What should the ideal online-archive documenting linguistic data of various (endangered) languages and cultures offer to interested parties? Some ideas of a technically naive linguistic field researcher and potential user

Paper to be presented at the LREC preconference workshop:
"Resources and Tools in Field Linguistics", Las Palmas, 26.-27.05. 2002.

Gunter Senft

MPI for Psycholinguistics
PB 310
NL-6500 AH Nijmegen
The Netherlands
e-mail: gunter@mpi.nl

What should the ideal online-archive documenting linguistic data of various (endangered) languages and cultures offer to interested parties? Some ideas of a technically naive linguistic field researcher and potential user

Gunter Senft

1. Introduction

It goes without saying that any archive documenting linguistic material is only useful if the available data can satisfy its users needs, demands and standards. An archive documenting material on (endangered cultures and) languages is of interest for many parties. This makes high demands on how the linguistic data have to be presented in the archive, how they have to be annotated, and how they have to be made available for its users. Among the users of such an archive are certainly not only scientists working in the various subdisciplines of linguistics, anthropology, sociology, demography, history, and other social and cognitive sciences but also members of the speech communities whose language and culture is documented there. Moreover, the archived data should also be useful for future generations and their potential interests in the materials.

As a (computer-) technically rather naive linguistic field researcher with interests in linguistic typology as well as in cognitive anthropology and linguistics I raise a number of questions in this paper that I think are of crucial relevance for the set-up of such an archive. Some of these questions have to do with past and present (anthropological-) linguistic research problems and interests in, for example, systems of nominal classification and serial verb constructions in various languages, some of these questions approach the problem of how much energy and time colleagues contributing to such an archive have to invest in order to document a language as adequately and exhaustively as possible. Some of these questions deal with aspects of language documentation that (to my mind) may be of interest for future members of the speech community in which I have been doing my research for the last 20 years – namely the Trobriand Islanders of Papua New Guinea, and some of the questions deal with ethical problems and with issues of intellectual property rights.

I hope that these questions will contribute to foster the cooperation and communication between field linguists, technicians and other experts developing linguistic resources and tools for optimal use of these resources by an as broad a group of interested parties as possible.

2. An outline of the ideal online language (and culture) archive

Let me start with the following practical example: Suppose I am interested in the following two linguistic topics: serial verb constructions and nominal classification systems. I want to get as much information as possible from those languages that are documented in an on-line archive. If such an archive would be ideal, I could do the following search and get the following kind of information and data.

I visit the web-site of the archive and find a "SEARCH" function for the archive. I type in "**serial verb constructions**" and "**nominal classification**". The search machine presents me the results of this search, listing

the languages and

the files

within the archive where I can get information on these topics.

However, the search function is constructed in such a way that it also offers me links to the following search entries:

With "**serial verb constructions**" I get links to the entries:

"types of serial verb constructions",

"complex verbs",

"multi-verb constructions",

"clause integration",

"speech formula", and

"culture specific rules for serial verb constructions".

And with “**nominal classification**“ the search function offers links to the entries:

“measure terms“,

“class terms“,

“numeral classifiers“,

“genitive (attributive, possessive, relational) classifiers“,

“noun classifiers“,

“verbal classifiers“,

“noun classes“,

“gender systems“,

“locative classifiers“,

“deictic classifiers“,

“nominal classification and culture“, and

“nominal classification and cognition“.

I print out this additional information and then go to one of the languages that were listed by the search function. Suppose this language is “Kilivila“. I click on the entry “Kilivila“ and the archive opens up its inventory list of files for this languages. There are basically two types of files:

“**metadata files**“, and

“**(audio/video/written or printed) data files**“.

In the **metadata files** I find the following kind of information:

typological classification of the language;

where the language is spoken and by whom;

anthropological and ethnographic information on the speakers and their culture(s);

information on the researchers who gathered and documented the language and culture data, where to contact them and how to get further information on their research, on these (and other related) data, and on their publications;

information on the theoretical basis the researchers chose for presenting, annotating and transcribing the archive data (including the justification of the orthographic system used for such a transcription) as well as a list of all abbreviations used in the (morpheme-interlinear) transcriptions;

information that is directly linked with each and every data file that describes the genre(s) to which the data belong, the variety used by the speaker(s), the relevant personal data of the speaker(s) or actor(s) like e.g. status, role, age, gender, name, relation to other speaker(s)/actor(s), place of living, etc., the documented scenes, the place where the scenes happened, and any other additional anthropological-linguistic facts that are necessary to understand the documented data in the file that presents audio- and/or video- and/or printed or written documents.

In the **data files** I find the actual data – either audio- or video-documents or documents with material printed and/or written in this language. The audio- and video-data will minimally come with an orthographic transcription (as mentioned above, the basis for this orthographic system is found in one of the metadata files), a morpheme-interlinear transcription, and a free translation into English. If the data file consists of documents with written or printed material, they should also be presented with a morpheme-interlinear transcription and a free translation into English (if the data file does not present something like, for example, the translation of the Bible or parts of it into the respective language documented in the archive).

Now – coming back to my example - I first want to find out whether there are publications on serial verb constructions and on nominal classification in Kilivila. Then, on the basis of the linguistic information which I found in the metadata files, I will open those data files where I know that I will find instances of, and examples for, serial verb constructions and nominal classification. In the most ideal of all worlds there would be another search function for the data files – and when I search for serial verb constructions and nominal classification the various examples for these phenomena would be selected (with the option: “find next“, of course). However, this may be a little too much to ask for. Therefore, I as the archive user, check these data files and the archive allows me to

download either whole data files or parts of such files that contain the data relevant for my research interests. Moreover, I can also print out the transcriptions and the translation of the chosen material.

Assume now that the data file in the archive presents the actual audio- and video signal together with the orthographic transcription, the morpheme-interlinear transcription and the free translation, but that the sentences of these transcriptions are not synchronized with the actual audio- and video-data. However, having loaded down this file I can now manipulate it, modify it, and work on it. That is to say, I can synchronize the audio- and video-signal with the transcription – and I can also do additional – new - analyses like, for example, intonation analyses of these data to find out whether it is really true for Kilivila that the intonational properties of a clause with serial verb constructions are those of a mono-verbal clause.

Researchers who put their data into the archive also have this possibility to revise the data, of course. However, they have to refer explicitly to such revisions in the respective metadata file that goes with any (such modified) data file.

The excellent information in the metadata files allow me to categorize the data on serial verb constructions and on nominal classification I found for Kilivila in such a way that I can distinguish between elicited and more natural, i.e. non-elicited data. Moreover, I find all necessary information in the metadata files with respect to which methods were used to collect the elicited data. With respect to the non-elicited data I can also differentiate the genres in which these data were produced. Moreover, the metadata files for the non-elicited data also provide me with all necessary etnographic information on culture-specific phenomena relevant for these data. If I am interested in the inventory and the use of classifiers in Kilivila, for example, I would find the information that a so-called zero-classifier refers to basketfulls of yams that are counted. However, if the data archived do not include recordings of the ritual during which the Trobriand-Islanders fill in their yams-houses and count the basketfulls of yams they carry to store the yams there, I will not find many tokens of this culturally so highly important classifier type. If this is so, the metadata description providing the information with respect to the inventory of classifiers in Kilivila must refer to this fact (for example under the search entry: “nominal classification and culture“).

Having finished these first search procedures, I now check all the other entries the search function had offered me to see whether I can find additional relevant material under those entries. If I have typological interests I would now go to the next language – say Lao – and repeat the procedure outlined above. Because the metadata files provide the relevant information, I will be sure that I collected data in the archive that are really comparable – that is to say, the metadata information allow me to check whether the various researchers‘ understanding and description of these phenomena agree with my own conception of serial verb constructions and systems of nominal classification.

Thus, such an archive would allow me to collect information on, and the actual language data with, serial verb constructions and nominal classification systems for a comparative, typological study within a relative short period of time.

The structure of the archive would provide other users with the same quick and effective information. Anthropologists interested in distribution ceremonies in various cultures, musicologists interested in songs in various cultures and languages, historians interested in “local histories“ in various places, cognitive scientists interested in the conceptualization of space in various languages and cultures – and also the members of the communities whose language and culture is documented in the archive and who are interested in these documents because of various reasons would all find excellent data extremely useful for their different purposes.

So far my brief sketch of the ideal archive in the best of all worlds especially for me as a user interested in anthropological linguistics and typology. But is this sketch not sheer Utopia? And what would the construction of such an archive imply for the technicians structuring and and building up such an archive on the one hand and for the researchers who are submitting their data to this archive and who are describing them for the archive on the other hand?

3 Consequences for technicians and data providers

Establishing such an ideal archive asks for an excellent constructive cooperation between the technicians who are responsible for the computer-technical aspects of the online-

archive and the data providers, the researchers who contribute to the archive, submitting the data they have been collecting during their field research. Moreover, this form of cooperation also means that the data providers are willing to work for, and find, at least a minimally common theoretical basis for the linguistic (and anthropological, and demographical, and historical...) description of the data that are documented in the archive. Any such cooperation requires much energy and time and an incredibly high level of a general good will for cooperation on all sides.

Linguists, anthropologists, historians, demographers and other researchers that are willing to submit their data have to tell the technicians which search entries may be relevant and related during a user's search procedure so that they can construct a search function (like the one briefly described above) which offers links to these other entries relevant for a specific search procedure. A good search function is completely dependent on the expertise provided by the data providers to the technicians.

The technicians – again in close cooperation with the researchers – have to come up with a platform or a framework that makes it as easy and effortless as possible for the data providers to put their data into the archive's data files and to come up with metadata descriptions as adequate and exhaustive as possible.

However, I am sure that even the best of such frameworks cannot diminish the high amount of time, energy and effort these metadata descriptions ask from the data providers. And here we are confronted with the problem whether younger researchers without tenured positions can afford to engage themselves in such a time consuming enterprise. They have to publish the results of their field research in dissertations, in scientific journals and in anthologies, they have to write new project proposals to get further research financed, some of them may have to teach at a university. Will the scientific peer group recognize the work that is invested in submitting data to such an archive in the same way and to the same degree as it acknowledges traditional forms of scientific publications like monographs and journal articles? Thus – can we really take up the responsibility to motivate students and post-docs to work in, and for, such archive projects? Or does the task to provide good data for such an archive address only the already established and the elderly scientists? I am afraid I have no answer to this important question – but maybe these worries will soon be proven out of date by the

incredibly fast technical development... Nevertheless, from my point of view as a linguist I still must say: any really adequate documentation of a language, be it endangered or not, finally has to come up with a grammatical description, a lexicon, and a good corpus of annotated, morpheme-interlinearized and well glossed actual speech data. And every project documenting linguistic data of various languages should keep this ideal goal in mind.

So far I have only talked about the technicians and the researchers that cooperate in establishing an online-archive documenting linguistic data of various languages and cultures. But what about the impact of such an archive on the actual producers of the the data – the speakers and the speech communities of the documented languages?

4. The impact of an online-archive on the actual producers of the data

For the last 20 years I have been doing anthropological-linguistic field research on the Trobriand Islands in Papua New Guinea. So far it is absolutely impossible for the Islanders to get access to the internet on the Trobriands. There is no electricity on the Trobriands and the climatic conditions (especially the high humidity) are not computer friendly at all. Thus, for the time being – and for the near future – I see no chances for the actual producers of the archived data to see and to criticize the archive and the way in which the data they had produced are presented worldwide to an interested public.

However, there are some Trobrianders living in Port Moresby, the national capital of PNG, and in the other big cities like Lae, Madang, Mount Hagen, and Goroka who have this possibility. What would they think about data on their language and culture in such an online archive and what would they like to have documented about their own language and culture in such an archive? This is sheer speculation, of course, but I think it is also necessary for every contributor to such an archive to keep this aspect in mind. I am almost sure that the Trobrianders are not at all interested in archived data elicitation sessions on, for example, serial verb constructions or on spatial reference. It may be funny for some of the Islanders to recognize familiar faces or even relatives – but I assume that they would like to see data in such an archive that document aspects of their

language and culture that they themselves value and would like to see preserved, like for example their oral literature – and there especially those genres that are rapidly vanishing. Among these genres there are the traditional myths, the various magical formulae, and the songs sung during the harvest and mourning rituals which document an archaic variety of the Kilivila language in which the eschatological knowledge of this ethnical group is encoded. Moreover, I can imagine that Trobrianders would like to see how decisions are (- or were -) made in village communities, how people argue in local village meetings and court cases, how distribution ceremonies are organised, how people build their houses and canoes, what kind of fishing techniques are used, how the people garden, what they do when they marry, when a child is born, and when someone dies.

Thus, even if I as a linguist think that it is absolutely interesting for my peer group if I document as many specific data on various linguistic phenomena as possible, I should keep in mind that I have also responsibilities with respect to my informants, the producers of the data I have been collecting and which I want to put into the online-archive. But there is yet another responsibility which I have not mentioned so far – there is the question whether on the basis of my ethics as a field researcher I can justify to make all these data available to the public – and if I do so, whether I respect my consultants' intellectual property rights. I will very briefly deal with these issues in the final section of this paper.

5. Ethical problems and the question of intellectual property

At the very beginning of my field research on the Trobriands in 1982 I tried to explain to my consultants that I want to use the data on their language and culture with which they provided me for publications – and whenever I came back to the Islands I gave the chief of Tauwema, the village in which I have been working, a sample of my publications. And all my consultants and friends in Tauwema – and on the Trobriands – have been appreciating this very much. However, I never published – and will never publish – any information which my consultants labelled as confidential. I am aware of the fact that I am in a privileged position with respect to the relationship I have with my consultants.

There are other situations where researchers have to think more than twice whether they can run the risk to present their data and their informants publicly without anonymizing their consultants. There are societies in which speakers of minority languages risk severe sanctions by a dominating and ruling majority if they speak their officially suppressed languages, and there may be topics in conversation or certain genres (like, for example, magical formulae and rites) that can be used against the speakers that produce them. If this is the case – but the researcher thinks that the data are extremely important and need to be documented, then the technicians responsible for the construction of the online-archive should provide safe and secure means to anonymize the producers of such data. If this is not possible, such data should not be archived at all – or, if they are archived, they must not be available to the public!

As mentioned above, another aspect which is very much important for any (online-) archive is the copyright question. The cultural and linguistic data archived are the intellectual property of the persons who produced them – or of their family, their lineage, or their clan. However, once these data are archived and publicly accessible, this property is no longer personal but public – and with the internet "public" in the broadest sense of the term. What about conflicts that may arise here? I must confess that I have no answer to this crucial question – at least not so far. However, I think this problem has to be solved again and again by every field researcher in every field situation – and I am convinced that it is impossible to come up with comprehensive guidelines for the solution of this problem. So far I have been sticking to my consultants' licence to publish the data with which they have been providing me for the last twenty years – and I am absolutely grateful to my friends and consultants on the Trobriands, especially in Tauwema, for their incredible hospitality, for their patient cooperation, and especially for their unlimited confidence and trust.