in a residential area (Lamarre, unincorporated) near Sainte Anne, Guadeloupe, France. Translation: "Slow down, children are playing here!". Guadeloupéen/Guadeloupean Creole French: "Lévé pié aw – Nì ti moun ka joué la!". Date: 30 March 2010, Photo: Kim Hansen

### Code switching and loan words

*Code-switching* is the use of more than one language, or language variety, in conversation. Multilinguals sometimes use elements of different languages in conversing with each other. Thus, code-switching is the use of more than one linguistic variety in a manner consistent with the syntax and phonology of each variety. Here is an example of code-switching using Indonesian, French and English: <http://www.youtube.com/watch?v=wgWQoZz6nEk>.

A *loanword* (or loan word) is a word borrowed from a donor language and incorporated into a recipient language. Donor language terms generally enter a recipient language as a technical term (in connection with exposure to foreign culture. The specific reference point may be to the foreign culture itself or to a field of activity where the foreign culture has a dominant role.

Examples which have come in English include:

- Australian aboriginal languages
  - *Billabong* (in Australian English meaning a water hole or small lake)
- African languages
  - *Jazz* from *Mandinka jasi*, *Temne yas* (meaning 'energy', 'drive')
- North America: from *Algonquian languages*
  - *caribou*
  - *moose*
  - *chipmunk*
  - *raccoon*
  - *muskrat*
  - *opossum*
  - *woodchuck*
  - *terrapin*
  - *skunk*
- Arabic
  - *Gazelle* from غزال *ghazāl*
- Dutch
  - *Aloof* probably from Dutch *loef* (= "the weather side of a ship"); originally a nautical order to keep the ship's head to the wind, thus to stay clear of a lee-shore or some other quarter, hence the figurative sense of "at a distance, apart".
- French
  - *Advertisement* (from *avertissement* [warning])

### ■ TRANSLANGUAGING

Language users may generally think of several linguistic features as belonging together, such as "words" in a "vocabulary". Typically the language users may also assign this group of features to a name, such as "German" – so that a vocabulary would be "the vocabulary of German." Thereby the language users have constructed and agreed upon the idea of a "language" which they call "German". "Speaking a

language” therefore means using features which are associated with a given language – and only such features. However, in real life speakers may use the full range of linguistic features at their disposal, in many cases regardless of how they are associated with different “languages”. Linguaging is therefore the use of language, not of “a language”. ‘Translanguaging’ or ‘polylinguaging’ is the phenomenon when speakers use all their communicative skills and some parts are associated with different “languages”, including the cases in which the speakers know only few features associated with a given “language” (Moller 2008, Jorgensen 2010).

In other words, the speaker in the following clip (<http://guthan.wordpress.com/?s=norman+maclean>) uses both English and Gaelic in telling his personal story. We could perceive it in terms of ‘switching’ between languages or we can view it as a speaker telling his story using both of the languages he knows well, namely *Gaelic* and English. Here is the transcript of the first few minutes:

Tormod MacGill-Eain: Tighinn Dhachaidh

**I said, “Och, I’ll get a smoke at Cill-Amhlaigh.”**

**“Oh, Dhia, you’re not going to Cill-Amhlaigh, a Thormoid.”**

**“What? Where am I going?”**

**“You’re going to the hospital.”**

So, sin far an do **land** mi mu uair sa mhadainn. Baile a’ Mhanaich, anns an ospadal.

**“Well, you can stay tonight. You’re obviously...”** Bha da bhata agam.

**“...You can stay tonight, but we must get in touch with Social Services in the morning.”**

**“Fine, fine.** Leig leam chadal.”

Chaidil mi. Sa mhadainn, thainig an te mhor a bha seo, is ars ise:

**“Right, Thormoid, ‘s e DP a th’ annad a-nist.”**

**“De tha sin a’ meanigeadh?”**

**“Displaced Person.”**

**“Displaced Person? Story of my life. Yeah, right, I am.”**

(Source: <http://guthan.files.wordpress.com/2010/12/gd73.doc>)

Sociolinguistically speaking, we can see this as a fully coherent story. I have **highlighted** the English words in his speech, which form over half of his utterances. Does this mean he is therefore speaking ‘bad’ Gaelic? Not at all. What we have here is an example of translanguaging or polylinguaging, which people who can speak more than one language do all the time. Linguists’ obsessions with labels means that often people want to separate out the two languages, and have the narrative either in *Gaelic* or in *English*, but this would not reflect social reality. The speaker in the clip, Norman, is conflating three different short exchanges, at least two of them being with other Gaelic-English bilinguals who are also codeswitching in the same way he does in the clip itself. Translanguaging is a very important aspect of many minority language communities at the present time and focuses upon communicative contexts, rather than focusing on the minority language itself. This is a natural process, and does not necessarily mean that the language is ‘endangered’, since other, non-linguistic factors are involved in language endangerment.

Acknowledgement: Many thanks to Gordon Wells, project officer at *Sabhal Mòr Ostaig*, for his comments on the Gaelic example in this chapter).

## LET’S REVISE! – CHAPTER 7

Go to the [Let’s Revise section](#) to see what you can learn from this chapter or test how much you have already learnt!

### References & further reading

- Baker, C. (2006). *Foundations of Bilingual Education and Bilingualism*. Clevedon: Multilingual Matters.
- Bernsand, Niklas. (2001). *Surzhyk and national identity in Ukrainian nationalist language ideology*. Berliner Osteuropa Info 17, 38-47.
- Bialystok, Ellen. 2001. *Bilingualism in development*. Cambridge: Cambridge University Press.
- Bialystok, E., Craik, F. I. M., & Ryan, J. (2006). *Executive control in a modified anti-saccade task: Effects of aging and bilingualism*. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 32, 1341–1354.
- Cenoz, J. & J. F. Valenica. (1994). *Additive trilingualism: Evidence from the Basque Country*. *Applied Psycholinguistics* 15, 195-207.
- CILT/InterAct International (2007). *Effects on the European Union economy of Shortages of Foreign Language Skills in Enterprise (ELAN)*.
- Cook, Vivian J. (2001). *Requirements for a multilingual model of language production*. Retrieved from [homepage.nttworld.com/vivian.c/Writings/Papers/RequirementsForMultilingualModel.htm](http://homepage.nttworld.com/vivian.c/Writings/Papers/RequirementsForMultilingualModel.htm)
- Cook, Vivian J. (ed.). (2002) *Portraits of the L2 user*. Clevedon: Multilingual Matters.
- Corsellis, A. (2005). Training interpreters to work in the public services. In M. Tennant (ed.). *Training for the new millennium*. Amsterdam: John Benjamins. Council of Europe (2007). From linguistic diversity to plurilingual education: Guide for the development of language education policies in Europe. [http://www.coe.int/t/dg4/linguistic/Guide\\_niveau2\\_EN.asp](http://www.coe.int/t/dg4/linguistic/Guide_niveau2_EN.asp) (accessed 17 June 2012).
- Cummins, James P. (1981). The role of primary language development in promoting educational success for language minority students. In Leyba, F. C. (ed.). *Schooling and language minority students: A theoretical framework*. Los Angeles, CA: Evaluation, Dissemination, and Assessment Center, California State University, 3-49.

- De Keere, Kobe, Mark Elchardus & Olivier Servais. 2011. *Un pays, deux langues*. Lannoo : Tielt (Belgique).
- Edwards, John. 1994. *Multilingualism*. London: Penguin Books.
- Jespersen, Otto. 1922. *Language: Its nature, development and origin*. London: George Allen & Unwin Ltd.
- Jorgensen, J. N. 2010. *Languageing. Nine years of poylingual development of Turkish-Danish grade school students*, vol. 1-2. Copenhagen Studies in Bilingualism, the Koge Series, vol. K15-K16.
- Kenner, Charmian. 2004. *Becoming Biliterate: Young Children Learning Different Writing Systems*. Stoke-on-Trent: Trentham Books.
- Linguistic Society of America website: About linguistics <http://lsadc.org/info/ling-fields-multi.cfm> (accessed 17 June 2012).
- Moller, J. 2008. *Polylingual performance among Turkish-Danes in Late-Modern Copenhagen*. *International Journal of Multilingualism*, 5 (3), 217–236.
- Ricciardelli, Lina A. 1992. *Bilingualism and cognitive development in relation to threshold theory*. *Journal of Psycholinguistic Research* 21 (4), 301-316.

[back to top](#)

# Language endangerment

Home > Book of Knowledge > Language endangerment

## ■ CHAPTER AUTHOR: MICHAEL HORNSBY

### Chapter contents:

Indicators of language endangerment

- Which languages are endangered?

How do languages become endangered?

- Factors
- Attitudes
- Language endangerment is not always language death

Why does it matter? How do speakers respond to language endangerment?

- How do majority language speakers feel about language endangerment?

References & further reading

According to the linguist [David Crystal](#) (2000), only 600 of the 6,000 or so languages in the world are 'safe' from the threat of extinction. According to one count, 6,703 separate languages were spoken in the world in 1996. Of these, 1000 were spoken in the Americas, 2011 in Africa, 225 in Europe, 2165 in Asia, and 1320 in the Pacific, including Australia. These numbers should not be taken at face value, because our information about many languages is lacking or outdated, and very often it is hard to distinguish between languages and dialects. But most linguists agree that there are well over 5,000 languages in the world. A century from now, however, many of these languages may be extinct. Some linguists believe the number may decrease by half; some say the total could fall to mere hundreds as the majority of the world's languages—most spoken by a few thousand people or less—give way to languages like English, Spanish, Portuguese, Mandarin Chinese, Russian, Indonesian, Arabic, Swahili, and Hindi. By some estimates, 90% of the world's languages may vanish within the next century.

- But what does it mean when we say a language is under threat or endangered?

## ■ INDICATORS OF LANGUAGE ENDANGERMENT

Three main criteria are used as guidelines for considering a language 'endangered':

1. The **number** of speakers currently living.
2. The mean **age** of native and/or fluent speakers.
3. The percentage of the **youngest** generation acquiring fluency with the language in question.

Thus, as a rule of thumb, a language is endangered when the children in a community are being spoken to in a language other than that of their parents. The children may understand their parents' language but will be unable to speak it fluently – they are passive bilinguals. The language is then lost to their children, as they will not be able to speak or understand it at all. This can lead to the situation where grandparents and grandchildren speak totally different languages and sometimes cannot effectively communicate with each other.

UNESCO's Ad Hoc Expert Group on Endangered Languages offers this definition of an endangered language: '... when its speakers cease to use it, use it in an increasingly reduced number of communicative domains, and cease to pass it on from one generation to the next. That is, there are no new speakers, adults or children' (<http://www.unesco.org/culture/ich/doc/src/00120-EN.pdf>). Thus a language with a relatively small number of speakers, such as Icelandic (300,000 speakers) can be considered very much alive as it is the primary language of a community, and is the first (or only) language of all children in that community. Yemba (spoken in the western province of Cameroon, Africa) likewise has 300,000 speakers but is considered endangered as people locally shift towards a linguistic variety known as Pidgin and towards English.

Of course, the above scale of endangerment is not a very sophisticated one. There are many factors which are involved in the endangerment of languages, not just the three "rules of thumb" mentioned above. A more complete scale would look something like that proposed by [Lewis \(2006\)](#), which includes seven parameters of endangerment:

Parameter 1:	Age
AGE-SF1	number of users by age group
AGE-SF2	age of the youngest known user

## BOOK OF KNOWLEDGE

Chapters: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#)

[List of all languages](#) referred to in the Book of Knowledge and other sections of the website.

[Glossary](#)

**DOWNLOAD** and print out the Book of Knowledge.

## LET'S REVISE! – CHAPTER 8

Go to the [Let's Revise section](#) to see what you can learn from this chapter or test how much you have already learnt!

**Comment:** *In other words, if the number of speakers is evenly spread throughout the speaker population, and the youngest children acquire the language as one of their first languages, then the language is not endangered. However, if most speakers are elderly, and children are not using/speaking the language from an early age, the language can be considered endangered.*

Parameter 2:	Demographics
DEM-SF1	number of L1 users
DEM-SF2	number of L2 users
DEM-SF3	number of bilingual L1 users
DEM-SF4	number of language users who report their ethnicity as associated with L1
DEM-SF5	regional population norm of L1 speakers

**Comment:** *The higher the number of native speakers, and the higher number of speakers who consider their language as an essential part of their identity, the 'safer' the language is.*

Parameter 3:	Language Use
USE-SF1	predominant language use in the home
USE-SF2	predominant use of the language in public encounters
USE-SF3	predominant language use in recreation
USE-SF4	predominant language use in the public market
USE-SF5	predominant language use at work
USE-SF6	predominant language use in religious gatherings
USE-SF7	predominant language use in commerce
USE-SF8	predominant language use in mass media
USE-SF9	predominant language use in formal education
USE-SF10	predominant language use in formal public functions

**Example:** *Whilst Breton (a Celtic language spoken in North-West France) fulfills many of the criteria for parameter 4 below, it fulfills few of the criteria for parameter 3 and thus can be considered endangered at this level, and indeed at many other levels.*

Parameter 4:	Language Cultivation, Development, Literacy and Education
DEV-SF1	ongoing transmission of oral literature
DEV-SF2	existence of a practical orthography
DEV-SF3	existence of standardization materials (e.g. dictionaries)
DEV-SF4	existence of literacy instruction materials
DEV-SF5	existence of a significant body of print literature
DEV-SF6	existence of mass media materials
DEV-SF7	existence of elementary education materials
DEV-SF8	existence of secondary education materials
DEV-SF9	existence of tertiary education materials

**Example:** *The lack of literacy among the remaining 34 speakers of Mavea on the Island of Mavea (Oceania) and that most of the island's 210 residents are literate in either English, French or Bislama means that Mavea can be considered endangered according to this parameter (Guérin 2008: 47).*

Parameter 5:	Status and Recognition
STA-SF1	any and all kinds of official and semi-official recognition

**Example:** *Yemba, mentioned above, has no official status in Cameroon and thus the pressure is on speakers to shift towards English and Pidgin. Icelandic, on the other hand, with the same number of speakers, is the official language of a state, which gives it much higher prestige.*

Parameter 6:	Language Attitudes
ATT-SF1	number of community members who positively value their own language
ATT-SF2	number of members of the most significant outside group who positively value the language in question



**Comment:**

*It may seem obvious, but when speakers have a positive attitude to their own language, the more likely they are to use it. The American linguist Nancy Dorian showed how the last speakers of East Sutherland Gaelic did not want to be 'conspicuous', and how this led to Gaelic dying out on the east coast of Scotland.*

**Parameter 7:****Amount and Quality of Documentation**

DOC-SF1	existence of a word list
DOC-SF2	existence of audio or video recordings
DOC-SF3	existence of phonological descriptions
DOC-SF4	existence of grammatical descriptions
DOC-SF5	existence of bilingual dictionaries
DOC-SF6	existence of text collections

**Note:**

*Whereas documenting a language may not actually help speakers of an endangered language (they might be unaware of the existence of such documents), documentation can help linguists and revitalisers who wish to maintain or revive the language in question.*

Thus, depending on how many of the above parameters are met, a language can be described as 'safe' or 'unsafe'. A language like English, German or French would fulfill all of the above criteria; much smaller languages such as [Belarusian](#) (400, 000 speakers), [Kurru](#) (300, 000 speakers in India) and [Rutul](#) (just over 29, 000 speakers in Dagestan and Azerbaijan) obviously would not on many levels. It is worth mentioning at this point as well that not everyone sees the decline in use of a language in the same way, least of all the speakers themselves of an endangered language. For example, while [Ukrainian](#) is the official language of Ukraine, Russian (spoken by a sizeable minority of the Ukrainian population) is often portrayed in terms of being 'endangered' in the country; this, however, is more a political than a sociolinguistic issue. Conversely, people on Singapore are pleased that English is starting to predominate and people are using Chinese languages (such as [Hokkien](#), [Teochew](#), [Hakka](#), [Hainanese](#) and [Cantonese](#)) less and less: 'In the Singapore context, the fact that the shift to bilingualism has happened at the expense of other Chinese languages has always been officially celebrated' ([Chin 2008](#): 74). In other words, language endangerment is sometimes as much a matter of opinion as it is of actual statistics. To take an example from the United States, the rise in awareness of the existence of languages other than English (for example, Spanish) has led to the formation of groups such as the English-Only Movement, who claim that English is 'endangered' because there is some recognition in local government and in schools of other languages. This is clearly not supported by demographic data, which show that the majority of American citizens are native English speakers (82% in the year 2000, according to The U.S. Census Bureau) (<http://www.census.gov/population/www/cen2000/briefs/phc-t20/index.html>).

**CLASSROOM ACTIVITY**

"When language endangerment is imagined" – in the section [Teaching Materials](#) (English as a foreign language)

**■ WHICH LANGUAGES ARE ENDANGERED?**

More than 200 languages have become extinct around the world over the last three generations (<http://www.unesco.org/culture/languages-atlas/>). For example, in the Atlas, the entry for Uganda lists 6 languages, of which 3 are now considered extinct, namely Napore, Nyang'i and Singa.

**LEVELS OF ENDANGERMENT**

UNESCO's Atlas of the World's Languages in Danger categorises 2,500 languages in five levels of endangerment:

- vulnerable,
- definitely endangered,
- severely endangered,
- critically endangered and
- extinct.

It is important to remember that even when a language becomes extinct, some people may in fact 'remember' words or phrases or indeed people speaking in the language when they were young. For example, the following video shows that the Wichita language, spoken in the U.S.A. by about 10 speakers, does have people who remember parts of the language but who don't speak it:

([www.colorado.edu/linguistics/faculty/rood-old/Witchita/movies/Last\\_Witchita\\_W.mov](http://www.colorado.edu/linguistics/faculty/rood-old/Witchita/movies/Last_Witchita_W.mov))

Here you may listen to two sentences in Wichita spoken by one of the last speakers: <http://www.mpi.nl/DOBES/projects/wichita/data>

UNESCO further distinguishes four levels of endangerment in languages, based on intergenerational transfer:

1. **Vulnerable:** Most children speak the language, but it may be restricted to certain domains (e.g. the home). Example: Ingush (spoken in the North Caucasus). Click this link to hear a song in Ingush: <http://www.youtube.com/watch?v=f5h5aaU1V4A>
2. **Definitely endangered:** Children no longer learn the language as the mother tongue in the home. Example: Pech (spoken in Honduras). Click here to see a list of 20 basic words in Pech: [http://www.native-languages.org/pech\\_words.htm](http://www.native-languages.org/pech_words.htm)
3. **Severely endangered:** Language is spoken by grandparents and older generations; while the parent generation may understand it, they do not speak it to children or among themselves. Example: Kaska (spoken in British Columbia, Canada). Click here to access a Kaska language website: <http://kaska.arts.ubc.ca/>
4. **Critically endangered:** The youngest speakers are grandparents and older, and they speak the language partially and infrequently. Example: Achumawi (spoken in California, USA). Click here to hear a folk tale in Achumawi: <http://www.youtube.com/watch?v=Jvth6S-qVnM>

(Moseley 2010)

For the full list of languages which UNESCO considers endangered, see <https://docs.google.com/spreadsheet/ccc?key=0AonYZs4MziZbdEFZOG13WXFLSU9rNkt6cWsxRWIaTUE&hl=en#gid=1>

## ■ HOW LANGUAGES BECOME ENDANGERED

### Factors

While nations all across the world strive to communicate with one another in the hopes of boosting their economy and national interests, they are forced to implement 'official languages' like English, Spanish, French, Russian, etc. to promote the high prestige of speaking an 'international' language. As Crystal (2000: 70) points out:

The full range of factors is fairly easy to identify, thanks to the many case studies which have now been made; what is impossible, in our current state of knowledge, is to generalize about them in global terms. The current situation is without precedent: the world has never had so many people in it, globalization processes have never been so marked; communication and transport technologies have never been so omnipresent; there has never been so much language contact; and no language has ever exercised so much international influence as English.

These factors include:

- **Intermarriage:** According to David and Nambiar (2003), marriages or partnerships where one parent speaks a minority language and the other only the majority language, can have a negative influence in the retention of the minority tongue by the children. The tendency is to adopt the majority language only. For example, Fulfulde (a language spoken in Nigeria) is under threat because of intermarriage with speakers of other languages in the state of Gombe (Baldauf & Kaplan 2007: 197).
- **Market forces:** Ridler and Pons-Ridler (1984) suggest that the choice of language reflects the workings of the market. People choose a language that will benefit them in the long run. In addition, Schiffman (1998) states that language shift (i.e. where people stop using one language and adopt another, more prestigious language) in the minority group is inevitable when the language of the minority is seen as a language which does not help the speakers to improve their socio-economy and social mobility. Thus, the minority group will shift to the dominant language. As previously mentioned, parents in Singapore are shifting toward English and abandoning Asian languages in the home because of the market value the English language has and the advantages it will give their children (Coupland 2011).
- **Migration:** Grimes (2001) notes that sociolinguists agree that migration, either voluntary or forced, is a cause of language shift. When members of a language community migrate, the remaining community decreases in size and thus they may be unable to maintain their language.
- **Assimilation:** Another possible cause of language shift in the family and community is when there is very little difference in terms of lifestyle, custom and culture between the majority and minority language community. It could be argued, for example, that the Welsh have maintained their language relatively more successfully than other Celtic languages because of their literary tradition, based upon the **Eisteddfod** (bardic poetry) festival, thus keeping their identity distinct from that of the neighbouring English. There is a children's version of the **Eisteddfod (Eisteddfod yr Urdd)** which encourages children to participate in traditional poetry and literature in a 'modern' way. Click on this link (<https://www.youtube.com/watch?v=sWJd8lspx84>) to see how younger speakers of Welsh are being encouraged to take up Welsh literary traditions. Of course, this is just one factor among many, since Ireland, Scotland and Brittany also have distinct cultural identities, but this has not prevented massive language shift among Celtic language speakers in these countries as a result.
- **National Education Policies:** According to Grimes (2001), nation-state building through the schools (by educating pupils in the national language) has contributed to language shift in several countries, although it does not cause universal shift of the language. This is because sub-ethnic languages are not given attention in all education policies drawn up by the government. For example, one of the major causes of language shift among regional language speakers in France has been the lack of recognition of these languages in the French educational system.
- **Modernization:** Grimes (2001) notes that modernization, among other things, is a factor which accompanies language shift. When industrialization comes to areas where minority languages are spoken, it is the majority language which is used to train employees in the new plants and factories, and the majority language which is used as a lingua franca.

## Attitudes

Another factor that might lead to languages becoming endangered is the views held by parents. Parents today encourage their children to learn languages of wider communication instead of their heritage languages due to the globalization of the world. Nowadays it is more likely for children to succeed if they are able to speak the popular languages of the world in order to obtain better jobs and prospects.

One major factor that affects the survival of minority languages is the attitudes of the majority language speakers with whom the minority language speakers co-habit on a given territory. One of these groups is the dominant language group (for example, English in Canada) and controls access to authority in the areas of administration, politics and the economy, and gives job preference to those applicants who have command of the dominant language. The disadvantaged language group (in this case, French both inside and outside Quebec) is then left with the choice of renouncing its social ambitions, assimilating or resisting. While numerically weak or psychologically weakened language groups tend towards assimilation, in modern societies numerically stronger, more homogeneous language groups possessing traditional values, such as their own history and culture, prefer political resistance. This type of conflict becomes especially prominent when it occurs between population groups of differing socioeconomic structures (urban/rural, poor/wealthy, indigenous/immigrant). Although in the case of French-speaking Canada, English appeared to be the necessary means of communication in trade and business, nearly 80% of the francophone population spoke only French, and were thus excluded from social elevation in the political/economic sector. It was a small French-speaking elite, whose original goal was political opposition to the dominant English, ultimately brought about socioeconomically motivated **language conflict**, as Nelde (1997) has termed it. (See the following clip from the 1970s, when the Parti Québécois instituted French as the official language of Quebec: <http://www.youtube.com/watch?v=c2Cr693XaFU> ). Such situations of language conflict are inevitably complex and not straightforward.

One writer has suggested that in today's modern world, there simply may not be enough room for too many languages. Harrison (2007: 5) says that 'languages do not literally "die" or go "extinct", since they are not living organisms. Rather, they are crowded out by bigger languages. Small tongues get abandoned by their speakers, who stop using them in favour of a more dominant, more prestigious, or more widely known tongue.' According to Crystal (2000: 77), this crowding out is facilitated by urbanisation, whereby rural populations move into the cities and the learning of the dominant language is more likely. Thus the three key factors in one language being replaced by another appear to centre on the associated power a language has (its status), on its association with elite groups in society (its prestige) and how widely spoken it is and by how many people (its distribution and its demography). Take, for example, the case of the Irish people massively adopting the English language as their vernacular, even though it is the language of the traditional 'enemy'. When the shift toward English began on a huge scale in the 19<sup>th</sup> century, Irish was associated with a rural, isolated and impoverished way of life in the west of Ireland, while English represented one of the most powerful empires in the world. The following extract illustrates the attitude towards Irish, noted in 1927:

Ni raibh aon tora ar Ghaoluinn an uair sin; agus nuair a théighinn go dí aonach Chathair Saidhbhín, n'fhéadfainn mo bhó ná mo chapall a dhíol gan cúnamh fháil ó fhear a' Bhéarla, agus ba bhristecruíoch an obair í sin, ná féadfá do ghnó a dhéanamh gan a bheith a braith ar a' bhfear thall.

Irish was of no use at that time, and when I went to the fair at Cathair Saidhbhín, I could not sell my cow or horse without getting help from the English speaker. It was heart-breaking not to be able to do your business without having to rely on the other man.

(Ó Duilearga 1977: xxviii).

The English language was seen as the only way for people to escape the misery of their impoverished existence and for them to improve their situation (if not them, then their children). That another option was available, namely bilingualism, simply did not occur to speakers of Irish seeking a better life for themselves and their children. Luckily, more and more language communities are now recognizing the benefits of bilingualism and are choosing to have their children educated in the local, endangered language, as well as a more widespread language.

## Language endangerment is not always language death

However, we should not see language endangerment in simplistic terms. Because there are so many factors involved, a language does not usually die out uniformly. It might be vanishing in one place but not in others, for a variety of different reasons. Population size, though important, is not always critical: a smaller group can dominate a larger one – as has been seen often with the European presence in Africa. Moreover, geographical proximity is not always critical for one culture to influence another. The Québécois variety of French (very distinct from the French spoken in France) is being successfully maintained in eastern Canada, despite being surrounded by hundreds of millions of English speakers (see below for an account of how this previously-endangered variety is now secure). And we should be wary of seeing English (or any other globalizing language) as a "killer language", since the spread of such international linguistic varieties are resisted successfully at local levels:

McDonald stores in non-Anglophone countries are not operating in English, just like their menus have not replicated the original American menu. Nowadays, Hollywood movies are usually released concurrently in many major languages (sometimes in editions adapted to local cultures), because the industry is more interested in making money than in spreading English and/or American values. The literature accompanying American computers in non-Anglophone countries is not exclusively in English. BBC and Voice of America radio programs also broadcast in a wide range of languages other than English, which suggests that even at such an ideological outreach level, the spread of English is not the main goal.

(Mufwene 2006: 126)

## ■ WHY DOES IT MATTER? HOW DO SPEAKERS RESPOND TO LANGUAGE ENDANGERMENT?

We should not expect a uniform response to language endangerment any more than we should expect to see a uniform process involved in the disappearance of languages. In many cases, there is cause for regret that particular world views are lost when smaller languages cease to be spoken, as documented by Harrison (2007: 4):

What does it feel like to speak a language with 10 or fewer speakers? For people like Vasya Gabov of Siberia, who at the age of 54 is the youngest fluent speaker of his native **Ös** language, it means to feel isolated and to rarely have an opportunity to speak one's native tongue. It means to be nearly invisible, surrounded by speakers of another, dominant language who do not even acknowledge yours. Speakers in this situation tend to forget words, idioms, and grammatical rules due to lack of practice. When asked to speak, for example, by visiting linguists hoping to document the language, they struggle to find words. **Ös** is now spoken by fewer than 30 individuals, as it is the daily, household language of just a single family. All other speakers reside in households where Russian serves as the medium of most conversations. In this situation, one shared by speakers of thousands of small languages worldwide, it becomes hard to be heard, hard not to forget, hard not to become visible.

### LISTEN AND WATCH

Vasya Gabov talks about the invention of script to write **Ös**: [http://www.youtube.com/watch?feature=player\\_embedded&list=PL77E9549A1F2DB148&v=N3AaMUaCmFw](http://www.youtube.com/watch?feature=player_embedded&list=PL77E9549A1F2DB148&v=N3AaMUaCmFw)

However, as Edwards (2010: 6) notes, 'the forces acting upon a minority-language community may be such that a shift to the overarching variety becomes inevitable'. In such circumstances, it may indeed make economic sense for minority language speakers to shift to the majority language, for the sake of future generations, if nothing else. As outside observers, it is important that we refrain from making value judgements about such decisions. Who does not want to see his or her own children benefit from modernization? Of course, culturally speaking, the loss of the communicative use of a minority language may weaken the sense of group identity a language community has, but it should be noted that 'a language that is no longer regularly spoken may yet have a role to play in the maintenance of group boundaries' (Edwards 2010: 6). Language shift may not always be viewed in such tragic terms by members of the language community in question as it is by outside commentators. If 'the price of original-language retention is geographical and cultural isolation' (Edwards 2010: 11), then in some cases, this may be too high a price to pay. Nevertheless, it should be noted that in many cases, speakers of endangered languages *do* wish that it could all be different. The following clip is by a speaker of Akélé (or Kélé, spoken in Gabon), explaining how she regrets the decline in the use of the language: <http://www.sorosoro.org/en/videos-in-akele-language-gabon#state-language-akele> or: <http://youtu.be/1emhmDX8Aa0>

### How do majority language speakers feel about language endangerment?

David Crystal, in a newspaper article published in 1999, presents a not uncommon view about language endangerment and death: 'Is language death such a disaster? Surely, you might say, it is simply a symptom of more people striving to improve their lives by joining the modern world. So long as a few hundred or even a couple of thousand languages survive, that is sufficient' (The Guardian G2, 25 October 1999). However, not all views are so restrained:

Welsh is an ugly, guttural language and Gaelic is not much better. Languages don't just die because a more powerful nation says it should be so (ask Estonians) but because they lack the means and the flexibility to actually express the subtleties of modern-day existence. English is a fantastically subtle language... and the Scots and the Welsh should consider themselves lucky to be exposed to it from an early age.

(Howard J. Rogers, Australia) ([http://news.bbc.co.uk/2/hi/talking\\_point/664149.stm](http://news.bbc.co.uk/2/hi/talking_point/664149.stm))

Such attitudes suggest a 'survival of the fittest' type of approach to language diversity, in that only some languages deserve to 'live' while others deserve to 'die'. Such talk is unscientific – as Harrison points out (see above) languages are not species which die out or become extinct. And furthermore, such analogies are unhelpful. Languages are systems of human communication which are closely bound up with emotion, affect and identity, which such an approach ignores.

In the past, various authorities have actually tried to engineer language endangerment for political ends. In the United States during the late 19th century, so-called 'Indian boarding schools' were founded to assimilate Native American children into Euro-American culture. In some areas, these schools were primarily run by religious missionaries. Especially given the young age of some of the children sent to the schools, they have been documented as traumatic experiences for many of the children who attended them. They were generally forbidden from speaking their native languages, taught Christianity instead of their native religions, and in numerous other ways forced to abandon their Native American identity and adopt European-American culture (Marr, Online).

A similar system was in operation in the USSR. In 1924, the USSR established the Committee of the North designed to administer the affairs of Northern minorities. Schools were established among the 26 indigenous peoples' groups in the North that included the teaching of indigenous languages. Thirteen alphabets were created using the Roman alphabet for indigenous languages. By 1926, eighteen residential schools were in place across Siberia, and five day-schools had been established. However, in 1937, Northern alphabets were

outlawed. After World War II, the USSR began the process of Russification. Northern groups were forcibly settled into mixed areas in order to assimilate and foster Russian unity. From the age of 2 years, Northern indigenous children were forced to attend boarding schools where they were prohibited from speaking their languages. By 1970, no indigenous used were as language of instruction in schools (United Nations Economic and Social Council 2010: 11).

Crystal (1999) gives very convincing arguments for encouraging linguistic diversity:

We should care about dying languages for the same reason that we care when a species of animal or plant dies. It reduces the diversity of our planet. In the case of language, we are talking about intellectual and cultural diversity, not biological diversity, but the issues are the same ... "Every language is a temple," writes Oliver Wendell Holmes, "in which the soul of those who speak it is enshrined."

Whereas the comparison with endangered species is an emotive metaphor (and again not strictly accurate), it is an understandable one – something *intangible* is lost when a language falls out of use. This is especially true when referring to the emotional and identity aspects of language use – while of course you can be Irish and not speak Irish fluently, be Breton and not speak a word of Breton, be Belarusian and not use the language regularly, most people recognize that at some level at least, identity is closed bound to the language you speak. And while the language continues to be spoken by at least some of the population, there is a reference point for the rest of the community who do not speak the language very fluently, or not at all. It is this emotional connection to language which, I believe, is one of the most compelling reasons to seek to maintain the widest possible spectrum of linguistic diversity

## LET'S REVISE! – CHAPTER 8

Go to the [Let's Revise section](#) to see what you can learn from this chapter or test how much you have already learnt!

## References & further reading

- BBC News: Talking Point – Should minority languages be protected? (Wednesday, 8 March, 2000). [http://news.bbc.co.uk/2/hi/talking\\_point/664149.stm](http://news.bbc.co.uk/2/hi/talking_point/664149.stm) (Accessed 07 June 2012).
- Baldauf, Richard B. & Robert B. Kaplan. 2007. Language planning and policy in Africa: Algeria, Cote d'Ivoire Nigeria and Tunisia . Clevedon: Multilingual Matters.
- Chin, Ng Bee. 2008. Linguistic pragmatism, globalisation and the impact of the patterns of input in Singaporean Chinese homes. In Peter Tan & Rani Rubdy (eds.). Language as commodity: Global structures, local marketplaces. London: Continuum, 70-88.
- Coupland, Nikolas. 2011. The handbook of language and globalisation. Oxford: Blackell.
- Crystal, David. 1999. Death sentence. The Guardian G2, 25 October 1999, 2-3.
- Crystal, David. 2000. Language death. Cambridge: CUP.
- David, M.K. & Nambiar, M. 2003. Exogamous marriages and out-migration: Language shift of Catholic Malayalees in Malaysia. *Multilingua: Journal of Cross-cultural And Interlanguage Communication* 2, 141-150.
- Edwards, John. 2010. Minority languages and group identity. Amsterdam: John Benjamins.
- Grimes, Barbara F. 2001. Global language viability. In Osamu Sakiyama (ed.). *Endangered languages of the Pacific Rim: Lectures on endangered languages 2*. From Kyoto conference 2000, 45-68. ELPR Publication Series C002. ELPR.: Osaka, Japan [http://www.sil.org/sociolx/ndg-lg-grimes\\_topics.html](http://www.sil.org/sociolx/ndg-lg-grimes_topics.html) (Accessed 07 June 2012).
- Guérin, Valérie. 2008. Writing an endangered language. *Language documentation and conservation*. Vol. 2, No. 1 (June 2008), pp. 47-67.
- Harrison, K. David. 2007. When languages die: The extinction of the world's languages and the erosion of human knowledge. Oxford: Oxford University Press.
- Lewis, M. Paul. 2006. Evaluating Endangerment: Proposed Metadata and Implementation. In Kendall A. King et al. (eds.). *Sustaining linguistic diversity: Endangered and minority languages and language varieties*. Washington DC: Georgetown University Press, 35-49.
- Marr, Carolyn. Online. Assimilation through education: Indian Boarding schools in the Pacific Northwest. Available at: [http://www.english.illinois.edu/maps/poets/a\\_f/erdrich/boarding/marr.htm](http://www.english.illinois.edu/maps/poets/a_f/erdrich/boarding/marr.htm) (Accessed 08 December 2012).
- Moseley, Christopher, ed. 2010. *Atlas of the World's Languages in Danger*, 3rd edition. Paris: UNESCO Publishing.
- Mufwewe, Salikoko. 2006. Language endangerment: An embarrassment for linguists. *CLS42: The Panels*, 111-140.
- Nelde, Peter H. 1997. Language conflict. In Florian Coulmas (ed.). *The handbook of sociolinguistics*. Oxford: Blackwell, 285-300.
- Ó Duilearga, Séamus. 1977. (ed.). *Leabhair Sheáin Í Chonaili*. Dublin: Comhairle Bhéaloideas Éireann.
- Ridler, Neil B. & Suzanne Pons-Ridler. 1984. Language economics: A case study of French. *Journal of Multilingual and Multicultural Development* 5:1, 57-63.
- Schiffman, Harold. 1998. Language shift. <http://ccats.sas.upenn.edu/~harolds/messeas/maltamil/node2.html> (Accessed 07 June 2012).
- UNESCO's Ad Hoc Expert Group on Endangered Languages. <http://www.unesco.org/culture/ich/doc/src/00120-EN.pdf> (Accessed 18 April 2012).
- UNESCO Atlas of the World's Languages in Danger. <http://www.unesco.org/culture/languages-atlas/> (Accessed 18 April 2012).





# Endangered Languages, Ethnicity, Identity and Politics

Home > Book of Knowledge > Endangered Languages, Ethnicity, Identity and Politics

## ■ CHAPTER AUTHOR: TOMASZ WICHERKIEWICZ

### Chapter contents:

[Introduction](#)  
[Ethnicity and language](#)  
[Language policy](#)  
[Language planning](#)  
[Language planning in practice](#)  
[Language revitalization and language maintenance](#)  
[Notes](#)  
[References & further reading](#)

## ■ INTRODUCTION

The native language spoken by a certain community serves predominantly as their internal means of communication (thus, the function used is **communicative**).

An equally important and salient function of the language is referred to as **symbolic**. A group of people identify as an ethnic community (ethnos) mainly because they speak the same language (variety), while other/neighbouring communities speak 'otherwise' (i.e. use other language varieties = ethnolects). The term **ethnos**/ethnic community is used here in reference to nations, nationalities, national or ethnic minorities, micronation(alitie)s etc.

Ethnolects spoken by such neighbouring/distinctive communities can be genetically unrelated or distant relatives – in such case they are unambiguously called languages. Frequently, however, they constitute closely related (similar) language varieties, therefore their status as language or dialect (language complex, dialect cluster, etc.) is disputed both inside and outside the community. The status is often determined by means of politics and language policy.

## ■ ETHNICITY AND LANGUAGE

**Ethnicity** is a term that denotes a subjective sense of community, meaning a shared identity based on common descent which results in a sense of group solidarity. Very often it is regarded by parts of the given community as an objective criterion. The term 'ethnicity' derives from the Greek *éthnos* which originally meant a large, relatively homogeneous group of warriors or a group of animals.

This ethnic sense of community is based on many factors, among which the following are considered the most important:

- the aforementioned and crucial sense of common descent, i.e. common lineage
- language, more appropriately the ethnolect – this is a term in use by linguists (such as Joshua Fishman and Alfred F. Majewicz) to denote a linguistic variety employed by an ethnic group, seen as a vernacular as well as constituting an indicator of the group's separate identity. In linguistic classification it can correspond to such varieties as: subdialect, dialect, a group of dialects, language and L-Complex (Majewicz 1989: 10-11). The term ethnolect often provides the possibility to avoid (fruitless) debates on whether a given linguistic variety (which can be identified as separate by the ethnic group) is a language or a dialect;
- a similar and subjectively common culture: both spiritual (immaterial) and material;
- Geographical location and the territorial continuity associated with it. Also references to such territories once owned in the past;
- The sense of distinctiveness from other groups (of this kind) as a means of group integration.

Being subjective and instrumental (by means of, e.g. politicians, sovereigns/leaders, the group itself or other groups), the nature of many of these factors contributes to the fact that many scholars on ethnicity more and more often are describing ethnicity and the ethnic group with the term **constructed identity**. Although language is usually considered as one of the most unequivocal determinants of ethnicity, the criterion of language – as well as the remaining criteria – is often employed in order to construct ethnicity. Thus, the aim of linguistic distinctiveness is to substantiate ethnic distinctiveness, and vice versa: ethnic identity is supposed to substantiate the autonomous nature of the linguistic variety used by the community.

This can be exemplified by the recent, on-going debate in Poland on the status of Silesians and their Silesian language. The distinctiveness of the Silesian language acts as a basis for their separate ethnic identity, while the ethnic identity in question is to serve leveraging the status of the Silesian ethnolect.

Follow this link to find the web portal which controls and coordinates the standardization processes concerning the Silesian language:  
<http://www.ponaszymu.eu>.

The discussions on the ethnic status of the Silesian people can be illustrated by the covers of books written by both sceptics and supporters of the idea.

### BOOK OF KNOWLEDGE

Chapters: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#)

[List of all languages](#) referred to in the Book of Knowledge and other sections of the website.

### [Glossary](#)

**DOWNLOAD** and print out the Book of Knowledge.

### LET'S REVISE! – CHAPTER 9

Go to the [Let's Revise section](#) to see what you can learn from this chapter or test how much you have already learnt!



'Are the Silesians a Nation?' 'The History of the Silesian Nation'

The case of Wallonia represents an example of a different kind of relation between the ethnic and linguistic processes concerning identity. Wallonia is the southern part of Belgium which has gained broad autonomy due to the long standing efforts of the Flemish-speaking inhabitants of North Belgium. Such autonomy was gained along with Flanders, the region of Brussels, and the German-speaking region of Eupen-Malmedy (East Cantons) in East Belgium. The ethnopolitical processes of decentralization and the creation of autonomous regions in Belgium has led Wallonia to seek and develop its own distinct identity; and even to postulate the recognition of its 'Walloon' languages (apart from Walloon per se, also the Picard, Lorrain and Champenois ethnolects) as being an L-complex which is distinct from the French language. This L-complex would be a mark of Walloon ethnicity – separate from both the Flemish-speaking and German-speaking Belgian people, and from the French or the immigrant French-speaking populace of Belgium. Below you can see two editions of *The Little Prince* (*Le Petit Prince*) translated into Walloon and Picard.



The Little Prince (*Le Petit Prince*) translated into Walloon and Picard.

Generally, making use of the mutual relationship between ethnicity and language/ethnolect results from the political aspirations of certain groups and/or their leaders. Often they hold expectations and require linguists to fulfil them by providing conclusive evidence on whether the linguistic variety in question is a fully-developed language or a mere dialect of another (national and official!) language. This matter cannot be resolved, however, by considering only the intralinguistic features (the lexical and grammatical systems) of language. It is the extralinguistic criteria that determine the high or low linguistic status of a variety: its geopolitical, military and historical background. This is best described by the well-known metaphor attributed to the Yiddishist Max Weinreich: 'a language is a dialect with an army and navy'.

Language does not constitute the sole determinant of ethnicity – this can be supported by examining examples of different ethnicity that are based on the same language or its similar varieties. For example, the ethnicity of Dutch people in the Kingdom of the Netherlands and of Flemish people in the Kingdom of Belgium; the ethnicity of Croats, Serbs, Bosnians, and the Montenegrins; the ethnicity of the British, USA Americans, Australians, New Zealanders and many others.

It is, thus, impossible to assign a separate language/ethnolect to every existing ethnic group – and vice versa. It is an even more daunting task with regards to national states. This means, as was said previously, that leaders/politicians and groups of interest will strive to raise the status and prestige of their own language variety in order to underline and emphasize their ethnicity and the distinctiveness from other ethnic groups they already possess. Opposite phenomenon/processes may occur as well: based on the differences between linguistic varieties, the group will want to create and strengthen their ethnic separateness. Many of such linguistic varieties are endangered in today's world. Raising their status and prestige may reduce this endangerment and/or halt the process of language death, at least according to the groups employing the given variety. Another notion that reflects the subjective nature of the bonds binding communities into ethnic groups is the notion of the imagined community. This term was coined by Benedict Anderson and introduced in the title of his book reissued in 1991. The notion can be most easily explained by understanding an ethnic group (or a nation in a broader context) as a community which has been socially constructed through means of a subjective belief of its members that they are, indeed, members of the said community. Yet again this can be exemplified by processes which develop ethnicity without the role of language as the decisive factor, as well as processes which construct ethnic identity hand in hand with language separateness.

The Hui people can act as an example of the first process. The Hui people are a community of nearly 10-million Chinese Muslims who are recognized by Chinese law and by the Chinese government as an ethnic minority. They, however, do not have their own language that could act as a distinguisher. Arabic, used in the whole of the Muslim world as a liturgical language, cannot be treated as such – it is not used as means of communication neither by the Hui people nor by many other peoples practicing Islam. The legal recognition as an autonomous ethnic group and, on top of that, the granting of the autonomous region of Ningxia in North China strengthened the ethnic identity of the Hui people. They can be, thus, treated as an example of an ethno-confessional identity (one based on religious separateness).

Follow this link to a video clip regarding the Hui people's identity. Pay attention to the standardized form of Chinese that the characters are using: <http://www.youtube.com/watch?v=DOWggGVfgQ>.

In the chapter devoted to writing systems (Chapter 5) the case of the Dungan people is discussed. The Dungan people are a community deriving their origin from the Hui people who settled in Kyrgyzstan and Kazakhstan. Given the diaspora, their language became autonomous – and their ethno-confessional identity developed into an ethnic one.

These linguistic and ethnic processes are employed also in the formation of collective identity of, for example, the Rusyns. The Rusyns are communities that use East Slavic language varieties in Central Europe, that live or lived in the Carpathian mountain range and, also, that adhere to the traditions of Eastern Christianity (and, thus, use the Cyrillic script – see Chapter 5 on Writing Systems). The **Rusyn L-complex** (often and by many treated as a group of dialects of Ukrainian) comprises of the following varieties: **Lemko** in Poland, **Pryashiv Rusyn** in Slovakia, **Subcarpathian Rusyn** in Ukraine, **Pannonian Rusyn** in North Hungary, **Rusyn in Vojvodina** (an autonomous province in Serbia), and also, according to some, the **Boyko** varieties in Poland and Ukraine, and **Hutsul** in Ukraine. The first Rusyn variety that created its own official literal standard form was Vojvodinan Rusyn – it was even one of the official languages there. It all happened in the times of the Socialist Federal Republic of Yugoslavia. Below can be found a photograph taken in Novi Sad (capitol of Vojvodina) showing a sign meaning “bus station” in: Serbian, Croatian (the first and second signs are at the same time the Cyrillic and Latin versions of the Serbo-Croatian language), Hungarian, Slovakian, Romanian and Rusyn.



Photography by Tomasz Wicherkiewicz

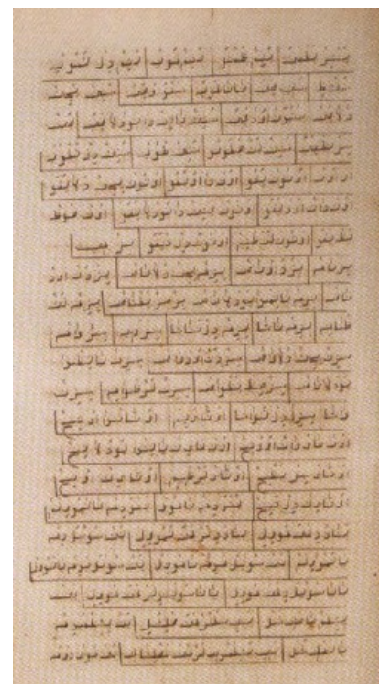
Although the Rusyn standard was well established in the Yugoslav Vojvodina, it cannot be treated as the norm of a common Rusyn language as it contained too many West and South Slavic features. Apart from that, the political status of Rusyns was different depending on the country they lived in. The recognition of Rusyns as a separate ethnos/nation took different amounts of time – in Ukraine it was only 2012 that a bill had been put forward that granted regional language rights to minority languages, Rusyn among them. The creation of a standardized form of the Rusyn L-complex is, thus, a matter of time – similar to the formation of a common Rusyn ethnic identity.



Follow this link to find a performance of the Rusyn 'national' anthem – such symbols (as anthems, flags), too, constitute essential factors in the process of forming an ethnic identity: <http://www.youtube.com/watch?v=KBh6BKgWugk>.

Yet another illustrative example of the transfer from a clearly religious identity to an ethnic (or ethno-confessional) identity and, then, a linguistic one is the case of Lipko Tatars (also known as Polish-Lithuanian-Belarusian Tatars and Tatars of the Grand Duchy of Lithuania; not to be mistaken for, e.g. Crimean Tatars or Volga Tatars). Deriving from different tribes of the Golden Horde, Tatars started settling in, or were resettled to, Lithuania by the XIV century and in Poland by the XVII century. Their descendants relatively quickly abandoned the Tatar language and formed their identity based on the traditions of Islam, although under heavy influence of local Christian traditions. The Arabic script was a source of Muslim values and was even employed by the Polish-Lithuanian Tatars to record the local Polish-Belarusian varieties. Below can be found a page from an Arabic-Polish dictionary which was a part of a *kitab* – a kind of a hand-written book preserved to this day by many Tatar families.

The Polish-Lithuanian Tatars for over three centuries functioned as an ethno-confessional community lacking a vernacular ethnic language. It was most probably due to the European revitalization movements that they decided upon stepping on the path of 'revitalizing' their own language – in their own understanding of the term. On the official website of the Tatar Union of the Republic of Poland (Związek Tatarów Rzeczypospolitej Polskiej: <http://www.ztrp.org/index.php/tatar-tele.html>) one can find basic information about the Volga Tatar language and its (Cyrillic) alphabet, along with exercises. It is around the Volga variety that the Polish Tatars want to construct their ethnicity – with no connection to past times, thus giving them a fresh start.



Source: Drozd, Dziekan & Majda 2000: illustration 29.

The intensification of ethnic movements around the world was most noticeable by the end of the 20th century. Since then, they have been accompanied by endeavours to raise the status of many linguistic varieties. In many cases these ethnopolitical movements aim to counter the European ideology behind language that pervades the world, and break up with the synonymy of:

#### **Nation=state=language**

Introduced after the French Revolution, this notion constitutes the building block of many countries over the last two hundred years, especially of the 20th century. It still is believed to be an acceptable model of how a (national) state should function along with its (national) language. This simplified correspondence can, however, be easily countered by comparing the number of countries in the world with but an approximant number of ethnic groups and the number of languages discussed in other chapters of this publication. The policies of many countries aim at ethnic homogeneity – that is, making the whole of a population of a country uniform and attributing them with a feeling of being a part of an indivisible nation. Languages may become 'victims' of such ethnic homogeneity policy; especially language that do not enjoy the political status and social acceptance of being national languages. These officially unrecognized minority groups have answered to this by launching numerous ethno-linguistic movements in the 20th and 21st centuries that strive to grant these minorities legal rights in different domains:

- the recognition as an ethnic minority – as in the case of the Rusyn people,
- the granting of a special territorial status – as in the case of the Kashubian language in the Pomerania region of Poland,
- the granting of politico-territorial autonomy – as in the case of the autonomy of Silesia,
- the recognition of the language variety as a separate language (raising its linguistic status),
- the recognition as an auxiliary language, second language, or an equal co-official language on a given territory.

As has been repeatedly emphasized in this chapter, the vast majority of the world's countries are multi-ethnic societies, which trait in turn is connected with their multilingual character. Among them one can list: Papua New Guinea and the Republic of Indonesia (it comes as no surprise that the island New Guinea, with around a thousand languages being spoken there, belongs to both of these countries); the island countries of the Republic of Vanuatu and the Solomon Islands in Oceania; the Republic of India, the People's Republic of China, Malaysia, Nepal, Myanmar (Burma), the Socialist Republic of Vietnam in Asia; the Republic of Cameroon, the Democratic Republic of Congo, the Central African Republic, the Republic of Chad and the Republic of Tanzania in Africa; also the United States of America and the Russian Federation.

Listing monolingual and, thus, mono-ethnic countries causes much more difficulty. These can be, for example: The Democratic People's Republic of Korea and the Republic of Korea (both considered being the most homogenous countries in the world). Countries that are almost fully homogenous with respect to language and ethnicity can be listed the following way: the Caribbean Republic of Haiti, the Republic of Cuba, the island state of St. Vincent and the Grenadines (the indigenous Indian populace has left the island by the most part), the Republic of Salvador in Central America, the Independent State of Samoa in Oceania, also the Republic of Rwanda and the Republic



of Burundi in Africa, as well as the Republic of Maldives in the Indian Ocean. The Vatican passes as the most monolingual country in the world (even though legally it has two official languages: Italian and Latin. The second language is not even used on the country's official website [www.vaticanstate.va](http://www.vaticanstate.va). It is worth mentioning that only the official status of the two languages is being discussed here; in reality, citizens of Vatican are multilingual due to the multinational nature of the Catholic clergy, administrative officers and service staff. A 'Vatican ethnic identity' hardly exists at all).

## ■ LANGUAGE POLICY

One of the definitions of language policy describes it as a set of any legislative acts that strive to shape the relations between society and the language or languages that exist and are used in this society (if at least symbolically) (Kaplan & Baldauf 1997: 12). Language policy is concerned with:

- the official language or languages (known also as state or national languages) that function either *de iure* (through legislative proceedings as French in France or Polish in Poland) or *de facto* (in practice as in English in the United Kingdom or the United States)
- regional languages
- (ethnic/national) minority languages
- sign languages (used by the deaf, deaf-mute and hard-of-hearing communities)
- immigrant languages
- foreign languages that are taught and spoken
- classical/dead languages existing in the education system and certain occupations (for example, Latin in medicine) or used in religious context (for example, Latin, Ancient Greek, Old Church Slavonic, Biblical Hebrew, Quranic Hebrew, Sanskrit, and the classical Grabar language of Armenia, and others).

In many countries and especially in Europe, the term regional language denotes an autochthonous (i.e. indigenous) language which is not a language of an ethnic minority (that is, its speakers do not feel ethnically/nationally separate from the majority of the society) and which is closely related to the major (official) language of the given country. It is so to the point of naming it a dialect of the national language, even though in many cases historically it developed not as a variety (as a dialect) but in parallel. Among such languages one might list [Kashubian](#) or [Silesian](#) in Poland, Low German in Germany and the Netherlands, Scots in Scotland/Northern Ireland, [Asturian](#) in Spain, [Latgalian](#) in Latvia, [Samogitian](#) in Lithuania, and [Võro and Seto](#) in Estonia.

Recently [sign language](#) has begun to be treated similarly to languages of ethnic minorities – mostly by sociolinguists but, in some countries (for example Finland), also by their governing bodies.

Usually, language policy constitutes an element or one of the areas of a country's ethnic and cultural policy. Apart from making decisions that regulate the relations between different ethnic groups in a country, the state may also take action in order to either support a given language or language variety, or quite the opposite; it may discourage their use or even ban the varieties in certain or all domains of language use. As a whole, these decisions and actions can be described as language planning. In extreme cases they can be called language engineering – a phenomenon which can be exemplified by the language policies of Lenin and Stalin in Soviet Russia, which were taken as an example in the People's Republic of China with regards to its ethnic minorities. The aim of its endeavours was not only to regulate the relations between ethnic and language groups in the country but also to manipulate the groups through their creation, merging, separation, transformation or removal. This was done by means of nomenclature, border change, ethnic propaganda and, of course, language policy.

Throughout human history the actions of different language policies have taken the shape of the persecution of one ethnic group and/or granting privileges to another. These actions can include assimilating whole groups or individuals into the major ethnolinguistic group, depriving the groups of the possibility to express their group ideals and values, relocating them due to their ethnicity, introducing inner colonialism (assuming that a certain group inhabiting one's country is inferior and treating it as such), or even, in extreme cases, committing ethnic cleansing and genocide.

Less violent examples of unequal language policy include stereotypization, i.e. the forming and spreading of bias towards one's ethnicity, nationality, race or religion, and ubiquitous ethnocentrism, i.e. considering one's own ethnic group as the privileged one and concentrating the efforts of ethnic policy on supporting it. An ethnic policy that supports different or all other ethnic groups inhabiting a given country encompasses, for one, the equal or, at least, similar consideration for every element forming an ethnic identity of a minority or a majority group. Also, it should provide minority groups the right of participation in making political decisions as well as it should ensure the equality of the rights of minorities with the rights of the majority. In the context of minorities which are threatened with extinction or full assimilation, language policy ought to take steps towards employing positive discrimination/ affirmative action (enforcing auxiliary laws that apply only to the given group in order to ensure their survival) and giving legal and political guarantees concerning the support and protection of minorities both on a national and international level.

## ■ LANGUAGE PLANNING

comprises any decision or action made by the government, or different organizations, institutions, groups or individuals which are to affect the presence, use and development of a language or languages.

These decisions are made to answer the demands of society and politics, when, for example, different groups of speakers of different languages or language varieties strive to ensure the unhindered and day-to-day presence of their languages in various domains, or when some of these domains are unavailable to certain (often minority) groups. State institutions and NGOs are faced, thus, with the

necessity/need to fulfill these linguistic needs, which are voiced by either the majority of the society (in which case it often concerns the linguistic uniformity of the whole country) or by minority groups (who often are in favour of a country being as multilingual as possible), and to fulfill them effectively and justly. In multilingual states/societies the ideal scenario consists of a situation in which state institutions and governmental bodies and their service are available at a similar level to users of different vernacular language varieties. In reality, however, the aim of language planning is to limit linguistic diversity by appointing a single, official language in a multilingual state or, for example, by ascribing the status of a “standard” language to one of the used varieties in order to promote linguistic uniformity. The first type of action plan may be exemplified by the language policy of the Republic of Indonesia, one of the most multilingual regions in the world. Across all of its territories and domains of public life, it actively supports the use of a unified Indonesian language *Bahasa Indonesia*. The second action plan can be represented by the People’s Republic of China which strives to consolidate the country and nation by means of encouraging the use of the Mandarin Putonghua, but one of the many linguistic varieties that exist in China.

## ■ LANGUAGE PLANNING IN PRACTICE

Languages, especially endangered ones, very often require decisive and informed language planning campaigns in the following areas:

- education, e.g. specialized curricula concerning the teaching of a language and the teaching in a language
- normalization and standardization of endangered languages
- support for endangered language’s publishing industry
- the legislative conditions concerning language
- raising the prestige and status of a given language in public life
- promoting bilingualism in trade and in the workplace
- promoting bilingualism in administration and society
- encouraging the increase of the presence of a language in the (local) linguistic landscape.

Initiatives taken up as a part of language planning may concern various areas of language use and language presence:

- corpus planning – meaning the standardization of language and its normalization (see below); the choice of one particular language variety over another to become considered the basic form, whether acting as an official language or literary language; the publishing of language guides, dictionaries, normative grammars, handbooks of orthography; the creation of research circles or associations (concerning literary studies or linguistics), publishing handbooks, and also organizing spelling competitions;
- status planning – is concerned with granting status, for example, an official, co-official, working or auxiliary status to a language on a national or regional level, especially in domains of education, administration, service, trade or media; sometimes also in the areas of regional legislature, justice system and relations with other groups. Status planning also applies to prohibitions imposed on certain areas of language use;
- acquisition planning – it encompasses all initiatives (legal instruments and normative acts included) that regulate the teaching of a language in the domain of education (teaching through a language, teaching as a national language, teaching as a foreign language, pre-school and adult courses, external courses) and language competitions that promote learning;
- language technology planning – often considered the newest dimension of language planning. It creates the possibility of language use in text editors, on-line dictionaries, cell phones, cash machines, etc.

Language planning, as understood from the above listed aspects, should be employed in a number of stages. The first stage usually is the analysis of linguistic behaviour in society (or a smaller community) that uses a certain language or a whole range of languages (like the previously mentioned official, regional, minority, community, taught, and foreign languages). The next stage is to choose which language variety will be liable for language planning and its following aims:

codification – the provision of criteria or features of a “correct” language variety,

standardization – the establishment of a uniform version of the language, the development of vocabulary and a variety of languages styles allowing the language to function in a broad array of domains,

normalization – the development and expanding of the aforementioned areas of language use. An example of this comes from Catalonia. For the past thirty years, the regional government of Catalonia has put in place policies and other initiatives to encourage Catalan speakers (both fluent and potential) to use the language in all domains possible and make Catalan the “default” language of communication. Click [here](#) to see some examples of posters produced, which help speakers make sure they know the correct term in Catalan for a wide range of subjects.

language cultivation – the establishment of language councils and academies, the encouragement to use the language and its codified variety.

The ultimate aim of language planning, thus, is to maintain the presence of a language in each domain of public life in which the language has the capacity to function. Another primary aim of language planning is to create the best possible environment for broad language development, even though at the turn of the 21st century most of the languages of the world have been deprived of such environments.

## ■ LANGUAGE REVITALIZATION AND LANGUAGE MAINTENANCE

In sociolinguistic and language ecology terms, strategies and actions that focus on the restoration of (at least some) functions and areas of

usage to endangered languages are known as language REVITALIZATION, REVIVAL and RECLAMATION. These strategies and actions usually occur after a certain period of limited or completely abandoned use of such languages. The term language MAINTENANCE, on the other hand, is employed to describe such strategies and actions that support and strengthen an endangered language that still functions and which is spoken by young users (the youngest generation still learns the language), though its use is growing weaker.

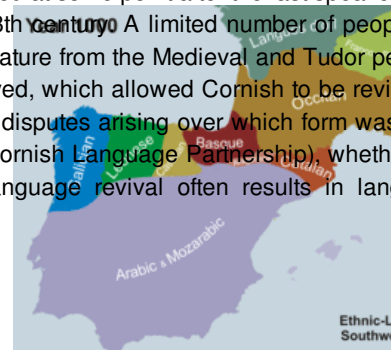
In popular terms, language maintenance, revival and revitalization are often used interchangeably to describe different linguistic situations. Strictly speaking, though, the terms refer to different processes in linguistic management. In case of the official languages, language maintenance occurs when some official body makes provision for the inclusion of new loanwords from other languages, as happens with the Académie Française in France, for example. It also refers to legislation which makes a language official – for example, when the majority of American states declared English as the official language (variously, between 1812 and 2008; see [here](#) for details), the aim was to maintain or enforce the dominance of English in these states. Language revitalization concerns the strengthening of a language that has suffered loss in the number of speakers. Evidence from placenames shows that Basque was once spoken much further eastward in the Iberian peninsula than it currently is, indicating that the language has been receding over the centuries (see the animated picture on the right for a visual representation of this shift)

However, the number of Basque speakers is in fact increasing – according to official statistics, there were 528,500 speakers of Basque in 1991, 665,800 speakers in 2006 and 714,136 speakers in 2011 (Gobierno Vasco 2012). This is largely due to revitalization efforts in the schools.

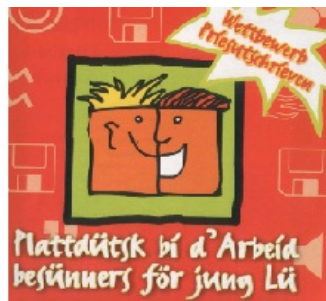
Language revival refers to a language which lost all of its speakers and has been reconstructed at some point after the last speaker died. Cornish was used as a community language in Cornwall, United Kingdom, until the late 18th century. A limited number of people did continue to use the language throughout the 19th and possibly the early 20th centuries. Literature from the Medieval and Tudor periods, and substantial fragments, including grammars, from the 17th, 18th and 19th centuries survived, which allowed Cornish to be revived in the early 20th century. However, the revival led to six forms of ‘revived Cornish’, with many disputes arising over which form was most ‘authentic’. In 2008 a ‘Standard Written Form’ was agreed upon by most users of Cornish (Cornish Language Partnership), whether this has settled all the disputes remains to be seen (a review was due in 2013). Thus language revival often results in language **transformation** as much as anything else, as reproducing the speech of speakers from the past is nigh-on impossible.

Language revitalization has two main aims:

- Teaching the language to those who do not speak it – this is fulfilled through different kinds of educational campaigns, that is actions within the scope of language acquisition planning. This subject will be discussed further on when examples of teaching endangered languages will be presented;
- Encouraging both learners and fluent speakers of a language to use it – in a more and more varied range of situations and areas of usage. Organizations, institutions and unofficial groups often take up campaigns to promote this notion. These kind of initiatives are carried out often in a form of posters that promote extensive and frequent language use. See below the posters concerning Frisian (*Praat mar frysk* [1] – which translates as “Just speak Frisian!”), Kashubian (*Ma tiż rozmiejemë gadac pò kaszëbskù* – “Here we speak Kashubian, too!”) which are placed in trade and service offices and encourage to use Kashubian out of home). Another example is a flyer prompting to take part in a competition – *Plattdütsk bi d’Arbeid – besünners för jung Lü* (“Low German in the working place – especially for the young”). [2]



Linguistic map. Southwestern Europe  
(Wikipedia commons)



Posters concerning Frisian, Kashubian, and Low German

\*\*\*

Leanne Hinton (2011: 292-293) enumerated a set of actions corresponding to the possibility of revitalizing/reclaiming an endangered language or a “sleeping” language (that which has no native speakers – one can only try to revive this kind of language). These actions include:

- Teaching a number of words/phrases such as greetings, short conversations and expressions allowing for establishing contact with

others;

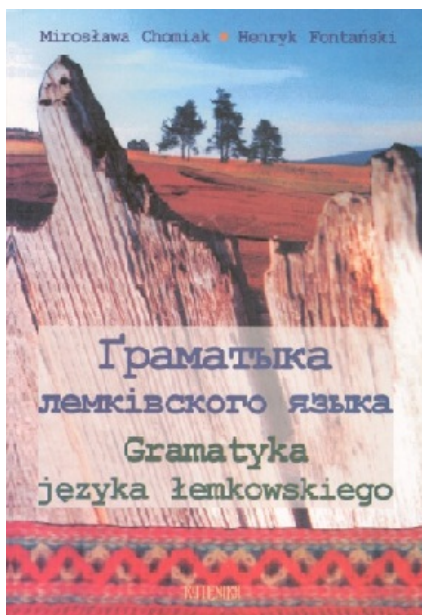
- Gathering publications in and about the endangered language, preparing notes and recordings in order to create an archive of the language;
- Creating/correcting writing systems, creating dictionaries (also illustrated dictionaries for children) as well as practical grammars and course books (also alphabet books) for language learning;
- Recording audio- and/or audiovisual materials of the living speakers of the language in order to create linguistically varied corpora so that examples of language use can be documented;
- Organizing language classes, summer schools and language camps;
- If possible, running community schools in the form of immersion schools for children (see below).

It is the teaching of language, whether in existing or in new educational facilities, that constitutes the most frequently employed tool of language revitalization/language maintenance. There are many ways to teach an endangered language:

**- regular language classes within the scope of a school curriculum.**

It was not until 2001 that the **Lemko** language was introduced to a number of schools in Poland. It is the vernacular language of the Lemko people which are acknowledged as an ethnic minority in Poland. In other countries they are known as Rusyns, Ruthenians or Carpatho-Rusyns. According to the General Censuses of 2002 and 2011 the number of people considering themselves Lemkian is about 5000-7000. Almost all of them declared that they use the Lemko language at their homes. The language is endangered mostly due to the assimilation by the Polish-speaking environment and also by the fact that Lemko was identified for many years as a dialect of Ukrainian. Today this language is taught from 1 up to 3 hours per week by 268 children in the whole of Poland: 5(!) children in 4 kindergartens, 110 children in 20 primary schools, 90 students in 10 junior high schools, 20 students in 2 high schools, 1(!) student in 1 vocational school and 42 students in 1 school complex.

More information and materials covering the Lemko people and their language can be found on [www.lemko.org](http://www.lemko.org); you may also listen to the Lemko radio **LEM.fm** on [www.lem.fm](http://www.lem.fm), watch an interview with a Lemko Orthodox priest A. Graban from Strzelce Krajeńskie (*Keeping the Lemko language alive* – <http://www.youtube.com/watch?v=JXSxZoZiOOk>), listen to Lemko songs by the young band Lemko Tower (<http://www.youtube.com/user/LEMKOTOWER1>), and also read the brochure about the teaching of Lemko (and Ukrainian) in Poland on [http://www.mercator-research.eu/fileadmin/mercator/dossiers\\_pdf/Ukrainian\\_Ruthenian.pdf](http://www.mercator-research.eu/fileadmin/mercator/dossiers_pdf/Ukrainian_Ruthenian.pdf).



A grammar and a course book for teaching Rusyn

**- bilingual teaching encompassing the learning of a number of subjects in the endangered language.**

The network of bilingual Sorbian-German schools in Upper Lusatia in Saxony, Eastern Germany, may serve as an example. The Sorbs are a small West-Slavic nation who uses two (standard) Sorbian languages: Lower Sorbian and Upper Sorbian.

The Upper Sorbian language is mainly used in the traditionally Catholic part of Saxony and, with around 18.000 people speaking it, is considered endangered (Ethnologue 2009). Lower Sorbian developed in the traditionally Protestant Brandenburg (former Prussia) and is in serious danger of extinction having less than 7.000 users. Bilingual schools provide the teaching of Upper Sorbian; Lower Sorbian is



being revitalized via immersion schooling in a network of schools and nursery schools known as **Witaj**.

The role of education in supporting the Sorbian languages in Germany is described in this brochure:  
<http://www.mercator-research.eu>



A bilingual Sorbian-German schools in Croswitz in Upper Lusatia



The cover of the same fable book published in Upper- and Low-Sorbian

Another example of this type of education focused on endangered language communities, in particular in lowly populated areas, are boarding schools. Here you can see children from a South-African school which runs, among many other, classes of the revitalized languages of the Khoisan language family.



Photo by. T. Wicherkiewicz

- **bilingual schooling encompassing the teaching of all (or most) subjects in the endangered language** – sometimes shifting between teaching in the minority language and in the dominant majority language. The photography below comes from a primary school in Baygamut – one of the two settlements of the Kazakh minority in Russia. The school is attended by less than 20 students and is located in a remote part of the Altai Krai. Despite the abject poverty of the area, due to the eager support of the school's head office, its teacher and the local education authorities, the Kazakh language, just like Russian, is a fully developed and frequently used language, which is by no means endangered.





Photo by. T. Wicherkiewicz

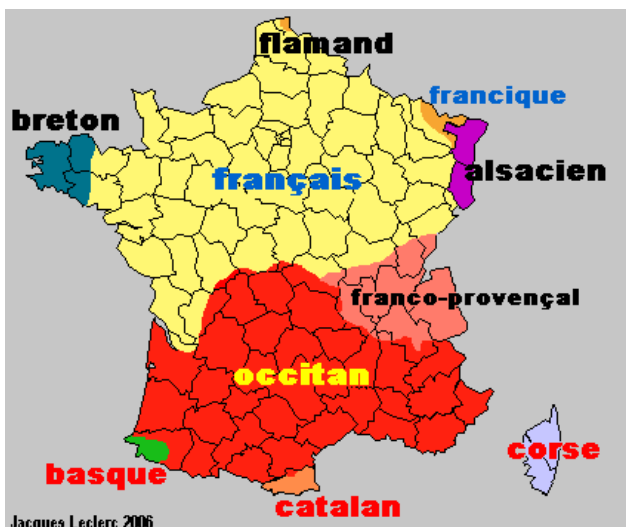
– **the aforementioned immersion schooling** – the language in which teaching is conducted in immersion schools and nursery schools is the endangered language itself. The dominant language of the majority is often taught as a foreign language. An exceptional example of immersion schooling are “language nests” (Spanish “nidos de lengua”) in which the roles of teachers in nursery schools (or the first years of primary education) are taken up by grandparents who are the last fluent speakers of the dying language. Such “language nests” have achieved great success in the revitalization of these languages, among others:

- the **Māori** language of New Zealand [3] (Te Kōhanga Reo – see <http://www.kohanga-reo.co.nz>)
- the **Hawaiian** language of Hawaii[4] (Pūnana Leo – see <http://www.ahapunanaleo.org>)
- the vernacular languages of Mexico, the **Aztec** and **Mayan** languages included (see the video clip on Nido de Lengua de Izamal [5] <http://vimeo.com/36603789>)
- the **Saami** languages of the Saami people [6]
- the **Lower Sorbian** language spoken by Lower Sorbs in Eastern Germany is being revitalized through the functioning of the network of immersion nursery schools and immersion schools Witaj (see: <http://www.witaj-sprachzentrum.de>; <http://www.witaj.de>).



Logo of the Lower Sorbian immersion schooling programme *Witaj*

– **school networks established on the initiative of the community itself** – they were set up in order to teach endangered minority languages on the initiative of parent groups in particular as a reaction to the hostile or indifferent attitudes/policies of state educational authorities. One should mention here the numerous schools for children belonging to the autochthonic language minorities of France. The French Republic's constitution does not acknowledge the existence of minority languages on its territory, naming them instead *langues regionales*. The legislation of the educational system does allow, however, for the establishment of such community schools which provide a wide range of minority language teaching methods.



source:<http://medialzas.wordpress.com/2008/05/27/17>

- the **Breton** *Diwan* school network ([www.diwanbreizh.org](http://www.diwanbreizh.org)) [7]
- the German-language schools *ABCM Association –Zweisprachigkeit* in Alsace (<http://www.abcmzwei.eu/sprachigkeit>) [8]

- the Catalan *Bressola* school network (<http://www.bressola.cat>) [9]
- the Occitan *Calandretas* school network (<http://calandreta.org>) [10]
- the Basque *Ikastolak* school networks in Spain and France ([www.ikastola.net](http://www.ikastola.net)) [11].



Promoting bilingualism in the ABCM school network  
(<http://www.abcmzwei.eu/sprachigkeit>)

– **classes organized outside the school system** – by institutions, individuals, summer language schools, etc. *Wilamowicean*, a language used in Wilamowice in southern Poland, may serve as an example of a language which has been revitalized using exclusively extra-institutional methods. The efforts to revitalize the language have been undertaken mostly by Tymoteusz Król, a teenager at the time he began. He runs private language lessons, he is in the process of creating a dictionary and a grammar of the language, archives it through making recordings and notes; he also writes in Wilamowicean, collects Wilamowian folk costumes and popularizes the knowledge about the Wilamowicean language and Wilamowian culture (see <http://www.youtube.com/watch?v=8nbJ9fW7WWk>). Another example: the teaching of a very endangered language, Karaim. The picture below shows prof. Éva Á. Csató Johanson, originally from Budapest, now working in Uppsala, assisted by a native speaker of Karaim, who are teaching the language to a group of children on a summer language course in Trakai in Lithuania. [12]

*[there will be two more video clips here]*

So far, examples of the areas of use of a language have been discussed, as well as the areas in which revitalization is possible. Now it is time for the illustration of ways that the presence of a language can be supported in a variety of its areas of use:

- underlining the presence of language in the **private linguistic landscape** – here is a sign placed in a car transporting little children. The text is in the South Estonian language *Võro*. Of course apart from cautioning other drivers, it is also a means of manifesting one's knowledge of the language and the fact of using it even in such a peculiar communicative context, although it is usually used explicitly in home and family situations.



photo by. T. Wicherkiewicz

- Underlining the presence and usage of the language in the **public linguistic landscape** – which can be exemplified by stickers that encourage to use lesser used languages, in this context the two varieties of Southern Estonian – *Võro* and *Seto* [13]



Photo by T. Wicherkiewicz

- Some endangered minority languages may use **institutional support** in order to become present not only in the public, sometimes official, linguistic landscape but also to sway the legal authority’s decision. Here is the Saami Parliament in the town of Kurina (which is Giron in Saami) in Lapland Sweden.



Photo by T. Wicherkiewicz

– Another way of supporting an endangered minority language is to promote it even in the most official domains used by the state. Below one can see a bilingual **Friulian-Italian** poster encouraging to participate in the Italian General Census 2011. [14]

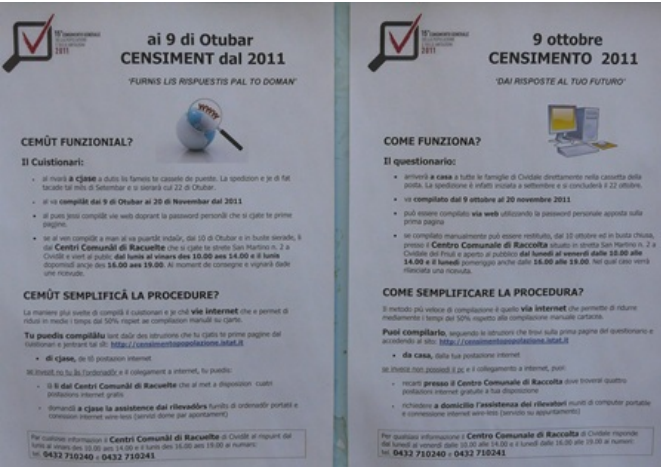


Photo by T. Wicherkiewicz

– One of the areas that endangered languages rarely engage in is **trade and commerce** – it is, however, an area of great importance for the correct functioning of every language. In the photography one can see a shop in the town of Vielha with the name and advertisements in the **Aranese language**. [15]





Photo by T. Wicherkiewicz

– Certainly, preparing and publishing attractive **educational materials** can serve a great purpose in strengthening a language's status, especially there where it is taught.



Primers and course books for learning Võro

- Popularizing and promoting the presence of an endangered lesser used language in mass culture – here is the South Estonian band *Neiokõsõ*, which achieved great success in the 2004 edition of the Eurovision music festival. They represented Estonia with their song sung in the Võro language – an action which lead to a great increase in the Estonian society's interest in this endangered language variety (listen here: <http://www.youtube.com/watch?v=8gOwmmxQaho>);



- **Microlanguages** are languages that are spoken by a small group, of a few or of a dozen, last speakers. These languages do, indeed, require a huge range of revitalization efforts, if revitalization is possible at all. That is why these efforts often employ the documentation of the endangered language in writing and recordings as comprehensively as possible, as well as the symbolical combining the language into the linguistic landscape.

The Livonian or Liv language of Latvia can be included in the category of “sleeping” languages. It is spoken by a handful of last speakers, most of which, if not all, have been framed within the picture below (the boy lying on the floor on the left is also on the cover of *the Alphabet Book for learning Livonian*). Right next to it see the cover of the newest (and probably last) alphabet book for learning Livonian – the boy on the cover holds an alphabet book published before the II World War when Livonian children were taught their language in schools in north-west Latvia. More information about the Livonian language can be found here: <http://www.livones.net>.



## Notes:

[1] The website [www.praatmaarfrysk.nl](http://www.praatmaarfrysk.nl) has been dedicated to this initiative, as well.

[2] More information about these languages (mainly their presence in the education system) can be found in these brochures:

- [http://www.mercator-research.eu/fileadmin/mercator/dossiers\\_pdf/frisian\\_in\\_netherlands4th\\_072010.pdf](http://www.mercator-research.eu/fileadmin/mercator/dossiers_pdf/frisian_in_netherlands4th_072010.pdf) (Frisian)
- [http://www.mercator-research.eu/fileadmin/mercator/dossiers\\_pdf/kashubian\\_in\\_poland.pdf](http://www.mercator-research.eu/fileadmin/mercator/dossiers_pdf/kashubian_in_poland.pdf) (Kashubian)
- <http://www.ciemien.org/mercator/bulletins/55-13.htm> (Low German)

[3] The Māori language is the indigenous language of New Zealand (the Māori for Māori is Aotearoa – see the entry in the Māori version of Wikipedia <http://mi.wikipedia.org/wiki/Aotearoa>); it belongs to the East Polynesian group of languages. Nowadays it is the co-official language of New Zealand being efficiently revitalized and supported by a varied array of campaigns, educational ones included. According to Ethnologue 2009 Maori is spoken by around 60.000 people, while a further 100.000 people are able to understand it.

[4] Hawaiian is also an East Polynesian language. It is, however, more seriously endangered than Māori as the number of its speakers spans from 1.000 to 8.000 (Ethnologue 2009) regardless of the fact that there are 240.000 native Hawaiians on Hawaii (around 100.000 more live in other US states). At the beginning of the 20th century, 37.000 people considered Hawaiian their first language.

[5] Itzamal is a town in the Mexican Yucatan Peninsula (The Mayan name is Itzmal). A large portion of its inhabitants speak the Mayan language Yucated as their vernacular language. The language is not considered endangered because it is transferred from generation to generation, and the total number of its users (according to Ethnologue 2009) amounts to 700.000. The teaching of this indigenous language is conducted through the system of language nests which is hoped to maintain it.

More on the "Language Nest" programme in New Zealand and Mexico can be found in the brochure:

<http://www.cneii.org/documentos/libros/nido-de-lengua-ingles.pdf>

[6] Revitalization initiatives (for example, the language nests project) are to halt the rapid death of the Saami languages, especially the extremely endangered Inari Sámi (300 speakers in Finland according to Ethnologue 2009) and Skolt/Koltta Sámi varieties (400 in Finland, 20 in Russia) – see <http://www.skr.fi/default.asp?docId=19214>

[7] Breton is the only Celtic language that is spoken outside the British Isles, on continental Europe. According to Broudic (2007), it is spoken by about 270,000 people. More about Breton in education can be found in this brochure:

[http://www.mercator-research.eu/fileadmin/mercator/dossiers\\_pdf/breton\\_in\\_france2nd.pdf](http://www.mercator-research.eu/fileadmin/mercator/dossiers_pdf/breton_in_france2nd.pdf)

[8] The Alemannic variety of German is still a second language for the majority (1.5 million) of the inhabitants of Alsace, the North-Eastern region of Germany with its capital in Strasbourg. More about the teaching of the German language in the French Alsace can be found in this brochure:

[http://www.mercator-research.eu/fileadmin/mercator/dossiers\\_pdf/german\\_in\\_france2nd.pdf](http://www.mercator-research.eu/fileadmin/mercator/dossiers_pdf/german_in_france2nd.pdf)

[9] Catalan is an example of a language which is by no means endangered in one country (The Kingdom of Spain in which it has over 11 million speakers along with its Valencian and Balearic varieties) and , at the same time, is threatened with a break in intergenerational transmission and with the shrinking number of domains of usage in another country (The French Republic where Catalan is spoken by around 100,000 users). More about the teaching of Catalan in France can be found in this brochure: [http://www.fryske-akademy.nl/fileadmin/mercator/dossiers\\_pdf/catalan\\_in\\_france.pdf](http://www.fryske-akademy.nl/fileadmin/mercator/dossiers_pdf/catalan_in_france.pdf)

[10] The Occitan language is, actually, a whole L-Complex encompassing the Southern-French varieties of Provençal, Gascon, Languedocien and Limousin. Despite a vast number of speakers (close to 2 million according to Ethnologue 2009) it is a language threatened with the break of intergenerational language transmission. More about Occitan in French education can be found in this brochure: [http://www.mercator-research.eu/fileadmin/mercator/dossiers\\_pdf/occitan\\_in\\_france.pdf](http://www.mercator-research.eu/fileadmin/mercator/dossiers_pdf/occitan_in_france.pdf)

[11] Basque is a non-Indo-European language and, similarly to Catalan, it is almost in no danger in the Kingdom of Spain (where it has over 500,000 speakers) and moderately endangered in the Republic of France (where Basque is spoken by about 75,000 people). More about Basque education in France can be found in this brochure: [http://www.mercator-research.eu/fileadmin/mercator/dossiers\\_pdf/basque\\_in\\_france2nd.pdf](http://www.mercator-research.eu/fileadmin/mercator/dossiers_pdf/basque_in_france2nd.pdf)



[12] The Karaims are the smallest of the officially acknowledged ethnic minorities in Poland (around 50 people) and Lithuania (270 people). The Turkic Karaim language that originated on the Crimean Peninsula is used by around 120 speakers (and an unverifiable number in the Ukrainian part of the Peninsula). The website [www.karaimi.org](http://www.karaimi.org) is dedicated to the Karaim people; more about the revitalization of the Karaim language can be found on [http://www.hrelp.org/events/workshops/eldp2007\\_9/resources/multimedia.ppt](http://www.hrelp.org/events/workshops/eldp2007_9/resources/multimedia.ppt) as well as [http://eprints.soas.ac.uk/6083/1/karaim\\_orthography.pdf](http://eprints.soas.ac.uk/6083/1/karaim_orthography.pdf)

[13] The **Võro** and **Seto** varieties are used in the south of Estonia (Seto also on the borderlands in Russia) by, accordingly, 70,000 and 10,000-13.000 people. Despite a separate history of development and a separate linguistic identity from that of Estonians (in the case of Seto, also a separate cultural-religious identity), these languages have no official status in Estonia which is a factor strongly impeding their vitality. Since some time ago, school and extra-school language courses have been organized.

Võro Institute website: <http://www.wi.ee>; more information on the Seto region can be found here [www.setomaa.ee](http://www.setomaa.ee) (the English version is quite poor, unfortunately). More about Võro in education can be found in this brochure:

[http://www.mercator-research.eu/fileadmin/mercator/dossiers\\_pdf/voro\\_in\\_estonia.pdf](http://www.mercator-research.eu/fileadmin/mercator/dossiers_pdf/voro_in_estonia.pdf)

[14] the Friulian language is one the minority languages of Italy; the picture was taken in the small town of Cividale del Friuli (which is Cividât in Friulian). Due to the support of the Italian state, this language is used by nearly 800,000 speakers and is, thus, less endangered than it was in the past. More on Friulian-Italian bilingualism can be found here: <http://www.youtube.com/watch?v=rXQXCGdrRk> ; the website of the Friulian Language Society is: <http://www.filologicafriulana.it>

[15] Aranese is a language variety of Occitan. It is used on a patch of land of the autonomous communities of Catalonia in the Kingdom of Spain by a mere 4,000-5,000 people. It has been granted, however, the status of a co-official language along with Catalan and Spanish in Val d'Aran (and since 2010, the whole of Catalonia). More about the language and the region can be found on [www.aranges.org](http://www.aranges.org)

## LET'S REVISE! – CHAPTER 9

Go to the [Let's Revise section](#) to see what you can learn from this chapter or test how much you have already learnt!

### References & further reading:

- Anderson, Benedict R. O'G. 1991. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. London-New York: Verso.
- Coulmas, Florian (red.) 1997. *The Handbook of Sociolinguistics*. Oxford-Cambridge: Blackwell.
- Drozd, Andrzej, Marek M. Dziekan, Tadeusz Majda 2000. *Piśmiennictwo i muhiry Tatarów polsko-litewskich*. Warszawa: Res Publica Multiethnica.
- *Ethnologue. Languages of the World* 2009 (16<sup>th</sup> edition). Dallas: SIL International.
- Fishman, Joshua A. 2000. *Can Threatened Languages Be Saved?* Clevedon: Multilingual Matters.
- Hinton, Leanne 2011. "Revitalization of endangered languages; w: Austin, Peter K. & Julia Sallabank (red.) *The Cambridge Handbook of Endangered Languages*. Cambridge University Press. 291-312.
- Kaplan, R. B. & Baldauf, R. B. 1997. *Language Planning: From Practice to Theory*. Clevedon: Multilingual Matters.
- Majewicz, Alfred F. 1989. *Języki świata i ich klasyfikowanie*. Warszawa: Państwowe Wydawnictwo Naukowe.
- Cornish Language Partnership. Standard Written Form. <http://www.magakernow.org.uk/default.aspx?page=346> (Retrieved 30 November 2012)
- Gobierno Vasco (July 2012). "V. Inkesta Soziolinguistikoa". Servicio Central de Publicaciones del Gobierno Vasco. [http://www.euskara.euskadi.net/r59-738/es/contenidos/noticia/inkesta\\_soziol\\_2012/es\\_berria/berria.html](http://www.euskara.euskadi.net/r59-738/es/contenidos/noticia/inkesta_soziol_2012/es_berria/berria.html) Retrieved 30 November 2012.

**English translation by:** Piotr Szczepankiewicz. **Translation update:** Nicole Nau/Michael Hornsby.

[back to top](#)

# Language documentation

Home > Book of Knowledge > Language documentation

## ■ CHAPTER AUTHOR: KATARZYNA KLESSA

### Chapter contents:

What is language documentation, how is it done and why is it important?

- More than just words and sentences
- Data and metadata

Some methods of language documentation

- Documenting communicative behaviour and language in use
- Documenting what speakers know

Elements of documentary practice

- Preparatory steps
- Recording equipment and session set-up
- Processing and analysing data
- Example on-line archives for endangered languages
- Legal and ethical problems

Appendices: More about the history of sound recording, data formats and structures

References

Useful links

## ■ WHAT IS LANGUAGE DOCUMENTATION, HOW IS IT DONE, AND WHY IS IT IMPORTANT?

One of the important needs of humans is the desire to preserve the memory of the most meaningful achievements of their lives and to pass on the knowledge about their times, cultures and civilizations to the next generations. Over the centuries, people have developed various ways of transmitting knowledge from generation to generation based on oral tradition (oral culture) and written texts. Because of this, it is nowadays possible to track back and explore even quite distant history, as well as the history of language. An exception is the sounds of language, since it was impossible to preserve the sound of speech of our ancestors until not fairly recently: the first recordings of a human voice that we can listen to are dated only to the second half of 19th century and are therefore very 'young' compared to the most ancient written documents reaching back many centuries in the past. In this chapter, we will look at issues related to the preservation and use of information about languages. We will particularly focus on endangered languages and on the possible reasons why they should be dealt with in a special way. Some of the suggested reasons are listed in the box below.

### WHY DOCUMENT ENDANGERED LANGUAGES?

- To preserve human cultural heritage in general;
- To keep memory of the facts important for the local communities, families, individuals;
- To better illustrate linguistic theories with real-life observations of languages in use;
- To study language contact [1].

**Language documentation** comprises the collection, processing and archiving of linguistic data – for example, texts, word lists, recordings of conversations, videos where people tell fairy tales, etc. While people interested in languages have carried out such activities for centuries, the new technologies of our times, but also advances in linguistics and neighbouring fields, have led to considerable changes. Today, documentary linguistics is a new, 'hot' branch of linguistics where researchers with very different profiles and interests are working together: some travel to remote places of the world to collect data from lesser known and endangered languages, others think of new ways to process and store huge amounts of multimedia data, still others use language documentations for revitalizing endangered languages, for example, by preparing dictionaries and teaching materials.

### THE ADVENTURE OF DOCUMENTING ENDANGERED LANGUAGES

In 2008, two American linguists active in documenting endangered languages made a film where they show how interesting, sometimes how dramatic and sometimes how funny their work can be [2]. Find information about this film ("The Linguists") and additional materials on their website! [www.pbs.org/thelinguists](http://www.pbs.org/thelinguists)  
The trailer is also on YouTube [here](#)

### BOOK OF KNOWLEDGE

Chapters: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) **10**

[List of all languages](#) referred to in the Book of Knowledge and other sections of the website.

[Glossary](#)

**DOWNLOAD** and print out the Book of Knowledge.

### LET'S REVISE! – CHAPTER 10

Go to the [Let's Revise section](#) to see what you can learn from this chapter or test how much you have already learnt!

The documentation of endangered languages is an especially important and urgent task if we want to at least preserve some of the wealth that these languages possess and that otherwise will soon be gone forever. Having said that, documentary linguistics is not only concerned with endangered languages, and many issues, for example, issues concerning recording speech, transcribing spoken data, or building **corpora**, are the same for all languages. However, smaller, less studied and especially endangered languages also provide special challenges – first and foremost with respect to the amount of available data. If you want to document one of the larger languages such as English, Chinese, German, Hungarian, Dutch, Polish, etc., you can rely on already existing data and quite easily find samples of written and spoken language from which you could build up your documentation: books, newspapers and other written documents from the past and the present, many of these already digitalized, television and radio shows that can be recorded or simply downloaded from the Internet, language used in Internet forums and other social media, and many more. Because computers, the Internet and recording devices are widely available, the amount of such data and its accessibility is growing rapidly. For endangered languages, the situation is often completely different. Many of these languages do not have a written tradition and written data may be completely unavailable or sparse, the languages are not used in the media, or their speakers do not use the Internet (and if they do, they often use another language). In such cases, linguists must start from scratch and collect as much data as possible by recording speakers of a given language. Ideally, language documentation contains representative samples from different speakers – representing different age groups, different professions, of both sexes, and different origins –, but in the case of endangered languages this may not be possible, because the number of speakers is too small and/or there are only elder speakers.

An important issue apart from the number of speakers and amount of data concerns the communication between the linguists or other researchers who want to document a language, and the language community. In the case of endangered or minority languages, the documenters often are outsiders, not members of the community. They may not be fluent speakers of the language in question and can communicate with the speakers in a second or a third language. This often leads to an unnatural use of the language that is to be documented. The documenters may not know all the customs and the culturally and socially right ways to behave. Without wanting to, they may act in a way that is considered impolite or patronizing to the community. It is therefore always better if the documenting team includes members of the local community. This will not only improve the quality of the language documentation, but it is also a question of principle – after all, it is their language! For the researcher, a recording of an old man speaking about his childhood may be “data”, but for a member of the community – for example, for the granddaughter of an old man, this recording may be something very personal, like a treasured family remembrance.

Thus, language documentation fieldwork is often a lengthy process during which documenters need to travel, establish new contacts, integrate with the local community, become familiar with their customs, habits, and culture, before they can begin the actual work.

#### DOCUMENT A LANGUAGE OR DIALECT – AN EXAMPLE

Go to the section [What Can you do? – Document a language or dialect](#), and listen to Tymoteusz Król talking about his experiences with documenting Wilamowicean, one of the smallest minority languages spoken in Poland. Can you think of a small language or dialect in your region? Have you heard of any recordings, videos, TV programmes or books about or in that language?

#### ■ MORE THAN JUST WORDS AND SENTENCES

In the preliminary definition of language documentation given above in this chapter, we mentioned three elements of language documentation: collecting (recording, taking pictures, gathering written documents, etc.), processing (analysing, systematizing, transcribing, translating, etc.) and storing (archiving) data. These elements can be thought of as three successive steps. For example, first we record words, then we translate and analyse them, and the result, for example in form of a word list or a small dictionary, is stored in print or electronic form. However, the three steps are more intertwined. There may be overlapping – for example, transcribing (writing down) spoken data can be considered an instance both of collecting and of processing of data, and even as a kind of archiving. Furthermore, it is often necessary to “look ahead” and already think about possible ways of analysis and archiving from the very beginning, before starting any work. For example, when researchers are interested in the sound system of a language, they will probably first collect a small sample, then do some preliminary analysis in order to learn about the basic phonetic rules, and then collect more data more purposefully. Researchers interested in phonetics may also have different expectations concerning the kind of data collected and the way it should be archived than researchers with a primary interest in cultural traditions. Nevertheless, it is regarded as important to maintain the distinction between data collection and analysis. The same set of source (or “raw”) data, when properly constructed, can serve as a resource for various types of analyses conducted by researchers specializing in a wide range of fields such as: linguistics, cultural studies, sociology, psychology, history or geography. One of the researchers might look for typical linguistic features (e.g. syntactic structures or phonetic characteristics), another would be interested in social relationships reflected by the material (e.g. the functions and roles of speakers within a community). They will apply different methods, but use the same set of data. Language documentation should therefore be seen from an interdisciplinary perspective.

This perspective leads to a broader view of what is the object of language documentation. Rather than collecting words and sentences, linguists have to document linguistic practices and traditions that exist and can be observed within a community. These linguistic practices and traditions are manifested by [3] [4]:

- **linguistic behaviour:** everyday conversations, language use in social contacts between community members, linguistic customs and traditions (cf. the section “Language is doing and culture is a verb” in Chapter 6 on language and culture);

- **the speakers' knowledge about their language:** what speakers know and can explain about the rules and structures of their language (after Himmelmann 1998: 161-195), another thing of interest might be the speakers' ideologies – what they actually think of their language, are they convinced that it is worth preserving, and what they actually do in order to maintain it (e.g. whether grandparents use the language while speaking to their grandchildren?).

Based on these definitions, the aim of documentation is not just recording the sounds of language as such, but recording the sounds of language as **communicative events** [3] [5]. A communicative event includes more than speech, and to understand what is going on we also need information, for example, about gestures, face expressions, the broader situational context, the presence of third parties, or artefacts used at the time of the recording.

## ■ DATA AND METADATA

Briefly speaking, metadata can be defined as data describing other data or even simpler: data about data.

For example, the basic type of data for a phonetician will usually be acoustic data derived from a sound file together with its transcriptions while the accompanying metadata may include various types of information about the speakers (such as their sex, age, region and community of origin, health condition, social and family status), recording conditions (environment, background noises), technical properties (equipment, software, quality), authors(s), etc. However, the distinction between data and metadata is not always obvious because metadata can in fact become data and vice versa, depending on the aims of the study. In our example, the phonetician could treat the region of origin as data and not as metadata in case when he/she wanted to study regional variations of pronunciation. Furthermore, if the same corpus were to be analysed by a culture anthropologist, then the focus would likely shift to the description of family relationships and social information which would consequently be treated as data rather than metadata.

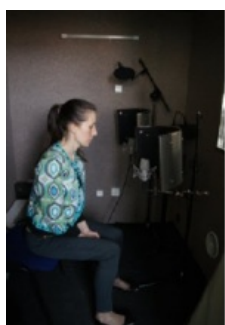
## ■ SOME METHODS OF LANGUAGE DOCUMENTATION

### Documenting communicative behaviour and language in use

Until about the middle of the 20th century, the idea most linguists had of language documentation was a dictionary, a grammar, and a collection of texts, preferably fairy tales or other narrative texts. Much documentation like this had been carried out already during the 18th and especially the 19th century, and the results are still useful for linguists today. As an illustration, you can **listen to a recent recording (2013)** [here](#) and look at **the transcript and translations** of a **fairy tale** first documented in **1895** by S. Ulanowska [6].

However, today we feel that something is missing in these older documentations, something that was difficult or impossible to document in an area where the only means of documenting was writing and drawing. As discussed above, language documentation assumes not only a description of particular elements of grammar and vocabulary, but its central issue is documenting the language in its natural environment, including the characteristics of the speakers, their mutual relationships and the situation in which they live.

When looking at the technical quality we must yet admit that the best recordings can be obtained in an anechoic chamber of a recording studio rather than in the language's natural environment. Noise can be controlled and minimized in the studio, and we may precisely adjust the types and positions of microphones or video cameras in advance, so that even the slightest details will be appropriately recorded. Such recordings are especially valuable for subtle analyses of the sound system of a language, and that is why anechoic chambers are usually situated in departments of phonetics at universities and research laboratories (see an example list of such laboratories and departments [here](#))



Studio recording environment: recordings in an anechoic chamber, the Laboratory team at work, equipment (the Laboratory of the Psycholinguistics Department, Adam Mickiewicz University in Poznań, photos: Agnieszka Czoska (left, middle), Maciej Karpiński (right)).

Language documentation thus faces a compromise between quality control and natural environment requirements. By definition, it is practically impossible to capture most real communicative events in an artificial surrounding of a studio. This may be particularly true for elderly speakers who sometimes are the only speakers left of a severely endangered language (cf. UNESCO's "Levels of endangerment" discussed in [Chapter 8](#) of the Book of Knowledge). On the other hand, recording speech outside the studio usually means difficulties in achieving good quality, even if we use excellent recording devices. When we sit in a room and chat, we usually don't pay attention to small background noises, but when we listen to a recording of that conversation we suddenly discover that there was a clock ticking or a fridge buzzing!

Listen to the following recording of utterances in the language [Teop](#): [click here](#)

What were the conditions under which this text was recorded? What kind of background noises can you hear?



Find [Teop](#) on the [Interactive Map](#) where you can listen and learn more about the language and the recordings you heard in the exercise above!

Speakers who are recorded (even in a very friendly environment) usually pay more attention to their way of speaking and consequently modify their linguistic behaviour: their speech can change in a quite unpredicted way. For example, it can become more/less formal, more/less polite or politically correct which can be manifested by changes in phonetic-acoustic parameters such tempo, intonation, intensity, timing patterns, pausing schemes, and others. This is related to the so called **Observer's Paradox**.

### OBSERVER'S PARADOX

The aim of linguistic research in the community must be to find out how people talk when they are not being systematically observed; yet we can only obtain this data by systematic observation. [7]

Many experiment set-ups have been designed with a view to obtain data on varying speaking styles and levels of spontaneity. In case of large and well-documented languages, a wide range of corpora have so far been collected using either the existing recordings or creating corpora from scratch. Existing recordings can include, for example, television shows or parliamentary speeches. However, as you can imagine, communication in front of TV cameras, lights, and microphones is quite specific and not always suitable for documentation or research needs.

A method half-way between collecting spontaneous speech and eliciting data after a fixed scheme is to set up a certain scenario for a dialogue between two speakers of the language to be documented. For example, one part of the Kiel Corpus of spoken language [8] was created by giving two speakers a different time table each and asking them – using a fake telephone line – to make an appointment that would fit both. The result was short spontaneous dialogues with a limited and largely predictable vocabulary.

An interesting example is also the JST/CREST database of spontaneous and expressive speech [9]. This database consists of several subsets one of which was collected by volunteering speakers wearing small portable recording devices and recording their own spoken communication during their daily activities (home, work, school) for a relatively long period of time (e.g. several months). The speakers could switch off the device at any moment as well as decide whether any part of the recorded material should be excluded from further analysis. The experimenters assumed that after some time speakers would become familiar with the recording equipment to the extent that they would stop paying attention to the fact of being recorded and thus their behaviour would become more (or even fully) spontaneous. However, a potential drawback in this case might be related to the fact that the recordings took place in varying locations characterized by unpredictable noise levels and thus cannot be fully controlled for quality. Another drawback concerns the metadata – it may be difficult to keep track of all the conditions of the recorded speech event (participants, context, etc.). And of course there are copyright issues, the question of personal data protection for all participants, not only the one who had agreed to participate in the experiment (see below the section on ethical and legal issues).

### STUDY QUESTION

Think of a recording scenario that would possibly enable producing good quality audio/video recordings of spoken communication of (a) children, and (b) elderly speakers, without losing too much of the spontaneity of speech.

Note: there might be many competing scenarios.

In the case of endangered or minority languages, the choice of materials is often very limited so almost any type of data might be a source of valuable information. It is worth remembering, however, that in order to make further analyses and descriptions easier we ought to work on the data thoughtfully and carefully rather than to “(mindlessly) collect heaps of data without any concern for analysis and structure”, as Nikolaus Himmelman put it [10]. In other words: if you have access to an endangered language (for example, a dialect your grandmother speaks), you may of course just make a quick recording with your smartphone and put it onto YouTube or other website (and if you surf the Internet you will see that many people do just that) – but this is not how professional language documentation is done.

### Documenting what speakers know

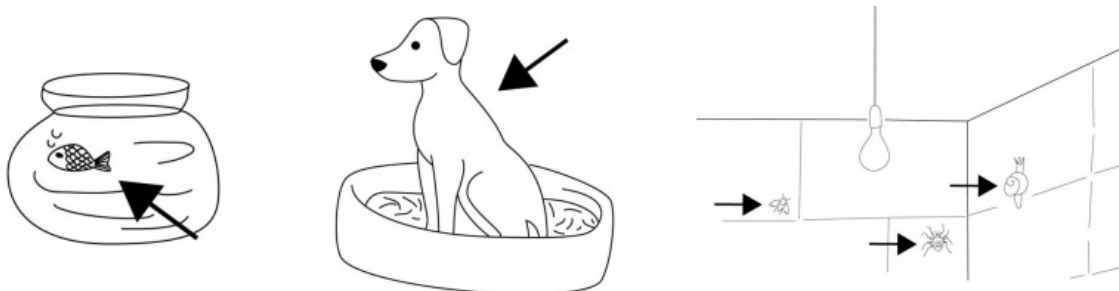




Documenting endangered languages, fieldwork conditions: *Yurakaré* (left, photo: Sonja Gipper & Consejo Educativo del Pueblo Yurakaré) and *Tahuatan* (right, photo: Gabriele Cablitz).

As important as the documentation of linguistic behaviour in a natural setting is, it is often not enough to get a comprehensive picture of the language. This is because speakers always know much more of and about their language than they show in their linguistic behaviour. In the course of natural conversations or interviews, the coverage of vocabulary items and structures depends on the topic of a particular conversation and usually reflects the items most frequently occurring in everyday use of the language. Less frequent words or structures may not come up at all even when many hours of spontaneous speech are recorded. If documentation is based solely on such material, the range of vocabulary and constructions is always a matter of chance. For example, if you happen to record a conversation about how many children there are in various families and how old they are, you may get many examples of numerals. In other recordings there may be no single numeral, but maybe many words for colours, and so on. Researchers interested in a special topic (numerals, colour terms, names of animals or plants, words for body parts, prepositions, etc.) will not wait until these items happen to come up in spontaneous discourse, but will use other methods of data collection, i.e. **elicitation** (elicitation is a term frequently used to designate the collecting of desired types of data directly from speakers, according to a previously designed scenario rather than using only what is available at the moment).

A common way to elicit vocabulary is to use a list of words that have to be translated into the language being documented. For documenting basic vocabulary, the **Swadesh list** is often used (see: Swadesh lists given in [Chapter 2](#)). Using translation as a method for elicitation however has many drawbacks. Not all words can be translated between two languages, and words that have no equivalent in the source language (for example, English) will not be discovered by this method. Therefore it is often better to use pictures, props or artefacts to elicit specific vocabulary items. Pictures and props are also very useful for eliciting grammatical structures, for example, to find out how spatial relations and motion are expressed (how you say things such as *The cat is on the map*, *The cat is climbing the tree*, *The apple fell from the tree*). There are already special sets of pictures and other stimuli available for several purposes. Story builder [11] is an example for such a set that is freely available on the Internet and can be used to elicit many different structures, especially constructions with verbs. Another example is a ready-to-use Field Manuals collection [12] where you can find pictures for eliciting vocabulary related to location of objects in space such as those shown in the picture below [13].



Mutual location of objects in space. Example TPRS (Topological Relations Picture Series, Bowerman et al., 1992, the complete source set of pictures at: [fieldmanuals.mpi.nl/volumes/1992/topological-relations-picture-series/](http://fieldmanuals.mpi.nl/volumes/1992/topological-relations-picture-series/))

When you download the pictures usually you also obtain suggested instructions for the recording scenario and terms of use (see for example materials for route description elicitation: [fieldmanuals.mpi.nl/volumes/1993/route-description-elicitation/](http://fieldmanuals.mpi.nl/volumes/1993/route-description-elicitation/) or a body colouring task: [fieldmanuals.mpi.nl/volumes/2003-1/body-colouring-task](http://fieldmanuals.mpi.nl/volumes/2003-1/body-colouring-task)). It should be taken into account, however, that not all pictures are universal and some of them cannot be useful because of cultural differences (e.g. dogs or representations of humans in Muslim cultures).

A method for eliciting coherent text after a given scheme is to show a mute film or, especially with children, a comic book or picture book without text and ask the speakers to retell the story in their own way. The most famous film in linguistics is **The Pear Story** [14] – this story has already been told in very many different languages, and the text thus collected can be compared and analysed for various aspects of grammar.

Linguists who are primarily interested in the sounds of language and not in collecting new words or constructions may also ask speakers to read aloud prepared and specially designed lists of words or short texts. An example for such a text is Aesop's fable: The North Wind and the Sun commonly used by phoneticians and phonologists to illustrate the sound of languages (cf. Handbook of the International Phonetic Association

[16]). (see [Chapter 2](#) and [Chapter 4](#)). Of course translating the text to some languages may be difficult because not all words or structures always have their direct equivalents in the language. Another problem might be the simple lack of a (commonly known) written form of the language making it impossible to design a reading task.

## LISTENING

Listen to the recordings of *The North Wind and the Sun* Aesop's fable in three languages. Pay attention to the technical quality:

- **Polish** – [click here to listen](#) (speaker: Ewa Sobczak, a professional actress [15])
- **Latgalian** – [click here to listen](#) (speaker: Evita Kozule, a student of a linguistic faculty, recorded by K. Klessa & N. Nau)
- **Halcnovian** – [click here to listen](#) (speakers: Fryderk Hanusz, Józef Jancza, source: [6b])

**Note:** **Halcnovian** is a critically endangered language: according to recent records it had only 8 speakers in 2013. Furthermore, it does not have a written tradition, so in this case the text of the recording was not read in Halcnovian, but translated from a model written in Polish.



Find **Latgalian** on the [Interactive Map](#)!

If one wants to document what speakers know about their language, it is of course also possible to ask them directly (though this can never be the only or the main method of gathering data). For this purpose it is useful to collect the terminology that speakers of a certain culture use to speak about language – do they use equivalents to English words such as word, sentence, syllable, past tense, and so on, or do they have other categories? This brings in the possibility of **speaking about a language in this particular language** instead of using a third language for descriptions.

## ■ ELEMENTS OF DOCUMENTARY PRACTICE



Analog 12-inch record (photo: Maciej Karpiński)

As defined above, language documentation comprises the activities of collection, processing and archiving of linguistic data. When we consider that language documenters often need to travel a lot in order to collect their data and then they have to safely store, process, and share that data, we can easily understand a strong link between language documentation and technology. Some years ago, when a standard recording device was quite a weighty and sizeable machine, the task of a documenter was much more challenging than today, when you can carry a high quality audio or video recorder in your pocket.

Thanks to technological developments the work on data has gained an unprecedented efficiency and speed. For example, it is now possible to search a piece of information through millions of vocabulary items over a time shorter than a few seconds or to store high-quality videos or sounds on a one-centimetre portable device (while a similar amount of data would once have needed a few rooms in a building of dozens or even hundreds of square metres of capacity). Moreover, people can make their data available to a wide audience practically at any moment. The Internet is abundant in various types of information.

## LEARN MORE

If you want to learn more about the history of speech recording, reproduction and storage, see the [Appendix 1](#) to this chapter.

However, together with this great potential, new challenges and questions emerge. For example, it can become more problematic than before to organize or even search through data, to avoid chaos and control data access by various users and as well as to account for all legal and ethical issues (see below).

## ■ PREPARATORY STEPS

The preparatory steps preceding the actual data collection may include contacting speakers, getting familiar with the so far available materials, planning recording scenarios, testing software and equipment, choosing file formats or defining file naming conventions. When you already contacted the speakers of the language to be documented, it is a good practice (and often a duty (read also about some [legal and ethical problems](#)) to ask for their formal **agreement** to the recordings, and to take care of their positive attitude and willingness to participate. Endangered languages are often spoken by small communities, where the family and other social relationships may play even bigger role than in larger and thus much more diversified language communities. The attitude of the local community can be really crucial in such case, and their rules and internal laws have to be taken into account. – Is it all right for a stranger to walk around, take pictures of houses and sacral places and record people's speech? – It is always better to ask first before doing such things. Some communities explicitly state that permission has to be granted before pictures are taken (see for example [\[17\]](#)).

### STUDY QUESTION

What could be the reasons if a speech community does not want external researchers to make video films of speakers and their homes? What could be done to resolve a conflict of interests between the researchers and the community?

Another useful step to be done in advance is to decide about ways to organize your data, e.g. think about where to store the data, how to make backup copies and how to name your files and folders. Making such plans is already a preparation to data analysis because afterwards it will help you to classify and describe your data.

### HOW TO ORGANIZE MY FILES / FOLDERS?

When you are about to work with a number of files, one of the crucial steps is to decide about file naming conventions, preferably before starting data collection. Think of the elements that should be included in the names of files or folders. Maybe some of the following:

- date of creation (although usually the date is encoded in the file header, it might be convenient to have it also in the file name for your convenience)
- speaker's ID
- type of data (speaking styles, registers, environment...)
- other?

What should be the order of the elements? Remember that you will probably wish to sort your files by names. You can use abbreviations or codes in the file names to make the names shorter. Otherwise, you may wish to name your files only with unique ID numbers and include information about the contents in additional information files.

- In case if you want to deposit your data to an existing repository such as DOBES [\[18\]](#), it would be best to first visit their website, check about recommended conventions and use them from the beginning of your work, e.g. [\[19\]](#)
- To explore some of the other existing file formats and standards, see e.g. [\[20\]](#).

In the case where you do not produce new data but rather deal with already existing archives (for example digitize historical data or transcribe old recordings), an important step will be to preserve the original naming conventions and other information related to the source materials. Keeping this information can be very useful in case you or someone else would like to go back to the very first version of the data.

## ■ RECORDING EQUIPMENT AND SESSION SET-UP

Before choosing from many available types of audio recorders, photo and video cameras, recorders or microphones, you should consider their parameters and prices as related to your specific needs. The influencing factors can be the type of recording scenario, recording modalities (audio/video), the location for the **recording session** ("session" is a term usually used to name all the recorded events) as well as the characteristics of speakers such as age, sex, social status, etc. A crucial technical question is whether the equipment will be used on a stationary basis or rather for fieldwork, requiring travels. In the latter case, the critical parameters are its size and weight but also the power supply solutions, the availability and type of batteries, charging options, the shake resistance etc. Microphones can be embedded in various types of devices (such as portable recorders) or used externally, and connected with cables.

## STUDY QUESTION

Think of a list of factors which could influence the choice of your fieldwork equipment, considering that:

- you need to travel by plane to the destination region
- your task is to document a language spoken in a couple of villages, not very distant from one another, and you will use a bicycle to travel between them (carrying your equipment)
- you will have a limited access to the Internet and will need to 1) backup your data in the meantime, 2) occasionally send samples of your data via a slow Internet connection



Dynamic (left), condenser (middle), condenser head-mounted (right) microphones (photo: Maciej Karpiński).

Two main types of microphones are distinguished according to their construction: dynamic and condenser microphones. Condenser microphones are often used in radio and TV studios, they are characterized by higher sensitivity and can capture a variety of complex sounds, including subtle background noises. A variety of condenser microphones is also used in cellular phones and the cheapest recorders but these are built with a different technology and their quality is incomparably worse than that in the more expensive models. All condenser microphones require an additional power source which may be a disadvantage in the case of fieldwork. Dynamic microphones can be used without any additional source of power which makes their usage simpler. They are characterized by a lower sensitivity as compared to the condenser microphones (which can actually be an advantage when the recording session takes place in a noisy environment) and are very useful for speech recordings, especially when the speaker speaks directly to the microphone, from a close distance. Dynamic microphones are also commonly used by singers during live concerts, while condenser microphones are usually used for recording vocals in an anechoic chamber.

## STUDY QUESTION

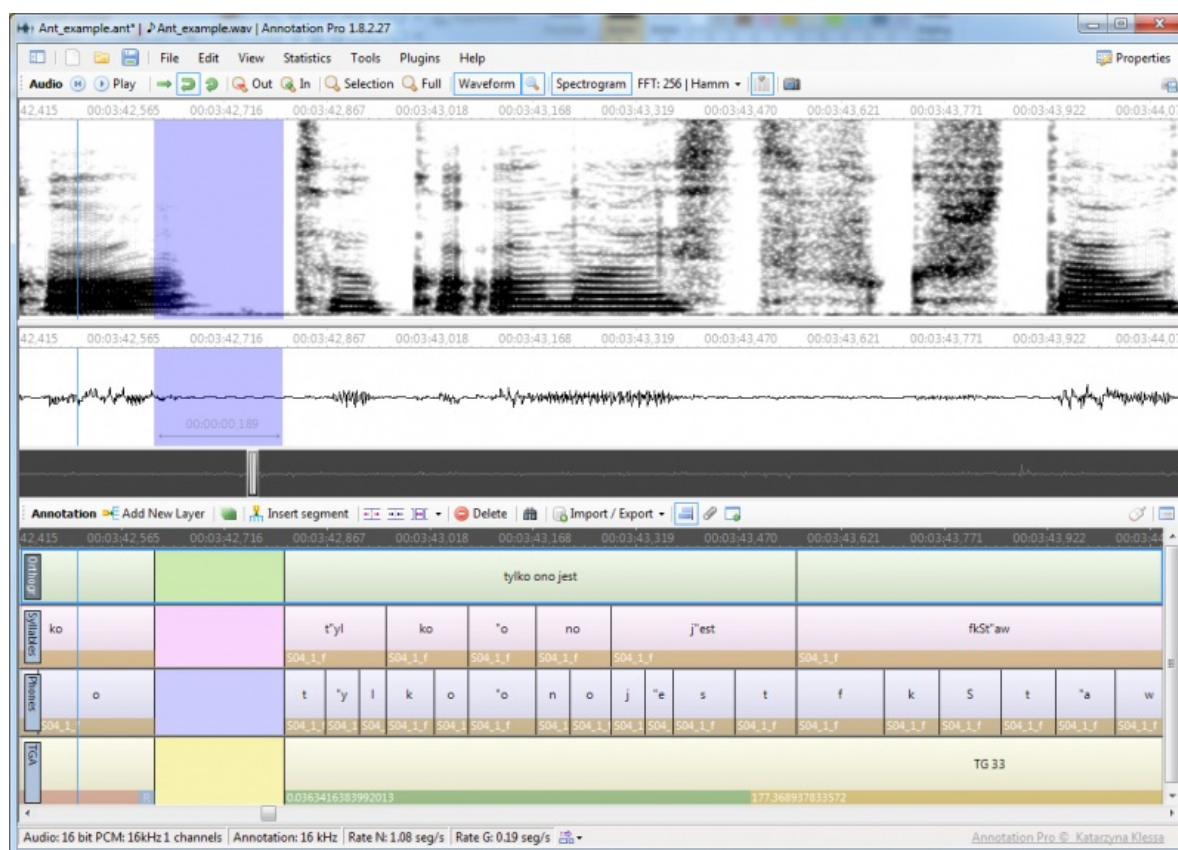
Think of 2 or 3 recording locations and scenarios for a recording session. Consider the environment and the number of speakers. Which type of microphones would be better for these sessions?

## ■ PROCESSING AND ANALYSING DATA

A step usually following data collection is creating a backup copy of the data in the original form, without any modifications. Backup copies can be stored on CD, DVD, blu ray discs, local or remote hard disks or small-sized portable storage media such as pendrives or memory cards. After ensuring that the backup copies are safely stored, the data can be analysed and/or further processed.

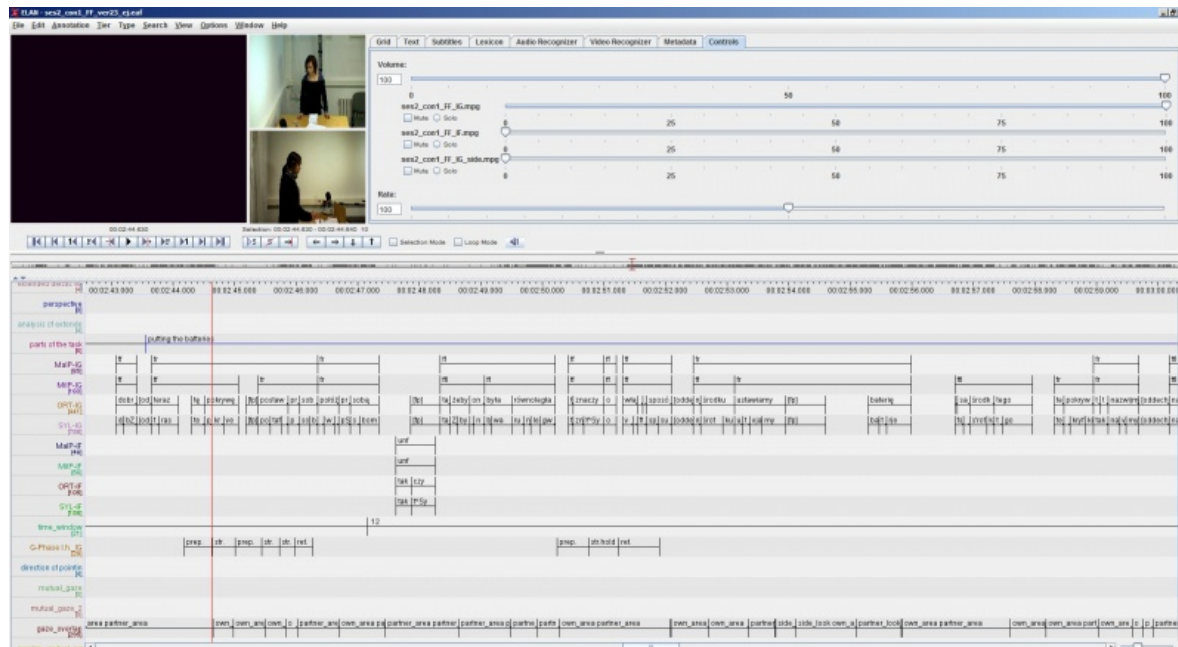
Data description usually is a multi-task process. In case of description of audio data, the tasks include annotation of the recordings, i.e. time-aligned transcription of the utterances so that one can afterwards follow the transcribed text and listen to the recording just like in case of film subtitles. Thanks to this, it is possible to analyse particular speech sounds, syllables, words or any other fragments of the signal (see the animation below). Depending on the needs, annotation can also include other information such as description of prosody, dialogue acts, pausing schemes, hesitation markers, pronunciation errors, individual features of speech or speakers.





An example multilayer annotation of an audio file (Annotation Pro).

A number of computer programs for annotation are currently available and many of them are free of charge for research and education purposes. Some of them enable only annotation of audio files while with others it is also possible to annotate video files (see some of the tools in the list [here](#)).



An example multilayer annotation of a video file (Elan).

In case when you wish to perform some more detailed phonetic analyses it is useful to choose a tool including a spectrogram to display your sound files (see [Chapter 4](#), especially the section on visible speech). Usually, recordings are first transcribed orthographically (using the official alphabet of the language if it exists) then, phonetically. The International Phonetic Alphabet (IPA) [21] enables a detailed phonetic transcription of speech. For the needs of computer processing, the Computer Readable Phonetic Alphabet (SAMPA) [22] is also used. Among others, SAMPA owes its popularity to the fact that it does not use any special fonts apart from those available in a standard Latin computer keyboard. For large languages you can find tools to automatize the annotation and transcription work (e.g. GTP – grapheme-to-phoneme converters automatically transforming orthographic texts to phonetic transcriptions, ASR – automatic speech recognition tools). Finding similar tools for endangered languages is more difficult although not completely impossible (see [Chapter 4](#), the section on: Less widely used languages and technology).



Although phonetic alphabets are most suitable for transcribing speech, it is worth noting that in certain cases it might be preferable to use transliteration (writing the text in one alphabet with the use of another alphabet, e.g. Cyrillic script with Latin letters) or a kind of quasi-phonetic transcription which might lack certain details but on the other hand, is easier for non-specialists. Such solutions are often chosen for corpora or dictionaries dedicated to the use of both scientists and for local communities.

Transliteration	Orthographic	Translation EN	Phonetic
ix bin avek fun varše in jor 1939, nox dem vi der dajč iz šojn dort geven cvej voxn.	איך בין אַװעק פֿון װאַרשע אין יאָר 1939 , נאָך דעם װי דער דײַטש אין שװין דאַרט געװען צװײ װאָכן	I left Warsaw in 1939. The Germans had already been in the city for two weeks.	ix bin avek fun varše in jor 1939   nox dem vi der dajč iz šojn dort geven tsvej voxn
Varše der iker dos jidiše varše – hot šojn demolt ojsgezen vi a košmar: fil hejzer farbrent, fil češotn.	װאַרשע - דער עיקר דאָס ייִדישע װאַרשע - װאָבן האָט שװין דעמאָלט אויסגעזען װי אַ קאָשמאַר: פֿיל הײַזער פֿאַרברענט, צעשטאָנען פֿיל	Warsaw - and more precisely, Jewish Warsaw - by then already looked like a nightmare - so many burned out houses, so many destroyed.	varše der iker dos jidiše varše hot šojn demolt ojsgezen vi a kajmar fil hejzer farbrent   fil tseshotn

An example record from the Polish Heritage Database: transliteration, orthographic script, English translation, and phonetic transcription for a text in Polish Yiddish (find more at: [inne-jezyki.amu.edu.pl/](http://inne-jezyki.amu.edu.pl/))

## ■ EXAMPLE ON-LINE ARCHIVES FOR ENDANGERED LANGUAGES

One of the most important archives for endangered languages is the DOBES (Dokumentation Bedrohter Sprachen) archive [dobes.mpi.nl/](http://dobes.mpi.nl/) – an Internet database of complex documentation for many endangered languages. DOBES is maintained within the Language Archive (TLA) located at the Max Planck Institute for Psycholinguistics in Nijmegen. This initiative involves not only archiving data and metadata related to endangered languages but also development of linguistic archiving and tools, as well as methods of documentation. Another example of a noteworthy archive for endangered languages is the ELAR [23] at SOAS (School of Oriental and African Studies, London), also specialising in preserving and publishing endangered language documentation. Apart from providing information and data, both of these archives offer the possibility of depositing and storing your own data on their servers.

The rules concerning the access to data in language repositories are often defined individually for each resource. Some data is publicly available for anyone while for others various limitations may apply. For instance, with some data, you will be asked to contact its authors or contributors in order to get permission, and in other cases you may have to explain the purpose for which you want to use the data before you get permission to download. This might seem complicated but it becomes understandable when we treat the data more as pieces of someone's real life or heritage than only as "words or sentences" needed for our studies, as was discussed above.

### TASK

DOBES is a program that has financed many language documentation projects since 2000. Find a list of these projects here: [dobes.mpi.nl/projects](http://dobes.mpi.nl/projects). Choose three of the projects and try to answer the following questions:

- What were the main aims of the project? Which researchers besides linguists were involved in the documentation or could be interested in the data?
- How were members of the community involved in the project?
- What kind of data was collected? How was it collected, processed and stored?

An interesting example of a website dedicated to language diversity and endangered languages, including support for indigenous people, where you can also learn about collecting, digitizing and describing data is the SOROSORO program's website [24]. The goals of informing and sharing knowledge about endangered languages around the world are also pursued by the Endangered Languages project [25]. An example localised for the region of Poland and the neighbouring countries is the Linguistic Heritage website [6] developed for endangered languages spoken in the territory of central-eastern Europe, once belonging to the so called Polish-Lithuanian Commonwealth (Rzeczpospolita), currently being the areas of several countries (Poland, Lithuania, Latvia, Belarus, and Ukraine). The website was created for the use of both researchers and native speakers of the languages or any other interested persons.

### A DATABASE SEARCH TASK

Visit a language database site on the Internet and search it for information about an endangered language(s) spoken presently or in the past in your region of the world.

Pay attention to the types of information provided (descriptions of speakers, culture, geography, sound or text resources in the language).

- You can try out one of the websites: [dobes.mpi.nl/projects](http://dobes.mpi.nl/projects), [www.endangeredlanguages.com](http://www.endangeredlanguages.com), [www.sorosoro.org/](http://www.sorosoro.org/), [inne-jezyki.amu.edu.pl](http://inne-jezyki.amu.edu.pl)
- Was the search task easy? What kind of problems (if any) can you see?
- Maybe you can find other similar websites?

## ■ LEGAL AND ETHICAL PROBLEMS

One of the main restrictions of using and sharing linguistic resources may be related to the protection of private data. Other restrictions might be a result of cultural, ethical, social or religious issues specific to the linguistic community. Another consideration is that in order to use the data and especially to publish it, it is usually necessary to obtain formal consent from the speakers recorded.

From the legal point of view, the written consent of each participant of a conversation is usually sufficient for recordings. However, legal solutions may vary between countries. For example, recording telephone conversations (even one's own) without explicit agreement of all participants is illegal in many countries (e.g. Poland, Germany), while some countries allow recording with the consent of only one party (selected states in the USA) or even do not provide any regulations in this respect and thus any recordings of this type are allowed (Latvia).

### EXERCISE

Search the Internet and try to find answers that are true for your country:

- Is it legal to record your own telephone conversation with another person?
- Is it legal to use (e.g. as a proof in court) a recording of your own telephone conversation with another person?
- What are the restrictions (if any)?

In practice, speakers are asked to give separate consent to the participation in the recordings, the use of the recording for particular purposes, and last but not least – the publication of the recordings. Most frequently, the consent is prepared in a written form (in case of audio recordings it may also be expressed verbally and recorded together with the remaining data).

### TIP

Audio/video recording consent

The text of the consent should be clear, free from specialized terminology. Write the text in a way that by signing it the participant:

- confirms that s/he voluntarily agrees to participate in the recording session

can give separate consent for:

- all modalities of the recording (audio / video)
- the usage of the data in research studies
- publication
- archiving of the data

Consider creating 2 copies of the consent form for yourself and for the participant.

Remember the signatures!

In the cases of certain local communities, the audio/video recording consent might be not only an individual decision but more a question of the general 'policy', customs or attitudes in the community. The documenters need to carefully consider maintaining good relations with the members of local communities, both during the design stage, the proper recordings, and afterwards, when the data are analysed, systematized and processed. The same affects the choice of archiving methods and the ways of sharing the data.

### FOOD FOR THOUGHT

Imagine that you have learned about an endangered local dialect and you would like to become involved in documenting the dialect and/or its revitalisation. What would be your first steps?

First think of your answers and then go to section [What can you do – Become a linguist](#) and see working examples of young people doing documentary work.

The interested reader can read more about data, corpora and databases in [Appendix 2](#) to this chapter. You will find there some details about data formats and structures, sharing and exchanging information, plus some more examples concerning the design and development of language resources.

## ■ APPENDICES: MORE ABOUT THE HISTORY OF SOUND RECORDING, DATA FORMATS AND STRUCTURES

To find out more about issues related to language documentation see two appendices to this chapter:

- **Appendix 1:** History of speech recording, reproduction and storage: selected facts. Read more [HERE](#)
- **Appendix 2:** Data formats and structures. Read more [HERE](#)

## LET'S REVISE! – CHAPTER 10

Go to the [Let's Revise section](#) to see what you can learn from this chapter or test how much you have already learnt!

### References

- [1] Seifart, F. (2011). Competing motivations for documenting endangered languages. In Haig, G.L.J., Nau, N., Schnell, S., Wegener, C. (Eds.) Documenting endangered languages. Trends in Linguistics, De Gruyter Mouton.
- [2] The Linguists: [www.pbs.org/thelinguists](http://www.pbs.org/thelinguists)
- [3] Himmelmann, N. P. (1998). Documentary and descriptive linguistics. Linguistics 36:161-195.  
(on-line e.g.: <http://fil.phil-fak.uni-koeln.de/fileadmin/linguistik/asw/pdf/Publis/1998a.pdf>)
- [4] Himmelmann, N. P. (2006). Language documentation: What is it and what is it good for? In Essentials of Language Documentation, Gippert, J., Himmelmann, N. P., Mosel, U. (Eds.), Trends in Linguistics, Studies and Monographs 178:1-30. Mouton de Gruyter, Berlin – New York.
- [5] Lüpke, F. (2010). Research methods in language documentation. Language Documentation and Description, 7, 55-104.
- [6] Poland's Linguistic Heritage website: [inne-jezyki.amu.edu.pl](http://inne-jezyki.amu.edu.pl)
- [6b] Poland's Linguistic Heritage website (Halcnovian recording): [inne-jezyki.amu.edu.pl/Frontend/TextSource/Details/40](http://inne-jezyki.amu.edu.pl/Frontend/TextSource/Details/40)
- [7] Labov, W. (1972). Sociolinguistic Patterns. Philadelphia: University of Pennsylvania Press, p. 209.
- [8] Kiel Corpus of spoken language [http://www.isfas.uni-kiel.de/de/linguistik/forschung/das\\_kiel\\_korpus](http://www.isfas.uni-kiel.de/de/linguistik/forschung/das_kiel_korpus)
- [9] Campbell, N. (2002). The recording of emotional speech: JST/CREST database research. Proceedings of Language Resources and Evaluation Conference (LREC), Las Palmas, Spain.
- [10] Himmelmann, N. P. (2012). Linguistic Data Types and the Interface between Language Documentation and Description. In Language Documentation & Conservation Vol. 6 (2012), pp. 187-207.
- [11] Story Builder: <http://www.story-builder.ca/>
- [12] <http://fieldmanuals.mpi.nl/>
- [13] Bowerman, M., Pederson, E. (1992). Topological relations picture series. In Stephen C. Levinson (ed.), Space stimuli kit 1.2: November 1992, 51. Nijmegen: Max Planck Institute for Psycholinguistics.
- [14] The Pear Story: <http://www.pearstories.org/docu/ThePearStories.htm>
- [15] Klessa, K., Wagner, A., Oleśkiewicz-Popiel, M., Karpinski, M. (2013). "Paralingua – a new speech corpus for the studies of paralinguistic features", in Vargas-Sierra, Ch. (Ed), Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions, Procedia – Social and Behavioral Science 95. (48-58), 2013.
- [16] IPA Handbook: <https://www.langsci.ucl.ac.uk/ipa/handbook.html>
- [17] Russian Old Believers: <http://www.alaska.org/detail/russian-old-believer-communities>
- [18] DOBES Project: <http://dobes.mpi.nl/>
- [19] DOBES – Deposit your data section: <http://dobes.mpi.nl/deposit-your-data/>
- [20] Gibbon, D., Moore, R., & Winski, R. (Eds.). (1997). Handbook of standards and resources for spoken language systems. Walter de Gruyter. Available on-line at: [http://sldr.org/SLDR\\_data/Disk0/preview/000836/?lang=en](http://sldr.org/SLDR_data/Disk0/preview/000836/?lang=en)
- [21] IPA Chart: <http://www.langsci.ucl.ac.uk/ipa/ipachart.html>
- [22] SAMPA Alphabet: <http://www.phon.ucl.ac.uk/home/sampa/>
- [23] The Endangered Languages Archive at SOAS, London <http://elar.soas.ac.uk/>
- [24] The SOROSORO program's website <http://www.sorosoro.org/>
- [25] Endangered Languages website: <http://www.endangeredlanguages.com/>

### Useful links

About language documentation:

- DOBES (Dokumentation Bedrohter Sprachen) – <http://dobes.mpi.nl/>
- L&C Field Manuals and Stimulus Materials – <http://fieldmanuals.mpi.nl/>
- SOAS (School of Oriental and African Studies) – <http://www.soas.ac.uk/>
- Endangered Languages – <http://www.endangeredlanguages.com/>
- Endangered Languages Documentation Programme (ELDP) – <http://www.hrelp.org/>
- SOROSORO – <http://www.sorosoro.org/>
- Poland's Linguistic Heritage – <http://inne-jezyki.amu.edu.pl>

Speech annotation tools:

– for annotation of audio and video files:

- Elan – <http://tla.mpi.nl/tools/tla-tools/elan/>
- Anvil – <http://www.anvil-software.org/>

– for annotation and phonetic analysis of audio files (include spectrograms):

- Praat – <http://www.praat.org/>
- Wavesurfer – <http://sourceforge.net/projects/wavesurfer/>
- Annotation Pro – <http://annotationpro.org/>

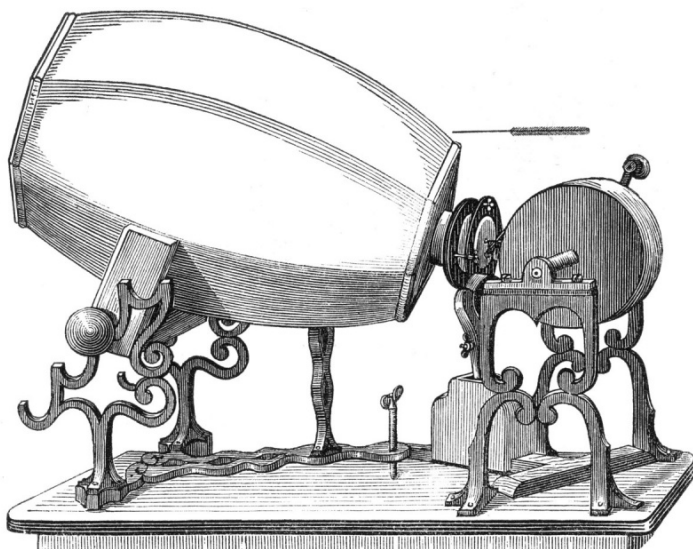
# Language documentation. Appendix 1

Home > Book of Knowledge > Language documentation > Language documentation. Appendix 1

■ CHAPTER AUTHOR: KATARZYNA KLESSA

## History of speech recording, reproduction and storage: selected facts

While the most ancient written archives reach many centuries of age, the first successful attempts of sound recordings date in the second half of 19th century. In 1857 Édouard-Léon Scott de Martinville invented the first sound recording device named a **phonoautograph**. The idea of the device was inspired by the structure of the human eardrum (also called the tympanic membrane) and the auditory ossicles (three small bones inside the middle ear). Phonoautograph recorded speech sound by writing the trace of sound vibrations to a rotating recording surface (the barrel in the picture of phonoautograph) in a form of a line. The duration of the recording depended on the speed of rotation (slower rotation allowed for longer recordings, however, only thanks to rapid rotation was it possible to capture the higher frequencies of the speech sounds). Phonoatographs enabled recording sounds but at the time of their development there was no possibility of playing the sounds back. The recordings had to wait almost 150 years to become available for listening. In 2008, sound historians used computer techniques (e.g. optical scanning) in order to re-play the recordings made with the first phonoautograph ([firstsounds.org](http://firstsounds.org)). These remain to be the first recordings of a human voice ever.



Phonoautograph by Édouard-Léon Scott de Martinville (1817–1879) (source: <http://www.firstsounds.org/press/032708/images/print/pisko-p73-phonautograph.jpg>)

## BOOK OF KNOWLEDGE

Rozdziały: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#)

[List of Languages](#) referred to in the Book of Knowledge and other sections of the website.

[Glossary](#)

[Download](#) and print out the Book of Knowledge.

## LET'S REVISE! – CHAPTER 10

Go to the [Let's Revise section](#) to see what you can learn from this chapter or test how much you have already learnt!

## PHONETIC EXERCISES

Do you wish to learn more about speech sound analysis and phonetic practice? Take a look at the exercises in the [Phonetic Exercises section](#).

## LISTENING TO THE FIRST RECORDINGS OF HUMAN VOICES

One of the first speech sounds ever recorded from **phonoautograph**: (recording of Au Clair de la Lune – By the Light of the Moon, recorded on 9 April 1860 by Léon Scott de Martinville, source: [www.firstsounds.org/sounds/scott.php](http://www.firstsounds.org/sounds/scott.php)).

Find more pioneer sound recordings at [firstsounds.org](http://firstsounds.org).

Recordings from phonograph, constructed in 1877 by Thomas Alva Edison, an example from The America's Library Web site: [www.americaslibrary.gov/assets/jb/recon/jb\\_recon\\_phongrph\\_1.wav](http://www.americaslibrary.gov/assets/jb/recon/jb_recon_phongrph_1.wav)

In 1877 Thomas Alva Edison constructed a phonograph which supported both recording and playback of sounds.

The material used for recordings in the first implementation of the device was tinfoil wrapped around a cylinder (see also the materials of a preservation and digitalization of cylinders project here: [cylinders.library.ucsb.edu/history-tinfoil.php](http://cylinders.library.ucsb.edu/history-tinfoil.php)).

Below you can watch a short video showing the optical imaging studies of Edison's tinfoil recordings: [youtube.com/watch?feature=player\\_embedded&v=uiNTFrMtCXs](https://youtube.com/watch?feature=player_embedded&v=uiNTFrMtCXs)



You can find more information at: [www.scpr.org/blogs/news/2012/10/25/10712/hear-thomas-edison-sing-rare-1878-audio-restored-f](http://www.scpr.org/blogs/news/2012/10/25/10712/hear-thomas-edison-sing-rare-1878-audio-restored-f).

In 1888 tinfoil was replaced in mass production with wax coating (patented in 1886 by Chichester Bell and Charles Sumner Tainter), which additionally enabled the creation of a limited number of copies of the recordings. Wax cylinders could include only about 2 minutes of recordings. Another progress came in 1908 together with the cylinders made of celluloid which could include 4 minutes of recordings and it was possible to create even a few thousand copies from one stencil. The sound quality of the earliest recordings was not perfect, and many noises were recorded along with the target recording. Moreover, with each playback the cylinder was worn down and the quality deteriorated.

#### ■ VIDEO

Watch an interesting video explaining and demonstrating sound recording using wax cylinders below and find more information at <http://www.phonographcylinders.com/about-us.php>:



Edison's phonograph. Photo: Leonardo Novaes  
(<http://www.freeimages.com/photo/680215>)

The last phonograph was produced in 1929 when the cylinders were replaced by a new invention – the gramophone disk-shaped record patented in 1887 by Emil Berliner, together with a device for recording and playback of sounds: the **gramophone**.



Gramophones and an analogue record (images: left – a picture first published in Spiegel May Stern Co. Universal Home Furnishers (1908), source: <http://olddesignshop.com/>, middle – a photo by Erik Araujo, [www.freeimages.com/photo/925626](http://www.freeimages.com/photo/925626). A 12-inch record, photo: Maciej Karpiński.

The earliest records used as storage devices before the vinyl record were made of (mixtures of) various materials such as wax, brass, copper or binder. Another type of sound data storage devices used and developed for many years were tapes of various types. The beginning of their popularity was related to the invention made by Valdemar Poulsen: the telegraphone (see more e.g. here: <http://cs-exhibitions.uni-klu.ac.at/index.php?id=220>). The initial application of the device was the first answering machine (listen to the speech of Emperor Franz Joseph (the oldest magnetic recording made with Poulsen telegraphone), 1900). After some time, metal tapes were replaced with less weighty and more stable tapes made of paper and plastic materials. Tapes were important not only for sound recordings but also for recording of motion pictures. In 1892, Thomas Alva Edison invented the kinetoscope, a machine thanks to which it was possible to watch motion pictures. Not long afterwards, in 1895, brothers Auguste and Louis Lumiere built the cinematograph which displayed silent motion pictures on a screen.

#### ■ VIDEO

The Jazz Singer (1927) the first motion picture with synchronized dialogues:



The Jazz Singer movie poster  
(a public domain image,  
source:  
[http://en.wikipedia.org/wiki/File:  
The\\_Jazz\\_Singer\\_1927\\_Poster.jpg](http://en.wikipedia.org/wiki/File:The_Jazz_Singer_1927_Poster.jpg))

The films were recorded on tapes, the earliest ones being very short (a dozen-or-so metres of tape which gave only several dozen seconds of recordings) and included no synchronized sound until the twenties of 20th century when a methodology providing sound together with the pictures was devised. The first film produced with synchronized dialogue sequences was *The Jazz Singer* publicly presented in 1927. Language documenters have always been interested in technological inventions potentially useful for the purposes of language documentation and archiving. Just one of many examples might be the work of the Polish documenter, ethnographer, and cultural anthropologist Bronisław Piłsudski (1866-1918) who thoroughly investigated and documented the culture, traditions and language of the Ainu people, the inhabitants of the Sakhalin Island at the time of Piłsudski but currently living mostly on the Hokkaido Island in Japan.

## MORE

A website devoted to Bronisław Piłsudski – [icrap.org](http://icrap.org)

Listen to one of Bronisław Piłsudski's early recordings of Ainu songs (over 100 years old):  
[panda.bg.univ.gda.pl/ICRAP/ainu.mp3](http://panda.bg.univ.gda.pl/ICRAP/ainu.mp3)

Ainu people – [www.ainu-museum.or.jp/en/study/eng01.html](http://www.ainu-museum.or.jp/en/study/eng01.html)



Ainu music performance in 2013, Hokkaido. Photo by courtesy of Ewcyna, see more at:  
<http://www.ewcyna.com/ajnowskie-klimaty/>



The recording equipment used for recording dialects spoken in Poland in 1948 and following years. Photo by courtesy of Jerzy Sierociuk.

An excellent example of an archive including a huge number of historical recordings of dialectal speech stored on several types of records and tapes since the 40ties of 20th century can be found at the [Laboratory of Dialectal Studies](#), the Institute of Polish Philology of Adam Mickiewicz University in Poznań (see one of the oldest recorders in the picture above). The resources are currently being digitized and copied to new storage carriers (hard or flash disks and CDs; cf. also Sierociuk (2014)). The photographs below show the team of the Laboratory as well as their vans used for transporting fieldwork equipment and staff. Considering the sizes and weight of recording devices available in the past, the logistics of fieldwork obviously must have been more difficult than it is today. But it should be noted that a useful and inexpensive means of transportation for a documenter, especially for shorter distances on sandy country roads between small villages mentioned already in the book by Prof. Zenon Sobierajski (1952) was simply a bicycle (considering the contemporary miniaturisation of recording devices, today a bike can become even more handy as a means of transportation).



The documentary field-working team of the Laboratory of Dialectal Studies, Adam Mickiewicz University in Poznań (left), and their

equipment vans, years 1951-1957. Photo by courtesy of Jerzy Sierociuk.

The mass production of reel-to-reel and tape recorders in the second half of the 20th century was an important step in the process of miniaturisation and accessibility of recording equipment. One tape could hold from a dozen or so up to even a few hundred minutes of recordings. However, a really great step forward as regards the quantity of stored data was the result of the “digital revolution” in the end of 20th century which led to the conversion of analogue audio archives to digital ones (e.g. wave or mp3). A similar transfer of technology occurred in photography, video, text messages (the use of e-mails and Internet communicators) thus enabling processing and storage of almost unlimited data quantities.



Audio (top left) and video (top right) tapes, a 1100 mm tape (bottom). Photo: K. Klessa, M. Karpiński.

The quality of analogue recordings is quite “conditions-sensitive” as it depends on various factors such as the cleanness of the recording head, its exact position in relation to the tape (or record), etc., while with digital formats, data recording can be practically lossless (when you choose high quality parameters of your recorder), error-resistant and the quality does not deteriorate during data transfer. Moreover, digital data can be stored and accessed much faster than those in analogue formats, and they can be compressed so that much less storage capacity is needed to hold them. Nowadays, you can easily buy small-sized storage devices (hard disk drives, CD or DVD records replaced more and more often by portable microdrives or pendrives) characterized by a high capacity enabling storage of substantial amounts of data, including large audio or video files. Another important feature of these devices is that they do not have any moving, rotating parts which makes them even more reliable and safer.

Although digital media and file formats prevail in mass production for practically all applications, the analogue recording format used in earlier devices has not been replaced completely by the digital successor. For example, in some recording studios it is still preferable to use analogue equipment, and the gramophone records are still favoured by many music fans and connoisseurs.

## LET'S REVISE! – CHAPTER 10

Go to the [Let's Revise section](#) to see what you can learn from this chapter or test how much you have already learnt!

### References

- Majewicz, F. M. (2010). Bronisław Piłsudski – globalny hit polskiej orientalistyki. In Majda, T. (Ed.) *Szkice z dziejów polskiej orientalistyki*. Tom V. Wydawnictwa Uniwersytetu Warszawskiego, Warszawa.
- Sierociuk, J. (2014). *Poznańskie archiwum dialektologiczne – zasób nagrań fonograficznych*, opracowanie (forthcoming).
- Sobierajski, Z. (1952). *Gwary kujawskie*. Prace komisji filologicznej. Tom XIV, zes. 2. Poznańskie Towarzystwo Przyjaciół Nauk. Poznań.



# Language documentation. Appendix 2

Home > Book of Knowledge > Language documentation > Language documentation. Appendix 2

■ CHAPTER AUTHOR: KATARZYNA KLESSA

## Data formats and structures

Many popular data formats are based on the XML (Extensible Markup Language, see e.g. [here](#)), recommended e.g. by the CLARIN consortium (Common Language Resources and Technology Infrastructure, read more about format recommendations [here](#), or refer to: Schmidt, Elenius, 2010:123-145). The XML is used as a basis for the output file format in many annotation tools. It enables storing multi-level annotation data and metadata together in one file.

In those cases when the data collected is of various types and formats (texts, photographs, speech or music files, video, etc.), an effective way to deal with them is to create a relation database. When XML files are stored with the use of a relation database (instead of being collected in a folder as a number of separate files), it is much easier and faster to search through them and to generate new files or statistics based on the contents of the original XMLs (see below for more information about databases).

## ■ RAW DATA, CORPORA AND RELATIONAL DATABASES

**Raw (primary) data** is the original or source data that has not been previously systematized, validated or extensively processed. Thus, before making any use of them there is a need to perform at least basic work consisting of checking the actual data quality, quantity and contents. In situation when the linguistic data collection has an explicit aim and is done in a systematic way (often including simultaneous creation of annotations and metadata) it is common to use the term **language corpus**. Language corpora can be organized in various ways, e.g. as collections of files (e.g., stored in folders on hard disk) or with the use of **relational database** technology. A relational database is a powerful tool making it possible to better control and use data. The data in such a database is organized into tables connected to one another which enables searching through all tables, looking for relationships between various types of data, automatically edit data (e.g. bulk operations done for all or selected annotation files) as well as to share the data among many users who can access them in a controlled way without the risk of losing information. Therefore, using a well structured relation database is advantageous over using collections of data files, especially when dealing with larger amounts of data. For example, benefits can be seen when several people want to annotate a number of sound files. In the case of working directly with sound and annotation files, quite a lot of manual work will be devoted to the distribution of files among annotators, controlling file versions, validating them etc., and it will be quite difficult to track the progress (timing, quality) of their work. Moreover, as with any manual work, it may happen that files are lost or mixed while exchanged among users and computers. A solution to this type of problem can be the use of a relation database (together with an appropriate file management software) enabling the control (distribution, inspection) over the annotations at any time, statistics of the annotators' work time, automatic backup copies and many others.

Many of the existing databases are stored on on-line servers and can be accessed via the Internet which makes their use even more efficient. The servers are powerful computers with large capacity hard disks. Users can access data from their own personal computers after connecting to the server. Depending on software solutions the access to data can be granted in various ways, and the same database can be made available with the use of various types of software tools.

One example might be a **client-server architecture** developed for the development of a lexical database used within a Polish automatic recognition system (Klessa, et al., 2010). All lexicon entries and their descriptions are stored in a server, and the annotators can access and edit them using their own computers, but the changes are always saved in the server and thus can be immediately viewed by other users. In this particular case, the database management software distinguishes several levels of access rights for the users: some of them are only allowed to access the data while working in a local network (computers in laboratory rooms), and some have also rights to connect from outside, via the Internet. This type of software solution can provide very powerful and fast options for data processing but you need to install specific software on both the server and client (personal) computers.

## BOOK OF KNOWLEDGE

Chapters: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) **10**

[List of all languages](#) referred to in the Book of Knowledge and other sections of the website.

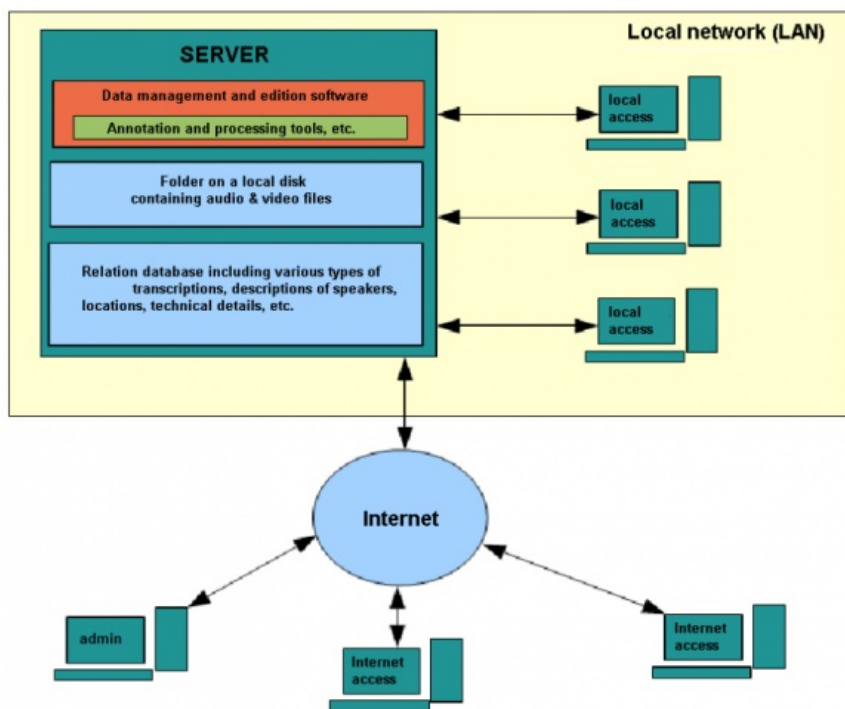
[Glossary](#)

**DOWNLOAD** and print out the Book of Knowledge.

## LET'S REVISE! – CHAPTER 10

Go to the [Let's Revise section](#) to see what you can learn from this chapter or test how much you have already learnt!





An illustration of client-server architecture.

A different example might be an on-line database architecture developed for the use with Internet browsers which does not require installation of any special software to access the data. It might be less efficient or slower than the client-server architecture with respect to processing certain types of data (e.g., management of sound file annotation tasks by a team of annotators), however, it might be extremely useful for sharing data included in language archives or for knowledge dissemination and is very often used for such purposes. Actually, most of the existing on-line language corpora make use of this type of technological solutions. An example might be the on-line database of Poland's linguistic heritage. All data are stored in an on-line server and the same data can be accessed from two websites. One of them is an on-line database editor, available via any Internet browser for registered users (the database creators or data contributors). The second one is a publicly accessible website. The two websites differ with respect to both their functionality and layout. While the editor is a more technical website, not always very convenient in use but providing all necessary editing options, the website for general public presents the same contents in a more clear and reader-friendly way but without any editing options.

#### ■ DATA SHARING, RE-USING AND INTEROPERABILITY OF FORMATS AND RESOURCES

Due to the fact that a variety of language archives co-exist and many new ones will most probably emerge, it is important to consider the issues of re-usability of language data as well as the possibility of flexible use of information from many various archives at a time. That is why one of the important contemporary technological and methodological challenges is to think of possibilities to fruitfully use data originally stored in various repositories. In order to meet the challenge, standards for sharing information and for dealing with data need to be established in a way enabling 'translation' of the formats from one to another. Another issue is the growing need for collaboration between the repository holders with a view to not only make the data accessible (which is comparably easy in the era of the Internet) but also to avoid unnecessary multiplication of the same data and to improve the styles of data presentation.

### LET'S REVISE! – CHAPTER 10

Go to the [Let's Revise section](#) to see what you can learn from this chapter or test how much you have already learnt!

#### References

- Klessa, K., Karpiński, M., Baldys, O., Demenko, G., (2010). Speechlabs ASR. Polish Lexical Database for Speech Technology: Design and Architecture. *Speech and Language Technology*. 12/13, 2009/2010, Poznań, 191–207.
- Klessa, K., Wicherkiewicz, T. (2014). Design and Implementation of an On-line Database for Endangered Languages: Multilingual Legacy of Poland. *Proceedings of the 6th International Conference on Corpus Linguistics (CILC 6)*, Las Palmas de Gran Canaria, Spain, 22-24 May 2014.
- Schmidt, T., Elenius, K. (2010). Multimedia Encoding and Annotation, Common Language Resources and Technology Infrastructure. *Interoperability and Standards*. CLARIN Deliverable D5.C-3 (pp.123-145).