Language survival kits

Jens Allwood

1. Should we save the languages of the world?

According to the survey of the world's languages published by Ethnologue, there are around 6 800 languages in the world. Most of these are spoken by very few persons. Approximately 96% of the world's languages are spoken by less than 4% of its population, implying also that 96% of the world's population speak only 4% of the world's languages. Since many of the speakers of "small languages" (in terms of number of speakers) often are older people, this means that languages are presently disappearing at a high rate. Crystal (2000) claims that the rate of language disappearance is two languages each month. Most of the ”small languages” are located in hot climate zones and are often spoken by people who are very poor in terms of economic and material resources. Because of the prevailing socio-political and economic situations/factors, a safe but sad prediction is that we will be losing between 1 000 to 2 000 languages over the next century. In contrast to that, around 100–200 "strong" languages are likely to maintain their position. Of these, around 10 languages, languages such as English, Chinese, Spanish, Arabic, Malay, Hindi, French, German, Russian, and Japanese are likely to strengthen their position. Of the other "strong" languages, more than half are spoken in Europe, within in economically affluent countries that have a comparatively high standard of living (see Matsumara (1998), Crystal (2000) and Romaine (1989) for a more general background).

Is the process leading to language loss good or bad? Although most linguists deplore language loss, there are those who are more positive to it. Let us therefore briefly summarize some of the arguments for and against linguistic diversity on Earth.

1.1. Arguments against linguistic diversity

1. The pressures of military, economic, technological, and social development push towards more integration in the world, with the use of fewer languages as means of communication. Languages which are tied to political, economic, and military power are likely to survive if the trend continues. Try-

ing to maintain linguistic diversity, in the end, may be a meaningless effort. It is thus an unwanted and hopeless quest to try to preserve “small human languages”.

2. The diversity of languages on earth is a hindrance to global trade, scientific cooperation and communication in general. Large multinational companies need to use immense economic resources to localize products to fit with “local” languages and cultures. These resources could be used more productively if one could avoid these “local adjustments”. From a commercial outsiders' perspective, linguistic diversity mainly benefits those who can use linguistic diversity to their advantage, such as the ”localization industry”, the growing community of translators and interpreters, and language teachers and language training institutes.

3. Linguistic diversity not only leads to a waste of resources, it also complicates interaction between different parts of the world and can lead to conflict through the misunderstandings that often arise when people do not share languages and cultures.

In other words, cooperation in the world would be considerably easier, more efficient, and less expensive if there were fewer languages (perhaps only one).

1.2. Arguments for linguistic diversity

1. Human languages are our greatest collective cognitive achievement . The development of human languages probably coincides with the development of human beings as a species. Through the mutual attunement of increased brain capacity and human languages, humans were able to develop not only individual information processing but also collective information processing and human cultures. Through languages, humans were able to coordinate thoughts and actions which enabled a survival capacity in diverse environments. Human languages maintain and, for new generations, preserve human conceptual and social development. Through human languages, we have access to diverse ways of classifying both the natural and the social environment, including artifacts as well as concepts having to do with cognitive states and abstract features of a world view. There is no better testimony to human intellectual effort over one hundred thousand years than human languages. Rather than letting go of this information, we should try to preserve as much as possible of it before it is too late.

2. The conceptual frameworks associated with different human languages

have developed under mutual influence, interaction, and competition between languages and cultures for a long time. For millennia, human languages have provided a basis for a cultural and conceptual competition between different communities which probably has been for the good of mankind. If this competition were to be diminished or disappear, the risk is that the population of Earth would have to collectively suffer the drawbacks which are usually associated with a situation of monopoly.

3. Multilingualism is a large part of the world's population. The majority are, and have always been, multilingual. We are perhaps not only genetically prepared for one language but for several. At any rate, research seems to show that multilingual people, by having to learn to live with the conceptual frameworks of several languages, become more creative (Allwood, Strömqvist and McDowall (1985)) and Skutnabb-Kangas (2000)). Multilingualism is thus good for our cognitive development. Learning several languages increases mental capacity, flexibility, and creativity.

4. There are also ethical arguments in favor of a multilingual world. The first one is of a more general nature: Should we human beings really give in to military and economic pressures in the shaping of our future world? If we believe that linguistic and cultural diversity is beneficial, should we not be able to create social and organizational structures which would make it possible to preserve this diversity? Should we not use technology as a tool in maintaining linguistic diversity?

There is also an ethical argument favoring linguistic diversity at a more individual level. What happens to the generation of people whose language disappears? Their language loss, in many ways, makes them a lost generation. By losing their languages, they also lose access to their conceptual, cultural, and social heritage. In most cases a language loss occurs by language shift, when speakers of a language start using another language in domains which they earlier used their own language for. This language shift brings with it the cultural values of this other group, both towards their own language (community) and to the outside world. This language shift thus not only means that speakers lose their mother tongue but also their socio-cultural values and traditions. So, loss of languages is usually connected to social and psychological suffering for those whose languages is lost.

5. The final reason for wanting to preserve human languages is related to the first point mentioned above. Human languages can potentially provide insight not only about the nature of human languages *per se* but also about human nature itself. Languages give us information about possible cognitive, social, and communicative structures both for and within human beings. If the diver-

sity of human languages diminishes, this source of information for insights into human nature also diminishes.

If we weigh the arguments for and against linguistic diversity, I suspect that the situation is still not entirely clear. Depending on one's goals and orientations, one will favor one of the two positions outlined above. Being a linguist with an interest in the nature of human languages and in the relation of language to human nature, communication and social organization, I tend to be more impressed by the arguments in favor of trying to maintain linguistic diversity than by the arguments against diversity.

2. Levels of survival

Even if one is persuaded that maintaining linguistic diversity is important, one is confronted with what one means by the survival, preservation, and maintenance of linguistic diversity, as these can be of many different types. Let us distinguish at least the following three levels.

Level 1: Possible extinction
Level 2: Preservation for the record (museum)
Level 3: Maintenance of full functional viability

To reach the first level, it seems that we do not need to do anything. This seems to be the end result of the present development for many of today's threatened languages. The second level implies that we recognize the importance of linguistic diversity, and that we want to do something about it before a language completely vanishes from the face of this earth. This, at the second level, can, for example, be done by collecting written, spoken, and gesture data of these languages and trying to store these data in such a way that they will be accessible for future study. It also implies that we should try to describe and explain as fully as possible, by providing contextual and cultural background for the data we have collected. One reason for this is that languages are not self-explanatory. They require interpretation based on use in context. We can see this by examining now-extinct languages like ancient Maya or Hittite. Even if we have been able to collect written data from these languages and even if we have been able, to some extent, to interpret some of the texts, our interpretation is limited by the fact that we do not have access to the speakers and cultures behind these languages.

The third level (i.e., maintenance of full functional viability) is a level where the language is used for all communicative functions which an individual encounters in his or her private or public life. If we also include in this the need to communicate with people from other language communities than our own, or to interpret documents and other remnants of the historical past, perhaps no language has been able to provide full functional viability in this sense. However, in a world which was less interconnected through uses of transportation and information technology than the world of today, many more languages came very close to the goal of full functional viability. In today's world the number of languages which provide full functional viability is diminishing, with one language—English, for example— taking over functions (like scientific writing) that were previously more distributed between languages.

The general picture, in fact, could be characterized as one of English plus (+). In the English-speaking countries only one language, English, needs to be learned. This will be sufficient both for internal domestic and very much, if not most, external communication. In a nation with a strong (official) dominant language, the situation is one of "English + dominating language". This would be the case, for example, for speakers of Hindi in India but also for speakers of most of the European languages other than English in Western Europe. For speakers who come from a minority group in a region with a more dominant language, the situation is one of "English + dominant language + minority language" and in some cases "English + dominant language + subdominant language + minority language". This is, for example, the case with a Kinnauri speaker in Northern India, who learns Himachali at an early age as it is the neighboring dominant language, and, then when this person starts going to school, learns Hindi, and then English. In other words, speakers of such minority languages often learn three, four, or more languages. Today, this is the situation for most speakers of minority languages. A main reason for this is the high penetration of dominant cultures into most areas of the world through the development of means of transportation and information technology.

Against this background, maintenance of functional viability means preserving a minority language in at least the functional domains it is used in today and, if possible, in regaining some of the functional domains that have been lost.

3. Means to prevent language extinction

We now turn to the question of what can be done to support languages that are threatened by extinction.

Depending on what our ambitions are, we can imagine different means to be used for aiming at different levels of language survival. The means can be, for example, analytical, legislative, educational, and/or technological. We will now consider some of the possible types. Some of these could possibly become part of the "language survival kit" which will be discussed in section 4.

A first step would probably be to discuss the kind of survival or functional viability that would be optimal, possible, and realistic in a particular language community. It should be fairly clear that what is possible might not be realistic and that what would be optimal might be neither probable nor realistic. At any rate, the contrast of the three concepts might help to clarify the situation.

As we have already noted, in most cases, in order to be realistic, this will probably mean support both for some type of individual multilingualism and for some type of societal multilingualism. A second step might therefore be to investigate what legislation exists in relation to both individual and societal multilingualism, what "individual linguistic rights" exist in the society, and what "community linguistic rights" exist.

Three important areas of support are education, the media, and increasingly the Internet. When it comes to education, perhaps the most basic question is: "what kind of education is available in the language?" Since education is one of the main ways in which children gain access to participation in a particular society, and this participation usually involves use of language, providing education in a particular language is one of the key means to maintaining a language. If for economic or practical reasons only some topics or subjects can be taught in a local language, it is useful to make an analysis of what topics would make most sense given the needs, interests, and beliefs of the local population. In fact, local participation in the educational process (based on local competence) will probably almost by necessity be the decisive criterion for what can, at least initially, be taught in a local language. This could, for example, be information about the culture and customs which the language encodes. Over time, however, also other topics should be taught locally.

In fact, irrespective of whether local education takes place in the local language or not, it is usually preferable to arrange the education locally. There is otherwise a considerable risk that education might be a main cause of a "brain drain" out of the local community. People go elsewhere for education and since conditions are often better than in their original home, they do not return to bring the benefits of their education to bear on local problems so that, in the worst case, the local community is deprived both of people with an entrepreneurial spirit and of native speakers using their languages in a variety of functional domains.

Besides education, the mass media are a second strategic area for use of a language. If a language is used in the media, it automatically acquires a sort of public/official status. The language becomes publicly recognized and its speakers usually experience this as a boost to their self-esteem and prestige. Access to as many media as possible is probably helpful. But there should be an analysis of which medium, (e.g., books, newspapers/magazines, TV or radio) is most beneficial in a given situation. Clearly, if there is no written language, only TV or radio can be considered. Further, for economic reasons, radio will probably often turn out to be most cost-effective. Local radio stations broadcasting in the world's non-written languages should therefore be high on the list of contents in a "language survival kit". It is also desirable that educational initiatives be combined with media use. In an "oral community", radio (and TV) combined with face-to-face interaction are clearly the most straightforward ways of providing education in a threatened language.

A third means that can be enlisted in empowering endangered languages is the Internet. Since so far most of the information on the Internet exists in written form, its use is mainly restricted to languages that have a writing system. A strong *desideratum* is therefore that we develop multimedial uses of the web (including sound and film) which are easy to use and are widely available. The Internet can, in this way, be used, for example, to provide multimodal literacy training, perhaps starting with illiterate speakers of a language that already has a writing system and later continuing with speakers of languages that are acquiring writing systems. In practice, not all languages that have writing systems can make use of the Internet since most uses of the web are based on the Latin alphabet and the ASCII code. Again, a strong *desideratum* is that standardized uses of Unicode for non-Latin writing systems be made more widely available. This would greatly improve the possibilities for these languages to make use of the internet to provide increased functional viability. (For some of the problems that need to be solved, see Baker et al. (2003)).

We now turn to a discussion of some of the ways in which language technology can be used to support the threatened languages of the world, cf. also Allwood (2001) and McEnery and Ostler (2000).

4. Language survival kits

A very basic level or goal of language survival is the maintenance or preservation of a language "for the record" on a "museum level". However, this goal is not incompatible with more ambitious goals of functional viability. In fact, meeting the requirements of the goal of documenting a language for the record is often a prerequisite for realizing the more ambitious goals. In what follows, I will briefly discuss some ways in which language technology can be used to create the contents of a "language survival kit" which has both the goal of preserving a language for the record and the goal of providing a basis for more active and functional use of the language. The contents of the kit correspond to three kinds of *desiderata*:

(1) general *desiderata*
(2) creation of language resources
(3) creation of useful applications

Below, we will now discuss these three kinds of *desiderata*.

4.1 General desiderata

Some of the general *desiderata* for a language survival kit might be the following:

- Both languages with and without writing systems should be covered
- Low-cost or free ware
- Open source and general standards
- Enable automatic computer-based analysis
- Enable reuse of technology and linguistic analyses

The ambition is to be useful both in relation to languages with a writing system and to languages that lack a writing system. Since speakers of "small languages" usually also have very small economic resources, another concern is that everything that is suggested should preferably be low-cost or free ware. Since we also want to invite participation from sympathizers around the world, open source programs like Linux are to be preferred. Similarly, to facilitate cooperation, general standards that can handle a variety of linguistic phenomena should be used. For languages with writing systems, perhaps the best alternative at the moment is Unicode.

A further *desideratum* is that we should be able to make use of automatic computer-based analysis if possible. One example of this is to use machine-learning techniques, i.e. algorithms which automatically learn linguistic generalizations from raw or annotated language data. We will need such techniques considering the fact that the number of endangered languages is high, time is short, and economic resources are limited.

Finally, it is desirable that Language Technological tools are resuable. Languages have a great deal in common and the components of a language survival kit should be sufficiently generic, standardized, and robust to be reusable when moving from one language to another. Similarly, in many cases, especially when languages are typologically and historically related, many of the features of an analysis for one language should be reusable in the analysis of a related language. Thus, the kit should facilitate structurally similar analyses for the phonology, morphology, lexicon, phraseology, syntax, and even pragmatics for related languages.

4.2 Creating linguistic resources

The first task of a language survival kit will be to establish what is often called "linguistic resources" for the language community. The following are some of the resources that should be created.

Basic resources

- A multimodal spoken or written database (corpus)
- Routines for recording, storing, and analyzing the data
- Transcription standards and transcriptions
- Standards for the creation of writing systems and their digital encoding (e.g. ASCII, Unicode, SAMPA)
- Annotation standards of various types
- Routines for automatic linguistic analysis (e.g. by machine learning)
- Guidelines for descriptions (and explanations) of the language, e.g. grammars and lexica

Thus, the first goal will be to establish multimodal spoken or written corpora for any given language. What this means is that a database for the language containing written, spoken, or gestural data in combination with cultural information in digital form should be created. The "raw data" will be texts, audio recordings, and video recordings. Everything should preferably be in digital form to facilitate later processing. The survival kit, thus, has to include good robust low-cost digital audio and video recording equipment or information about such equipment. It has to include the means of storing recorded data (tapes or digital storage). In cases where digital print files are available, there should be routines for organizing them into a corpus. In the event that printed digital data is not available but there is printed material on paper, scanning equipment has to be included.

Creating the corpora should be seen as an incremental process, where multimodality might not be reached initially or goals of size might only be reached after some period of time. For this reason, it is not meaningful to give absolute quantitative goals for the size of the corpus. However, a reasonable ultimate goal for a spoken language corpus might be between 50 and 100 hours of recording, which corresponds to between 500 000 and 1 000 000 words. For written language, the initial goal might be a corpus of between 1 000 000 and 3 000 000 words.

In collecting the corpora, the recordings which are made should strive for optimal sound and video quality, and at the same time also strive for "ecological validity". This can be achieved by obtaining recordings of a representative sample of societal activities and speakers, that is, drawing from the life of the community. Since the goals of optimal sound quality and "ecological validity" are not always compatible, deciding what to record will sometimes involve a process of "suboptimization". Another important concern, related to recording, is that the video recordings provide a constant view of the interaction between as many speakers as possible. Only this type of recording allows for a study of communication as an interactive process. Focusing and cuts are, thus, to be avoided since they remove a clear view of the interaction.

Once records have been made, it is essential to agree on standards for storage and data classification (metadata). In the worst case, large amounts of data are recorded which can never be used, since it was not classified and stored in such a way that it can be retrieved. The language survival kits should therefore include suggestions for a system of data classification and storage. The system should be rich enough to allow searches on, for example, the date of recording, what was recorded, who was recorded, what transcriptions were made, and what kinds of analysis are connected to the recording. For more detail, cf. Allwood et al. 2000.

The next step will probably consist in some form of annotation (coding) and/or transcription (if there is a writing system). If the data are not transcribed and annotated in a standardized way, no consistent patterns will be found, even if they exist. Standardization is, thus, a clear prerequisite for further analysis, especially when automatic analysis is used. Since some languages are connected with several written variants and there is often a fairly large distance between spoken and written language, it is essential to agree on a standard for annotation and/or transcription, in order to avoid much effort at a later date.

Often it is desirable that the standard be such that more features of spoken language can be included than are reflected in normal standard written orthography. The spoken language features can be divided into two types. Some of the features, like overlap, stress, pausing, and the use of gestures, occur in all spoken languages and can thus be standardized independently of language. Other features of spoken language concern pronunciation. For these features, one alternative would be phonetic or phonemic transcription. Since this is rather labor-intensive, it might be more desirable to use standard orthography with some modifications for spoken language (cf. Nivre 1999). The corpus will also be more valuable if audio recordings, video recordings, and transcriptions are aligned temporally, so that it is possible to simultaneously read the transcription and to hear and view the relevant recorded passages on which it is based.

A special and intriguing problem is connected to languages that have no system of writing. Either these languages must somehow be processed and analyzed directly on the basis of audio and video files, or they have to be provided with a system of writing.

Regarding the first option, there is hardly any available language or speech technology that does not require written language as input or output. For example, writing is the normal input for speech synthesis and the normal output of speech recognition. One could imagine having pictures or diagrams as input or output instead but, by and large, we are still lacking language (or speech) technology that can work without a writing system. To aid in the preservation and analysis of languages without a writing system, it would therefore be important to develop such technology. Some examples of what could be done are use of recorded samples, use of concatenated synthesis and use of multimodal interfaces with graphics, photorealism, streaming etc. A future goal would be to use speech recognition to build interactive dialogue systems of different types, e.g. public information and educational systems.

The second option is therefore perhaps equally or even more realistic—providing a language with a writing system. One *desideratum* here would be to combine automatic analysis of speech with phonemic manual analysis. Automatic analysis based on speech recognition would provide suggestions for sound units, which would then be subjected to manual phonemic analysis. The end result might be a writing system based on IPA, e.g. in its ASCII compatible form SAMPA. Since IPA is an extension of the Latin alphabet, this solution might, for cultural and historical reasons, be undesirable in some parts of the world (e.g. India). Here, instead, a writing system based on the local tradition of writing, e.g. some extension of Devanagari or Arabic script, might be a better alternative.

Once a corpus has been established (or even before), it should be described and explained. The description should include the standard linguistic types of analysis, i.e. phonological, morphological, lexical, phraseological, syntactic, semantic, and pragmatic analysis. In the interest of time and re- sources, as much as possible of the analysis underlying the descriptions should be done automatically. If the language has a writing system, or if a writing system can be established for the language, one of the earliest products of analysis could be a frequency dictionary of morphemes, words, or collocations. This can later be elaborated and refined through the addition of manual or automatic analysis yielding outputs like a lemmatized word frequency list, concordances, part of speech tagging, or morphological analysis. In performing such analyses, results from related languages should be reused and attempts should be made to gradually convert manual routines into automatized routines, for example, by including them in a machine learning program.

4.3. Useful applications

In order to have functional viability rather than mere preservation for the record, the language must be usable. From a language technology point of view, this means that it must have some useful applications.

Some of the applications that might be useful are the following. As above, the list is not to be taken in an "all or nothing" sense, but can be gradually extended as resources develop.

Useful applications

- Multimodal and text interfaces
- Speech synthesis
- Authoring support (word processing)
- Multimodal tutoring support
- Support for Internet use
- Support for information retrieval
- Support for translation
- Support for generic dialogue tools
- Speech recognition

For languages without a writing system, multimodal interfaces with icons and recorded speech should be created. For these languages, good handling of recorded speech and speech synthesis based on concatenated phones, diphones, or triphones is essential. Such multimodal interfaces (containing combinations of icons, animated cartoons, and recorded or synthesized speech) can then be used to create multimodal tutoring systems that can be used to distribute information about health care, agriculture, or low-cost (solar) energy production. Multimodal interfaces could also be used to develop voice mail systems that can serve as alternatives to e-mail based on writing. Future developments of speech recognition might even make an integration of voice-based and writing-based email possible. Further, many household appliances could be equipped with systems for speech control, the use of which could perhaps be extended to some of the "small languages" of the world.

Similarly, for languages that have a writing system, as already mentioned, one of the most basic things is to provide users with interface texts to the computer. Since levels of literacy might not be high, a combination with a multimodal interface using icons and recorded speech might also be useful in this case. Given a writing system, perhaps the most basic application is a tool providing authoring support (a word processing system). This can be incrementally extended as components become available. In other words, basic functions, like delete, copy and paste, are perhaps more needed than hyphenation, spelling, and grammar correction. A special challenge might here be the construction of a speech-based word processing system, where an example of a question requiring an answer might be: What kind of (graphical) means could be used in a speech-based system to give the kind of overview that is normally associated with the reading and editing of a text?

As we have also briefly mentioned above, another area of functional use will be the Internet. Just as with radio broadcasting, speakers of a language will receive a boost in self-esteem and will experience a heightened prestige for their language if it can be used on the Internet. There should therefore be support for Internet use, in the form of email programs and tools for the creation of homepages. But besides being used for private personal communication, the Internet should of course also be used to provide public information and to create the educational programs discussed above.

Further development should lead to support for information retrieval, for classification in a database, and for translation. Finally, various interactive applications utilizing generic tools for dialogue can be created. At first, these will probably involve written language, but over time attempts should be made to create systems for speech recognition, so that interactive systems for illiterate users can also be established.

5. Conclusions

Many of the world's languages are threatened by extinction. After having discussed arguments for and against interfering in this process, I conclude that it would be desirable to interfere. I then go on to briefly discuss some means to prevent language extinction, and suggest that one of the ways to do this would be to use present day language and speech technology to create a "language survival kit". Such a kit could be used not only to preserve samples, descriptions, and explanations of the language for future generations, but also to make the language more functionally useable for its present speakers. The paper provides a discussion of what the contents of such a kit should be, and of some of the challenges that have to be faced in putting it to use.

Bibliography

Allwood, Jens
2001 Language Technology as an aid in preserving linguistic diversity. *ELSNews* 10.1. <http://www.elsnet.org>.

Allwood, Jens, Sven Strömqvist and Monica MacDowall
1982 Barn, språkutveckling och flerspråkighet. Government Report SOU 1982: 43. Also in *Gothenburg Papers in Theoretical Linguistics S6*, 1–290. Department of Linguistics, University of Göteborg.

Allwood, Jens, Maria Björnberg, Leif Grönqvist, Elisabeth Ahlsén and Cajsa Ottesjö
2000 The Spoken Language Corpus at the Department of Linguistics, Göteborg University. *FQS – Forum Qualitative Social Research* 1(3): 1–22.

Baker, Paul, Andrew Hardie, B. D. Jayaram and Tony McEnery
2003 Corpus Data for South Asian Language Processing. Unpublished manuscript.

Crystal, David
2000 *Language Death*. Cambridge: Cambridge University Press.

Grimes, Barbara (ed.)
2000 Ethnologue: Languages of the World. Dallas: SIL International.

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

McEnery, Tony and Nicholas Ostler
2000 A new agenda for Corpus Linguistics—Working with all of the World's languages. *Literary and Linguistic Computing* 15: 401–418.

Nivre, Joakim
1999 Transcription Standard Version 6. Department of Linguistics, Göteborg University.

Romaine, Susan
1989 Pidgins, creoles, immigrant and dying languages. In *Investigating Obsolescence: Studies in Language Contraction and Death*, Nancy Dorian (ed.), 369–383. Cambridge: Cambridge University Press.

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