

# Linguistic Diversity and the Digital Divide

Jens Allwood, Dept of Linguistics, Göteborg University

## Abstract

The paper starts by discussing the phenomenon of linguistic diversity, going into some of the arguments for and against preserving diversity. We then briefly turn to the problem of the digital divide and its relations to linguistic diversity. This section is followed by a brief discussion of some of the means to preserve diversity, ending up with a discussion of the role of language technology and a suggestion for the development of “language technological survival kits” or “language survival kits”.

## 1. The Digital Divide and Linguistic Diversity

Let me start by making a few observations concerning linguistic diversity (Cf. also Romaine 1989, Skutnabb-Kangas 2000, Crystal 2000 and Matsumara 1998.). Briefly speaking, there are four to eight thousand languages in the world and a very large number of them are threatened by extinction. The reason that it is hard to give a more exact figure for the number of languages is that the notion of language is vague in a number of much discussed ways. For present purposes, we may avoid the problems of giving a definition of language by saying that we have around 6000 languages in the world. Around 96% of the languages are spoken by about 4% of the population of the Earth. Furthermore, many of the languages included in the 96% are spoken by a low number of persons who are often fairly old. Sometimes there's only one speaker who is over 65 years old which means that when this person disappears the language will also disappear. It has been claimed that, at present, two languages are disappearing every month (Crystal 2000). Whatever the correct rate is, the small languages of our planet are disappearing. Most of these “small” languages are located in the hot climate zones of the earth. There are around 2000 small languages in Africa, about 800 languages on New Guinea and about the same number of languages in India. There are also many small languages in Australia and in the Americas. So where we have a lot of plants and animals, we also have a lot of languages. It has been suggested that there might be relationship between the number of animals, plants and languages but I'm not so sure about this myself. Possibly there is a link through the fact that warmer climate zones often have been more densely populated than colder climate zones.

The remaining 4% of human languages, i.e. roughly 240 languages, are spoken by 96% of the population of the Earth. So now we have the reverse relationship, 96% of the population speak 4% of the languages (240 languages). This means that there are a few languages (like English, Chinese, Spanish, Arabic, Malay, Hindi, French, German and Russian) with very many speakers. About half of these languages come from Europe. It is for these languages with a few additions from other parts of the world that digital language technology is primarily being developed.

Unfortunately, I think we can now make the prediction, that over the next hundred years( i.e. the next century)we will be losing more than a thousand small languages. Although I don't think there's much we can do about this, I don't think we should just passively let it happen. There are today several initiatives and associations that are

trying to struggle against "language death". I think we should support and join these efforts, but unfortunately the trend seems to be that in spite of our efforts, in 100 or 200 years from now, will probably only have a few hundred languages left. Do we want it like this? Most linguists would probably say "no", we don't want it like this. We need to preserve the languages of the Earth and we should join the initiatives that exist to support endangered languages. But there are many people who do not think like linguists, who, in fact, think we should not worry too much about language extinction at all.

## **2. Arguments for and against preserving linguistic diversity**

Since this fairly strong negative opinion exists, it might be of some interest to try to summarize some of the arguments for and against maintaining linguistic diversity on Earth. Let us first look at the arguments against maintaining linguistic diversity. The persons who believe in these arguments are probably mostly to be found in the business community or in international politics, usually from nations not strongly threatened by "language death". Their main attitude is that preserving human languages is a hopeless quest. In the end, we all know that what languages will survive will be decided by money and guns. Those languages survive that can be connected with political, economical and military power. There is nothing we can do about this. If you look back in history at what languages have been widely adopted, those languages have all at some time or another been associated with military and/or economical power. This is the reality of how languages spread. As the Germans say, "it is all realpolitik", i.e. the politics of the real world.

A second argument against linguistic diversity is to say that it is a hindrance for free communication, trade and cooperation. All these crazy languages are just in the way, and mean that we have to spend a lot of money and energy on localizing our products so that they will fit local languages and cultures. Another consequence is that linguistic diversity necessitates language learning and language teaching, which is a real chore. We can see a reflection of this in the fairly low number of people who are learning foreign languages in English-speaking countries. In smaller and less powerful nations, like Sweden, virtually everybody is learning at least English. But the enthusiasm for learning other languages over and above English does not seem to be very high. In any case, if we didn't have linguistic diversity, we could spend our time and energy on other things than language learning or translation between languages.

A third line of argument against linguistic diversity is to say that it is a source of conflict. It constantly creates misunderstandings that sometimes lead to conflict. If everyone spoke the same language, things would be much simpler. We would have a great harmonious community, in which we would have much less misunderstandings and conflict.

Let us now turn to the arguments for preserving linguistic diversity. We should try to preserve linguistic diversity because, after all, human languages are our greatest collective invention. We human beings have probably existed for at least 100,000 years and probably the development of language was one of the factors that made our brain, specifically the cortex, expand in order to hold an increased number of concepts collectively sharable through language. Increased brain capacity and the development of language, probably together had survival value for our species.

Through the development of language, we learned how to coordinate information with other people, a kind of collective information processing, which could then be used for coordinating not only information but also action, in order to survive more successfully in diverse environments. So long before we had writing, which after all may be only three or four thousand years old, we spent many thousands of years developing our spoken languages, storing in these languages a great deal of information concerning natural phenomena like animals and plants but also concerning ways of conceptualizing psychological states and social relations. In general, the vocabularies of human languages contain a wealth of conceptual information, which could be of value to us. For example there might be concepts related to herbs and medical treatments or concepts which have to do with different forms of social interaction, guiding us to how we can better exist together harmoniously or concepts teaching us how we can better understand different states of our minds, for example, through combinations of meditative activity and linguistic articulation of what we are experiencing. Human languages are storehouses of information that we have spent thousands of years developing. It would indeed be sad to see all this great human effort just disappearing.

A second argument says is that linguistic diversity through the conceptual frameworks that are housed in different languages, provides a kind of peaceful conceptual and cultural competition which will be beneficial for the human species in the long run. Having only one language with only one conceptual framework can never lead to the conceptual and cultural diversity that can be provided by several languages and cultures competing with each other.

A third type of argument might be that linguistic diversity leads to a necessity for multilingualism and that there is research (Allwood, Strömqvist & McDowall 1982, Skutnabb-Kangas 2000), which seems to show that people who have become multilingual, actually become more creative. In conceptualizing the world, they have access to several parallel conceptual frameworks which they can potentially make use of. Coming to terms with several languages and several conceptual frameworks seems to be good for your mind. It seems to be healthy to learn several languages. It increases your mental capacity, flexibility and creativity.

Fourthly, there are ethical arguments for linguistic diversity. Since there is a strong connection between a form of life, a way of thinking and a language, the people who lose their language are in a sense a lost generation. They will, especially if they are adult or elderly, usually never be able to participate as fully in social life, or communicate as well in a new acquired language, as they could, had they been able to maintain their first language. In other words, language extinction is connected to social and psychological suffering for the people who lose their language.

This argument then gives rise to a more general ethical reflection of the following sort: It is true that money and guns usually run the world (including the life and death of human languages) but should we really give in to this? Should we, humans not sometimes be able to develop and cultivate something which is a little better than the mere results of a “realpolitik”. We need some ideals over and above money and guns. Perhaps we should not give in, but try to preserve our linguistic diversity in spite of the pressures of the so-called real world.

Finally, there are also we, the linguists who say: “ Well, after all, languages are interesting in themselves. We need increased insights into the nature of human language”. This will be much harder if the diversity of human languages is no longer there. But, of course, working with human languages is what gives linguists their bread and butter (realpolitik) so we are to some extent biased in this matter.

Let us now try to weigh the arguments pro and contra diversity against each other (and I should confess that since I am a linguist, I am perhaps not totally neutral). However, I think the arguments for preservation are slightly stronger than the arguments against preservation but I can understand if the people who think they (and perhaps everyone else as well) have something to gain by diminishing linguistic diversity will not share this opinion. In any case, whatever conclusion we arrive at, it does not seem like the “real politics” of the world is supporting linguistic diversity. As I said initially, we’re moving towards a situation where soon many of the languages of the world will have disappeared.

### **3. The Digital Divide and “oral” languages**

The so-called digital divide to a large extent coincides with the division between “haves” and “have-nots” in the world. The division can be applied to single individuals, ethnic groups and perhaps even nations. Some individuals (ethnic groups, nations) have more access to resources than others. These resources include such things as education, material goods (including computers), knowledge of world language(s) and knowledge of the use of computers. In general, the old, the poor and to some extent children and women have less access to resources while adult rich males have most access. On the level of ethnic groups, dominant (often large or majority) groups have more access than dominated (often minorities) groups. On the level of nations, this means that nations in North America, Europe, Australia, New Zealand and Japan have more access to resources than the rest of the world.

Unfortunately, the areas of the world where there are many “have-nots” also tend to be areas where there are many languages used by few persons. Thus, they are areas of low education, low material standard and low levels of knowledge of computers and of “world language(s)”. Added to these problems is the fact that many “small languages” are only “spoken” or “oral”, i.e. they have no associated system of writing. According to <http://www.yourdictionary.com/languages.html>, which is a continuously updated overview of the world’s languages, 2261 languages at present have some form of writing. This would point to a figure of about 2/3 of the world’s (around 6000) languages as being without a system of writing. A fair number of these languages are found in south Asia. Another figure which is of interest is that the number of illiterates in the world is 862 million people or 16% of the world’s population, (according to the UNESCO statistics website 2003). Thus, even in cultures where there is a system of writing many people live “orally”. Let us therefore briefly consider some of the consequences of not having a written language:

The preservation of an oral (only) language is totally dependent on speakers’ memories. When memory fails there is no more record of the language. There are no word lists, no grammars, no texts at all. This also means that cultures and societies which are totally dependent on an oral language, similarly have to depend on memory of individual speakers. In such cultures we have no contacts, no peace

treaties, etc. All agreements will be based on oral language and basically be a matter of trust.

This means that all forms of control are made more difficult. It also means repeatability and long collective memories are harder to establish. An example of the methods that might be used are the verse forms of originally oral poems like the Iliad or the medieval regional laws of Sweden.

In general, life in oral (only) culture and oral (only) society is in many ways very unlike what we are used to from societies which have relied on written language for a long time. At this point it should be noted that there are few such societies today since most of them are part of larger nation states in which there are languages connected to a form of writing).

Thus, in working with problems of how to support continued linguistic diversity, we simultaneously very often will be faced with problems connected with the division between “haves” and “have-nots” in the world. In addition, we will also in many cases have to deal with problems of providing support in societies which are oral only and therefore are characterized by different conventions and methods of communication and social control than those which are characteristic of societies which for long periods have made use of writing systems.

#### **4. Levels of preservation**

The question of what to do to preserve the world’s languages is to some extent dependent on what level of preservation we have in mind. Let us therefore distinguish 3 such levels.

- A. Possible extinction
- B. Preserved for the record
- C. Full functional viability

The first level means that we engage in no action and accept the fact that the language might become extinct. The second level means that we would like to preserve a record which is as complete as possible of the language for future generations even if there are no more active users of the language, i.e. a kind of museum preservation. The third level of “full functional viability” means that the language is maintained in full active use. The crucial factor for achieving this is that users of the language have to be able to do something sensible with the language. something that counts in their life. The language should serve as an instrument for need satisfaction from basic material and social needs in everyday life, to needs which arise in medial, political, intellectual or religious settings. If the language cannot serve as an instrument in one or more of these settings, it will eventually disappear.

#### **5. What can be done**

So what should and can we do? In order to answer this question, we need an analysis of the situation. Let us take a brief look at the following five factors:

Multilingualism

Legislation and organizational support

Media access  
Education  
Language technology

One of the results of the analysis will probably be that the preservation of small languages requires that their speakers become multilinguals. In the final analysis, it is only possible for people who have English as their first language to remain monolingual in the world today. Speakers of nationally dominant languages, other than English, have to become bilinguals in their own language and English. Speakers of minority languages have to become at least trilinguals, in their own language, the nationally dominant language and English. Depending on the local situation the number of languages may in fact be higher. Thus, realistic support for small languages should also be related to support for multilingualism among the speakers of the language.

Another result of our analysis might be that we find that one thing we can do is to support and create organizations that support the struggle to preserve endangered languages, for example, by pushing policy making in favor of linguistic diversity. This has to be done internationally, in bodies like the European Union or the United Nations, but it also has to be done nationally in all the different nations of the world. One example of this would be to create legislation which promotes several official languages in a country rather than just one. Recently, this has happened quite generally in Europe. Today, there are actually five official languages in Sweden (even though most Swedes are not aware of this). Same (or Lappish) has become an official language. So has Mienkelä (a Finnish dialect in northern Sweden) as well as Romani (the language of the Gypsies) and Yiddish. This kind of legislation is a step forward and we can only hope that it will be more common in the world at large. Normally, this legislation can be used to provide requirements that particular languages must be used, or be available for use in all or several public institutions, like those concerned with medical care, justice or education.

In general, we may note that use of a language in the media gives the language recognition and prestige. It is therefore very desirable to provide space for all languages in the media. This can be done through local radio and TV stations. It could also be done on a worldwide scale by an organization like UNESCO which is committed to support for all human cultures and languages. Concretely this could take the form of an international multimodal TV channel continuously presenting the languages and cultures of the world in short programs.

The Internet, constitutes another interesting area for the preservation of small endangered languages. About 10 years ago, 80% or more of all information on the Internet was expressed in English. Two years ago, the share of English on the Internet had gone below the 50% mark and it seems still to be going down. So today English seems to have only about 45 percent of the information on the Internet. What this seems to indicate is that many people around the world, speaking many different languages, are beginning to learn to apply the new technology using their own languages.

Education is a third important area. Two basic questions connected to “functional viability” are “What kind of education is available in which languages?” and “How

should we best teach and provide learning opportunities for different human languages?”. The first question is related to the question of whether children can be socialized into participating in society through “a threatened language”, thus, maintaining its use in a societal context. The second question is the question of how we provide access to the language for external users who want to extend their multilinguality.

But education should also be related to the fact that many small languages are languages of “have-nots”. In other words, education for people who have a great need to change their situation. How should such education best be organized? Should it be brought to the places where it is needed or should the people who need the education move to places where the relevant education is provided.

A well-known problem with the latter model is that it tends to promote a “brain drain”. A typical case might be sketched as follows. a person leaves a fairly impoverished rural environment in order to receive education in a city or in another richer country. After finishing he/she chooses not to return but rather to stay in the new environment, thus depriving their background environment of their much needed new competence. It can therefore be argued that educational efforts should be brought to the place where they are needed in order to increase the likelihood that they will be of local benefit.

This brings us to a discussion of the role of language technology. Language technology is one of the crucial factors that can help bring about “functional viability” of the local languages and thus increase the probability that there will be transference of education and technology to the benefit of local populations. Today if this is to happen, the language preferably should be “digitally compatible”, i.e. it must be possible to use the language in various computer supported applications like word processing systems, internet links, etc. In other words the “functional viability” of a language will increase if the language can be provided with the fruits of language technology (cf. also McEnery and Ostler 2002 as well as Allwood 2001).

However, language technology will also be valuable even if our goal is primarily to preserve the language for the record. In fact, the goal of recording a language to some extent harmonizes with a goal of functional viability. in the sense that most language technological tools presuppose basic linguistic resources in the form of corpora. Thus, the creation of digitalized corpora is a very basic task in the effort to preserve the world’s languages. The corpora can be multimodal, spoken or written, depending on what type of linguistic material and recording equipment is available.

Since so many of the world’s languages do not have writing systems, it is essential that we develop methods for efficient creation of multimodal (including pictures and video clips) and spoken language (including sound files) corpora. It goes without saying that such corpora are especially desirable for languages that are threatened by extinction. For a language that has no writing system, a multimodal or spoken language corpus can serve as a basis for the development of writing systems and subsequently for such resources as dictionaries and grammars. Wherever written language material is available this should be used even if the difficulties concerning different standards of writing are considerable, cf Baker et al. 2003. The methods should be such that they give guidance on recording (microphones, cameras),

scanning, sampling (types of genres, activities, speakers, regions, etc) and organization of collected material into a digital database.

Unfortunately, the tools provided by language technology are almost entirely based on the existence of a written code, which should preferably be ASCII compatible. Thus, there are problems as soon as we want to use Roman letters, like Swedish å, ä, ö which are not ASCII standard and even greater problems if we want to use non-Roman systems like Devanagari or Chinese. See Baker et al 2003. In order to apply language technology to oral (only) languages we are, thus faced with the double task of creating a writing system for the language and the task of then using this system to provide language technology.

It might be thought that technologies such as speech recognition and speech synthesis should enable us to work directly on spoken language without the intermediary of a written code. However, speech synthesis presupposes written language as input and speech recognition presupposes written language as output. There is, in other words, at present no good way to provide language technology without the use of a written language.

A natural next question will then be – which written language. Should we in South Asia choose an extension of Devanagari in keeping with cultural traditions or should we choose an ASCII compatible variant of Roman script? Reasons of economy and expedience at present point to the latter choice.

Another question that arises is the question of how language technology can be provided efficiently and inexpensively. Probably a key factor here is reuse of technology for similar languages. Using this method, we can either extend technology from one language to another related language or in some cases present similar solutions for a group of related languages. This would, in fact, also facilitate introduction of a written code. We could perhaps envisage a “language technology localization kit” or “language survival kit” which would contain methods for the creation of multimodal, spoken or written corpora and for introducing an ASCII compatible writing system if this is required, and on the basis of this, then introducing other technologies like word processing and automatic translation through the use of aligned texts. Access to the web and information retrieval can then be gradually provided, either directly or through translation. Finally, speech synthesis and speech recognition would be valuable since they could perhaps provide a short cut to the use of various digital services in a predominantly “oral culture”.

## **6. Conclusions**

In this paper we have surveyed some of the factors affecting linguistic diversity in the world. In doing so we have considered some of the arguments for and against preserving linguistic diversity, drawing the conclusion that perhaps arguments for preserving diversity are stronger than arguments against. We have also considered how linguistic diversity is related to the digital divide. Finally, we have discussed different levels of preservation for a language and taken a look at some of the means that are possible to use in order to preserve linguistic diversity ending up with a suggestion for the development of a “language technology localization kit” or “language survival kit” to help preserve the world’s languages and to make them digitally compatible.

## Bibliography

Allwood. 2001. Language Technology as an aid in preserving linguistic diversity, *ELSN*ews (10.1) Spring 2001, webversion at <http://www.elsnet.org>

Allwood, J., Strömqvist, S. & MacDowall, M. (1982) Barn, språkutveckling och flerspråkighet in Government report SOU 1982:43. Also in Gothenburg Papers in Theoretical Linguistics S6, University of Göteborg, Dept of Linguistics.

Baker, Paul; Hardie, Andrew; Jayaram, B:D; McEnery, Tony (2003). "Corpus Data for South Asian Language Processing".

Crystal, D. (2000). *Language Death*. Cambridge University Press, Cambridge, UK.

Matsumara, Kazuto (Ed) (1998). *Studies in Endangered Languages (Papers for the International Symposium on Endangered Languages, Tokyo 18-20 November 1995)*, Ititzu Syobo, Tokyo.

McEnery, Tony and Ostler, Nicholas (2000), *A New Agenda for Corpus Linguistics - Working with All of the World's Languages*. In *Literary and Linguistic Computing*, 15, 401-418

Romaine, S. (1989) *Pidgins, Creoles, Immigrant and Dying Languages*: In Dorian (ed.), *Investigating Obsolescence: Studies in Language Contraction and Death*. Cambridge University Press, Cambridge, p. 369-383.

Skutnabb-Kangas, Tove (2000) *Linguistic Genocide in Education – or Worldwide Diversity and Human Rights?*. Mahwah, NJ & London: Lawrence Erlbaum.

UNESCO Institute for Statistics. "Literacy" <[http://portal.unesco.org/uis/ev.php?URL\\_ID=4926&URL\\_DO=DO\\_TOPIC&URL\\_SECTION=20](http://portal.unesco.org/uis/ev.php?URL_ID=4926&URL_DO=DO_TOPIC&URL_SECTION=20)> Updated 14 February 2003.

<http://www.yourdictionary.com/languages.html>