Issues in the design, construction and use of Language Resources (LR) for Endangered Languages (ELs)

Monica Ward

School of Computer Applications
Dublin City University, Dublin 9, Ireland
mward@compapp.dcu.ie

Abstract

Growth in the development of Human Language Technologies (HLT) means that it is easier to document and archive languages than has been the case in the past. This is especially important in the Endangered Language (EL) context where it is imperative to document the language while its remaining speakers are still alive. This paper outlines the additional constraints that prevail when documenting languages in the EL context and how Computer Assisted Language Learning (CALL) development can help in language documentation exercises. It also highlights the importance of the management of the Language Resources (LR) once they have been procured, including the need to provide different access rights to the material depending on the EL community requirements. A forward looking, flexible technology is essential to ensure that current LR are not made obsolete by changes in technology and XML technologies offer a suitable platform in this regard. The paper presents a case study of the development of CALL materials for Nawat, an EL of El Salvador and the ensuing language documentation benefits that arose from the project.

1. Introduction

Growth in the development of Human Language Technologies (HLT) means that it is easier to document and archive languages than has been the case in the past. This is especially important in the Endangered Language (EL) context where it is imperative to document the language now while its remaining speakers are still alive. This paper looks at the benefits of language documentation efforts in general. It considers the potentially conflicting goals of linguists, who want to document a language, and the EL community who may wish to develop language teaching material. The development of CALL materials is one possible solution. Language documentation in the EL context is more challenging than the non-EL one. This paper considers the additional constraints that prevail in the EL situation and the extra factors that must be considered when procuring Language Resources (LR) in the EL context.

Once the resources have been obtained, it is important to manage them correctly. This involves using appropriate storage technology and implementing an agreed upon access strategy to the resources, which will vary from community to community and also on the nature of the material. Finally, a case study of the development of CALL materials for Nawat, an EL of El Salvador, is presented. This example demonstrates that using good data management techniques, combined with careful planning can result in a system that goes some way to satisfying the needs of both linguists and the EL community.

2. Human Language Technologies (HLT), Language Resources and Endangered Languages (ELs)

Growth in the development of Human Language Technologies (HLT) means that it is easier to document and archive languages than has been the case in the past. Audio files can be digitally recorded and stored. Written texts can be stored and accessed in various different ways. Hand-written manuscripts or drawings can be scanned and preserved. Technological advances also make it easier to gather language material. Audio recording equipment is light, small and eminently portable. Field trips can be more flexible and spontaneous as the equipment can be carried by one person rather than a team of support staff. They can be less stressful as recordings can be "tidied-up" later if mistakes, gaps or other unwanted items occur. Furthermore, the pressure to get a recording perfect before leaving a field site is decreased as items can be recorded several times as the cost of the raw material (e.g. mini-discs) is usually not significant.

2.1. Language Documentation and Endangered Languages (ELs)

These improvements in HLT hold for most of the world's languages. In the case of the world's Endangered Languages (ELs), there are more than just documentation and archiving possibilities. An Endangered Language is one that is in danger of disappearing, usually because it has been replaced as the language in common use within a community (Unesco, 1993). Most of its speakers are elderly and the language will disappear with the death of the last remaining speaker. There may be potentially conflicting goals between a (field) linguist and the EL community. The linguist is interested in documenting the language before it disappears and would like to record the language as comprehensively as possible. Hale (1992) refers to the onus on linguists to do precisely this and Crystal (2000) remarks on the changed emphasis in the fieldwork under taken by linguists today. Gerdts (1998) refers to the role of the linguist in language revitalisation programmes (see FEL (1998) for other related articles). However, the EL community members may prefer to avail of the linguist's services to develop language teaching material (rather than scholarly documentation) for the community, to enable them to pass on the language to their children. There is a middle ground and that is to use HLT to develop Computer Assisted Language Learning (CALL) materials for these ELs.

Developing CALL materials means that the linguist can document the language while at the same time reusing the material in a CALL program. Granted, this is not as simple as it seems, but with careful planning, it is possible to work on mutually beneficial language documentation and archiving projects. Issues such as the development of a writing system or the selection of one amongst several writing systems and dialectal differences will have to be dealt with but these can all contribute to the comprehensiveness of a language documentation project. CALL programs provide many benefits for EL communities and are especially useful in communities where human teachers of the language are scarce or non-existent. Section 4 illustrates an example of fieldwork with the Pipil community in El Salvador and the possibilities for both language documentation and the development of CALL resources.

2.2. Acquisition of Language Resources in the Endangered Language (EL) Context

While ELs share many of the language documentation issues that arise with non-ELs, they face additional constraints in the design, construction and use of their LR (Maffi, 1999).

2.2.1. Additional Constraints for ELs

The process of acquisition may have to deal with the fact that no standard dialect exists. Dialectal differences are hard to document even for non-ELs (see Milroy el al. (1994) for a dialect framework). If the language has not been previously documented, minimally documented, documented a long time ago or if the linguist is not very familiar with the language, it may be difficult to detect and determine dialectal differences. Furthermore, the dialect/language distinction is never straightforward, especially in the case of lesser documented languages. Bender (1997), for example, observes that in Micronesia a chain of dialectal connections can be established between very closely related languages.

The writing system is a further consideration. If it is the first time a language is being documented, a writing system must be developed or borrowed and modified. Political, pragmatic and cultural issues will contribute to the system chosen. Mülhlhäusler (1996) points out that introducing a writing system can have dramatic effects on the language and Day (1985) cautions that it may actually kill the language Similar decisions must be made in the situation where several different writing systems are in use (e.g. the case of Nawat, see section 4). It may be the case that over time, different linguists have used different writing systems or alphabets for a given language, with minor variations between them. The members of the EL community may not be literate in any language and this can further complicate the selection process. It is imperative that the EL community is involved and that the writing system is explained to them. Obviously, the writing system must be selected with their approval.

2.2.2. Acquisition Environment

Another issue to consider is the environment in which the language acquisition (e.g. audio recording) takes place. Although it may be desirable to record speech in an environment with minimal outside (unwanted) noise, it may not always be feasible to do so. EL speakers may not be accustomed to working in an "office" setting and may feel confined if the recording takes place in an office situation. Furthermore, they may not be used to sitting down during the day and may prefer to stand when doing the recordings. For example, Dr Lemus at the Universidad de Don Bosco (personal communication on the Pipil people of El Salvador) maintains that it was better for his informant to stand while recording as opposed to sitting down. This was because she was likely to fall asleep if she sat down as she normally stood at her market stall during daylight hours.

Each EL community has its own particular characteristics. Some EL communities may be city based, while others may be more rural (e.g. communities in Papua New Guinea). It may be the case that EL community members are unfamiliar with the equipment used in the language documentation process (e.g. audio recorders and computers). This may cause a degree of stress among the informants (which may also be the case for non-EL informants). The field linguist should aim to minimise this stress, perhaps by demonstrating the equipment and letting the informants hear what has been recorded.

2.2.3. Content Verification

When LR are being created, it is important to check and recheck the contents with the speakers. This can be very difficult if the is very limited information available about the language, but even in the cases of relatively well documented languages, case must be taken to ensure that the material has been correctly interpreted. There may be no dictionaries and grammar books available to check the "correctness" of the language being recorded. The speakers may be the only available authority on the language. It is important to write down the meaning of what is being said when still out in the field, especially if the linguist is not very familiar with the language. Intended nuances and even the meaning of the utterance may be forgotten by the time the linguist leaves the field and there may be no backup available.

2.2.4. Speaker and Linguist Comfort

Care must also be taken to ensure that the speakers are comfortable with the contents of the LR that are being documented (and also its future dissemination - see Whalen (2001)). They may feel comfortable with recording only certain types of information (e.g. non-sacred chants) or may wish to re-record items that they feel were not correct or free-flowing the first time. It is important to allow the informants to hear what has been recorded if they wish to do so and to give them the opportunity to re-record if necessary.

If the field linguist has limited time to work with the informants (which may often be the case), it adds to the pressure of the project. The linguist may be exposed to a number of new linguistic features or logistical obstacles that were unanticipated. For example, I was on a field trip to El Salvador in January 2001 when an earthquake of 6.8 on the Richter scale hit the country. In some cases, the community may not perceive these issues to be a problem and may expect the linguist to be able to deal with them. Not all of the above mentioned items will pertain in all situations, but the field linguist in the EL situation is bound to come across some of them and probably encounter some other challenges as well.

3. Managing the Linguistic Resources

Once the LR have been obtained, it is important to ensure that the data is correctly managed. This is doubly important in the EL context as it may be hard (or even impossible) to repeat the LR acquisition process. For example, if the EL community lives in a very remote region it may be very difficult to re-record the material and certainly, if the last remaining speaker dies, it will be impossible to do so. Good management practices must be followed in order to ensure all of the resources are stored correctly and have suitable access rights assigned to them. There are several simple yet effective techniques (e.g. from the field of Software Engineering) that can be adopted. For example, the use of a standard naming convention for audio and text files can help in this regard. If files are correctly named, it will be easier to determine where they should be stored in the overall system and also what type of information they contain. On older systems, there may have been a limit of 8 characters for the file name. However, most modern systems allow longer file names and care must be taken to ensure consistency.

3.1. Obsolescence Avoidance

It is frustrating for linguists to have LR available to them but not the technology with which to access these resources. This can happen when audio recordings were made using now obsolete technologies, or when text files were saved in a now-defunct format. To avoid this problem in the future, LR should be stored in using a flexible technology to help ensure that the LR are accessible for future generations. Although future compatibility cannot be guaranteed, using a technology that enables the resources to be converted from one form to another may extend the life of the LR. XML technologies (XML, 2001) are flexible and ensure that the LR are easy to use and update. The use of XML technologies also offers the option of adapting to future standards in the field of linguistic databases (LEW 2000, LEW 2001).

3.2. LR Access Issues

Some EL communities have particular issues with access to LR. They may be happy to share written material but not spoken material. For cultural reasons, some EL speakers may not wish to share their LR with people from outside the community (Whalen, 2001). Other EL communities (e.g. some indigenous people in Australia) consider it taboo to keep images and audio recordings of someone who has died. Issues of different access rights to different types of data may also arise. For example, EL community members may agree to share LR that pertain to everyday situations but may not want religious or culturally sensitive topics made publicly available. In other situations, access may be read textual information or hear audio materials but not to copy the materials.

All of this implies that the issue of security and access rights has to be an integral part of the LR management system. If the information is being placed on a public forum, such as the Internet, sensitive materials must have stringent password protection. Different levels of security can be arranged, depending on the requirements of the EL community. Implicit in all of this is that the field linguist has obtained agreement from the EL community and the informants as to what restrictions (if any) they wish to place on access to the material. This is not a straightforward task, as Sherzer (reported in Whalen (2001)) points out. An informant may give permission for the material to be used, but may be unaware of the potential of the Internet. Descendants of an informant may claim rights to their ancestor's material and want to withdraw it. It is not obvious as to what should be done in these situations but the field linguist must consult the dissemination possibilities with the informants.

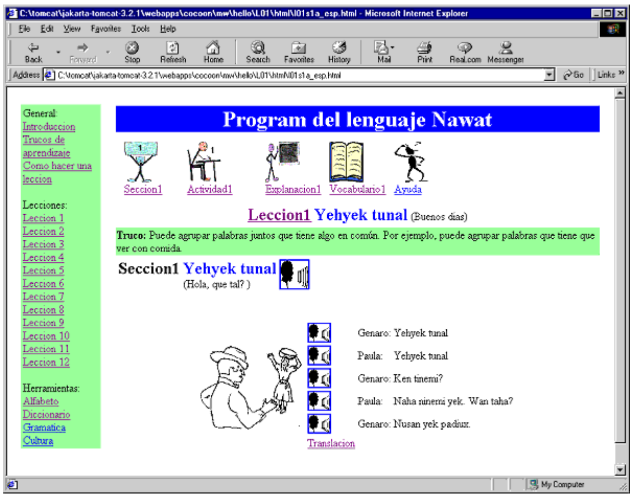
Notwithstanding the access issues raised above, some communities would like to see their LR being widely distributed. One such case is the Pipil people of El Salvador. Their language (Nawat or Pipil) is being replaced by Spanish since the early part of the 20th century and was further blow in 1932 (Byrne, 1992) and subsequent years when indigenous people were persecuted and their language banned. They consider this a way to help preserve their language. With this in mind, it is important to offer different distribution means. This may include different distribution channels including community based and state based channels. Putting the material available on different media (e.g. web, CD and printed documents) also increases the distribution options. Obviously, particular cultural values and economic and social conditions of the EL community members must be considered in the distribution strategy. For example, in the case of the Pipil in El Salvador, most community members do not have access to a computer. Thus, it is vitally important that a printed version of the material be made available to the community.

The use or deployment of the LR of an EL must be considered at the planning stage. How will community members be able to access and use the LR (especially the online resources)? How will other interested parties be able to use and analyse or enhance the data? Obviously, not all uses and users can be anticipated in advance but it is important to offer the LR in different modalities for different types of end user.

4. Nawat Case Study

Thus far, general issues in the design, construction and use of LR for ELs have been discussed. This section illustrates the points from a case study of the development of CALL materials for the Nawat (Pipil, see Campbell (1985)) language of El Salvador.

Nawat is typical of many ELs in that its remaining speakers are elderly. Grimes (2000) reports that there are approximately 20 remaining speakers, while based on my field trips to El Salvador, I estimate that there are slightly more, but still less than 100. Nawat is spoken by the Pipil people in western El Salvador, mainly in Santo Domingo de Guzmán, just north of Sonsonate. Various anthropologists and researchers have worked with the Pipil to document aspects of their language and culture but


C:\tomcat\jakarta-tomcat-3.2.1\webapps\cocoon\mw\hello\L01\html\l01s1a_esp.html - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Real.com Messenger
Address C:\tomcat\jakarta-tomcat-3.2.1\webapps\cocoon\mw\hello\L01\html\l01s1a_esp.html
Go Links
My Computer

General:
Introduccion
Trucos de aprendizaje
Como hacer una leccion

Lecciones:
Leccion 1
Leccion 2
Leccion 3
Leccion 4
Leccion 5
Leccion 6
Leccion 7
Leccion 8
Leccion 9
Leccion 10
Leccion 11
Leccion 12

Herramientas:
Alfabeto
Diccionario
Gramatica
Cultura

Program del lenguaje Nawat

Seccion1 Actividad1 Explanacion1 Vocabulario1 Ayuda

Leccion1 Yehyek tunal (Buenos dias)

Truco: Puede agrupar palabras juntos que tiene algo en común. Por ejemplo, puede agrupar palabras que tiene que ver con comida.

Seccion1 Yehyek tunal
(Hola, que tal?)

Genaro: Yehyek tunal
Paula: Yehyek tunal
Genaro: Ken tinemi?
Paula: Naha ninemi yek. Wan taha?
Genaro: Nusan yek padiux.

Translacion

Figure 1 Sample screen from the Nawat courseware

up till now no one has worked on the development of an online, interactive CALL system for Nawat. The Nawat CALL project was one of the few occasions that the community had a chance to work symbiotically with an outside researcher.

The development of CALL material is a non-trivial task (and a comprehensive summary is outside the scope of this paper, see Levy (1997)). CALL usually requires a multi-disciplinary team that includes a linguist, a pedagogical expert and a technical specialist. In the EL situation, often none of these specialists are available. Thus, use was made of a software template (developed by Ward (2001)) which allows the development of CALL materials for ELs. It takes into account the additional constraints faced by ELs. In the case of Nawat, there were up to 7 different alphabets that had been used to write the language (Campbell, 1985; Lemus, 1997). One informant (informant A) was literate in one alphabet, while the other informant (informant B) was semi-literate (but struggled with written material). Informant A was used to working in an office, so the recording environment presented no difficulties. However, informant B was not used to sitting down in an office and this had to be taken into account during the recording sessions by providing breaks and allowing the informants to chat and walk around if they wanted. The informants were also given the chance to listen to the recordings to make them more comfortable with the process.

Once the information had been gathered, it was processed and prepared off-site. Subsequently, back in the field, the material was revised by the informants to check its correctness and its Spanish translation. The software template stores language data in XML data files and the audio files in mp3 format, while image files are stored in .jpg and .gif formats. Each file (be it text, audio or image) follows a standard naming convention. Text files are organised into lessons and each lesson has three sections. The template automatically generates language lessons, with links to entire conversations or single phrases from each conversation. The template also enables a CD version and a printed version of the materials to be produced.

Figure 1 shows a sample screen from the Nawat language learning courseware. It shows the conversation and the links to the audio files for the entire conversation and each phrase of the conversation. The bar on the left-hand side of the screen provides links to pages for the other lessons, the alphabet, a multi-media dictionary, grammar notes and cultural information. The courseware was developed in both Spanish and English.

In summary, the development of CALL materials for Nawat meant that audio conversations (and their textual equivalents), a guide to the alphabet, an online dictionary and grammar information were procured and placed online for the first time. This is an example of how a CALL project can result in positive outcomes for the competing needs of a linguist (language documentation) and the EL community (teaching materials).

5. Conclusion

Improvements in Human Language Technologies (HLT) mean that it is now easier to document and archive Language Resources (LR). This is especially useful in the context of Endangered Languages (ELs), where the limited numbers of remaining speakers means that LR must be procured and stored now, rather than at some unspecified date in the future. One interesting option to marry the potentially conflicting goals of linguists and EL communities is to develop Computer Assisted Language Learning (CALL) materials for the ELs.

ELs present additional constraints on LR documentation efforts. These include lack of previous knowledge or documentation on the language, the lack of a writing system or the proliferation of different writing systems, the acquisition environment and the familiarity of the informants with the technology being used. Another issue is the need to verify the material with the informants after it has been processed and transferred to a different medium (e.g. on the computer). An important issue in the EL context is that informants may wish to place access restrictions on the LR (e.g. to restrict access to members of the community or to common rather than sacred texts).

Given that many ELs will have disappeared in the near future, it is imperative that the LR are managed correctly and with a view to future use. Thus, storage technologies that are flexible and can adapt to changing formats (e.g. XML technologies) should be used to avoid obsolescence of LR. It is important that the LR be produced in different formats to cater for different audiences and it must be remembered that EL communities may not have computer access and thus the materials should also be made available in a printed format. Finally, a case study is presented of the development of CALL materials (and inherent language documentation) for Nawat (or Pipil), an EL of El Salvador.

6. References

Bender, B., 1971. Micronesian Languages. Current Trends in Linguistics 8: 426-65.
Bland,G., 1992. "Assessing the transition to democracy." Tulchin, Joseph S. with Gary Bland, eds. 1992. *Is there*

a transition to democracy in El Salvador? Boulder: Westview Press (Woodrow Wilson Center current studies on Latin America)

Byrne, Hugh. 1996. *El Salvador's civil war: a study of revolution*. Boulder: Lynne Rienner Publishers.

Campbell, L., 1985. *The Pipil language of El Salvador*. Berlin: Mouton.

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

Day, R., 1985. The Ultimate Inequality: Linguistic Genocide. In Wolfson and Manes (Eds). *Language of Inequality*, Berlin: Mouton

Gerdts, D. B., 1998. The Linguist in Language Revitalization Programmes. In: *What Role for the Specialist?* Foundation for Endangered Languages II Proceedings (Edinburgh 1998)

Grimes, B., 2000. *Ethnologue: Languages of the world*. SIL International. Available at: http://www.ethnologue.com/web.asp [Accessed 16 October 2001]

Hale, K., 1992. On Endangered Languages and the safeguarding of diversity. *Language*, V68. No 1.

Lemus, J., 1997. Alfabeto pipil: una propuesta. *In*: *Estudios Lingüísticos*. El Salvador: CONCULTURA.

Levy, M., 1997. *Computer-Assisted Language Learning (Context and Conceptualization)*. Oxford: Oxford University Press.

LEW, 2000. *Linguistic Exploration Workshop: Web-Based language Documentation and Description*. Available at: http://www.ldc.upenn.edu/exploration/expl2000 [Accessed 20 November, 2001]

LEW, 2001. *Linguistic Exploration Workshop: Linguistic Databases.* Available at: http://www.ldc.upenn.edu/annotation/database/ [Accessed 20 November, 2001]

Maffi, L., 1999. Report on Language Maintenance and Death: reports from the field and strategies for the new millennium, Workshop, University of Illinois, 1999. Available at: http://www.terralingua.org/uillinois.html
[Accessed 20 November, 2001]

Milroy, J., Milroy, L., Hartly, S. and Walshaw, D., 1994. Glottal Stops and Variation and Change in British English. *Language Variation and Change* 6:327-357

Mühlhäusler, P., 1996. *Linguistic Ecology*. London: Routledge.

Unesco, 1993. *Unesco Red book on Endangered Languages*. Available at : http://www.helsinki.fi/~tasalmin/nasia_index.html [Accessed 27 March, 2002]

Ward, M., 2001. Nawat courseware. Available at:
http://www.compapp.dcu.ie/~mward/nawat.html
[Accessed 31 October 2001]

XML, 2001. World Wide Web Consortium. Available at: http://www.w3.org [Accessed 27March, 2002]

Whalen, D., 2001. Report on the SALSA Special Colloquium on Archiving Language Materials in Web-Accessible Databases: Ethical Challenges, Sunday, 22 April, 2001. Available at: http://www.ling.yale.edu/~elf/ethics.html [Accessed 28 March, 2002]