THE MULTIMEDIA DOCUMENTATION OF ENDANGERED AND MINORITY LANGUAGES

A thesis presented in partial fulfilment
of the requirements for the degree of

Master of Philosophy
in
Linguistics

at Massey University, Palmerston North,
New Zealand

Robert Graham Petterson

2002

ABSTRACT

This thesis examines the impending loss of linguistic diversity in the world and advocates a change in emphasis in linguistic research towards the documentation of minority and endangered languages. Various models for documentation are examined, along with some of the ethical issues involved in linguistic research amongst small groups, and a new model is proposed. The new model is centred around the collection of a wide variety of high-quality data, but includes the collection of other related materials that will be of particular use and interest to the ethnic community. The collected data and other materials are then structured as an internet-ready multimedia documentation designed for use by the ethnic community as primary audience, while still catering for the needs of linguistic researchers worldwide. A pilot project is carried out using the model.

ACKNOWLEDGEMENTS

I particularly wish to thank the following: my wife, Debbie, for forgiving me whenever I woke her up coming to bed at 3 o'clock in the morning; my school friend John MacLean for stirring and annoying me when I was learning to speak Maori in the 1970s by insisting that it was a dying language; Auni, Kenau, Itupi, Makiru and their fellow villagers for delighting in teaching my family and me to speak the Rumu language; Minoru Kasuya, Ute Walker, Grant Klinkum and other members of the Research Committee of the International Pacific College for showing an interest and approving time and financial support to pursue this study; Katsuya Idemaru for advice on technical matters; Dr John Newman of Massey University for pointing out some interesting and relevant sources of information, for keeping me from digressing too far down some other highly interesting but irrelevant leads, and for carefully reading through many drafts; my family for patiently listening to a "read through" of the less technical parts; and the Creator, who made his creation such an interesting place so full of variety, and who, in spite of humankind's tendency to wantonly obliterate large pieces of it, has the redemption of it all in his plan, and who has given me the desire to work in support of some of its small and neglected parts.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

In June 2001, as this thesis was being written, a headline appeared on the CNN web site: "Half of the world's 6,800 languages could die by 2100" (Associated Press, 2001). Such warnings are not new, having been made at least since 1992 (e.g. Krauss, 1992; Wurm, 1996), but, as Crystal pointed out in his aptly named book, *Language Death*, "we are at a critical point in human linguistic history, and most people don't know" (Crystal, 2000, p. ix). Perhaps, as a result of that headline, many more people have become aware of what can only be termed a crisis. If the prediction comes true, then as languages disappear, a rich source of knowledge about a significant side of human nature is about to be severely and irreversibly reduced, and sources of knowledge about our world expressed in these languages may well be lost or degraded. Moreover the rate at which languages are disappearing, about one language a fortnight,[1] is on a scale no less worrying than the reported rate of extinction for living species.[2] This thesis is, in part, a considered reaction to this impending crisis.

1.1 Minority and endangered languages

The focus of this thesis is on properly documenting minority and endangered languages before they disappear. An **endangered language** has been defined as "any language of a community which is not learned any more by children, or at least by a large part of the children of that community" (Wurm, 1996, p. 1). Such a language is very likely to become extinct within about 50 years, when the parents of those children die.

A **minority language** is a language spoken by a minority group, which has been defined as "a group which is smaller in number than the rest of the population of a State, whose members have ethnic, religious or linguistic features different from those of the rest of the population, and are guided, if only implicitly, by the will to safeguard their culture, traditions, religion or language" (Skutnabb-Kangas, 1997, para. 1). The reference

---

[1] Crystal (2000, p. 19) supports this estimate, with a reported potential loss in diversity of from 50% to 90% within 100 years.

[2] According to von Bieberstein Koch-Weser (December 22, 1999, para. 3) "it is now estimated that more than 20,000 species are lost every year, and that this loss is between 1,000 and 10,000 times greater than it would naturally be." A quarter of all mammal and reptile species, and 30% of all fish species are now endangered or vulnerable (IUCN, September 28, 2000, paras. 8-14).

to "a State" in this definition is debatable, because it assumes a political organisation that may be resented by some minority people groups, such as the Kurds, who are spread over the borders of more than one state; on the other hand I think it is useful to include the concept of statehood because political, economic and educational factors that can profoundly affect the status of minority languages are largely under control of the state.

Minority languages are also potentially endangered languages, because, as noted in the preamble to the manifesto of the Foundation for Endangered Languages: "a small community, isolated or bilingual, may continue for centuries to speak a unique language, while in another place a populous language may for social or political reasons die out in little more than a generation" (Baker, 2001, sec. 1.1). The manifesto also notes that 52% of the world's languages are spoken by fewer than 10,000 people, and most of these would presumably be minority languages. However, numbers can be deceptive: Irish is a minority language spoken by as many as 790,000 people (Lyovin, 1997, p. 47), but the fact is that this number is in decline and the language that was once spoken throughout Ireland is now spoken on a daily basis by only 120,000 people in rural areas in the west of the country. On the other hand, Rumu - a language studied by the author - is a minority language spoken by a mere 800 people in Papua New Guinea, but at this stage it is still being learnt by nearly all children of the community and so can still be considered viable. On the whole, the situation for languages with small numbers is more precarious, because a change in economic or social circumstances could affect the whole group and the status of the language could change within a generation. For that reason, any language spoken by fewer than 1,000 people should certainly be considered potentially endangered and given special attention.

Endangered languages are found in all continents. Some examples are the Celtic languages of Europe, the Turkic and Mongolian languages of western and southern Siberia, the Khoisan languages of southern Africa, the Aboriginal languages of Australia, most of the indigenous languages of Canada and the United States, and a large number in South America (Wurm, 1996, pp. 18-24).

In this thesis I shall also use the term **language community** to refer to a minority or endangered language group, and **ethnic community** to refer to such a group *or their descendants* if the language becomes extinct.

1.2 Documentation

The OED (1989) defines **documentation** as "the accumulation, classification and dissemination of information". When something is both valuable and fragile it behooves us to document it thoroughly so that if it is degraded or lost, a study of the documentation can provide at least an understanding and an appreciation of it, and may even put its restoration in the realm of possibilities.

For example, a thorough and useful documentation of a Stradivarius violin would be surely much more than a photograph and a valuation in a museum catalogue. It would be an accumulation of information on the instrument organised under various headings, such as its manufacture, history, physical and acoustic properties, and, presumably, recordings of it being played by competent musicians that demonstrate the full gamut of its sound. If this documentation is suitably published it will be possible for others not just to recognise it and admire it as an example of fine craftsmanship, but also to appreciate it as an instrument for the creation of beautiful music, and even to restore it or craft instruments of a similar nature for the benefit of mankind. In the same way, if a language is lost, a previously completed documentation of it will be the only way that the people of future generations will be able to seriously appreciate it, and the only hope they will have of reviving it.[3]

Documentation of some objects of interest is not straight-forward, because the objects concerned involve a complex interaction of a number of systems. For example, an archeological site being documented by a paleontologist may have intersecting layers with relics of several succeeding civilisations; a social revolution being documented by a historian may have economic, social, political and military factors contributing to it; the expertise of an engineer being encoded in an expert system involves the accumulated knowledge and experience that has come from dealing with many problems and projects;

[3] There have been attempts at reviving Cornish and Miami from documentation alone (see Hinton, 2001, p. 416).

similarly, the communicative practices of a language community to be documented by a linguist are manifested through the interaction of sub-systems of the minds and bodies of its members, and enmeshed in a complex web of socio-cultural relationships between them.

Even if the systems involved are not yet well understood, and this is important to the present thesis, a documentation carried out as *comprehensively* as possible will enable others to study it later and reach at least some useful level of understanding of those systems. Therefore it is important to establish a sound methodology that will enable such a comprehensive language documentation to be carried out.

1.3 Language Documentation

Language in its fullest sense involves a complex interaction of a number of human systems: anatomical, psychological, social and spiritual. There are a number of fields of study where the connection between language and these systems is explored, and the theories for these fields are now well developed, including phonetics, phonology, syntax, lexicon, semantics, cognitive science, language teaching and acquistion, literature, socio-linguistics and anthropology. Historically the focus of language documentation has been on the publication of a *description* of its elements and structure, but a vocabulary and grammar alone are only as useful to the appreciation of a language as a list of components and a blueprint are for a Stradivarius violin (to continue the analogy above). A complete language documentation, in the context of this thesis, should therefore entail "the accumulation, recording and dissemination of information" (OED) that each of these fields of language-related study can make use of.

This expanded view of language documentation is shared by an increasing number of linguists today. Mithun (1998, pp. 190-191), a specialist in North American languages, points out that a maximally effective documentation must record all kinds of communication situations from daily life, not just a grammar and vocabulary - what they say, not just how they say it. Himmelmann, in what is perhaps a seminal article on language documentation, differentiates documentation from description. He writes:

> The aim of a **language documentation** is to provide a comprehensive record of the linguistic practices characteristic of a given speech community.

> Linguistic practices and traditions are manifest in two ways: (1) the observable linguistic *behavior*, manifest in everyday interaction between members of the speech community, and (2) the native speakers' *metalinguistic knowledge*, manifest in their ability to provide interpretations and systematizations for linguistic units and events. This definition of the aim of a language documentation differs fundamentally from the aim of language descriptions: a **language description** aims at the record of A LANGUAGE, with "language" being understood as a system of abstract elements, constructions, and rules that constitute the invariant underlying structure of the utterances observable in a speech community. (Himmelmann, 1998, p. 166; various emphases are Himmelmann's)

The inclusion of native-speaker metalinguistic knowledge in a documentation is important because there has been a resentment towards the traditional Western academic approach to research among some non-Western tribal groups. Jahnke (1998), for example, reports that there is a mistrust of Western methods amongst Maori, because independent Western researchers interpret reality in a different way, and misrepresent Maori as they see themselves. She adds that, in fact, many have not found truth or new knowledge - they have missed the point. A case of representation more sympathetic to the language community is the elegant classification of Maori parts of speech achieved by the Maori scholar, Biggs (1973), which appears to fit the rules of syntax of Maori much more satisfactorily than the usual verb-noun-adjective-adverb classification that one often encounters in analyses by European linguists.[4]

In accordance with Mithun (1998), Himmelmann (1998) and Jahnke (1998), we can conclude that a full documentation of a language, then, will be not just about the language as a decontextualised system, but about the language in the context of its speakers and their culture. It will include not just example sentences to illustrate a grammar point, or word-lists to illustrate phonology or semantic fields. It will contain a large number of recordings of discourses in the language on matters of importance to its speakers, together with demonstrations of the language in use by the community of speakers in as many and various contexts as can be recorded, whether at play or at work, negotiating or

[4] Biggs complained that "in the traditional view of Maori grammar, there are a large number of parts of speech (at least eight), and a given word may, at times, be any one of several parts of speech." In his analysis "a much simplified system of classification results. All words are divided into two classes, bases and particles...Bases divide into five classes...determined by the constructions into which [they] can enter. There are no overlapping classes" Biggs (1973, p. 51).

arguing, entertaining or teaching, chatting or carrying out formal rituals. It will also contain, wherever possible, commentaries on and analyses of the data by native speakers.

1.4 Multimedia Language Documentation

The communication events which are to comprise the bulk of the proposed documentation of a language are, as Landow and Delany (2001, p. 212) have pointed out, "complex physical and intellectual experience[s], engaging all five senses." The traditional process of documentation in text form flattens these five dimensions into a one-dimensional "schematic visual code". In order to preserve as many of these dimensions as possible I therefore propose that modern language documentation be comprised of multimedia data. The term **multimedia**, of course, refers to the integration of text, sound, and still or moving images, especially using digital technology. Such data will be useful, not only for linguistic research, but also for studies in many other areas connected with language mentioned at the beginning of the last section.

Undoubtedly, written texts have been and will continue to be the most important part of documentation. Since writing systems developed in various places in the world as long ago as 3000 BC, especially in area between the Mediterranean and China, the written records uncovered by archaeologists have provided the only documentation we have of now extinct languages once spoken in Crete and Sumeria. The invention of phonemic alphabets in particular, e.g. by the Greeks about 500 B.C., has been of special benefit to our understanding of a number of extinct languages, including their sound systems. Being easier to learn than logographic writing, phonemic writing enabled more people to write material that can now be uncovered and analysed. Also, as understanding of the phonemic writing principle spread from language to language, we now have invaluable written records of many European and Asian languages and insights into their cultures that date back many centuries. The wealth of texts we have in languages such as Greek, Hebrew, Latin and Sanskrit, together with some remarkable grammatical descriptions (e.g. Panini on Sanskrit about the 6th century B.C.) have meant that these languages remain objects of study and learning by scholars, and, in the case of Hebrew, for example, have even been revived.

Studying a language from a book is a scholarly pastime, however, and unlikely to be so interesting to the *other* potential users of the documentation: speakers and descendants of speakers of the documented languages. In the last few decades electronic multimedia tools have been invented that are capable of being used even in thatch-roofed houses in a rain forest of Papua New Guinea or the Amazon. These inventions have changed the potential audience for language documentation.

The tools relevant to multimedia documentation are:

1. *Sound recording devices.* The invention of the magnetic tape recorder, especially the audio cassette tape recorder, has provided not only the means to easily record spoken text, but also the means of easily distributing such recordings amongst the people of the language community.
2. *The electronic digital computer.* The text-processing capabilities and cheap unlimited storage of modern computers have revolutionised the compilation of large corpora of literature and automated many of the tedious tasks involved in extracting data from them.
3. *The video camera.* Although the movie camera has been available since long before the computer, it has not been a significant language documentation tool. Costs and complex technologies have meant that text and audio tape remained the media of choice. The invention of the cheap hand-held video cassette recorder, however, has made video recording of language data much easier to achieve. Moreover, video players are becoming more common even in language communities where the technology has only recently been "stone age", so documentation on video can be distributed back to those communities.
4. *Digital multimedia technologies.* The World Wide Web and the newly popular compact disk writer are now making multimedia documentation of language not just an attractive possibility, but a burgeoning phenomenon.[5] This can be attributed to the technical feasibility of easily combining video, sound and text in the same document, and the low costs of widely publishing such documents.

[5] See Buszard-Welcher (2001) for a survey of endangered language Web sites.

It is ironic that the inventions that have made multimedia language documentation possible have come out of revolutions in communications that have perhaps precipitated the very crises that now make language documentation urgently necessary. Languages that have had writing systems developed for them and have developed a body of literature clearly make more desirable languages for education than those that do not have these advantages; the printing press has made both literature and the languages of literature even more widespread; and the electronic revolution of the last century has done much to enable the process of globalisation that has affected minority language groups all over the world through the dominance of world languages, especially English. Nevertheless, the electronic revolution has provided exciting tools that we as language documentalists need to exploit before much of value is lost.

I propose a documentation that utilises these electronic multimedia tools in a framework built on the general principles of language documentation discussed in section 1.3. Such a documentation would entail the following:

1. the accumulation of examples of linguistic behaviour (otherwise termed "texts" and "discourses") of all genres recorded in text, and, where appropriate, audio or video;
2. the recording of commentaries by members of the language community on these discourses, and translations of them in a suitable "language of documentation" (such as English, Spanish or Indonesian);
3. the accumulation of other data on the language and language community, such as scanned field notes of linguists and anthropologists, drawings and diagrams of linguistic and cultural interest, and databases and descriptions of the language;
4. the organisation of this data so that it can be accessed easily by those interested;
5. the dissemination of this organised data to those interested.

1.5 Stakeholders and ethics

I have already suggested that there are two main groups of people who can be expected to have an interest in the documentation of a minority or endangered language: language researchers on the one hand, and the speakers or descendants of speakers of the language on the other. The former may be interested in the documentation as a source of

data for their research; the latter may be interested because the language is their heritage. In this thesis both of these groups will be kept in mind. Other stakeholders that must be considered are the documenting linguist and the members of the language community who contribute directly to it.

Traditionally, a linguist researching a minority language publishes materials for other linguists to read, such as papers on aspects of grammar, phonology or semantics. Because of the broader scope of data collected in a comprehensive multimedia documentation such as has been discussed in sections 1.3 and 1.4, a much wider academic audience will find an interest in it - linguists, anthropologists, sociologists, historians and so on. There are already a number of archives on the World Wide Web that focus on the needs of the researcher. For example, the Archive of the Indigenous Languages of Latin America (AILLA) is developing a Web-accessible database of audio and textual materials, the goal of which is "that the breadth of materials held by AILLA will facilitate new research and understanding of the richness of indigenous Latin American language in use" (Sherzer, n.d.).

The expectations of the language community as an audience for the documentation are very different. As in any community, minority or majority, most of the people will not be interested in the technical aspects of their language, but they will be interested in what other people have to say about them and their words, and they will expect an expert to help them with their perceived linguistic needs. Crowley (1999b, p. 1) points out that in places like Vanuatu, the Solomons and Papua New Guinea, dictionaries, rather than grammars, are seen as important outputs that the researched language community expects in return for their cooperation. Mithun (1998, p. 185) indicates that where language curricula are being developed, a "strong documentation" of the language is essential, and that writing, rather than just audio or video, should be central to this documentation for future generations who do not control the language. As for content, she recommends that exemplary material of a wide variety should be chosen so that a good appreciation for the language at its very best can be obtained.

It is clear that there are tensions between what highly-trained linguists expects to achieve for themselves in terms of their own aspirations, what the academic community

expects them to produce, and what researched communities expect them to achieve. Crowley reminds linguists of their ethical obligations to researched communities:

> ...the days should be long gone when linguists could arrive in a community, gather and analyse data, and then publish, without providing some means for the dissemination of the results of the research back into the community in a form that is both useful to, and valued by, the community. (Crowley, 1999b, p. 1)

This kind of obligation has now been encoded as official policy in some countries. For example, one of the guiding principles in the Vanuatu Cultural Research Policy states:

> The knowledge and dissemination of the kastom[6] and history of Vanuatu should be directed firstly, if the subject is a particular culture to the people of that culture, secondly to ni-Vanuatu and lastly to non-citizens. (Hiscock, October 28, 1997, sec. 2.v)

There is also a tension between what is traditionally expected of field linguists by the academic community of linguists on the one hand, and what is now needed for the advancement of the science of linguistics on the other. Grammars or papers that use isolated snippets of material from a language as evidence to support a theoretical insight have received more "brownie points" (as Crowley puts it)[7] from the universities that provide the funds to do the research and the opportunities for advancement based on such demonstrations of intellectual prowess.

In this time of crisis, however, I believe that we need to take an approach to research that focuses on documentation rather than theory. This is an approach that was broached already in 1991 when the Linguistic Society of America (LSA) resolved that it should "respond to [the] situation [of widespread language endangerment] by encouraging documentation, study and measures in support of obsolescent and threatened languages..." (Linguistic Society of America, March 1991, cited in Craig, 1992, p. 4). Since then, the attitude of linguists has, I think, been changing, even though, in my own recent experience, a proposal to compile a dictionary or a full documentation

---

[6] *Kastom*, as used in the original document, is a term in Bislama which refers to the "traditional political, social, religious and economic structures, and their associated practices, systems of knowledge and material items" (Hiscock, October 28, 1997, sec. 1).

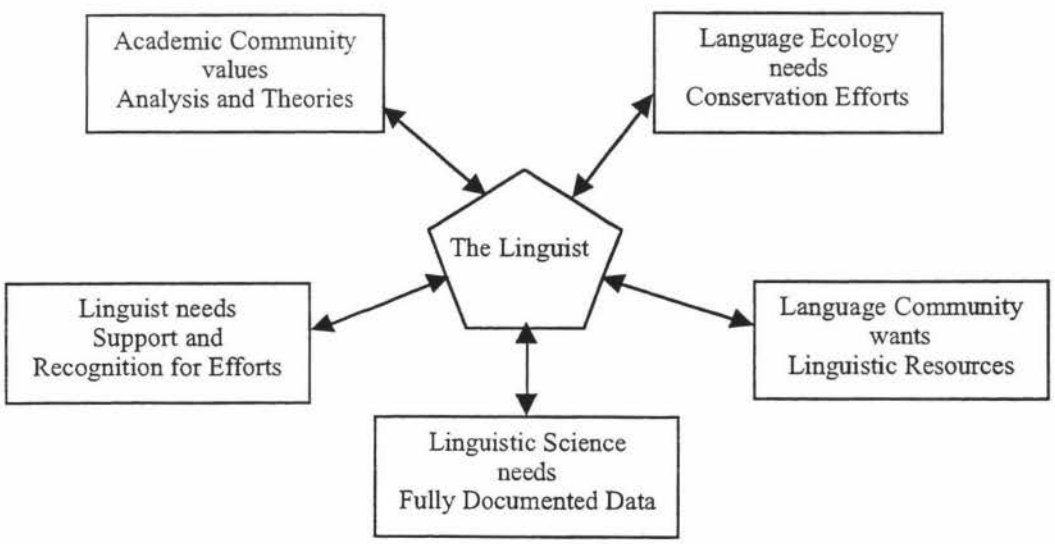
[7] Crowley (1999b, p. 5-6) discusses a negative reaction by a member of the academic community to a published work that was oriented towards a language community.

of a minority language has been seen as unworthy of academic credit because of the pressure to advance theory over data. But without careful documentation, the world's linguistic heritage may be lost to us with implications of a future of stunted growth in linguistic theory due to a dearth of data!

The LSA resolution raises another issue - the role of linguists in the conservation of the linguistic ecology. This issue has been raised again in more recent years. Mühlhäusler, for example, argues for an engagement in action research that gives the linguist a role in language maintenance and the preservation of linguistic diversity rather than the mere preservation of linguistic data that mirrors the well-outmoded "pickled squirrel approach to wildlife preservation" (Mühlhäusler, 1998, p. 222). The need for cultural and linguistic survival of the researched community can then become another tension affecting the linguist. The conservationist approach has political implications, however, and in this thesis I shall restrict my discussion of this issue to the support that documentation can offer the conservationist cause.

Finally there may be a tension between the academic aspirations of the linguist on the one hand, and the political needs of a struggling ethnic community on the other, a community who may want help with the conservation of their culture and language. These tensions that I have discussed can be diagrammed as in Figure 1.1:

Figure 1.1. Tensions affecting the linguist



The tensions can be eased as the academic community fully recognises the value of the linguist's support of the rights of the language community while carrying out his or her research. One way to do this is to recognise the Universal Declaration of Linguistic Rights (UDLR) which was formulated at the World Conference of Linguistic Rights in Barcelona in 1996.[8] The UDLR takes the stance that people groups speaking endangered languages have rights that need to be granted to them, because they are often marginalised and disadvantaged. These rights include the right to use their language (article 13), to receive an education in it (article 29), and to have access to local communications media (section iv). Of relevance to this thesis are articles 30, 43 and 46. Article 30 states:

> The language and culture of all language communities must be the subject of study and research at university level.

Interestingly, this takes more the form of a demand rather than a right, perhaps echoing the consensual view of the formulators of this declaration that the need to study these languages is urgent, and none should be neglected.

Articles 43 and 46 state, respectively:

> All language communities are entitled to access to the works produced in their language.
>
> All language communities have the right to preserve their linguistic and cultural heritage, including its material manifestations, such as collections of documents, works of art and architecture, historic monuments and inscriptions in their own language.

A properly conceived multimedia language documentation can meet many of the goals of these rights (excepting of course the preservation of art objects, architecture and monuments).

---

[8] The convocation of this World Conference was an initiative of the Committee for Translation and Linguistic Rights (International PEN) and CIEMEN (International Escarré Centre for the Ethnical Minorities and the Nations), with the moral and technical support of UNESCO, and was attended by representatives of 66 NGOs, 41 PEN Centres and 41 experts in linguistic legislations from all the world (CIEMEN, 1996).

To support these rights and resolve the tensions between the researcher and other interests, there needs to be an urgent and substantial weighting of the scales of intellectual effort and research funding in favour of the traditionally more despised work of language documentation.

There are other ethical issues that need to be considered too, for linguistic research, like other forms of social research, is not neutral; the research process can have an impact on the community - an impact that needs to be controlled.[9] Jahnke (1998) reported a strong desire by Maori to have control over the research process, so that the research properly represents and respects them, and leads to an enhanced community and not just an enhanced reputation for the researcher. For example, some recorded discourses may contain knowledge that is esoteric, and should only be entrusted to elders; some data may contain ritual wording that contributors are anxious to have recorded absolutely correctly; some data may contain private or sensitive knowledge that should not be published at all. In order to respect these concerns, a continual consultative process in research is necessary.

Many of the communities speaking languages in need of documentation today are less sophisticated in their awareness of the world outside their area, and in the common practice of publishing research findings for all to see. Documentalists may find themselves in the role of educators as they explain the purposes for their being there, the potential audiences for the documentation, and the advantages that publishing may have in promoting a language community.[10] At the same time the documentalist must strive to respect issues of privacy and safety by seeking informed consent for the dissemination of any recorded discourses, for, as Himmelmann (1998, p. 172) notes, "the interests and rights of contributors and the speech community should take precedence over scientific interests."

---

[9] Cameron, Frazer, Harvey, Rampton and Richardson (1993) discuss the other non-research roles of the researcher that affect the researched community, and argue that these should be exploited, rather than denied, for the good of the researched community, through negotiation and consultation with them.

[10] The idea of promotion for cultural advancement and recognition of rights needs to be carefully differentiated from the idea of linguistic research as a potential money-making venture, either for the documentalist or the language community - which, in my experience, it definitely is not.

Some discourses may have special restrictions. Some Rumu traditional stories, for example, are not told in the presence of those who have not undergone initiation. Hale (1971, p. 472) deals with this kind of issue in one publication by explaining the restrictions in the prologue to a discussion of the Warlbiri language, so that the reader will know that he or she should refrain from discussing the material with uninitiated Warlbiri men or with Warlbiri women and children. Other methods of restricting access may be to keep the published copies of certain texts in a "tabu-room" of a cultural archive or library.

In conclusion, a research work of documentation done in consultation with, and in a spirit of service to, the language community, with them as the primary intended audience will have several, perhaps unexpected, benefits to both academic and researched communities:

1. The output of the research will be readable by the community at large, and be useful for educational and other cultural efforts of the language community's own initiative.
2. It will also be a mine of data for researchers of many disciplines (depending on the range of data included in the documentation).
3. Being largely independent of the ebb and flow of transient theoretical fashions, it will be a more enduring and valued work.
4. It will allow the development of the theory and methodology for language documentation. Once matured, they will render documentation an even more useful and prestigious form of scholarship.
5. It may cause the language community to take on a sense of pride in their language which may make all the difference to the survival or revival of the language and culture.

1.6 The scope of this study

The main thrust of this study is to consider what a thorough documentation of a language can and ideally should consist of, now that we have a powerful new set of tools with which to do it. Firstly, I examine the nature of the language crisis in chapter 2;

then, in chapter 3, I discuss the different types of data that need to be accumulated; in chapter 4, I survey the options for formatting the data, and in chapter 5, for structuring the data; in chapter 6, I propose an instrument for the documentation of languages and illustrate it with examples from a pilot documentation project for one minority language in chapter 7. A concluding chapter, chapter 8, summarises the results of this thesis.

The thesis is supplemented with two appendices: Appendix A gives information about a selection of online organisations concerned with endangered languages and multimedia language documentation; Appendix B gives examples of various kinds of multimedia language documentation that can be found on the World Wide Web.

Because this thesis makes much reference to fast-changing technology and ideas in linguistics that are being discussed and promulgated over the internet as I write, there is a need to rely heavily on the World Wide Web for up-to-date information, and this is reflected in the large number of online references.

A compact disk (CD) is included with this thesis that includes not only the pilot documentation project, but also the reference and appendix sections with "live" links.

Bold type will be used in this thesis for key terms introduced for the first time.

2. THE NEED FOR LANGUAGE DOCUMENTATION

2.1 The current state of the languages of the world

The fourteenth edition of the Ethnologue (Grimes, 2000), a recognised source of information on the languages of the world, purports to have records for 6,809 human languages, including sign languages and languages considered extinct. The exact number of *living* languages in the world today is very difficult to determine precisely, for several reasons. One reason is the difficulties in deciding whether two related language varieties are really distinct languages or dialects of the same language. In many places in the world there are dialect chains where neighbouring dialects are mutually intelligible, but where dialects at either end of the chain are not. Such chains are found, for example, in the Germanic and Romance language areas of Europe.

A second reason is the confusion caused by the multiple names that many languages have. For example, the Rumu language, which features in the Sample Documentation of chapter 7, is also known as Kairi, as are two other related but mutually unintelligible neighbouring languages; Serbian, Croatian and Bosnian are essentially the same language, formerly called Serbo-Croat, but given separate names for socio-political reasons (Crystal, 2000, p. 8).

A third reason is that there are difficulties in determining whether a language known to have only a small number of speakers is still alive, or whether it has already disappeared with the death or dispersion of those last speakers. Other reasons are that there may still be some undiscovered languages, and new languages may even come into being. An example of the latter is the development of Nicaraguan Sign Language since 1979 (SBE Science Nuggets, n.d.).

In short, much language survey work would need to be done to determine precisely the number and status of languages still being spoken, but, in the absence of such survey work,[11] there is a common consensus amongst linguists that the number of living

[11] Crystal (2000, p. 5) observes that over 3,000 languages listed in the Ethnologue (13th edition) have the appended comment - 'Survey needed'.

languages in the world today is about 6,000 (e.g. Crystal, 2000, p. 4; Crystal, 1997, p. 287; Krauss, 1992; Woodbury, 1993, sec. 0; Wurm, 1996, p. 1).

As for the status of these 6,000 languages now being spoken, it has already been pointed out in the first chapter that 3,000 languages, that is *half* of the world's languages are reported to be endangered or dying (Wurm, 1996, p. 1). This figure would presumably include some 417 languages listed in the fourteenth edition of the Ethnologue as being "nearly extinct". Hale (1998, p. 192) goes further and predicts that not only will all of those 3,000 languages perish in the next century, but there are 2,400 more languages that will come close to extinction. If we accept these figures, this means that we can expect that by about 2200 AD there may be only as few as 600 languages still being spoken - just 10% of the number being spoken today.

Since it is a Papuan language that I use to illustrate language documentation in this thesis, it is appropriate to say something about the situation in the Pacific region, for it is in this region that some linguists are predicting the heaviest losses. At present, the Australia/Indonesia/Pacific region, with 2,000 languages spoken by a relatively small population, is the most linguistically diverse region in the world. In Papua New Guinea, for example, over 800 languages are spoken by some 5 million people, i.e. an average of about 6,000 people per language. Vanuatu has an even more complex situation. With a population of nearly 200,000 and just over 100 languages, there is an average of only about 2,000 people per language.

Dixon (1991, p. 253) has warned that, in spite of the rich diversity of the Australia/Indonesia/Pacific region, only 200 languages can be predicted to survive in the long term.[12] Certainly the situation of the Aboriginal languages of Australia appears to lend weight to this prediction, for, according to Crystal (1997, p. 326), in the 18th century there may have been 500 languages spoken by 300,000 people. At an average of only 600 people per language, this would have been a situation of even finer

[12] This estimate is controversial, and Crowley (1999a) disputes a similar claim made in Dixon (1997), saying that vital languages such as Samoan cannot be predicted to suffer the same fate as moribund Australian languages such as Dyirbal. Mühlhäusler (1996), too, has made claims concerning the loss of language diversity in the Pacific, which Siegel (1997, p. 227) has criticised, claiming that the situation in Australia cannot be compared with the vast majority of Pacific Islands languages which are "still alive and well".

granularity than Vanuatu. Nevertheless, today 90% of Australian languages are **moribund**, i.e. on the verge of extinction, being spoken by only a few elderly people (Crystal, 2000, p. 87).

In this chapter I will argue that the impending loss of language diversity is a loss for humanity which demands an urgent response. I will consider various possible responses, including documentation, and show how the urgency of the current language situation must be taken into account when determining the methodology and shape of the documentation proposed in this thesis.

2.2 Language death

When the last speakers of a language die, the language dies too. In fact, language death may take place even before then, as Crystal points out: "if you are the last speaker of a language, your language - viewed as a tool of communication - is already dead" (2000, p. 2). Although that last speaker's knowledge of the language is retained like an archive while he or she lives, "the moment the last speaker of an unwritten or unrecorded language dies, the archive disappears for ever. When a language dies which has never been recorded in some way, it is as if it has never been" (ibid.).

Language extinction has been a part of history, but what is most disturbing in these present times is the huge increase in the rate of extinction. Until about 1700 AD only a relatively small number of languages ever became extinct suddenly (that is, in one or two generations). It could happen when a tribe was wiped out by a natural disaster or an epidemic, or when they were overwhelmed by a more powerful nation. In the last 300 years, however, hundreds of languages have become or are in the process of becoming extinct, and the rate of extinction may be increasing. These trends have been pointed out by Wurm, and he attributes the present rapid extinction of languages today to globalisation:

> With the upsurge of communication, our own period seems to have created more situations of conflict between languages of the world than ever before, by the same token causing more and more languages to disappear at an accelerating pace (Wurm, 1996, p. iii).

The specific causes and processes of language loss today have been discussed at length by both Wurm (1996, pp. 1-2) and Crystal (2000, chap. 3). In summary, there are two main causes: (a) the destruction or dispersion of a language community, and (b) the assimilation of the language community into the community of a more aggressive culture with a different and usually metropolitan language.

In the last 200 years destruction and dispersion of language communities has occurred for a mixture of political, economic and natural reasons that can be connected with globalisation. In particular, the takeover of lands and the non-traditional exploitation of natural resources by colonising nations in North America, New Zealand, Australia, the Amazon and many other places has had a deep impact on indigenous language communities, many of whom have moved away from their homelands to towns and cities. New diseases such as measles or influenza have severely reduced the populations of indigenous communities; within a hundred years from the arrival of the Spanish in Mexico, for example, it is estimated that the native population plummeted to less than 7% of its original number because of disease (Crystal, 2000, p. 72).

The assimilation of language communities and their cultures into another culture may take place as a result of dispersion, or it may insidiously take place in the heartland itself. Crystal gives three stages:

1. there is a pressure to learn and speak the dominant language;
2. the language speakers become bilingual;
3. the ethnic language is gradually dropped from use.

These stages may take place over several generations, or even in as short a time as a decade (Crystal, 2000, p. 79). The most critical point in the process is that where a language is said to be "endangered" - where the children, on the whole, are no longer learning it.

The intriguing question is: what brings a language community to the point where their children abandon the learning of the language of their parents in favour of another one? This is a much discussed question, and the most general answer seems to be that it

happens when children have a low esteem for their language and culture. As Dorian (1998, p. 3) expresses it: "languages are seldom admired to death but are frequently despised to death." Perhaps the most significant circumstances in which a language becomes so despised are:

1. The language community is in constant face-to-face contact with another language group for whom they have a high esteem because of a perceived political, economic or cultural superiority. When this superiority cannot be matched by some particularly admirable quality of their own, the ethnic group begin to doubt the value of their own culture and the language that marks them as belonging to that culture. Television/video is a powerful medium for transmitting cultural values today, and monolingual television, it can be argued, conveys a message that minority languages are not important:

   > With television young people realised that outside the community people spoke a completely different language. Their own language was only spoken by some fuddy-duddies at home. (Mithun as reported by Itkonen, 1998)

2. When there is a pressure and opportunity to learn the other language, the result is bilingualism. As noted by Crystal (2000, p. 78), this pressure could be top-down in the form of educational policy, or bottom-up in the form of peer pressure. It should be noted that bilingualism is the norm in many parts of the world, and does not of itself lead to language shift. Language shift can occur when the ethnic language speakers perceive their language to be useless for discussion in more and more domains, either because the vocabulary for these domains has not been enriched by education in the ethnic language, or because there are too many people involved in these domains who do not understand the language. Then,

   > ...the younger generation becomes increasingly proficient in the new language, identifying more with it, and finding their first language less relevant to their new needs. (Crystal, 2000, p. 79)

The new language then becomes a lingua franca for the community and gains prestige, while the ethnic language loses value in its speakers' eyes.

> It is easy to see how a language could eventually die, simply because, having been denuded of most of its domains, there is hardly any subject-matter left for people to talk about, and hardly any vocabulary left to do it with...It lacks prestige. (Crystal, 2000, p. 83)

3. There is a perception that the other group have a low esteem of the people of the ethnic community or their language, whether through persecution or neglect. For example, children have been punished for speaking their own language at school in America. Mithun, for example, is reported as saying:

> No one thought that the Indian languages had their own importance and [that] it was valuable to speak them. The ability to speak an Indian language was not appreciated at all; in fact, children were punished if they spoke their own languages. Many parents thought later that they didn't want their children to suffer the same way and spoke English to them. (Itkonen, 2001)

In New Zealand too, I have spoken with people who reported that they were punished for speaking Maori at school.

Eventually, all value for the language is lost, and, as Dorian (1998) observes,

> ...it's fairly common for a language to become so exclusively associated with low-prestige people and their socially disfavored identities that its own potential speakers prefer to distance themselves from it and adopt some other language. Parents in these circumstances will make a conscious or unconscious decision not to transmit the ancestral language to their children, and yet another language will be lost. (p. 3)

From my own observations and personal communication with colleagues in Papua New Guinea, I suggest that the factors typically involved in shifting people's attitude to their language in that country are:

1. a high esteem for English or Pidgin as the language for those seeking an economic or social advantage (e.g. a job in the town or for a multinational corporation exploiting a nearby resource, or commercial or social interaction with people traveling a nearby main road);

2. an educational policy that has promoted universal education in English throughout the country;[13]

3. the influx into a language community of a sizeable proportion of families who do not speak the language, and find that they can get by without having to learn it. The language of interaction in the playground and for community as a whole then begins to shift to a common language (such as Pidgin or English).

To illustrate how these factors can work, a typical scenario is: a young man goes to town to get work; while there he marries and has children; the children are educated along with children of many different language backgrounds, and who therefore interact using Pidgin or English. Later, the young man feels a need to return to his home village with his family. If the number of such returning families is small, and the local children are not comfortable with Pidgin or English, the returning children do eventually learn the vernacular. I have seen this first hand at Kopi village in the Rumu language community.

If, however, there is a sizeable number of returnees, or the village children are already comfortable with speaking English or Pidgin, then a shift in language use will become noticeable in the village. I have seen this only with regard to expatriate children (my own and friends') who arrived in a community, but did not know the language of the community. If they arrived before education in English was firmly established they learnt the village language; if they arrived after a community school had been running for several years they were spoken to in English, and might never learn the village language effectively.

This language shift tends to take place in localities close to towns in what some colleagues in Papua New Guinea informally call "the main road phenomenon". For example, one colleague reported that the language that he was studying in the Madang Province was viable only in remote villages; in villages on the main coastal road a language shift was taking place where New Guinea Pidgin was becoming the lingua franca, and parents were no longer speaking in the vernacular with their children (Johann Lottermann, personal communication, ca. 1993).

---

[13] In the last decade there has been a move to include vernacular literacy in elementary schools.

It should be said, however, that this path towards language loss is not inevitable. In many parts of the world, including Papua New Guinea and Continental Europe, there is a long history of healthy bi- or multi-lingualism. The ethnic language is used to express the identity of the speakers as members of their community, to foster family ties, to maintain social relationships, and to preserve historical links (Crystal, 2000, p. 81).

2.3 The Value of Languages

On the 1st of February 1999, it was reported that the use of the Morse code SOS signal was being officially replaced by the Global Maritime Distress and Safety System, which uses digital satellite communications equipment to pinpoint location (Wired Digital, February 1, 1999). It seemed that there would no longer be any need for the code as an emergency communication system. The attitude of many people to the minority languages of the world is similar, and the sentiment could be expressed in a form such as: "why bother with the language of a remote jungle village that has no function in the modern global village?" I believe there are important differences between an obsolete man-made code and a human language that make the latter very valuable.

It is crucial to an understanding of the need for documentation of minority languages that there be appreciation of their value as objects of beauty, as an essential part of culture and cultural knowledge, for giving insight into the workings of the mind, as vehicles to express spirituality, and as keys to understanding more about ourselves.

2.3.1 Languages as objects of beauty

For many languages, the speakers themselves see their language as beautiful when spoken or written well. It is to encourage the use of and safeguard the purity of their languages that societies such as the Italian *Accademia della Crusca* and the French *Académie Française* were established. The Rumu that I have talked with speak of the fine language of the older generation in contrast to that of the younger people who "mix" many words "stolen" from other languages so that it does not "enter beautifully into the ear." The Maori too are renowned for their practice of oratory, the Welsh for their singing, and the English of course have a wealth of poetry and prose that they are proud of. Often, it seems that the value of beauty in a language is only accorded to those

finest examples of language - the oratory, the songs, the poems and the prose - as artifacts for performance, while the language as an everyday means of communication disappears.

While purity of vocabulary, diction and composition are readily appreciated by the native speaker, there are other beauties of language that the native speaker is likely to miss, especially the complex patterning and harmony that can be observed in its phonology, morphology, syntax and semantics. For example, the Rumu paradigm of possessive adjectives shown in Table 2.1 is perfectly regular (tones, all level, are not shown):

Table 2.1. Rumu possessive adjectives

	Singular	**Dual**	**Plural**
1st person	*na*	*nati*	*namë*
2nd person	*ka*	*kati*	*kamë*
3rd person	*a*	*ati*	*amë*

In a paper in which she compares features of the North American languages Central Pomo and Mohawk as an illustration of the beauty and variety of languages, Mithun (1998) makes an appeal for an appreciation of this beauty as a motivation for the documentation and teaching of these languages:

> There is not a language in North America that fails to offer breathtakingly beautiful intricacy. For descendants of speakers to discover this beauty can profoundly enrich their lives, much like the discovery of music, literature, or art, if not more. (p. 189)

The discovery of such "beautiful intricacy" is unlikely without the assistance of someone who will document it well so that it can be studied and described for the language community. This also means that these matters need to be expressed in a way that can be appreciated by the native speaker - not necessarily an easy matter for a linguist entrenched in theories pervaded with latinate terminology, hellenic symbolisation and mathematical formulations. Moreover, for a people whose language

is endangered and who have no history of writing, the oratory is likely to be lost and the poems, songs and proverbs are likely to be degraded and eventually lost unless documented.

2.3.2 Language as an essential part of culture

Without language, culture is impossible. It is language that is the most important means for one generation to pass on to the next their understanding of life, people, history and the environment, and the way to deal with each of them. It is sometimes said that language is simply a code, and that knowledge is code-independent, so that any language would be sufficient to express a culture. But this ignores some fundamental interconnections of language and culture:

> Each language reflects a unique world-view and culture complex, mirroring the manner in which a speech community has resolved its problems in dealing with the world, and has formulated its thinking, its system of philosophy and understanding of the world around it. In this, each language is the means of expression of the intangible cultural heritage of a people, and it even remains a reflection of this culture for a while after the culture which underlies it decays and crumbles, often under the impact of an intrusive, powerful, usually metropolitan, different culture. However, with the death and disappearance of such a language, an irreplaceable unit in our knowledge and understanding of human thought and world-view has been lost forever. (Wurm, 1996, p. 1)

As an example of the complex connection between language and culture, Woodbury shows how the complex demonstrative system of the language of the Yup'ik of Alaska reflects the way they build their houses and live inside them while at the same time reflecting the way villages are placed in river valleys; he finds that the ways of expressing the demonstratives in English comprise a "pale shadow" of the Yup'ik system, and argues that:

> Codes are really not interchangeable: individual codes, and the ways they are practised in individual communities, **are** linked, indirectly or directly, to essential cultural content. Language preservation is therefore a crucial part of the maintenance of cultural diversity. (Woodbury, 1993, sec. 4)

Not only does language express culture, it is itself a part of the culture. "Language represents," states Mithun (1998, p. 189), "the most creative, pervasive aspect of culture." In his defence of the proposition "When a language dies, a culture dies", Woodbury (1993) brings together arguments that language is part of culture in three ways: (a) language as identity, (b) language as a store of knowledge or intellectual wealth, and (c) language as constitutive of verbal art.

As a source of identity, Krauss (1992, p. 11) goes even further, asserting that "language is identity". People of different cultures, even when they look and dress similarly, can realise their cultural identity by the language or variety of language that they speak. The importance of language to culture is a source of controversy, however, and some would argue that non-linguistic features such as clothing or certain ritual or celebratory activities, are sufficient markers of culture. It is common for people who speak the language to say that the language is an important part of the culture, while those who do not speak it say that it is not important. Dorian has found this to be the case in Scotland, for example:

> I found that when I asked speakers of Scottish Gaelic whether a knowledge of Gaelic was necessary to being a 'true Highlander', they said it was; when I asked people of Highland birth and ancestry who did not speak Gaelic the same question, they said it wasn't. (Dorian, 1998, p. 20 as cited in Crystal, 2000, p. 120)

There is, however, a one-to-one correspondence between minority or local languages and the cultures of their speakers, so that, as Woodbury argues, "ancestral languages are more natural and viable as emblems of cultural identity" (1993, sec. 4).

As a store of knowledge, language is used to pass on a culture's accumulated understanding of themselves and their environment from generation to generation in the form of literature, whether written or oral. The semantic structure of the language reflects the cultural taxonomy used to interpret the world, so that when a language and cultural system are lost, there is, "irretrievable loss of diverse and interesting intellectual wealth, the priceless products of human mental industry" (Hale et al., 1992, p. 36).

As a constitutive of verbal art, the very structure of the language is exploited to the full. Skilled composers of poetry, songs and chants employ features of morphology, syntax, phonology and semantics rhyme, rhythm, metaphor, plays on words and other literary devices. As Hale et al. (ibid.) have said, "the art could not exist without the language."

2.3.3 Language as a reflection of the mind

As an expression of the mind, language reflects the functioning of the mind. Different world-views mean different foci for thought, and these foci may well be reflected in areas of complexity in the language related to their expression. The complex demonstrative system of the Yup'ik mentioned in the previous section, for example, seems to reflect the cognitive focus on the cultural implications of the position of a person within a house. In a similar way, for the Rumu who live in a rainforest in Papua New Guinea, canoe travel is an extremely important part of life, and this is reflected in a highly developed morphology which is based on directions with respect to one's position on a river: *ka* "across the other side", *kea* "up stream", *roa* "down stream", *nu* "towards the bank or up a side stream", *hoi* "towards the main stream" are part of a set of directionals that can be applied to verbs as suffixes, e.g. *kërë-hoi* "push out (into the river)", and also together with a distance morpheme, form the basis for a set of demonstratives, e.g. *hoi-ko* "way out there (in the main stream)".[14]

There has been a common misconception that a technologically advanced people have more advanced cognitive abilities, and therefore will have developed a more complex language. But Dorian (1998, p. 8) points out that Europeans, "unable to conceive that a people who lacked a rich material culture might possess a highly developed, richly complex language, ... wrongly assumed that primitive technological means implied primitive linguistic means." Documentation and description of language can help counter this view, because, as Wurm has pointed out, there is a "complexity and high level of thought inherent in each language, including those spoken by people

[14] A Rumu man actually taught this to me by taking me out in his canoe on the river and pointing in various directions. The demonstratives are also applied by analogy to other long spaces such as paths and houses.

whom speakers of languages of general or international currency are very much inclined to regard as 'primitive'" (Wurm, 1996, p. 6).

The loss of even one language means, as Wurm expresses it, "a contraction, and reduction and impoverishment of the sum total of the reservoir of human thought and knowledge" (ibid). The loss of a large number of languages then, especially without good documentation, will leave us even poorer, and will mean that "we will never even have the opportunity to appreciate the full creative capacities of the human mind" (Mithun, 1998, p. 189).

2.3.4 Language as a vehicle to express spirituality

The spiritual side of people tends to be discounted in modern scientific literature, but Zepeda and Hill (1990, p. 1 as cited in Woodbury, 1993), in speaking of spoken language as a source of identity for its speakers, point out that this includes spiritual identity, and that *spirit* is "a key term in the discourse of indigenous peoples." The use of language in worship or communication with spiritual beings is a universal human phenomenon, and set forms such as prayers, chants or spells become meaningless if the language is lost.

For Muslims, for example, maintenance of the Classical Arabic language is important so that the Koran can be read in its original form. Likewise for Jews, the Hebrew language of their scriptures has been maintained for millennia. For Hindus, the importance of a perfect language for religion and science motivated Panini to compose his grammatical description of Sanskrit.

For Christians too, the connection between language and the spiritual side of life is very important: Christ is even known as *ho logos*, "the Word" (John 1:1). Furthermore, for the Christian, the diversity of language is another evidence of the richness of creativity that the creator has put in his creatures, and God's approval of this diversity is shown, for example, in the Biblical report (Acts 2) of the events in Jerusalem on the Day of Pentecost seven weeks after the crucifixion of Jesus, when the new Christian church was said to be born. The report depicts the Holy Spirit empowering the Aramaic-speaking disciples to speak in the languages of people from all

over the then-known world. This use of any and every language for worship is also depicted by the Apostle John who wrote of seeing a vision of an enormous crowd of people gathered before God's throne: "They were from every race, tribe, nation and language...They called out in a loud voice, 'Salvation comes from our God...'" (Revelation 7:9, Today's English Version).

2.3.5 Linguistic diversity as a key to knowledge about ourselves

There is a huge diversity in the languages of the world today. This diversity can be seen at every level of language. For example, the 1989 edition of the IPA phonetic chart contains as many as 140 symbols for differentiating sounds which are produced in the world's languages. No language produces all the possible sounds, and some sounds occur in only a few known languages. The famous click sounds, for example, are most prolific in the Khoisan languages, most of which are now considered endangered.

The phonemic inventories of languages vary tremendously. Crystal (1997, pp. 167-170) uses data from the UPSID (the University of California, Los Angeles Phonological Segment Inventory Database) to illustrate the variety. For example, Rotokas, an Indo-Pacific language, has just seven consonants, while !XU, a Khoisan language, has 97, including 48 clicks. (Many of the !XU consonants have to be written using two or three IPA symbols.)

There is diversity in the way languages combine meanings. Typologists have traditionally classified languages into four morphological types: analytic/isolating languages such as Chinese where each morpheme is a separate word; agglutinative languages such as Turkish where words are composed of strings of easily identified morphemes; fusional/inflected languages, where there are complex paradigms with suppletive forms and portmanteau affixes; and polysynthetic/incorporating languages such as Yup'ik Eskimo which, like agglutinative languages, have long words with many morphemes that have characteristics found in fusional languages. A careful study of many languages has shown that most languages cannot be neatly classified as belonging to just one or another type. English, for example, has characteristics of isolating, fusional and even agglutinative languages (Lyovin, 1997, pp. 14-26).

A more recent way in which language typologists classify languages is to observe the normal order for the main parts of a sentence, Subject (S), Object (O) and Verb (V). The three commonest orders are SVO (e.g. English), SOV (e.g. Japanese) and VSO (e.g. Maori). There are a few well-known languages with VOS order (e.g. Malagasy), but until recently Object-first languages were thought not to exist. However, some languages spoken in the Amazon area which have an OVS order have been found (e.g. Hixkaryana), and a few may be OSV languages (e.g. Apurina) (Crystal, 1997, p. 98). Without a rich diversity of languages to study it would not be known that all six possible permutations of S, O and V could exist for human languages.

There is diversity in semantics too. It is well-known amongst translators that a word in one language never really has an exact equivalent in another language. This is because ways that people of different cultures and languages categorise their world is very diverse, and this is reflected in the way semantic space is divided up in the lexicon. A common way to illustrate this is to look at colour terms. For example, the Rumu have four main colour terms: *hokore* covers the colours known as "red" or "brown" in English; *yaëa* covers "pale yellow" and "pale green"; *kihopuri* "black", "dark green" and "dark blue", and *paruho* "white" and "pale grey."

The diversity of languages that I have illustrated here is important because finding universal similarities within all the diversity can help us understand how our minds work and how we acquire language, and the differences show how flexible the mind can be in expressing thought.

Another reason why a diversity of languages is important is that clues to the pre-historic movement of cultures and peoples over the earth can be found through the comparison and establishing of relationships between languages. For example, using comparative methods it has been postulated that the Austronesian languages of the Pacific (and Madagascar) originated in the southern coastal area of China (Lyovin, 1997, pp. 249-250).

Woodbury (1993) reminds us, however, that

> we are about to lose most of the linguistic diversity that has developed, largely independently, over the course of human history. If we do, our hope of reconstructing linguistic prehistory around the world, or of meaningfully testing precise theories of how the languages people speak can and cannot differ, will be limited catastrophically. (para. 3)

Documentation of this linguistic diversity is perhaps the only way to maintain the data for this kind of scholarship.

2.3.6 Language as a part of our heritage

A good heritage gives us a head-start in being able to enjoy a rich life. Our forebears have passed on to us not only our genetic makeup, but also the environment, social organisation, culture, technology and knowledge of our roots and our language that we grow up with. Typically, we appreciate what we have received but are sad at what our ancestors have foolishly lost - whether it is a once flourishing and clean environment now depleted and polluted, a bird or animal hunted to extinction, or knowledge not properly passed on. This sadness that we experience should be a warning to be sure to preserve and pass on those things that make life rich to our children. Most linguists would hold that language is one of those things. SIL International, for example, is an organisation whose stated purpose is "to work with language communities worldwide to facilitate language-based development through research, translation, and literacy" (SIL, 2002). They hold, as one of the tenets of their creed, that "all languages are worthy of preservation in written form by means of grammars, dictionaries, and written texts. This should be done as part of the heritage of the human race" (ibid.).

2.4 Responses to Language Extinction

There are three ways to respond to the impending loss of language diversity. The first is to allow it to happen, or even to promote a policy of assimilation to majority cultures and languages in a belief in "a linguistic survival of the fittest, a social Darwinism of language [which assumes] a correlation between adaptive and expressive capacity in a language and that language's survival and spread" (discussed by Dorian, 1998, p. 10). In the light of the arguments I have presented in this chapter such a

response can only be described as arrogant or negligent, and I shall say no more about it in this thesis.

A second response is to promote conservation of these endangered languages, and a third is to promote the documentation of minority and endangered languages on a wide scale. Conservation, while not the topic of this thesis, will be examined briefly in the next section in order to further justify devotion of effort to documentation. The last option, documentation, is perhaps the more desperate-sounding response, but it is also, I believe, the more achievable, and in some cases, it may be the means to achieving conservation.

2.4.1 Conservation

As has been mentioned in chapter 1, conservation of species has become a major international concern, for, through over-exploitation, the introduction of stronger species, islandisation, destruction of habitat and pollution, species are being lost at an alarming rate. Proposals to halt this loss are being implemented in some places, but ignored in many other places through greed, ignorance or desperation. Perhaps conservation of such non-tangibles as languages, also in the same desperate plight, seems a much less important issue.

The most important factor in language conservation is the adoption of an attitude that one's language is something of value, and in this chapter I have presented many arguments to support this. This attitude can be brought about with the raising of the self-esteem of the group speaking an endangered language. Minority languages such as Irish, Welsh, or Maori have faced endangerment because they were considered the languages of rural and uneducated or neglected peoples. Now they are reported to be undergoing relative revival as they have become associated with attributes of vitality such as nationalism, regional pride and cultural identity, but only, it would seem, after a strong and sometimes radical leadership have taken up the cause.[15]

[15] For accounts of Irish, Welsh and Maori language language revitalisation, see Quinn (2001), Morgan (2001), and King (2001) respectively.

Other approaches include (a) the promotion of bilingualism and multilingualism as an attractive alternative to language shift (Crystal, 2000, pp. 79 & 112), (b) the setting up of a Universal Declaration of Linguistic Rights to counter factors such as "invasion, colonization, occupation and other instances of political, economic or social subordination" (CIEMEN, 1996, preliminaries section) that have disadvantaged minority languages, and (c) the revival of dead or dying languages, such as the revivals of Cornish in England (Grimes, 2000), and of Hebrew as the national language of the state of Israel (Quinn, 2001).

Each of these approaches originates in a belief that the language in question is of value. There are many people groups today speaking endangered languages that are not valued. In order that such endangered languages survive, their communities must be pro-active; procrastination will mean irrevocable loss for future generations.

2.4.2 Documentation

If language conservation strategies such as those outlined above do not work, then documentation is the only alternative. Krauss (1992, p. 8 as cited in Woodbury, 1993) asserts that "...without [extensive] documentation the language [of a people whose language is dying] must irrevocably disappear into oblivion, and very likely so also the national identity in the long run."

At the same time, as discussed previously in section 1.5, to think of the documentation of languages as a task worthy of a linguist's devotion is a relatively new idea. Wurm (1996) points this out in the introduction to his *Atlas of World's Languages in Danger of Disappearing*:

> The systematic study [of the disappearance of languages] at a world level is very recent and the task of describing and recording languages before they disappear is only just beginning. (p. iii)

Until recently, language documentation has been seen as more of a task for a librarian or a drudge; it has not been seen as being as valid a topic for those engaged in serious research such as description or theorising, even though these latter may, in the near future, depend on documentation efforts carried out now while many endangered and minority languages are still alive. In the past, the tools for documentation have been

poor, the interest level in the academic community has been low, and the costs of publishing large bodies of data prohibitive. Grenoble & Whaley (2002a) illustrate this in a case study of Charles Furlong's attempt at documenting the endangered languages of Tierra Del Fuego from 1907 using wax cylinders to record sound and an inadequate transcription system. A few years later, John Harrington made more successful efforts from 1928 to 1942 to document Californian languages, some of which are now extinct. He had the benefit of a better technology for recording sound utilising aluminium disks. He had also developed a very detailed transcription system. Although publishing his wealth of data was a problem, his efforts have resulted in a valuable legacy for today's linguists and Californian ethnic communities (Hinton, 2001, pp. 269-270; Yamane, 2001, pp. 430-432). Now we are in a new era with much better tools and means of publishing, and it is time to promote documentation as a valid academic exercise and valuable humanitarian service.

3. THE SCOPE OF LANGUAGE DOCUMENTATION - TYPES OF DATA

3.1 The nature of data

Textbooks on information systems commonly begin by differentiating data, information and knowledge. For example, Laudon & Laudon (1998, p. 16) define **data** as "the raw facts that can be shaped and formed to create information", **information** as "data that have been shaped by humans into a meaningful and useful form", and **knowledge** as "the stock of conceptual tools and categories used by humans to create, collect, store, and share information." For the linguist, data is "observable linguistic behavior" (Himmelmann, 1998, p. 166). The focus of this chapter is on what types of linguistic data should be included in a documentation, and how much shaping and forming of that data is appropriate in order to make the documentation a source of useful information that will contribute most widely to the body of human knowledge.

3.2 The traditional view of documentation

The traditional view of language documentation is as an "edited version of the field notes" (Himmelmann, 1998, p. 165). This is not the same as a publication of a carefully composed description of the language, but the editing of the raw language data (recordings and transcriptions) and any pertinent notes and partial or completed analyses, simply to make them available to interested people. This is a reasonable form of documentation when further data gathering is, or may be, no longer possible. Linguists with SIL in Papua New Guinea, for example, are expected to prepare an archive of language data and up-to-date field notes when they are about to leave the country for an extended time. Although such an archive is not readily available to other linguists, it is intended to be passed on to any linguist who may take over the field work of the departed linguist, and it could form the basis for a published documentation.

As an example of the accumulated wisdom of many years of field linguistic research, it is instructive to examine the kind of documentation that an organisation such as SIL expects its members to produce, and we turn to this in the following section.

3.3 SIL language documentation

SIL encourages linguistic publications of their members, and, in fact, requires their members who are doing linguistic work to produce certain documents at various stages of their language-based development projects. The particulars of these requirements change from time to time and from place to place, but, as an example, the Technical Studies Handbook for the Papua New Guinea Branch (Loving, 1983) expected the following documents to be produced at various stages:

1. dialect surveys (including Swadesh list) (sec. 3.7);
2. Anthropology Sketch (after about 2 years) (sec. 6.7). This was later divided up into an initial Background Study or Social Organisation paper to be followed by a Culture paper at a later stage (TSD, 1985);
3. Grammar Essentials (sec. 7.4.1) initially, to be followed by a higher level grammar paper called a Grammar Sketch (sec. 7.4.2) based on transcribed and analysed texts. These were to be based on authentic text, and the amount of text was later specified: the Essentials was to be based on 40 pages of double spaced text (or 72 kbytes), and the Sketch was to be based on 100-150 pages (180-270 kbytes) (TSD, n.d. , pp. 7 & 14);
4. Phonemic Statement (sec. 7.2.2). This was later divided up into an initial Organised Phonological Data paper to be followed later by a Phonology Essentials paper (TSD, n.d. , p. 2);
5. Approved Orthography Status write-up (sec. 7.2.3);
6. primer series for literacy (sec. 8.4) and post-primer materials (sec. 8.5.5); 100 pages before orthography approved and 300 pages after;
7. translation of parts of the Bible (sec. 9);
8. dictionary (sec. 7.6).

Other documentation was also encouraged:

9. recordings of indigenous music, noting that "it helps preserve for the world a cultural heritage which could be lost forever" (sec. 6.7);
10. ornithological observations and folk taxonomy (sec 6.8).

It can be seen that in an SIL language project, which may last for 15-20 years, substantial amounts of audio and transcribed text data of various kinds are collected for the purposes of analysis and literacy, and substantial amounts of other language literature is authored (primers, a dictionary and considerable translated materials). There was no requirement to integrate the materials, although modern computer software such as SIL-produced Shoebox and LinguaLinks, by their very nature, encourages integration. The versions of the two SIL applications mentioned, Shoebox and LinguaLinks, that I have access to do not appear to be capable of directly producing internet-ready material, but there is an interest in some quarters in doing just that (see section 3.4 of this thesis).

A program focused on language documentation rather than translation could be achieved without complete linguistic analysis. Primers and translated materials take considerable skill and effort to produce, and are probably of minimal interest to researchers, so they would not be the focus of an academic language documentation project. However, such materials may well be of interest to the language community, and if they exist there is no reason to exclude them, as long as there is some qualification as to their nature and, if possible, some commentary on their quality by native speakers.

The orthography used for the documentation is very important if it is to be of use to the language community, and I see the obtaining of an orthography approved by the community as an important task for the language documentalist.

3.4 Simons' suggestion

Documenting hundreds or thousands of minority languages will have to be the work of many institutions rather than just one. Gary Simons, an information technology specialist working with SIL, has published a suggestion that a framework be developed for the World Wide Web with which the language documentation work being done throughout the world could be coordinated (Simons, June 12, 2000). This coordination would include a metadescription (i.e. an abstracting and indexing) service and a resource for relevant software and standards information for linguists doing the work. These services would be maintained at single, centralised sites (ibid., The Major Components of a Solution section).

Simons suggested that the framework for software resources and standards information could be accessed through an index in the shape of grid, as shown in Table 3.1. This table could be displayed in a Web browser, for example, and the cells of the table would be buttons. A click on one would take you to the resources available for carrying out a given function on a given data type, so, for example, if a linguist wanted to know what software resources were available to allow look-ups for his dictionary, he or she could press the button at the junction of the Query column and the Lexicon row.

Table 3.1. A framework for the repository, from Simons (June 12, 2000)

Function > Data type v	**Store** *Formats for storing data*	**Display** *Stylesheets for displaying data*	**Query** *Procedures for querying data*	**Convert** *Procedures for converting data*	**Create** *Programs for creating data*
Meta-description					
Word list					
Writing system					
Annotated speech					
Lexicon					
Field notes					
Description					
Common					

Of particular interest for this chapter is the list of data types down the left hand side of the grid. Below I have listed Simons' detailed description of their purpose. I have appended lists (in braces) of the sorts of documents produced by SIL linguists that would fit into each category:

1. *Metadescription.* The description of what is in a data resource that can be used in the [single, centralised] online catalog as an aid [for users on the World Wide Web] to find the resource;
2. *Word list.* A list of word-forms in the language indexed by reference glosses (for example, a Swadesh word list). This is not just a simplified lexicon; unlike a lexicon,

the indexing against a list of universal reference glosses provides a data structure for cross-linguistic comparison {Swadesh lists from dialect surveys};

3. *Writing system.* A description of the writing system used to express text in the language {Orthography write-up};
4. *Annotated speech.* Samples of speech (in audio or video recording) that are annotated for transcription and various kinds of analysis. Interlinear text can be treated as a special case of annotated speech in which the base recording is absent {Analysed texts};
5. *Lexicon.* A listing of the lexical items in a language with descriptions of their phonological form, morphosyntactic function, and semantics.{Dictionary};
6. *Field notes.* The initial observations a linguist makes in the field;
7. *Description.* Any work of prose that describes some aspect of the language (for instance, a grammar sketch or a workpaper on phonology) {Organised Phonological Data, Phonology Essentials, Grammar Essentials, Grammar Sketch, dialect surveys};
8. *Common.* This row is for resources that are common to all data types (like fonts or character-code conversion tables).

(Adapted from Simons, June 12, 2000, A Framework for the Repository section)

The first and last of these data types are to do with indexing and cataloguing, and technical software resources, respectively, so I postpone discussion of these until chapters 4 and 5. We will turn now to consideration of his other six data types.

Simons implies that his list of data types is not exhaustive, but it is a useful starting point. Of the documents produced by SIL linguists, most fall into either the Description or the Annotated Speech category, but it is difficult to find categories in his list for documents such as primers and post-primer materials, translated materials, and anthropological papers. As argued in the previous section, while these uncategorised materials may not be of much interest to linguists, they are language-related materials, and are likely to have a very high interest value to the language community. If a language documentation is going to be relevant to the language community, and in this thesis I

argue that it should be (see section 1.5), then Simons' categories will need some adjustments.

3.5 Himmelmann's proposal

For Himmelmann, the core of documentation is raw data, and the place for analysis is as a commentary on the data, or a minor section of the documentation. In simple terms, a language documentation would be a "radically expanded text collection" (1998, p. 165), and there are two main reasons for promoting data above analysis. Firstly, the data can give rise to different analyses, depending on the theory in fashion when the analysis is done. As theories become dated, so analyses based on them become dated, but that will never happen to the data itself. Secondly, raw data can be used for types of analysis and application which the data gatherer never had in mind; a language description, on the other hand, is too focused and the quantity of data collected too small to be of much use to anyone other than a linguist. Large quantities of raw language data collected in an appropriate way can be the basis not only for linguistic description, but also descriptions and analyses in a variety of other disciplines: socio-linguistics, discourse analysis, anthropology, oral history, language acquisition, language learning, literacy, literature, and so on.

Of course, some analysis is important to data collection: an efficient and maximally useful transcription system requires a good orthography, which has to be based on sound phonetic and phonological analysis; segmentation into words, intonation units, sentences or paragraphs requires a certain amount of phonological, morphological and syntactic analysis. Himmelmann (pp. 162-163) acknowledges this, and observes that "a good and comprehensive documentation will include all the information that may be found in a good and comprehensive descriptive grammar", although, as he notes, this "will not be accessible in the usual way since language documentations are not organized by grammatical chapters" (p. 170); instead it would be scattered through the documentation as commentaries on the data.

It can be seen that such a data-focused documentation should better serve the needs of most target language communities who are interested in what is said in their language

(the data) rather than how it is encoded (the analysis). Furthermore, using digital media with hypertext, a skilled documentalist could work at overcoming the problem of the scattered nature of the analysis by compiling meta-documents. These would use hypertext links and indexes to collect and organise comments and examples of various topics scattered all over the documentation. The result could be a dynamic, hypermedia phonology paper, for example.

The kind of documentation that Himmelmann proposes would contain an introduction and three main sections for filing three major categories of language data or document:

1. *General Introductory Commentary.* This gives information on the speech community, language, field work, methods, contents and scope of the documentation (p. 170).
2. *Communicative Events.* "The core of a language documentation", this section consists of a "comprehensive and representative sample of communicative events as natural as possible" (p. 168).
3. *Lists.* These are paradigms, folk taxonomies and other "list-like linguistic phenomena" best obtained by elicitation (p. 169).
4. *Analytic Matters.* Write-ups of other linguistic or cultural matters elicited or discussed are kept in this section (p. 169).

Moreover, Himmelmann sees that each document or piece of data filed in the Communicative Events section will, in turn, have three main components:

i. *Raw data* - audio, or (ideally) video, together with a transcription;

ii. *Translation* - word-by-word and free;

iii. *Commentary* - additional information about the data and the data collection process.

There is no need to deviate from Himmelmann's focus on data in a documentation. From the point of view of the target language community, however, it can be argued, firstly, that some of the materials included should be rearranged into different categories according to their interest to the community, and secondly that there are other related materials that should also be included, again because of their interest to the community.

In the first case, cultural notes, for example, should be separated out from more technical studies such as phonological and grammatical analyses. In the second case, background materials from other researchers (historical diaries, survey notes, etc) are often accessible to the researcher in archives in distant centres or other countries. Members of the ethnic community may well want to have them included in a full documentation. They may also want to include material in the language which cannot be considered "raw data" in the sense of Himmelmann's proposal - material such as translations and educational or literacy material, even if only for archival purposes. It could be argued, of course, that such material is, in fact, "raw data" for educational and translation theory research.

3.6 Data sampling approaches

Because language documentation is envisaged as having multiple purposes, it is important that the communication events recorded for the corpus comprise an extensive sampling of the sorts of communicative activities carried out in a language community. In order to ensure a sample that properly represents all of these activities, it is necessary to set up some parameters by which they can be characterised.

If we regard communicative events as a term roughly equivalent to "discourse", then there are three approaches to classifying them that I have found in the literature, and which can be termed the anthropological, linguistic and functional approaches, respectively.

3.6.1 Anthropological approach

Pike calls discourse "a verbal behavioreme", that is, a stretch of language recognised by the culture as an entity; it has a beginning, an end, and an internal structure (in Grimes, 1975, p. 21). Himmelmann recommends that the documentalist select a sample of communicative events taken from as many *native speaker classifications* as possible. This approach has the advantage of allowing the language community some input into deciding what a good specimen of a given classification would be.

3.6.2 Linguistic approach

This approach has been put forward by Himmelmann (1998, pp. 176-183), and is based on two parameters that can be applied to any communicative event: spontaneity

and modality. The **spontaneity** parameter can be used to place communicative events on a continuum from unplanned (spontaneous) to planned, for example, from exclamations (very spontaneous), to directives (i.e. demands for action), to conversations, to monologues, to ritual speech events (very planned). The theoretical basis for choosing spontaneity as a useful parameter is "the assumption that spontaneity is correlated with aspects of linguistic structure" (p. 180), i.e. the more planning that is put into an utterance, the richer and more complex the language used will be. The **modality** parameter covers linguistic practices related to the medium of communication, and is categorial rather than continuous: speech, song, writing, hand-signing, etc.

A comprehensive documentation will "provide specimens from each modality in as many degrees of spontaneity as possible" (p. 183).

3.6.3 Functional approach

This approach uses classifications of communicative events (as discourses or texts) used in text-linguistics. Longacre (1972, pp. 134-136), for example, utilises four categories of discourse: **narrative** (stories), **expository** (explanations), **hortatory** (advice and rebuke) and **procedural** (how things are done). Each of these shows organisational characteristics based on different combinations of the two features **contingent temporal succession** (time sequence as opposed to logic or theme) and **agent orientation** (attention on who did or should do something as opposed to what is done or how something is) (Edmondson & Burquest, 1994, p. 89). Another common classification system has five categories: narrative, expository, argumentative, instructive and descriptive (Werlich, 1982, as cited in Renkema, 1993, p. 9).

Himmelmann (1998, pp. 176, 183) suggests data should be selected using both anthropological and linguistic approaches, because of a perceived difficulty in arriving at a "cross-linguistically applicable definition of genres [i.e. text types]." This sounds reasonable, especially if the categories of a functional approach are used as a cross-check in making certain that a reasonable analysis of a native speaker typology of discourse has been made. It should also be remembered that a Himmelmann-type documentation is intended for a much broader research base than that of the pure linguist, and so would

include raw data such as mother-child and learner-native speaker dialogues for studies in language acquisition.

3.7 A user-oriented documentation

The data-focused approach of the type Himmelmann proposes is very suitable for documentation of endangered languages, especially where, for reasons of obligation or urgency, analysis and thorough descriptive work must be postponed till a later time. Where this latter work is already in progress or completed, however, such as in a typical SIL program, it is clear that the Analytic Matters section of a Himmelmann-type documentation would be much more substantial. In Table 3.2 I show how an SIL-type documentation and Simons' categories (rearranged) relate to Himmelmann's data-focused approach.

Table 3.2. Comparison of Simons' and SIL's categories with Himmelmann's

Simons' Categories	**Himmelmann's Proposal**	**SIL Requirements**
Metadescription Writing System Some Field Notes Common	General Introduction	Parts of Background Study
Annotated Speech	Communication Events	Transcribed and analysed texts
Word list Lexicon	Lists	Dialect Survey Dictionary Paradigms of Grammar papers
Bulk of Field Notes Description	Analysis	Grammar papers Phonology papers Orthography papers Anthropology papers
(no category)	(no category)	Primers Post Primer materials Translated materials

With their emphasis on communicative *events* as language data and researchers as the main prospective users of the data, I cannot see that Himmelmann or Simons have given any real place to extant written literature and educational or translated materials, nor anthropological descriptions in their proposals for documentation. These are language-related data items that are likely to be of high interest value to the ethnic community, and there should be a place for them in a documentation oriented towards use by the ethnic community.

In this thesis I therefore argue for what could be termed a "user-oriented" documentation - a documentation accesible to as many potential users as possible. I have already mentioned the two main groups of potential users: the ethnic community and the research community. A researcher-oriented documentation (such as the kinds of collections of data, lists and analyses discussed in this chapter so far) will be difficult for a more general audience (such as would be found in an ethnic community) to appreciate. However, a documentation oriented towards the ethnic community would not only include them in the group of potential users, but would still remain just as useful for the research community, because all the data, lists and analyses would still be there. Such an orientation is justified, then, for two broad reasons: firstly, a larger potential audience; secondly, consideration for the minority language rights discussed in chapter 1.

For a documentation that is oriented towards the language community or their descendants as potential users in the first place, a modification of the organisation of the data will be necessary. It would be simplistic to divide the materials into two categories - "for the ethnic community" and "for the researcher" - especially considering that researchers may be members of the ethnic community at some stage. A more appropriate solution would be to have a series of categories ranging from "general interest" to "researcher interest", so that any user could delve into the documentation to whatever depth he or she felt comfortable with. My proposal then is that the documentation consist of documents categorised under the following headings:

1. General introduction
2. Historical notes

3. Cultural notes
4. Literature, readers and translated materials
5. Instructional materials
6. Language data
7. Lists
8. Analyses

The order of these headings is *not* crucial, because everything can be made equally accessible using a hypermedia design. The principle for whatever ordering is used is that material which the general user is more likely to want to access should be towards the top of the list, whilst material of a more esoteric nature should be relegated to the bottom.

The actual headings are those that I have found convenient from experience after categorising language and language-related data in a pilot documentation project (see chapter 7). I will now discuss the various types of data included under each heading.

1. *General Introduction.* This will contain information about the location and history of the language community, the classification of the language and its dialects, and an explanation of the writing system.
2. *Historical Materials.* This will include accounts by the ethnic community of their own origins, and accounts by explorers and other outsiders who have had contact with the community: old field diaries, patrol reports, memoirs, etc. Language documentalists are very likely to come across these materials in libraries and archives as they do their background research, and, from my experience with the Rumu and Koriki people of Papua New Guinea, they are likely to be of great interest to ethnic communities. When commentaries or even translations of such materials are made in the target language, they can become valuable linguistic resources as well.
3. *Cultural Materials.* These will include anthropological observations and descriptions that are likely to be of interest to and readable by the ethnic community. They will also include photographic records and videos and of

traditional activities of any kind, but especially those involving language, such as singing, dancing and oratory. Other cultural activities that involve less language use, such as art, instrumental music, hunting, gardening, etc are still valuable to record in a language documentation, because any commentaries recorded with them will be a valuable linguistic resource giving insights into the semantic organisation that is part of the world view of the ethnic community.

4. *Literature, Readers and Translated Material.* This section includes post-primer materials, newspapers, newsletters, hymnbooks, translated materials (e.g. for educational, community development or religious purposes), and recorded texts in readable form, that is, edited, formatted and in the accepted orthography, rather than raw text or phonetic transcriptions. Audio recordings of these materials may also be accessible from this section. This is a collection of materials that are likely to be useful for and used by the language community in general, and commentaries on them could be obtained from general readers of the community.

5. *Instructional Materials.* This will be a collection of materials likely to be used in the class-room by teachers, rather than a general readership. It will include primers for literacy classes, language-learning courses, and teacher-training materials specifically designed for the language community.

6. *Language data.* This will, in many cases, contain the bulk of the language documentation, being recordings and transcriptions with glossed and free translations of communicative events sampled using techniques outlined in section 3.6. There will also be commentaries on these communicative events. The commentaries will certainly give insights into the culture, and may also give insights into many different aspects of the language, depending on the skills of the commentator, or, more likely, the skills and promptings of the fieldworker.

7. *Lists.* This section will contain anything of a list-like nature, including the lexicon, survey word lists, paradigms, numbers, measures, and bibliographies. In the pilot project, I found it useful to include in this section a database of references that coordinated various versions of the same material that might appear in different parts of the documentation. For example, a recorded legend could appear in edited

form in the Literature section, and again as a glossed text in the Language Data section. I also found it useful to keep a Who's Who of authors, with biographical notes and photographs, and references to the materials they had produced.

8. *Analyses*. Notes and papers about the language, including grammars and phonologies. Because of their technical nature they are likely to be of interest only to linguists. This section is also a good place to keep a database of hyperlinks to commentaries on specific linguistic topics that occur throughout the documentation.

As mentioned before, it does not really matter in which order the material in the documentation is organised, but my ordering is an attempt to prioritise according to anticipated ethnic community needs and interest. It is quite within the philosophy of this thesis to put the Lists section before the Language Data section, for example, or even to take the lexicon out of the Lists section and make a separate section for it. With efficient flexible computer storage systems utilising menus and hyperlinked indexes, it is not difficult to reorganise this kind of data.

4. THE MEDIA AND FORMAT OF DATA

Documentation, as a discipline, emerged in the early 20th century in response to the growing number of scientific reports published in journals, and the consequent need to maintain indexes and compilations of abstracts so that the information in the reports could be evaluated and found easily by other researchers. At first, of course, documentation was on paper (or index cards), but with the development of computers from the 1950s on, the storage and retrieval of documentation on digital media added an important new dynamic, and the discipline became known as **information science.** The new **information theory** which developed alongside this science tended to focus, however, on the problems associated with the communication of information by cable and wireless and other channels. Information science, on the other hand, focused firstly on the storage media (i.e. the hardware and low-level formats), and then on the management of information on that media (i.e. the data structures and the software for accessing them) (see Information science, 1989, and Information theory, 1989).

In this chapter I shall consider the application of information science to language documentation, especially the importance of digital and other media, and the variety of methods for formatting and structuring digital data.

4.1 Media for the data

The selection of the physical media depends in part on the intended audience for the data being documented. If the intended audience cannot readily access data on a given medium, then the goals of the documentation have not been met. The selection of the physical media also depends on how well the goal can be met for a comprehensive and multipurpose documentation. In this chapter I will consider the suitability of the two broad classes of medium, analogue and digital, with regard to the following factors:

1. the ease with which it can be used for recording data;
2. the ease with which it can be used to edit and manipulate data;
3. the ease with which text, sound, and still and moving images can be integrated;
4. the ease and cost of reproducing multiple copies of documents;

5. the accessibility of the data by various audiences for various purposes;

6. the suitability of the medium for archiving.

Then I will consider how and where selected forms of the media can best be used in a documentation project.

4.2 Analogue media

An **analogue medium** is characterised by a straight-forward correspondence between the characteristics of an object and its physical representation on the given medium. For example, variations in amplitude of a speech sound wave are represented by variations in magnetic intensity on a tape. The main analogue media suitable for use in a language documentation are printed text,[16] photographs, audio cassette tape, and VHS video tape.

4.2.1 Recording.

All data types are now easy to record using relatively affordable analogue recording equipment - pencil and paper, typewriters, 35mm film cameras, cassette tape recorders and video cameras. Of these, the typewriter has been made relatively obsolete by the high-quality computer printer. Indeed, other analogue devices, even the pencil, now have digital devices that perform the same function and may eventually replace them.

4.2.2 Data manipulation

Editing of analogue data requires much copying and patience. For video and photography, considerable technical skill and complicated equipment is required, unless the data is converted to digital beforehand.

4.2.3 Integration of media

When analogue media are integrated, the media are usually physically separate items which are stored together in the same container. Language text books, for example, have often been published with accompanying cassette tapes, and educational video productions often have printed notes included in the cover. An example of a more

[16] Strictly speaking, printed text is, in principle, more like a digital representation of speech than an analogue representation, because of its use of a limited code set (i.e. alphabet) for discrete elements (phonemes). However, because it is a non-binary non-electronic representation that is directly interpretable from the physical medium (ink on paper) it is included with the analogue data types.

closely integrated approach is the video production which includes subtitles. This is rather unsatisfactory if there is a lot of text to show, or if the screens are small. In either case, tape data is sequential and rewinding or fast-forwarding to the right spot is often slow and somewhat hit-and-miss on all but the most sophisticated tape players.

4.2.4 Reproduction

With photocopiers, twin cassette tape decks, and video recorders connected together, copying of analogue data is quite straight forward. The printing of good-quality coloured photographs from negatives is a well-established technology, and the facility should continue, even if the printing technology becomes digital. Reproduction of large numbers of copies of documentation on analogue media tends to be time-consuming and/or costly.

4.2.5 Access

Texts and illustrations in book form have the obvious advantage that they are useable anywhere by anyone who can read or see. Dry-cell powered cassette tape players for sound recordings are common and sturdy. Access to analogue image technology is changing rapidly. For example, in the 1980s in the remote Rumu village of Kopi in Papua New Guinea, it was a great attraction when 35mm slides were projected onto a sheet using a small projector powered by a 12-volt battery; in the early 1990s, however, villagers were acquiring video players powered by small petrol-fueled generators.

4.2.6 Archiving

Paper documents, vinyl recordings, black and white photographs and coloured paintings have proved to be long-lasting if protected from destructive agencies like vermin, vandals, moisture, ultraviolet sunlight, flood or fire. Multiple copies in diverse places reduce the chances of loss. The Wellington Corpora, for example, was accumulated and digitised at Victoria University of Wellington, but a backup copy is archived at Massey University, over 100 km away.

Longevity is a concern for any important work of documentation. Choice of acid-free paper and enduring marking media (such as pencil or indian ink) can afford longer life to archived image and text. For extreme longevity, the Long Now Foundation's Rosetta Project (see Mason, n.d.) has an interesting solution. The aim is "to create a unique platform for comparative linguistic research and education as well as a functional linguistic tool that might help in the recovery of lost languages in unknown futures". The foundation intends to do this by creating an archive that will be publicly available in three different media, two of which are analogue: a micro-etched nickel disk with a 2,000 year life expectancy, and a single volume monumental reference book.[17] The disk (see Figure 4.1) is to be three inches in diameter, and will contain 27 page images of representative data for each of 1,000 languages, reduced to a scale requiring a 1000X optical microscope (a relatively simple technology) to read them. In order that the disk be recognised for what it is, the initial text of the archive of a few of the languages is readable by the naked eye, and quickly tapers down to microscopic level. It is conceivable that this technology could be used to safely archive a full language documentation (or any other sort of documentation, for that matter) that would last through any imaginable technological dark age to come.

Figure 4.1 A Rosetta disk, from Mason (n.d.)

Magnetic recordings are likely to degrade in the long term, both because of fragility of the physical tape, and also because of the lack of stability of the magnetic recording, whose patterns can, over time, imprint upon adjacent layers of recordings; they can also be accidentally erased by exposure to a magnetic field. As an example of the problems

[17] The third, an online archive, is digital.

we can face, in the 1970s I transferred onto cassette tape precious recordings of Rumu stories made by a missionary on reel-to-reel tape in the 1950s and early 1960s. After 15 years the original machine was still working and available, but the original tape had the magnetic material peeling off the plastic backing in places, and the pitch was slowly rising where the recorder's battery had been slowly running flat during the original recording. This meant that I had to use a smaller capstan whose size was adjusted by peeling off successive layers of selotape as I played the tape back.

Coloured photographic film is subject to colour change over time. Another example from my own experience: photographic slides I obtained of the Rumu taken in the 1950s and 60s were losing their colour and had patches of mould on them. In an attempt to preserve them I had them copied to prints in the 1970s. Later, in the 1990s, I was able to scan the prints and digitally adjust them for the shift in colour balance.

Technological obsolescence is a continuing problem for the preserving of sound and picture data. I have already mentioned the conversion I had to make of reel tape recordings onto cassette tape. Another example from my own experience is a "Regular" 8 mm movie film which I had taken on a clockwork movie camera on my first visit to the Rumu in 1976. It has yet to be converted to another medium - my first attempt in 2001 failed when I found that the drive belt of the projector had perished.

Another problem is data degradation as copies and upgrades on new media are made. Flaws in one generation are copied to the next generation, and new flaws accumulate from generation to generation.

In conclusion, analogue media, are excellent for making initial recordings, for publishing for use by members of the ethnic community, and, if stable media are chosen, for archiving. Furthermore, as I have illustrated above, it is quite feasible to attempt to recover the data on damaged or outdated analogue media. The greatest disadvantage of analogue media is that they are not at all easily manipulated and combined.

4.3 Digital media

A **digital medium** is characterised by the representation of discrete elements of data as binary numbers generated by electronic computer hardware and stored in highly compressed form as off/on blips of some physical characteristic of the medium, such as magnetic field direction or optical transparency. Thus, for example, a map (or any picture) is digitally scanned as thousands or millions of dots of various colours, each of which is represented by a one or more numbers that give information about intensity, hue, etc; letters that make up a text each have a binary character number code; sound waves are sampled at regular intervals and the amplitudes recorded as binary numbers. Digital data are commonly stored on media such as hard and floppy magnetic disks, magnetic tape, and various kinds of optical disk.

4.3.1 Recording

A number of field linguists I know now record text data directly into their laptop computers, rather than paper notebooks. At the time of writing, digital sound recorders, still cameras and video cameras have been reduced in price enough to be an affordable alternative to the equivalent analogue devices for linguists on a reasonable budget. Indeed, it may not be long before some of the analogue devices mentioned above become obsolete.

4.3.2 Data manipulation

A wealth of software now exists for manipulating digital data. For linguists applications such as Shoebox and LinguaLinks have revolutionised lexicography, text analysis and data organisation. Software such as Praat (see Boersma, 2001), Speech Analyzer and FindPhone can assist with phonetic analysis. (See Antworth & Valentine, 1998 & 1999, for information on these and other software.) For picture and video there are a number of commercial and shareware applications.

4.3.3 Integration of media

Of special interest to this thesis is that digital text, sound and visual data are easy to integrate using a variety of software tools. Multimedia teaching materials can be

produced using, for example, Macromedia Director, multimedia presentations using Microsoft Powerpoint, and multimedia web pages using an HTML editor.

4.3.4 Reproduction

Once data is in digital form, reproduction is relatively easy and inexpensive. Furthermore, digital data can be easily converted into a number of typographical formats for publishing in analogue media. In fact, most modern publishing and printing is carried out from material stored in a digital format, and, at the time of writing, the latest photocopiers take digital images rather than analogue. An important difference between analogue and digital data is that quality is not degraded by the copying process.

4.3.5 Access

Today the serious language researcher will prefer digital media, and the integrated interactive multimedia document is much more attractive than a paper one. Many of the minority language speakers, however, live on the other side of the "digital divide" at the moment, and the best current option for distributing literature in language communities in developing countries is still the printed book. This situation could well change as the digital revolution touches these countries too.

Indeed, an examination of internet growth figures is very exciting. Since the internet started, the majority of users of the internet have been from the United States. Since 1998 I have periodically observed the internet usage surveys published by a survey organisation called NUA (Internet Surveys, 2001), and it is interesting to see how the proportions have changed. I have tabulated some of the collected statistics in Table 4.1. The bottom row shows that the proportion of users from Canada and the US dropped below 50% of the total for the first time in 2000, and the right-hand column shows that growth in those countries is now the slowest in the world. It can also be seen that the fastest growing regions are now Asia/Pacific and Africa, the very regions where most of the undocumented languages of the world are spoken.[18]

[18] This trend is not the same in all countries. According to Reuters (July 17, 2000), the Taliban banned internet usage in Afghanistan, although this situation changed during the writing of this thesis.

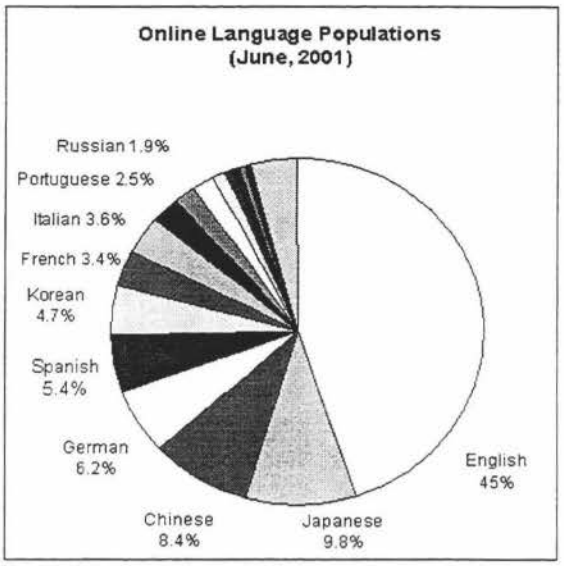
Table 4.1. Numbers of people (millions) using the internet by region of the world

Year / Region	1998	1999	2000	2001	% increase from 1998-2001
Canada/US	70	102	157	167	130%
Europe	33.25	42	94	113	240%
Latin America	4.5	5.3	13	16.5	260%
Middle East	0.75	0.9	2.4	2.4	220%
Africa	0.8	1.1	3.1	3.1	290%
Asia Pacific	22	27	89	105	370%
World	134	179	360	407.1	200%
Proportion Canada/US : World	52%	57%	43%	41%	

Another way to look at the importance of the internet to non-English language communities is to consider the proportion of speakers of different languages now using the internet as a result of these shifts in proportion. Figure 4.2 shows that in 2001 there were more people "online" in primarily non-English speaking countries than in primarily English-speaking countries.

The language used for web page content seems to lag behind the change in usage, but it is also changing. Cairncross (1997, p. 246), observed that "one study, in 1996, found that almost all the scientific material on the Internet was in English; overall, the proportion of English to other languages is around 70 to 80 percent." Another study has shown that this proportion had dropped to 68.4 percent by 2001 (Global Reach, 2001b, Chart of Web Content by Language).

Figure 4.2. Online language populations, from Global Reach (2001a)



Although internet access is presumably available mainly in the cities, this must surely change. For example, in visits to Vanuatu in 1997 and 2000 I was intrigued to see recently-constructed small solar-powered microwave telephone aerials on simple steel poles that linked even very small islands such as Uripiv and Tangoa with the national telecommunications infrastructure. It can be reasonably expected, therefore, that internet access will become available in the foreseeable future. As another example, a colleague of mine on returning from a visit to his home country of India in 2001 related to me how amazed he was to find that internet access was now available even in small villages wherever he travelled.

Nevertheless, digital devices are easy to damage and expensive to repair. It will be a long time before they will reach as many people on a daily basis as a simple and cheaply-produced book. Even here, though, digital data provides an advantage, for it can be converted very simply to a variety of typographical formats for publishing in analogue media.

4.3.6 Archiving

Digital media do have problems in terms of durability or longevity: handwritten parchments, engravings on bone and rock and impressions in clay have lasted for millennia, as have printed books since the time of the invention of printing. However,

even though digital media have only been in use for a few decades, there is already much evidence for the ephemeral nature of the data stored on them. As in the case of analogue image and sound technology, new developments quickly make older methods obsolete, and the recordings made with them unusable. This has serious implications for long-term archiving.

As an illustration from personal experience, in 1982, in an attempt to document the Rumu language, I collected all the manuscript and printed materials relating to the language that I could find, and entered selections of them into a mainframe computer system so that I could edit them and make concordances. Before I left to carry out fieldwork in Papua New Guinea, I backed up all that digital data onto a standard 10.5 inch tape reel, there being no other portable digital storage medium available to me. (In those days we posted our data around the world on these reels wrapped in corrugated cardboard packages.) I expected that it would be accessible in any situation where there was a mainframe computer, and that it would last for years. In 1985 in Papua New Guinea I at last acquired a computer of my own, a TRS-80 laptop. It had a floppy disk drive with a proprietary (Chipmunk) format. Technically, it was beyond me to find ways to transfer the tape data to it, and the paper printouts I had made of the data were more useful to me. Moreover, I learnt that magnetic tape data needs to be recopied about every two years in order to avoid the corruption of data from the magnetic "imprinting" of one layer of tape onto adjacent layers.

This same scenario was nearly repeated five years later, when all my field data on those 365 kilobyte 3.5" floppy disks with a Chipmunk format had to be hastily copied to 720 kilobyte 3.5" disks with a DOS format before my laptop computer's temperamental memory chips failed (as they were wont to do). Unlike magnetic tape data, magnetic disk data does not degrade from imprinting, but it certainly becomes unreadable if the disk surfaces grow mould in a tropical climate. I had to write special utility programs to recover as much data as possible off such disks.

Other people have had similar experiences with transferring their data on 5.25" floppy disks to 3.5" disks or Zip disks in the 1990s, before the hardware to read them became no longer available or serviceable. Now already the 3.5" floppy disk is no longer

standard equipment on some popular computers (e.g. the iMac), and at the time of writing CD writers are becoming the favoured means of backing up data, especially multimedia data.

It is ironic that perhaps one of the very earliest digital media, punched paper tape, will prove to be the most enduring and recoverable digital medium. Even if it is torn it can be mended and data may still be recovered, although it might be difficult to find a teletype machine to read it. As a last resort it can be read character by character by the unaided human eye, as I can personally attest! This is certainly not true for modern magnetic or optical media. Needless to say, the paper tape option is rather ridiculous, as it is slow and bulky, and the technology has been discarded. For now we will have to keep upgrading our digital data until a medium is ever invented that will endure.

There are already a number of organisations which archive language data, many of which have been catalogued by the *Linguistic Data Consortium* (LDC) at the University of Pennsylvania (Bird, n.d.). One example is the *Archive of the Indigenous Languages of Latin America* (AILLA), a project based at the University of Texas at Austin. Their stated aims are:

- to preserve irreplaceable materials, in particular fragile analog audio recordings, that have been made over the course of decades of research by anthropologists and linguists;
- to render these materials accessible to a wider audience, to promote greater understanding of, and further original research using, these unique materials. (Sherzer, n.d., About AILLA page, para. 2)

AILLA have also considered the question of longevity, and aim "to archive these materials as digital records that can be stored and maintained in perpetuity" (ibid.).

In conclusion, while the digital medium has some problems with access and longevity, it has overwhelming advantages over analogue media in the ease with which it may be manipulated and presented in integrated multimedia documents. Digital equipment will also probably become the preferred equipment for recording and reproduction in the near future.

4.4 Summary - preferred media for documentation

I have already argued that language documentation will be for a fairly diverse audience with fairly diverse needs: in the first place, it should be for the language community and their descendents, and in the second place, for researchers in various disciplines with an interest in linguistic data. At this point in history, although most linguists have good access to both digital and analogue media, a large number of target language communities still have only limited access to analogue media, especially books. The analysis of the advantages and disadvantages of analogue and digital media are summarised in Table 4.2:

Table 4.2. Analogue vs. Digital Media

	Analogue	**Digital**	**Future Technology**
Recording	Easy	Easy	Digital
Manipulation	Difficult	Easy	Digital
Multimedia Integration	Slow access	Fast access	Digital
Reproduction	Easy Some degradation	Easy No degradation	From digital to either
Access	Excellent	Limited	Both
Archiving	Paper & engraved media endure Magnetic media degrade Technology becomes obsolete.	Nothing endures Magnetic media degrade Technology becomes obsolete	Both, using managed archiving systems.

The advantages of using digital technology for a multi-purpose multimedia documentation are clear. In order to meet the needs of both categories of audiences with proper priority, I therefore propose the following:

1. the data be recorded using digital devices, or else on analogue devices with conversion to digital data at a later date;

2. the data be assembled on *digital computer media* so that it can be processed and manipulated in many different ways;
3. the primary publication of the integrated multimedia documentation be in *internet-ready digital form* (that is, browsable using a web browser);
4. some thought be given to the convenient reproduction of digital text and picture files on *paper*, sound files on *cassette tape*, and video files on *video tape*;
5. the internet-ready documentation be distributed initially on *compact disk*, so that hard copies can be conveniently made of selected materials by (or for) the language community;
6. the whole or selected parts of the internet-ready documentation be uploaded to a web server for access by the language community and the world-wide academic community over the internet at an appropriate time;[19]
7. both digital and analogue copies of the documentation be archived in at least two separate sites, and also in a site in a different country if there is danger of war or disaster. As the media deteriorate or become obsolescent over the decades, new analogue versions for archiving should be made from digital versions in order to avoid the accumulation of flaws.

4.5 The format of the data

I shall now examine various alternatives for the different data types of an internet-ready documentation: images, sound, video and text.

4.5.1 Images

The use of images is well established in the World Wide Web, where there are three different formats commonly used:

- GIF (Compuserve's Graphic Interface Format) is a popular non-lossy compression format which handles small images of symbols and drawings with flat colours well, and which can be decompressed quickly.

[19] If the internet is not available in the community, the documentation can still be accessed from any computer with a suitable CD drive and internet browser software installed. If computers are not available in the community, then a properly-designed internet-ready CD should also allow easy reproduction of parts of the documentation on analogue media at centres where there are computers.

- JPEG (Joint Photographic Experts Group), a popular format good for large photographs for display on screen. It uses a lossy compression that results in a very low file size, with only a small and smooth apparent loss of picture detail. Pictures intended for archiving or further editing should not be compressed in this way, as they will lose quality each time they are opened, edited and saved again (Chastain, 2002).
- PNG (Portable Network Graphics) is a fairly new format developed after some legal issues connected with a patent arose with the GIF format. Like GIF, it uses non-lossy compression, and like JPEG it can handle coloured photographs well, although the file size may be larger. It is suitable for archiving photographs that are intended for printing or that may need to be edited in the future (LeMay, 2001, pp. 174, 236).

Different kinds of images need to be treated differently. In my experience, the best results are obtained by scanning in colour (or greyscale if there is no colour) at a high resolution, such as 300 dots per inch (dpi). Software such as Adobe Photoshop or ThumbsPlus can then be used to adjust image parameters. If the image is a photograph, the brightness and contrast, and any colour balance problems can be adjusted; if it is a line drawing, statistical "curve" tools can be used to eliminate white-space smudges and darken up the lines, or the lines can be sharpened up using "unsharp mask" or other special tools for that purpose. The image can then be saved at that high resolution in the PNG format. This is the image that is suitable for printing and archiving, and is likely to be around 100-1000 kbyte in size.

After that, the image can be resized to suit a browser window, and the resolution changed down to 72 dpi, which is enough for a clear view on a computer screen. If the image is a photograph, it can be saved in JPEG format using medium or low quality settings. If the image is a drawing, or picture with simple colouring, it can be saved in GIF format. Sometimes conversion to bitmap or another image coding system first gives good results. An image reduced in these ways is likely to be around 10-90 kbyte in size.

4.5.2 Sound

Sound cards are standard equipment with modern computers, and there are a number of commercial programs that will record sound data. Sound files can be very large (in the order of megabytes of data per minute of recording), so this will be significant for internet access to recordings of more than a few seconds. An alternative is that it be delivered as it is played by streaming software on the internet server.

Typically sound files consist of either 8-bit or 16-bit digital values of sound amplitude sampled at rates of 8 kHz (telephone quality), 11 kHz, 22 kHz or 44 kHz (CD quality). The higher rate may be preferred for the best phonetic data, but since the frequencies of the important vowel formants are less than 5 kHz, a sampling rate of 22 kHz will be adequate for most analytic purposes, and the lower rates for "listening along". According to LeMay (2000, p. 437) 16-bit data is always better than 8-bit, which can have some hissing noise.

There are a number of application programs for handling sound data, including some for the linguist, for example, Speech Analyzer (SIL, 2001) and Praat (Boersma, 2001). I have used Praat for recording, editing and analysing phonetic data, and for recording sentence-length text segments, and even discourses lasting several minutes.

There are a number of different sound file formats which a program like Praat can output, and which are normally playable with web browsers on any platform:

- WAVE (suffix .wav) was developed originally as the default format for Windows operating system;
- AIFF (Audio Interchange File Format) was developed originally for the Macintosh operating system. It has a version with compression called AIFC;
- Mu-law (suffix .au) was the format used originally by Sun and NeXT, it is an older and more widely used format with files of smaller size but lower quality than WAVE or AIFF. It would be useable for sound data not needed for careful phonetic study.

Other sound file formats are useful for long sound files to be delivered by streaming. Proprietary software may be needed to produce them:

- MP3, gives "CD quality" with excellent compression;
- RealAudio (suffix .rm), offers small file sizes by compression with a lossy compression algorithm, so the quality is not so good;
- Windows Media Audio (suffix .wma), can be compressed to varying degrees (with varying loss of quality), depending on requirements.

Finally sound may be formatted using just the sound track of one of the video file formats discussed in the next section.

4.5.3 Video

Video files (movies) are created by playing the video from a camera or tape player into the video capture card of a computer while using appropriate software. There are two kinds of capture card: analogue and video. It must also be borne in mind that there are several analogue formats available: PAL, NTSC and SECAM. My own experience is limited to downloading PAL video to an Audio/Visual equipped Macintosh computer using Apple Video Player or Apple QuickTime Pro software.

Raw digital video files (movies) are very large, in the order of megabytes per second for even a small window of video. After editing is complete, they need to be compressed using a **codec** (compression/decompression algorithm) before they can be included in a multimedia documentation. This can take hours for a few minutes of movie. Even when compressed, movies, are not very satisfactory to watch over the internet; a more reasonable expectation is to play them from a CD copy of the documentation.

There are a number of popular formats:

- Apple's QuickTime (suffix .MOV), can be played in Macintosh, Windows and Unix operating systems;
- Video for Windows (suffix .AVI), which is used mainly only on Windows computers (LeMay, 2001, p. 448);
- MPEG Video, which requires expensive hardware for compression;
- RealVideo, a streaming option;

- Windows Media Advanced Streaming Format (ASF), another streaming option.

4.5.4 Text

Computers were largely developed in England and America, and designers in the early 1960s gave thought primarily to representing the English language with 26 letters (just capitals initially) and left-to-right directionality. This has had a lasting effect on the ability of people in non-English-speaking countries all over the world to represent their languages digitally, and for linguists to represent IPA data. Solutions had to be developed to the problems of representing the diacritics used in European languages, the right-to-left directionality for Hebrew and Arabic, and the contextual complexities of Devanagari and Thai, for example. An even more difficult problem was the representation of Chinese with its thousands of different characters.

4.5.4.1 ASCII

One early decision was to allot a hardware memory unit called the octet or byte to represent one character of type. The byte, with 8 binary digits (bits), can handle 2^8 = 256 different codes, so that after reserving some 35 codes for control purposes, there should be coding space for over 200 different characters. In 1963, however, the *Amercian Standards Authority* (ASA) reserved one of the bits for "parity checking" and settled on a 7-bit code standard, ASA X3.4-1963, called ASCII (American Standard Code for Information Interchange) for the remaining 2^7 = 128 possible different code values (Jennings, 1999, ASCII-1963 section; Searle, 1999, Introduction section, para. 5).

ASCII included only the 26 English upper case letters at first, but later that year the *International Organisation for Standardisation* (ISO) and an associated organisation, the *European Computer Manufacturers Association* (ECMA) developed a standard, ECMA-6, which included the lower-case letters and set aside eleven characters (#, $, @, [, \, ^, `, { | } and ~) for "national use" by alphabets with more than 26 letters, and also seven characters for "national graphics", i.e. overtyped diacritics such as the double-quote mark (") for an umlaut (Jennings, 1999, ASCII-1967 section).

In 1967 ECMA-6 was adopted by the ISO as ISO-646, and by the ASA (which had became the *American National Standards Institute*, or ANSI) as X3.4-1967. It is also known as ASCII-1967 or US-ASCII (ibid.).

In my experience, up to the time of writing, ASCII is the *only* standard for text data that works without inconsistencies across all operating systems.

4.5.4.2 ASCII Extensions

There are now three main 8-bit extensions to the ASCII code in which linguists of today (myself included) can find their data encoded: the Extended Character Set (ECS) used in by the DOS operating system of the first IBM Personal Computers, the so-called ANSI character set used by Microsoft Windows, and the Apple Character Set used in Macintosh computers.[20] All three ASCII extensions have the same character allocations for the first 128 codes - this is the ISO-646 standard. The second (or "upper") 128 characters are almost completely different, as can be seen by inspection of Table 4.3.

4.5.4.3 ANSI

Of the three extensions to ASCII, the ANSI set has been adopted as an international standard, ISO-8859, for use on the internet. ISO-8859 is actually a family of standards designed to encode many different alphabets; for example, ISO-8859-1, or Latin-1, has been adopted as the alphabetical code for English and Western European languages, ISO-8859-2 is the standard for Hungarian and other Eastern European languages, ISO-8859-5 to -8 are for Cyrillic, Arabic, Greek and Hebrew alphabets respectively. (The first 128 codes are still the ASCII set with the Roman alphabet, and the non-roman alphabet is assigned to the second 128 codes.) (Searle, 1999, Accented Latin section).

[20] The ISO 2022 standard was belatedly established in an attempt to give some order to the allocation of the 7- and 8-bit encodings to other characters (Searle, 1999, Accented Latin section).

Table 4.3. The upper 128 characters of some common 8-bit ASCII-based codes

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
128																
ECS	Ç	ü	é	â	ä	à	å	ç	ê	ë	è	ï	î	ì	Ä	Å
Mac	Ä	Å	Ç	É	Ñ	Ö	Ü	á	à	â	ä	ã	å	ç	é	è
ANSI	€		‚	ƒ	„	…	†	‡	^	‰	Š	‹	Œ		Ž	
144																
ECS	É	æ	Æ	ô	ö	ò	û	ù	ÿ	Ö	Ü	¢	£	¥	₧	ƒ
Mac	ê	ë	í	ì	î	ï	ñ	ó	ò	ô	ö	õ	ú	ù	û	ü
ANSI		`	´	¨	¨		–	—	˜	TM	š	›	œ		ž	Ÿ
160																
ECS	á	í	ó	ú	ñ	Ñ	ª	º	¿	⌐	¬	1/2	1/4	¡	«	»
Mac	÷	°	¢	£	§	•	¶	ß	®	©	TM	´	¨	≠	Æ	Ø
ANSI		¡	¢	£	¤	¥	¦	§	¨	©	ª	«	¬		®	¯
186																
ECS	▒	▒	▓	│	┤	╡	╢	╖	╕	╣	║	╗	╝	╜	╛	┐
Mac	∞	±	≤	≥	¥	µ	∂	∑	∏	π	∫	ª	º	Ω	æ	ø
ANSI	°	±	2	3	´	µ	¶	·	¸	1	º	»	1/4	1/2	3/4	¿
192																
ECS	└	┴	┬	├	─	┼	╞	╟	╚	╔	╩	╦	╠	═	╬	╧
Mac	¿	¡	¬	√	ƒ	≈	Δ	«	»	…		À	Ã	Õ	Œ	œ
ANSI	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
208																
ECS	╨	╤	╥	╙	╘	'	╒	╫	╪	┘	┌	■	▄	▌	▐	▀
Mac	–	—	"	"	'	'	÷	◊	ÿ	Ÿ	⁄	€	‹	›	fi	fl
ANSI	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
224																
ECS	α	β	Γ	π	Σ	σ	µ	τ	Φ	Θ	Ω	δ	∞	φ	ε	∩
Mac	‡	·	‚	„	‰	Â	Ê	Á	Ë	È	Í	Î	Ï	Ì	Ó	Ô
ANSI	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
240																
ECS	≡	±	≥	≤	⌠	⌡	÷	≈	°	∙	·	√	n	2		
Mac		Ò	Ú	Û	Ù	ı	ˆ	˜	¯	˘	˙	°	¸	˝	˛	ˇ
ANSI	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

4.5.4.4 IPA

Of great concern to linguists, the encoding and display of symbols of the International Phonetic Alphabet on computers has been problematic, and many solutions have been attempted. For example, in the days of 9-pin and 24-pin dot-matrix printers my colleagues and I in Papua New Guinea would design our own character

matrices for the IPA symbols (and other orthographic oddities) that we required, and then assign them to unused character codes. In the 1990s SIL made a TrueType IPA font available for use on both Windows and Macintosh computers (SIL, 2001, Encore IPA Fonts page). It uses one or two or more character codes to encode each symbol. A colleague and I who work on both Windows and Macintosh computers both found, however, that a text composed using the font on one kind of computer would not display properly on the other kind. It was not until 2001, with upgrades to operating systems that this problem resolved itself.

Another option for linguists has been to use various ASCII characters or character combinations to represent IPA symbols. One such system is the Speech Assessment Methods Phonetic Alphabet (SAMPA). The system is explained in Wells (2000). SAMPA assigns various ASCII symbols to the representation of IPA symbols that they somewhat resemble, for example, B for β (beta), E for ε (epsilon), and T for θ (theta).

4.5.4.5 Unicode

The use of two bytes per character theoretically allows codes for 2^{16} = 65,536 characters. Various systems using two or more bytes have been developed to cope with languages with large symbol sets, especially the thousands of Chinese, Japanese and Korean (CJK) characters. For example the JIS X 0208-1990 is a standard widely used today which includes ASCII together with the Japanese katakana and hiragana syllabaries, over 6,000 kanji characters, and also the Greek and Cyrillic alphabets. Others standards used for Chinese are GB 2312-80, HZ, EUC-CN and Big-5. (Searle, 1999, Chinese, Japanese and Korean Character Codes section).

Since 1988 the *Unicode Consortium* has set out to devise a coding system that could potentially cope with all the world's writing systems in a unified way under the motto "...a unique number for every character, no matter what the platform, no matter what the program, no matter what the language." (Unicode, n.d., para. 1). Since 1992 the consortium have been working in conjunction with the ISO, whose standard, ISO-10646, matches the Unicode standard.

Unicode uses a two-byte (16-bit) code, allowing for the encoding of over 65,000 different symbols. (More bytes can be used to extend it even further.) At the time of writing, number values have been assigned to 49,194 characters from the world's scripts, including all of the 27,484 Han character repertoire used in China, Japan, Korea, Taiwan, Singapore and Vietnam (Unicode, 2002a, sec. 1.1). Unicode is still based on ASCII, and when the first 8 of the 16 bits are set to zero, it is essentially ASCII (as ISO-8859-1). Table 4.4 shows the 8-bit binary codes of some ASCII characters on the left, and, on the right, some 16-bit codes of some of those same characters in Unicode, and also characters from Arabic, Chinese and IPA subsets of Unicode (see Unicode, 2002b). The 8- and 16-bit codes shown are divided into groups of 4 bits for ease of reading.

Table 4.4. Examples of ASCII and Unicode

ASCII/ISO-8859-1		Unicode		
A	0100 0001	A	0000 0000	0100 0001
B	0100 0010	B	0000 0000	0100 0010
C	0100 0011	C	0000 0000	0100 0010
1	0001 0001	1	0000 0000	0001 0001
2	0001 0010	2	0000 0000	0001 0010
3	0001 0011	3	0000 0000	0001 0011
	0010 0000	ش	0000 0110	0011 0100
A	0100 0001	ج	0000 0110	0010 1100
S	0101 0011		0000 0000	0010 0000
C	0100 0011	僩	0011 0100	0110 1101
I	0100 1001	善	0011 0101	1001 0110
I	0100 1001		0000 0000	0010 0000
	0010 0000	ɧ	0000 0010	0110 0111
	0010 0000	ʤ	0000 0010	1010 0100

Disadvantages of Unicode are that (a) text files in languages with alphabetical orthographies will be twice the size of files using an 8-bit code; (b) a comprehensive Unicode font would be enormously large; (c) fonts for various subsets of Unicode that work on all platforms are still being developed. (For instance, a font such as Lucida

Sans Unicode that uses the Unicode IPA encodings has been developed for Windows - see Wells, 2000 - but there is nothing at the time of writing for Macintosh.)

These disadvantages can be worked around: (a) file size can be reduced using a transformation called UTF-8 that drops the first unused byte when the second byte contains a plain 7-bit ASCII code;[21] (b) it is unlikely that a document will need more than a few subsets of Unicode, and fonts encoding those subsets can be made available with the documentation; (c) "image fonts" always work across all platforms, even if they are a rather cumbersome solution. This is a system where each character has its own tiny image file, usually GIF format; the text can then be transformed (using search-and-replace software) to display the GIF files instead of the characters. Unfortunately, if the fonts require bold and italic renderings or different sizes, then separate sets of images need to be made for each requirement. I use image fonts (one size, one style) to display IPA symbols and also Rumu text with full tone diacritic markings in the sample documentation in this thesis.

4.5.4.6 Text on the internet

The Unicode system, as ISO-10646, has been adopted by the World Wide Web Consortium for use with the markup languages HTML and XML, and a number of modern web browsers are now "Unicode-aware."

Because of the official recognition and support of Unicode as a system that will be able to handle all the orthographies of the world's languages, it is sensible to recommend that Unicode be used for language documentation. However, as there are other coding systems that have been developed for special purpose fonts, language documentalists must either (a) convert them to Unicode, (b) ensure that all the software/fonts for interpreting and displaying them is included in the documentation, (c) use images fonts, (d) use images of blocks of text in a peculiar font, or (e) use the document or page

[21] If the first byte is used, or the high-order bit in the second byte is used, the code value is transformed in a more complicated way into a 16-bit or 24-bit value, so there is no shortening of the code for "extended" ASCII. For example, the letter A has a Unicode value of 00000000 0<u>1</u>000001. Using UTF-8 it would be reduced to 01000001 because the first byte is all zeros, and the high-order bit (underlined) of the second byte is zero. On the other hand, the big letter Ash (A-E ligature) has a Unicode value 00000000 1<u>1</u>000110, and although the first byte is all zeros, the high-order bit of the second byte is 1, so it would not be shortened (Unicode, 1999).

imaging PDF format that works across all computer operating systems, but needs proprietary software, Acrobat Reader, to display it.

In the future, Unicode will be the better option for display on the internet than special fonts, because Unicode is independent of computer system, while fonts and software tend to be system-dependent. Nevertheless, since Unicode is still being developed, the use of images may be the best option for some character sets in the meantime. The use of PDF is good for publishing material now, but has the disadvantage that it is a proprietary format that may become obsolete in the future. It should be noted that the e-journal, Linguistic Discovery (which was launched while this thesis was being written), publishes documents as both PDF files and HTML files with image fonts (Grenoble & Whaley, 2002a).

5. THE LOGICAL STRUCTURING OF THE DATA

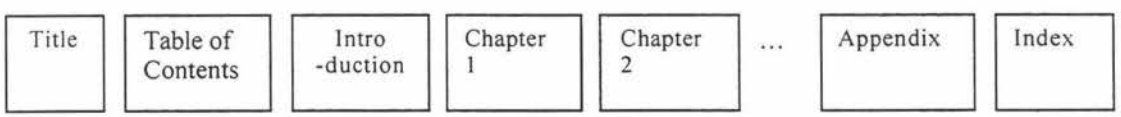
In the last chapter the focus was on how the *data* for text, sound and images could be encoded and recorded physically. In this chapter the focus will be on *information* – how "the data [can be] shaped by humans into a meaningful and useful form" (repeating Laudon & Laudon's definition given in section 3.1). In particular, this chapter looks at the two main aspects of how documentation consisting of an incorporation of sound, image and text data could be structured logically: both the structure of content within and between documents, and also the filing of those documents within the documentation as a whole. Finally, various methods of marking up content structure are evaluated.

5.1 Content structure

Structure can be thought of as units linked together into something larger. For a document, the units are blocks of text, which Landow and Delany (2001, p. 213), referring to the work of Barthes, term **lexia**. The most well-recognised lexia are the word, the sentence and the book. For a multimedia document, non-text objects, such as sound-bites and graphics, must also be included as units of structure, all of which can be referred to simply as **information blocks**.

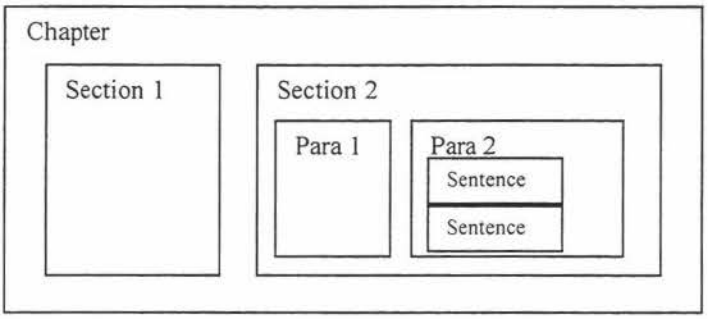
The structure of a book is largely a combination of linear and hierarchical structures: there is a linear sequence of title page, table of contents, chapters, appendices and an index, as depicted in Figure 5.1. The sequence has a definite beginning and end - it is bounded.

Figure 5.1. Linear structure



There is also a hierarchical structure of chapters, sections, paragraphs, sentences, and words, as depicted in Figure 5.2.

Figure 5.2. Hierarchical structure



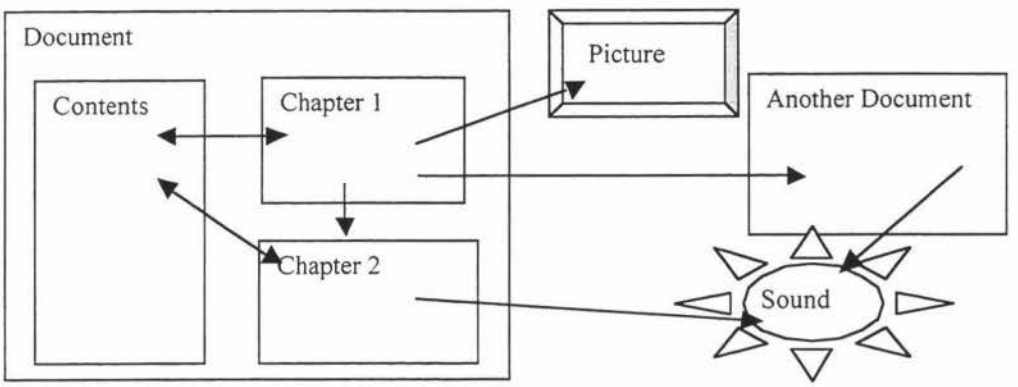
Furthermore, once the book is accepted for publication, it cannot be altered for that edition - it is fixed.

Landow and Delany (2001, p. 208) observe that generations of scholars and authors have internalised these linear and hierarchical structures with their limitations as the rules of thought. Berners-Lee (2001), in his proposal of 1989 for the design of what was to become the World Wide Web, points out these misconceptions about information in strong terms, blaming them for the problem of information loss in a complex scientific organisation. He maintains that "the method of storage must not place its own restraints on the information" (p. 193). Some information will not fit well into a hierarchical or "tree" structure, and even if it does, such a structure is either difficult to rearrange or else involves some data redundancy (i.e. extra copies of the same data). Data redundancy involves not only extra storage space, but, more importantly, can lead to inconsistencies in the data, where one copy of an information block is updated or revised in some way, but the other copies remain unchanged and are therefore out-of-date.

As an example of multiple uses for the same data from a language documentation, a description of marriage rites might fit into an article on rites of passage, into another article on kinship, and also serve as encyclopedic information for entries on "marriage", "husband" and "wife" in a lexicon, and as a reference from a traditional story involving a wedding. As an answer to this complex inter-relatedness of data, Berners-Lee advocates a structure consisting of a "web of notes with links (like references) between them" (ibid.).

Documents with a web (or network) structure are commonly known as **hypertext**,[22] which Berners-Lee defines as "human readable information linked together in an unconstrained way" (ibid. p.194). For such documents that are not constrained to text, but also include picture and sound, he uses the term **hypermedia**. Figure 5.3 shows illustrates such a hypermedia structure.

Figure 5.3. Hypermedia structure



It should be pointed out that traditional printed documents also have some link structures: tables of contents, footnotes, indices and inline references to other sections or other works. Hypertext, as it is understood today, includes these linkages, but more importantly it involves instant and dynamic linkages to related information that is available at the click of a button, and that may be different between recalls because of author updates.

5.1.1 Filing structure

So far I have discussed the logical or semantic aspect of structure of the information within and between documents. Now I shall turn to the organisational structure of the documentation, that is the arrangement of the documents comprising it. The normal arrangement of documents is once again hierarchical, whether in a filing cabinet or a computer filing system. Individual documents (files) are stored in a hierarchy of nested folders. This arrangement is very satisfactory for many kinds of documents. It seems reasonable for all the files connected with a documented communication event to be

[22] A term coined in the 1960s by Ted Nelson (Berners-Lee, 2001, p.194).

stored in one folder - the raw transcription, and perhaps an edited version for printing, a translation, an interlinearised version, commentaries, and an audio or video tape.

A problem arises when an item seems to belong to more than one folder. For example, in a traditional filing system, an audio tape may have recordings for seven or eight stories on it. The tape could easily get lost if it was stored in the folder of one of those stories. In a computer filing system, an image file of a man cutting leaves for thatch could be referred to by a description of house-building and also by a bibliographic reference to the man as the teller of a certain story. A hierarchical system would demand that two copies be made of the file, one for each document, giving rise to the data redundancy problem referred to in the previous section.

An alternative arrangement is the combination hierarchy-network model, where a careful note of the location (folder address) of the image file (to continue the last example) could be noted in any other referring documents. There are two problems with this model: (a) remembering which particular folder from a number of appropriate folders contains the original document, and (b) the invalidation of the address if the folders containing it are rearranged.

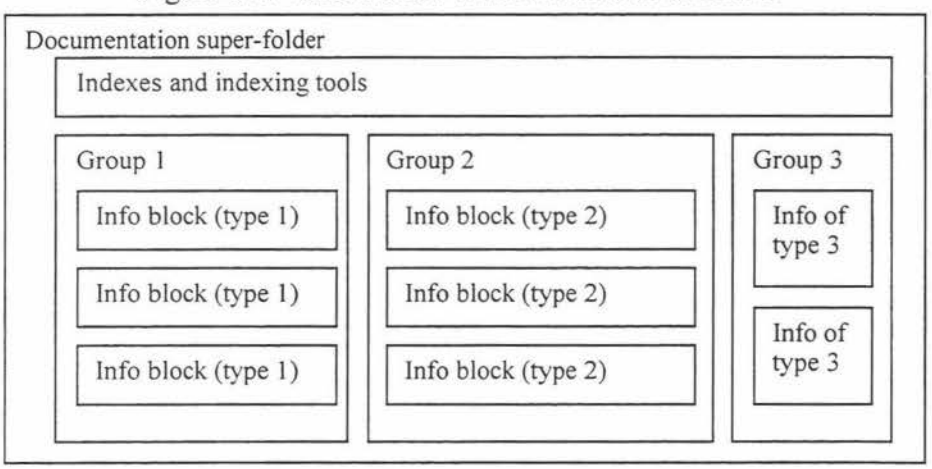
A good solution to these problems comes from modern database theory.[23] There are three well-recognised models for databases: **hierarchical**, **network** and **relational**. In the first two models, records are arranged in ways analogous to the hierarchical and web-like arrangements of information blocks in internal and external document structure discussed above. The relational model, on the other hand, involves grouping together all records of the same internal structure into a table, and then using powerful indexing tools to show the logical relationships between different tables. The relational model is very successful and is the favoured model for modern database systems.

By analogy with the relational database model, all the files/documents/information blocks of the same format type or structure type could be grouped together into one folder, so there would be a folder of image files, another of sound files, another of transliterated texts, and so on. All the folders would be in the same all-embracing

[23] A succint summary can be found, for example, in Laudon and Laudon (1998, pp. 219-226).

documentation super-folder, and there would then be no need to rearrange the files any further according to logical considerations. Logical connections between files would be carried out by the creation of links to "sister" files in the same folder or "cousin" files in the other folders. The hierarchical and semantic arrangement of documents in the documentation would be shown by the generation of a series of table-of-contents files or indexes (which are also a kind of information block that would all be stored in their own folder.)

Figure 5.4. Information with a relational structure



Such a structure means that the indexes are effectively meta-texts. Other meta-texts in the form of entirely new arrangements of the old material would then be possible by simply developing new indexes. Moreover, a new use of such material would be achieved by a simple reference to its location in an information block grouping (folder), thus avoiding data redundancy.

In summary, for a multimedia documentation it is appropriate to structure information blocks as hypermedia, using a combination of linear and hierarchical structures interconnected by a web of linkages. The whole documentation is organised by grouping together information blocks of a similar kind into a small number of folders, so that each information block will remain in a single constant location, facilitating multiple uses through hyper-linkages that remain valid.

5.2 Marking up content structure

Having argued for a documentation that has its primary implementation on digital media and with a relational-type filing structure and a hypermedia-type content

structure, it is now necessary to examine how this could be implemented. I have also argued for a use of plain ASCII/Unicode text files rather than proprietary formats with special binary encodings. There are two families of ASCII text-based markup methods that should be considered: the Standard Format (SF) markup systems, and the markup languages based on Standard Generalised Markup Language (SGML).

SGML has a number of important derivatives relevant to this thesis: HTML used in the World Wide Web, TEI used in text archiving, and the more recently developed XML and XHTML. SF is widely used by SIL linguists for documents and databases for many purposes, especially linguistics, anthropology, and translation; RSF is an interesting development of SF that makes it more like SGML markup.

It is very important at this point to note that the term **markup** has traditionally been used to refer to typographical instructions - font types, sizes and styles, margin management, etc, and both SF and HTML have a tradition of use as typographical markup languages. Just as has been said about data storage (section 5.1), the method of display must not place its own restraints on the information. It is the markup of the logical structure of the content that is the focus of this section, not the display format.

I shall now discuss the usefulness of each markup language for a language documentation project. I will do this by starting with the most widely known and popular markup language, HTML.

5.2.1 HTML

When what was to become the World Wide Web was developed by Berners-Lee in 1990, he decided to implement a hypertext solution to the documentation problems he experienced at CERN, a famous nuclear particle research laboratory in Switzerland. To do this he chose to structure the text by inserting mnemonic markup tags into the text itself, and defined a set of these tags following the specification of the general markup language, SGML (see next section). He called his tag set and their definitions HyperText Markup Language (HTML), and wrote browser software to interpret the tags and format the documents in a very readable way. (That browser software was later developed by others into the famous web-browsers, Mosaic and Netscape.)

HTML works very well for the usual kind of document that has a logical structure consisting of a hierarchy of chapters, sections and paragraphs. With HTML this hierarchy is indicated by the use of a hierarchy of headings, just as it is within a modern word processor. Within the hierarchical structure other sub-structures may also be included, such as bulleted and numbered lists and tables (which may also have a hierarchical structure), and embedded sounds and images. Most importantly, HTML documents allow explicit linkages to be made with other documents or parts of the same document, giving the documentation its famous hypertext structure.

Figure 5.5 exemplifies an HTML encoding of a document (including headings of various levels, a list, and some links) along with an example of how it might appear when displayed in a web browser.

Figure 5.5. HTML and web browser display of the document fragment

```
<html>
<body>
<h1>Rumu Phonology</h1>
   <a name="Cont"></a>
   <h2>Contents</h2>
      <ul>
         <li>Introduction</li>
         <li><a href="#OView">Overview</a
         . . .

      </ul>
   <a name="Intro"></a>
   <h2>Introduction</h2>
      <p>Rumu is spoken by 700-800 people
   <a name="Oview"></a>
   <h2>Overview</h2>
      <h3>Phonemes</h3>
         <p>Rumu has nine consonants and
      <p><a href="#Cont">[TOP]</a></p>
```

Rumu Phonology

Contents

- Introduction
- Overview . . .

Introduction

Rumu is spoken by 700-800 people ...

Overview

Phonemes

Rumu has nine consonants and ...

[TOP]

The HTML tags are enclosed in angle brackets. The tags used in this example are:

<html> identifies the markup

<body> the main part of the document (after the header)

<h1> heading level 1

<h2> heading level 2

<h3> heading level 3

	paragraph
	un-ordered list
	list element
<a>	anchor for a hypermedia link (or "hyperlink")

The tags are often applied in pairs that enclose a block of text, and the name of the closing tag is always preceded by a forward slash (/). For example, the closing tag of the paragraph tag, * *, is * *.

While HTML can indicate the logical structure of an ordinary text document, for the linguist, an important weakness is that it cannot show well the logical structure of other documents that are produced by linguists, especially dictionaries and interlinear texts. Certainly it can be used to format these documents for display, and an example of this can be seen in Figure 5.6. Some tags are specifically designed for glossaries, e.g. the *<dt>* tag is a definition-*term* tag, and the *<dd>* tag is a definition-*definition* tag. Other fields of a dictionary record must use standard list and style tags which provide only typographical information about the entry.[24]

Figure 5.6. A dictionary entry marked up using HTML

```
<dl>
  <dt>
    <b>ahini keka </b> <i>phr-v</i>
  </dt>
  <dd>
    <i>1 </i>to sting, smart
    <ul>
      <li>Aitini nane te yaratima po
          ahini keka po.
       <i>When iodine is put on a sore
          it stings.</i>
      </li>
  </dd>
  <dd>
    <i>2 </i>to speak sharply to,
          disdain, despise
    . . .
  </dd>
</dl>
```

ahini keka *phr-v*

1 to sting, smart

- Aitini nane te yaratima po ahini keka po. *When iodine is put on a sore it stings.*

2 to speak sharply to, disdain, despise

[24] Some practitioners user the *class* attribute within the ** and * * tags to indicate logical content structure.

For a more flexible markup system we must return to HTML's roots: SGML.

5.2.2 SGML

SGML (Standard Generalized Markup Language) was developed by a team led by Charles F. Goldfarb based on earlier work he had carried out for IBM with Edward Mosher and Raymond Lorie between 1969 and 1973 when they invented a markup language called GML (an acronym based on the names of its inventors as well as its official name, Generalized Markup Language). There were two main motivations for this invention: (a) "that markup would be useful for more than one application or computer system," and (b) that markup be restricted "to identification of the document's structure and other attributes", i.e. the separation of *content* markup from media- and software-dependent *formatting* instructions (Goldfarb, 1996, section 2, para. 15). Goldfarb's team added new concepts to GML, such as "short references" and "link processes", and developed SGML as ISO-8879 by 1986 (SGML Users' Group, 1990, sec. 2, para. 4).

Unlike HTML, SGML allows authors to invent their own tags for marking up document structure. For example, with HTML a chapter is only *implicitly* indicated by the heading tag applied to the chapter title. With SGML, a chapter or other block of information can be bracketed by its own tags. For example, the introductory chapter in Figure 5.5 could be marked up:

```
<chapter>
  <title>Introduction</title>
  <para>Rumu is spoken by 700-800 people in an area of rain forest and sago swamps...
  <para>Rumu has nine consonants and seven vowels...
</chapter>
```

(Note that SGML tag labels are usually (but not necessarily) inside angle brackets, and that, as in HTML, the tags may come in pairs, where, for example <*chapter*> is an opening tag, and its corresponding closing tag is </*chapter*>.)

For a linguist, there is a freedom to develop meaningful tags for grammatical analysis of texts or for dictionary databases. For example, the dictionary entry displayed in Figure 5.6 could be structured logically as follows:

```
<entry>
 <headword>ahini keka</headword>
 <ps>phr-v</ps>
 <senseset>
  <sense n=1>
   <def>to sting, smart
   <xv>Aitini nane te yaratimapo ahini keka po.
   <xe>When iodine is put on a sore it stings.
  </sense>
  <sense n=2>
   <def>to speak sharply to, despise
   . . .
</entry>
```

While there is a freedom for authors to invent their own tag labels based on the semantic structure of a document, the structure has to be carefully defined. The definitions are kept in a separate section of the document, or even a separate document altogether, called a Document Type Definition (DTD). The declaration of a tag includes a list of other tagged elements that can be included within the element bracketed by the tag. For example, a simple DTD for a book could look like this:

```
<!DOCTYPE Book [
. . .
    <!ELEMENT chapter  - - (title?, (para | table) *)>
. . .
]
```

The parentheses around *title*, *para* and *table* indicate that these are the kinds of element that can occur within a chapter; the question mark after *title* indicates that the title is optional; the comma between *title* and the other elements indicates that the other elements come after the title; the vertical bar separating *para* and *table* indicates that either element may come after a title, and the asterisk after the brackets indicates that any number of paragraphs or tables can occur in a chapter.

SGML has another important advantage over HTML in that it allows more than one structure to be applied to a document. For the linguist this means, for example, that a document can be tagged for both syntactic and semantic structures.

Once the structure(s) of a document type have been defined in a DTD, the formatting of an SGML document for display purposes needs to be described in style sheets written using Document Style Semantics and Specification Language (DSSSL).

For example, the format for the *para* tag of the chapter example given above could be defined in a style like this:

```
(element para (make paragraph
    font-family-name: "Times New Roman"
    font-size: 12pt
    line-spacing: 13pt
    space-before: 6pt
    start-indent: 6pt
    quadding: start))
```

DSSSL style sheets can be processed by software such as Jade or Seng (see Clark, n.d.), but most Web browsers at the time of writing do not handle them.

The reason that HTML is so easy to use is that it is an SGML application whose DTD and style sheet have both already been specified. Web browsers use these specifications when interpreting and formatting web pages.

While SGML is very useful for marking up certain kinds of language document, the disadvantages of using SGML instead of HTML for language documentation are (a) a DTD will need to be written; (b) SGML documents are not pretty to look at, and software applications that can interpret DSSSL style sheets are not universally available, while HTML documents are formatted automatically by a web browser; (c) SGML has a lot more features and options than a ordinary working linguist is likely to need, and applying them efficiently takes considerable technical expertise; (d) because of all those features and options, developing applications to process or display an SGML document is relatively complicated.

A response to the first problem is to use a well-defined SGML application developed for language documentation, such as TEI. A restricted subset of SGML called XML has been invented in response to the other problems.

5.2.3 TEI

The Text Encoding Initiative (TEI) is a project that came out of a meeting of humanities scholars in 1987 to develop guidelines for the digital encoding of texts used in humanities research so that they can be exchanged between institutions and processed using different software applications. The initiative decided to use SGML for the encoding, and have written a modular DTD so that a document DTD can be built up in

complexity as required. Some 400 elements (tags) have been defined that can be used for linking, cross-referencing, analysis and interpretation of texts such as prose, verse, drama, spoken texts and dictionaries.[25]

A TEI encoded text is preceded by an extensive header containing information about the title, authorship, source, languages, writing systems, speaker profiles, publication or distribution, editorial principles, and revision history of the text. In the text section, extensive use is made of the *div* tag to analyse the text into various units which are further specified using SGML attributes, e.g. * *.

The TEI is a useful archival format for language texts when there is a lot of cataloguing and structural information to be included, and when the texts are being worked on by many scholars over a long time, because there is an excellent provision to record this information in the header. It is used by the Oxford Text Archive, the British National Corpus and the Humanities Text Initiative, amongst others. TEI encoded documents have to be converted to HTML before they can be displayed on the Web.

In a workshop on web-based language documentation in 2000, Hockey did concede, however, that "people who do not have expertise in knowledge organization or library and information studies have not found headers easy to deal with" (Hockey, 2000, sec. 4), and while the header is intended to be both human readable and machine processable, "it has several problems which tend to place it somewhere between these two" (ibid.).

5.2.4 XML

XML (Extensible Markup Language) is a variety of SGML designed to be "straightforwardly usable over the Internet" (Bray & Sperberg-McQueen, November 14, 1996, section 1.1, point 1) whose specification was first proposed by a working group of the World Wide Web Consortium in 1996 (W3C, 2002, Current Situation section, para. 1). Some esoteric features of SGML and have been omitted and there are strict limitations to its syntax. For example, while tag and attribute names may still be defined by the author, the names are case-sensitive, all tagged items must be explicitly closed, and if attributes are included in a tag then both name and value must be included and the

[25] See Burnard & Sperberg-McQueen, (1993, What is TEI? section), and Hockey (n.d.).

values must be a literal (i.e. in quote marks). For example, in Figure 5.7 the SGML dictionary entry example given in section 5.2.2 is repeated on the left, and an XML equivalent is given on the right with the differences underlined:

Figure 5.7. A comparison of fragments of SGML and XML

```
<entry>
 <headword>ahini keka</headword>
 <ps>phr-v</ps>
 <senseset>
  <sense n=1>
   <def>to sting, smart
   <xv>Aitini nane te yaratimapo
        ahini keka po.
   <xe>When iodine is put on a sore
        it stings.
  </sense>
  <sense n=2>
   <def>to speak sharply to, despise
   . . .
</entry>
```

```
<entry>
 <headword>ahini keka</headword>
 <ps>phr-v</ps>
 <senseset>
  <sense n="1">
   <def>to sting, smart</def>
   <xv>Aitini nane te yaratimapo
        ahini keka po.</xv>
   <xe>When iodine is put on a sore
        it stings.</xe>
  </sense>
  <sense n="2">
   <def>to speak sharply to, despise</def>
   . . .
</entry>
```

Stylesheets for XML documents are written using a subset of DSSSL called Extensible Stylesheet Language (XSL), and one Web browser, Internet Explorer 5, can use these to transform some XML documents into HTML for display.

The restrictions on the syntax of XML have the advantages of ensuring consistency of markup, and making it much easier to write software to parse and process the marked up data. Furthermore XML also supports rich linking structures borrowed from TEI.

Another advantage of XML is that it is not necessary to undertake the rather technical tasks of writing a DTD or a style sheet for a document to be a valid document. Even those tasks can now be performed automatically by software tools now available, such as XML Spy for Windows (see Altova, 2002) or Emilé for the Macintosh (see Media Design in-Progress, 1999).

These advantages of XML over both SGML and HTML have made XML very attractive to linguists concerned with the documentation and analysis of languages (e.g. Bell & Bird, 2000; Bird & Simons, 2000; Jacobson & Michailovsky, 2000; Ide & Romary, 2000; Rempt, 1999). It is to be expected that XML standards (defined by DTDs and equipped with stylesheets) will be become established for various forms of

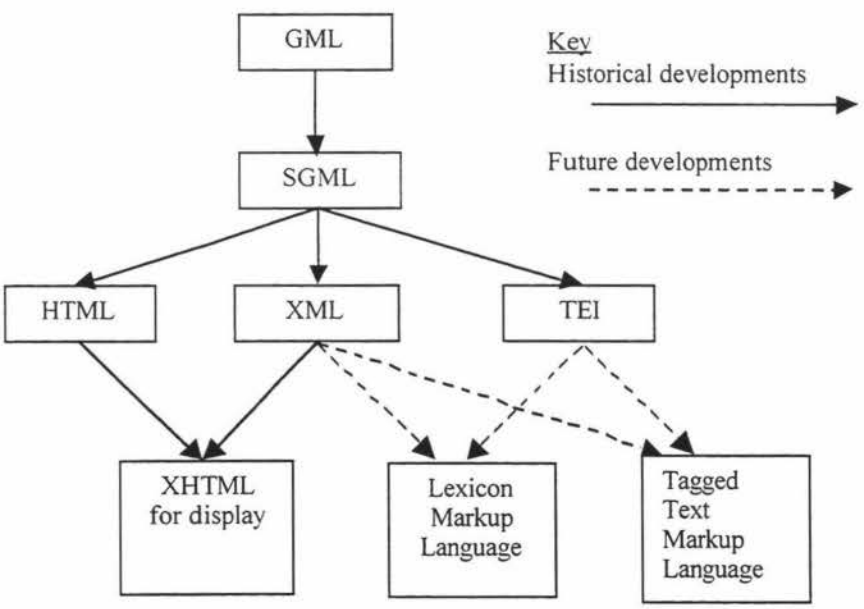
linguistic data, and linguistic software tools that produce and interpret XML documents will become common.

5.2.5 XHTML

Although XML documents are not yet (at the time of writing) easily prepared for viewing on web browsers, there is a version of HTML that conforms to XML, called XHTML (Extensible Hypertext Markup Language). XHTML will display in a web browser, and can be checked for syntax consistency, should their authors want them processed by other software. They conform to the XML specification with the restrictions that all tag and attribute names are in lower case, all attribute values are in quotes, and all tags must have closing tags (or else, in the case of empty tags, be self-closing, e.g. HTML's line break tag <*br*> in XHTML is <*br* />). This means that it will be much easier to write parsing software to process XHTML than HTML files.

The relationships between the markup languages discussed so far are diagrammed in Figure 5.8:

Figure 5.8. Relationships between selected members of the SGML family



5.2.6 SF

Standard Format (SF) uses tags that consist of a backslash (\) followed by a user-defined label of up to 5 characters. This form of markup has been used by SIL linguists for data management and printing for several decades, and is the markup language used in programs such as Manuscripter, Shoebox and the Text Analysis software package. The dictionary entry of Figure 5.6 as developed in Shoebox would appear as:

```
\lx ahini keka
\_no 00006
 \ps phr-v
   \sn 1
     \de to sting, smart
     \xv Aitini nane te yaratimapo ahini keka po.
     \xe When iodine is put on a sore it stings.
   \sn 2
     \de to speak sharply to, despise
     . . .
\lx ...
```

Data marked up with SF has a major limitation in that any hierarchical structure of the data is not marked explicitly within the data.[26] Instead, the program interpreting the data imposes that structure on the data. In the case of a dictionary database, for example, the user can specify, in the Shoebox program, that the *\lx* tag, which marks the beginning of a lexical data item should also define the beginning of a whole dictionary entry, and that the English definition tag, *\de*, comes under this tag. Shoebox then stores this information in a separate file, and uses it to access and display the data correctly.

5.2.7 RSF

With the more recently developed program, LinguaLinks, dictionary data for importing is marked up unambigously using what is called Revised Standard Format (RSF). With RSF both the beginning and the end of data is marked up using opening and closing tags in a manner similar to XML. As with XML, the closing tag is simply a modification of the opening tag. For example, a major dictionary entry could be enclosed in *\major* and *\-major* tags, as follows (adapted from LinguaLinks Ver 4.0r Help at *Revised Standard Format*):

---

[26] A certain amount of low-level hierarchy can be indicated when embedding certain kinds of material, such as stretches of text in a different language or font style, using tags based on braces, e.g. in *abc*\b{*def*}*ghi*, the text segment *def* is marked to be in bold, while surrounding letters are not.

```
\major
 \base ahini keka
 \pos phr-v
 \sense
  \def <ENG>to sting, smart
  \examp
   \vern Aitini nane te yaratimapo ahini keka po.
   \trans <ENG> When iodine is put on a sore it stings.
  \-examp
 \-sense
 \sense
  \def <ENG> to speak sharply to, despise
    . . .
\-major
```

5.3 Conclusion

In summary, a documentation using a text-based markup will be the most widely accessible and will be the most likely to endure. A document structure that retains maximal information about the data will be easiest for its authors to manage and update, and for users to make sophisticated queries of. Of the four markup languages useful for structuring linguistic data (TEI, XML, SF and RSF) XML is likely to be the preferred language of the future. For an internet-ready documentation that is browsable by a wide range of users, such as is the focus of this thesis, the markup needs to be converted to HTML, or, preferably, XHTML. In the future this will become straightforward for XML documents through the use of XSL transformations.

If language texts are in plain text format it is a simple matter to put them into XHTML format, but if they use special character fonts then characters will need to be converted to Unicode or image fonts using search and replace functions in a word processor, or using special-purpose programs. Linguistic descriptions and analyses that are authored in a number of modern word processors may now be saved in XHTML format. The documents that will give the most problems are tagged texts, dictionaries and databases. These can be displayed successfully in XHTML, and examples of interlinear text (a kind of tagged text) and dictionaries are given in chapter 7 and appendix B.

6. THE INSTRUMENT

In this chapter I set out a proposal for language documentation that integrates the principles and options considered in earlier chapters. The proposal integrates the types of data needed to make up a comprehensive documentation and the most useful and endurable formats available today for each of those data types.

The proposal is for a language documentation that consists of an organised collection of many documents of diverse natures linked together in the nature of a web site, and viewable using standard web browsers. It should be publishable both on Compact Disk and on the World Wide Web, and it should be committed to the long-term care of a professional language archiving body.

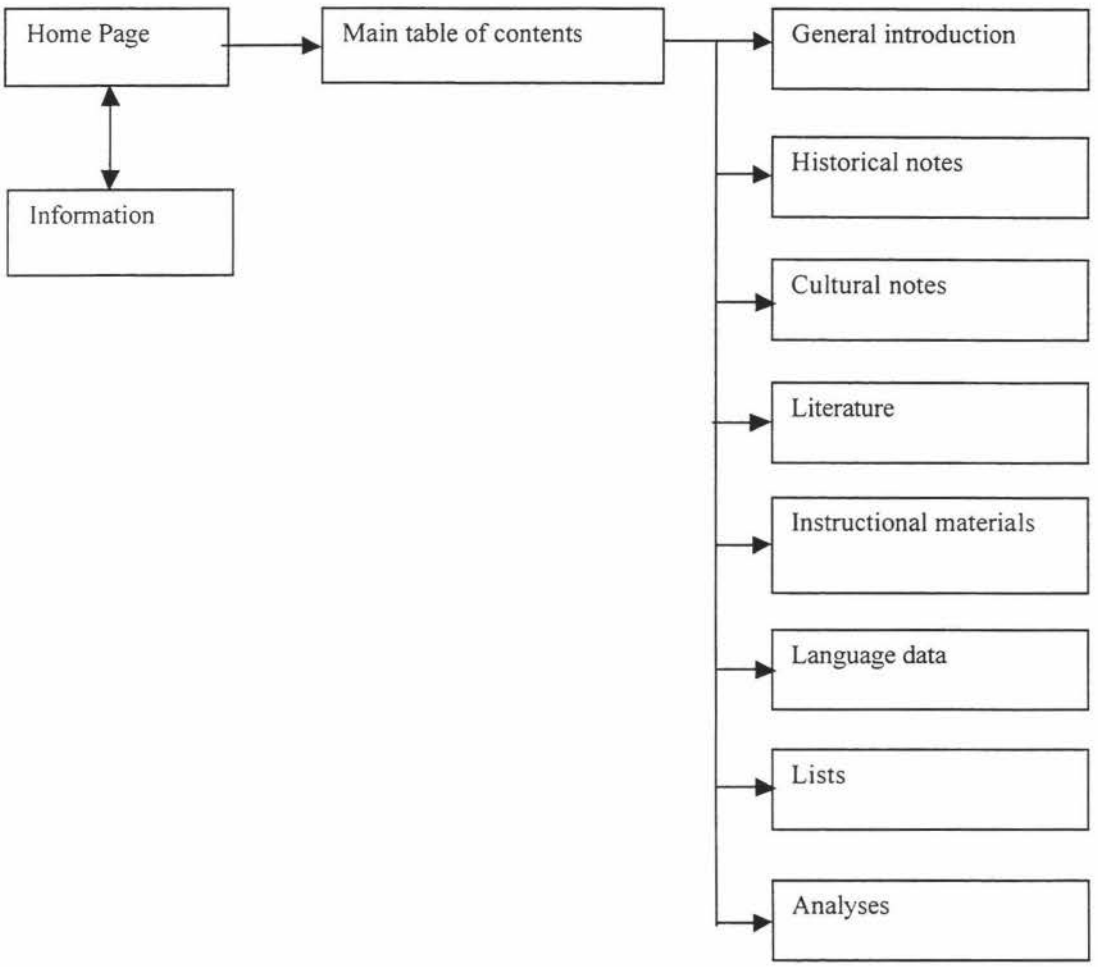
The proposal does have some limitations; certain formats proposed for use may well be superceded by more effective solutions not long after this proposal has been published. Nevertheless, I believe that the principle of selecting currently available technology that can be reasonably expected to be widely accessible for a long time in the future, will always hold true. Adhering to this principle should allow a documentation to be easily upgraded when the anticipated future technologies become generally available, especially browsers and other software that can handle Unicode and XML well.

6.1 Organisation of the documentation

From the point of view of the user, the documentation will be entered through a "home page" that points him or her to a series of tables of contents, which form the backbone of the documentation. It also points the first-time user to a page giving information about how to use the documentation.

The tables of contents will have a hierarchical structure, organising all the documents into the major topic areas at the first level (see section 3.7), and allowing easy access to every document relevant to each topic. This structure is illustrated in Figure 6.1.

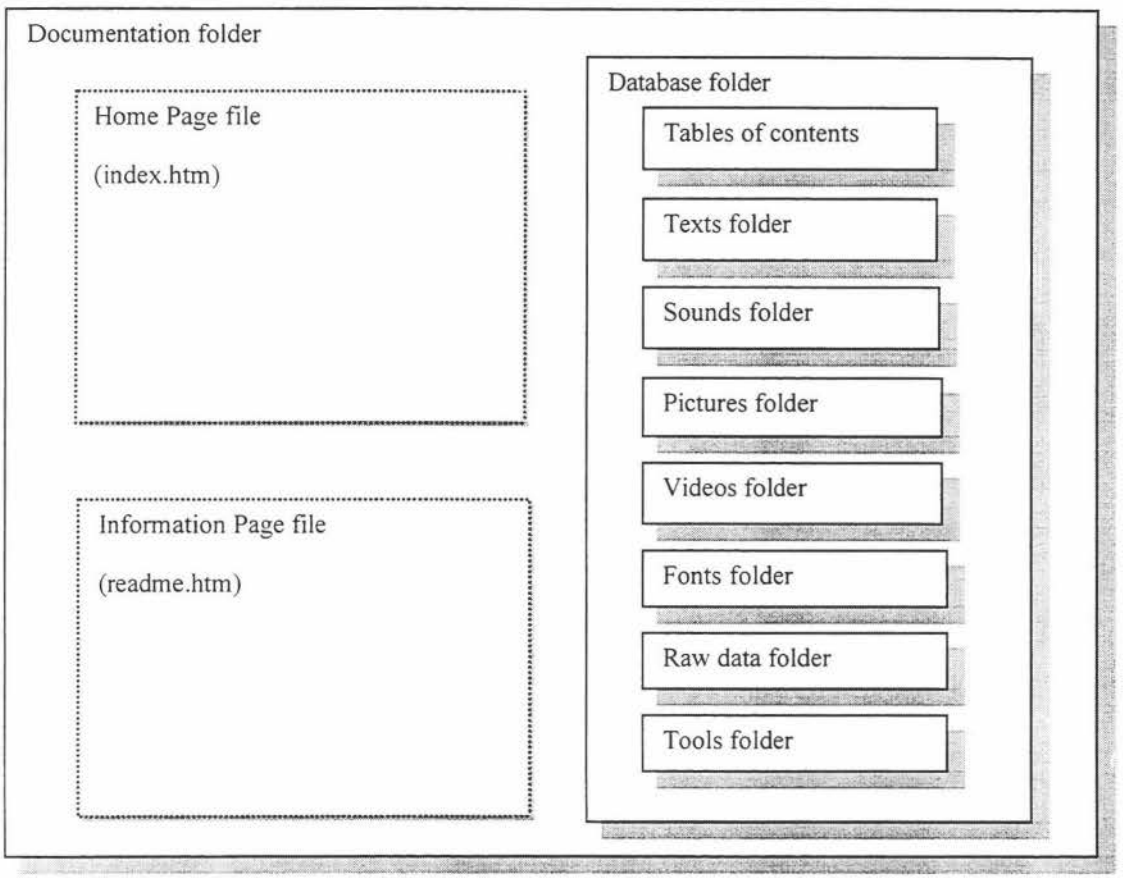
Figure 6.1. The user's view of the documentation



Topic areas with many documents will have further tables of contents organised in a hierarchical fashion so that a user can "drill down" quickly to the information he or she is looking for. In addition to this hierarchical organisation, there will, of course, be a network of hyperlinks that provide cross-referencing within and between documents, including those listed in different topic areas (as discussed in section 5.1).

From the point of view of the documentalist, the main body of the documentation database will be organised in a "relational" manner where documents of a similar format or structure are placed in the same folder (as discussed in section 5.1.1). Thus the documents are *not* filed in the hierarchical manner portrayed by the tables of contents; instead, they are filed at the same level in a small number of folders so that indexing and referencing can be kept simple. The filing structure is illustrated in Figure 6.2

Figure 6.2. Organisation of folders and files



6.2 General formatting recommendations for each file type

6.2.1 Markup

The Home Page, Information Page, Tables of Contents, and the files in the Texts folder should all be marked up in XHTML. Utilising XHTML rather than HTML has the dual advantages of being usable by current web browsers because of its compatibility with HTML, and also of being usable by XML data manipulation software being developed currently in many fields with an interest in data management, both scientific and commercial.

6.2.2 Original documents

Complex documents and databases in their *original* proprietary or non-HTML formats are kept in the Raw Data folder. Although they may not be viewable using a browser, it is important to include them in the documentation so that they can be used

with the relevant software by those who have it, or so that they can be converted to a browser-accessible form at a later stage.

6.2.3 Text

Literature and articles and that have been prepared for publication on a computer can be readily converted to PDF format for downloading and reprinting in the future. A version in (X)HTML should also be prepared.

Phonetic. For phonetic writing Unicode should be considered, especially once Unicode fonts with IPA symbols for all major computer platforms are available. An alternative for the meantime is a GIF font that uses a naming system that is easily convertible to Unicode. In the pilot project illustrated in chapter 7, I have, for example, the symbol beta (β) in a file called *P_3B2_P.gif*, where *3B2* is the Unicode hexadecimal number for that character. Grenoble & Whaley (2002b) have a different system for the e-journal, *Linguistic Discovery*. They digitise articles submitted in Microsoft Word format with phonetic text using SIL Doulos IPA93 fonts by converting each IPA symbol to a GIF image whose name includes a code number and the dimensions of its image. For example, the symbol β has a GIF image representation named *P42--x9x19.gif*. This means that the dimensions can be automatically included in the tag: *<IMG class=phonetic SRC="/symbols//P42--x9x19.gif" width=9 height=19>*. This allows the images to be displayed more quickly in a browser.

Another alternative is a single GIF image of the whole word/sentence/passage of phonetic transcription, although this has the disadvantage of being unsearchable using automatic search software. This disadvantage can be reduced if an image alternative is included in the XHTML image tag that displays the phonetic text using an ASCII phonetic text system such as SAMPA (see section 4.5.4.4). For example, an image file called *word9.gif* containing the image [βenɛ] could be displayed using the tag **, where *[BenE]* is the SAMPA equivalent.

Orthographic. For orthographic writing, if ASCII is insufficient then Unicode should be considered, unless the language uses an orthography that is compatible with that of a major language which has fonts that are already freely available and usable on

all major computer platforms. If neither of these options is satisfactory, then, as for phonetic writing, a good alternative is to generate a GIF font for the language with a grapheme naming convention that will allow easy conversion to Unicode when the time is ripe. If a GIF font is out of the question, then a GIF image of each word or line or passage may be the only option. In that case, an alternative display that uses an ASCII coding like SAMPA would make the data more useful to researchers.

6.2.4 Sound, picture and video

The choices for sound, image and video clips are independent of language, and so it is easier to make general recommendations about them than it is for text documents.

Video and sound. For long video and sound documents, MP3 is a reasonable choice if the means to produce them is available, as it is popular for the transmission of video and music over the internet with a very satisfactory "CD-like" quality and reasonable file size. For an internet-based documentation, such files can be played as they are downloaded using streaming software on the web server. For shorter video clips an alternative such as QuickTime, MOV or AVI would be satisfactory, and for sound, WAV or AIFF. Given that all of these formats maintain a reasonable data quality, and that it is to be expected that they can be played on most modern computer systems, the main criterion is a rapid response to a request to play.

Images. Images can be GIF for diagrams, computer-generated drawings, image fonts and thumbnails; PNG for photographs and scanned artwork where it is desired to preserve quality; and JPEG for other drawings and photographs where size is more crucial than data quality.

6.3 Recommendations for specific sections

The recommendations for the eight major sections consists of a list of data types to be included, with an indication of any special formats that may apply in square brackets, and locations of some relevant illustrations in round brackets . After the list of data types I provide further explanatory notes.

6.3.1 General Introduction

This should contain information on the following topics:

- location and size of language community with maps [JPEG] (see Figure 7.14);
- history of the language community and the language (where known);
- explanation of the writing system [Unicode/GIF-fonts] (see Figure B6).

This information would usually be documented in the first instance using a word processor. It is a simple matter to save such a document as an HTML (or XHTML) file from a modern word processor. Letters and words in phonetic and orthographic characters may need extra attention as outlined above.

6.3.2 Historical Materials

This could contain information of the following kind:

- old field diaries, scanned as digital images [PNG] and/or transcribed (see Figure B9);
- patrol reports from old governmental archives;
- published material (travel accounts, etc) likely to be of interest to the language community (see Figure 7.15).

It is quite satisfactory if hand-written accounts are stored as digital images, but if they can be transcribed this will make the material they contain more accessible to electronic searches. Sketches in such accounts should certainly be considered for scanning as digital images which can be incorporated into transcription files structured with XHTML.

6.3.3 Cultural Notes

This could contain information of the following kind:

- Cultural observations (see Figure 7.12);
- Ethnographies and anthropology papers [PDF];
- Photographic [PNG] and video records of cultural activities (See Figures 7.8, B10).

If there is an extensive collection of photographs, then an indexing system using small thumbnail images (JPEG or GIF) will be an advantage.

6.3.4 Readers, literature and translated material

This should be material in a recognised orthography that the language community is likely to want to publish or republish in bulk when the need arises. Relevant details about each document (such as author, date, copyright status, whether original or translated, and if translated, the source of translation) need to be included, together with any native-speaker commentaries and reviews:

- post-primer readers (see Figure 7.20);
- newspapers and newsletters;
- native-speaker authored materials such as histories, legends, novels;
- song and poetry collections, hymnbooks;
- translated materials (e.g. for religious or community development needs);
- sound and video programmes in the language;
- bibliography [XHTML/XML] (see Figure 7.6).

It would be useful to have the text-based material in this list in two formats: (1) either PDF files or else high-quality digital page images for publishable materials that will allow the creation of new masters for reprinting, and (2) transcriptions marked up with XHTML that will allow editing and reformatting before (re)publishing on the one hand, and searching and processing by linguists on the other.

If the published literature of the language community is extensive, then a representative corpus of text transcriptions would be more appropriate for a language documentation than large numbers of very bulky page images. (The latter would be more appropriate in a separate publisher's archive.) Other literature not included in the language documentation could then receive a mention in the bibliography.

Sound and video productions in the language are not very likely to exist for endangered languages,[27] but if they do they may need to be stored (in digital form) on separate media because of their bulk.

---

[27] The Jesus film may be an exception. According to The Jesus Film Project (n.d.) it has been translated into 723 languages, many of which are minority languages.

6.3.5 Instructional materials

This should be a collection of vernacular educational materials including:

- literacy primers (see Figure 7.16);
- language-learning materials (see Figure B5);
- other educational materials developed for vernacular education (e.g. for numeracy, economics, health, culture, games, forest lore or agriculture).

If these have been produced digitally, then the documentalist could seek to obtain copies of the original files and convert them to PDF, although for an enduring documentation they would need to be converted to XHTML, or even XML if they have a complex structure. If they have not been produced digitally and the formatting of these materials is complex, digital page images are probably the best option.

6.3.6 Language data

In many endangered language documentations this will be the largest section, and will consist of a corpus of raw and processed data of all sorts of communicative events in the language community which linguists or documentalists have obtained for analytical or archival purposes, and which might not be in a form that is publishable for the general purposes of the language community. It should include:

- raw texts in phonetic, phonemic or orthographic transcription [Unicode/GIF-fonts], with or without sound/video sources [AIFF etc], together with commentaries on each;
- translated and interlinear texts [XHTML tables], with or without sound sources [hyperlinked], together with hyperlinked notes and cross-references on each (see Figures 7.21, 7.23, B1, B12);
- tagged texts [XML/TEI], with or without sound/video sources, together with commentaries on each;
- an archive of any untranscribed digitised sound and video recordings of various types of communication events, together with commentaries on each.

Where the sound recording has a transcription such as an interlinear text, the documentalist can break the recording up into a number of small WAV or AIFF files corresponding to lines or even individual words in the transcription, and these can be linked to the transcription using standard mechanisms available in XHTML.

If interlinearising software cannot export the interlinear text output in an XHTML format, then it should be converted to XHTML using tables to keep columns of glosses and analyses in alignment. The original file should also be archived in the Raw Data folder (see Figure 6.2).

Tagged texts are commonly created using a format of the SGML family, e.g. the TEI system. Because the software used to process this kind of data at present expects that format, it is probably wise to keep it in that format. However, I do suggest that it may be useful to have this format made to conform with the XML standard so that it will be accessible using an anticipated large variety of software tools that will use this format.

Some researchers produce tagged files that coordinate video or sound recordings with their transcriptions. If possible these should be exported in an XML format for the sake of endurability of a resource into which a lot of effort would have been poured.

6.3.7 Lists

This collection should consist of:

- wordlists collected for dialect surveys [Unicode/GIF-fonts/SAMPA], together with commentaries (see Figure 7.18);
- grammatical paradigms in table format, together with commentaries (see Figure B7);
- lists of important semantic categories, such as numbers, measures, trees, birds, tools, etc, together with illustrations [JPEG] and commentaries;
- the lexicon [XML *and* XHTML] (see Figures 7.19, 7.27, B11).

If dialect survey wordlists are not already in digitised form, a scanned image of each list is satisfactory.

The lexicon is a very important document in which the linguist probably has invested huge amounts of time, and which is in a special format for access by database software. For the documentation, every effort should be made to store it in XML format so that the detailed structure is preserved, and so that it can be accessed using future generations of XML database software. It should then be converted (or transformed using XSL) to XHTML for the user to browse. If it is very large, it can also be divided up into sections that can be downloaded separately.

A number of linguists who have made their lexicons accessible on the World Wide Web have provided access in the form of a search engine. I have found this rather unsatisfactory, as unless one knows the culture it is impossible to know what are the interesting categories or words to search for. I find a browsable dictionary far more useful, and web browsers have inbuilt search facilities if one needs them.

6.3.8 Analyses

Of special interest to linguists and other researchers, these include

- papers on phonology [Unicode/GIF-fonts, hyperlinked sound] (see Figure 7.25);
- papers on grammar;
- papers on orthographical matters [PDF, ASCII/Unicode/GIF];
- dialect surveys;
- papers on theoretical aspects of anthropology [hyperlinked sound/image/video].

If feasible, these papers should be converted to both PDF and XHTML forms. Links to sound clips (and even video clips) of examples can also be easily incorporated for a truly multimedia linguistic experience.

7. SAMPLE DOCUMENTATION

7.1 Introduction

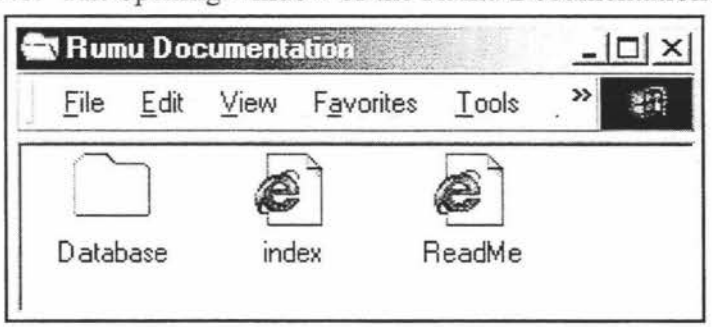
The instrument for documentation described in chapter 6 was tested and refined on a pilot documentation project for the Rumu language, which is spoken by a small group of just 800 people, and so certainly constitutes a minority language.

For the pilot project, a relatively small number of sample documents of many kinds were integrated in the form of an internet-ready multimedia documentation. This sample documentation has been stored on a compact disk, which is included with this thesis. This chapter contains a series of illustrations of various kinds of document included in the documentation on this CD.

7.2 Folders and Files

The documentation is organised in a form suitable for browsing using common Web browsers, and is intended for distribution both on a compact disk and via the internet. The sample documentation in the CD included with this thesis is in a folder labelled *Rumu Documentation*. In order to keep things very simple, on opening the Rumu Documentation folder, a single folder and just two files are visible: the slightly obscurely named *index* file, and the more invitingly named *ReadMe* file:

Figure 7.1. The opening window of the Rumu Documentation sample

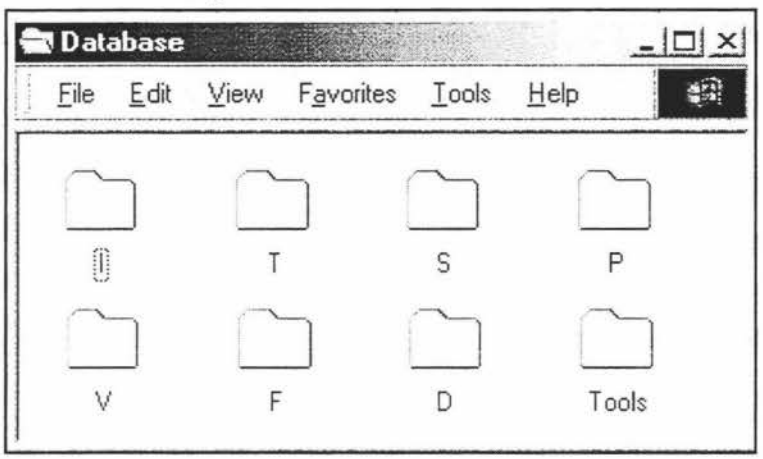


Opening either the *index* file or the *ReadMe* file will allow the user to browse the documentation. The *index* file is a title page or 'home page' (see Figure 7.3), and the *ReadMe* file contains important information about using the documentation (see section 7.3). Both files are hyperlinked to each other.

The reason for having the title page called *index.htm* is convention; when an internet access is attempted on a web page address (URL) and the file name is not specified, the browser will try to open a file named *index.htm* or *default.htm* by default.

The Database folder also visible in the opening window contains the rest of the documentation, organised (as described in section 6.1) into folders according to file type (see Figure 7.2):

Figure 7.2. The database folder



The folders have been given short names in order to keep references between files in different folders as short as possible. Furthermore, since the folders are not designed to be browsed by naïve users, I have not tempted them by giving the folders names that might encourage this. The actual contents of the folders is explained clearly in the ReadMe file (see the next section).

Files have names that conform to the old DOS filename specification (up to 8 letters etc followed by a 3-letter extension) because if I use longer names the CD writer software that I use alters them in undesirable ways to something that does conform to that specification.

7.3 Using the documentation - the *ReadMe* file

Information on how to use the documentation is explained in the *ReadMe* file, and the main text content of that file is given overleaf:

About the Documentation

Click here to get to the title page.

This multimedia documentation is being compiled from notebooks, audio cassette tapes, VHS-C video tapes and computer files collected over eight years of field work, and then mulled over and worked on for a long time off the field. It is being prepared in HTML (or XHTML) format for access over the internet and from a compact disk using a web browser. This documentation is an experimental sample of different kinds of document organised into the structure that is planned for the full documentation.

Navigation

The files in this documentation are standard (X)HTML files that can be viewed using browsers such as Internet Explorer 5 or Netscape 4. Many of the index files have a navigation bar at the top of the page to major sections. If pages are naturally linked together in a series, or in a hierarchy, appropriate links can be found at the bottom of the page. There are also cross-referencing hyperlinks between and within files.

System

This documentation has been composed using a Macintosh computer running MacOS 9 and an IBM-compatible computer running Windows ME, each with about 128Mbytes of RAM. I made extensive use of SimpleText, Notepad and BBEdit to compose XHTML files, and also Perl for Win32 and MacPerl to write programs to do conversions. Some documents have been "saved as" HTML files from software such as MicroSoft Word. The documentation has been tested on older Macintosh and Windows computers using Internet Explorer and Netscape.

Older computers with smaller memory sizes may have problems loading pages with multiple embedded sound files, and videos may play in a jerky manner.

Settings

Picture files for screen viewing are optimised to 72 dots per inch, and view well with screen resolution set to 640x480. Where appropriate, some pictures have larger versions for viewing or printing. These can be accessed by clicking on the picture itself, or by clicking on a nearby link.

In order to display all characters properly, the browser's text encoding should be set to Unicode (e.g. UTF-8). In Netscape Navigator 4.74, this can be achieved from the Characters option in the View menu; in Internet Explorer 4.0, use the Fonts option of the View menu.

The background is best set to "white", as the IPA characters and characters with multiple diacritics are GIF images with white backgrounds. (They have white backgrounds because I found that transparent backgrounds printed with an unwanted checkered pattern.) With Netscape 4.74 this can be achieved after going to the Edit menu and selecting Preferences, Appearance, Colors; with Internet Explorer 4.0, go to the View menu and select Internet Options, General, Colors.

Multimedia Plugins

Sounds are WAV or AIFC format, and can be played using, for example, Apple QuickTime. Check the browser settings to ensure these file-types are catered for.

Videos can also be played using, for example, Apple QuickTime.

Apple QuickTime version 5.0 was used in the testing of this documentation.

Organisation

The files are organised into a "database" consisting of several folders containing the following file types:

I
Index files - tables of contents for sections and subsections, having lots of links to other files

T
Text files - files consisting mainly of text and embedded sounds and videos

P
Picture files - JPG, PNG and GIF formatted files

S
Sound files - WAV, AIFC, MP3 etc

V
Video files

F
Font files, including small GIF files for IPA symbols and Rumu letter symbols that could not be displayed using Unicode on both Macintosh and Windows computers at the time of writing

D
Original documents. These could be in RTF, SF or other formats.

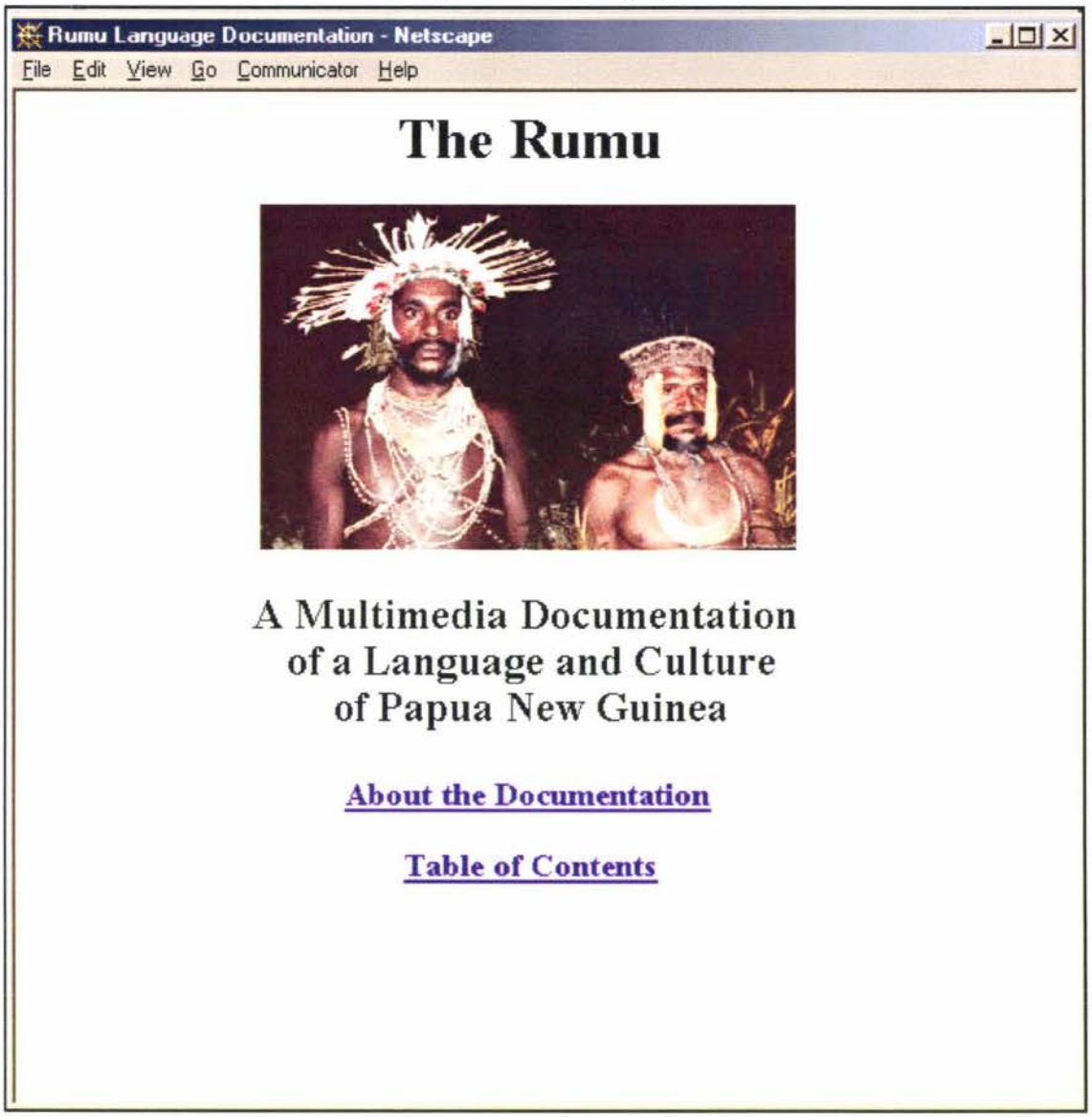
Tools
Software tools for making the documentation. (Templates, fonts, analysis and conversion programs.)

The *ReadMe* file also contains acknowledgements and contact addresses.

7.4 Sample views

Because the documentation is designed to be viewed on a screen using an internet browser, the following illustrations consist of a browser window view of the page together with a short commentary on special features illustrated by the page, and the path of links by which the page can be accessed.

Figure 7.3. Title page



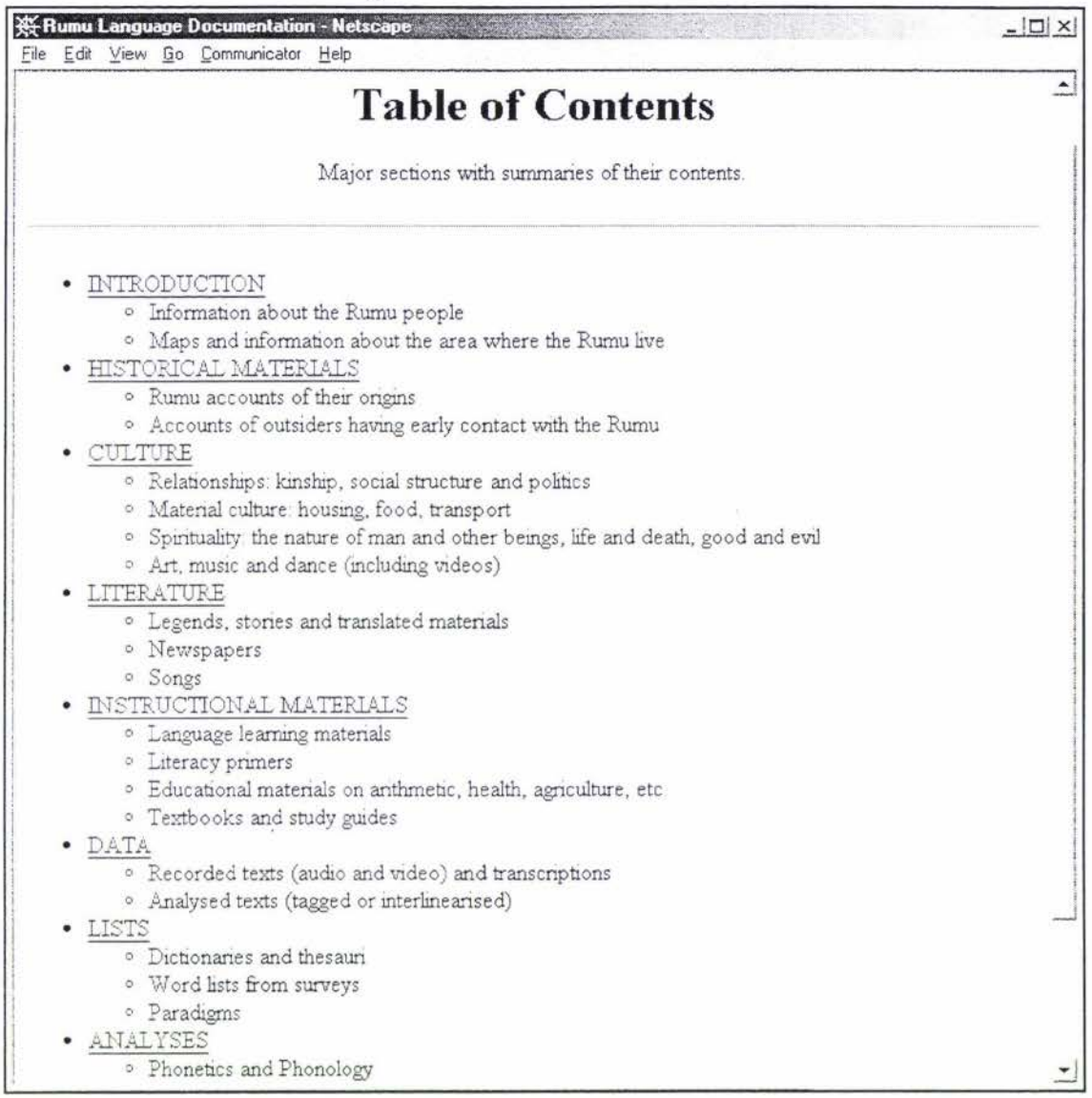
Comment

The title page has a simple layout: title, picture, explanatory subtitle, and two links. The picture shows some Rumu people dressed for a celebration of their traditions. The picture file has been optimised so that it will load reasonably quickly. The first link is to a page of information important for the first-time user to read, and the second link leads the user into the rest of the documentation.

Access path

index.htm in the opening window

Figure 7.4. Main table of contents



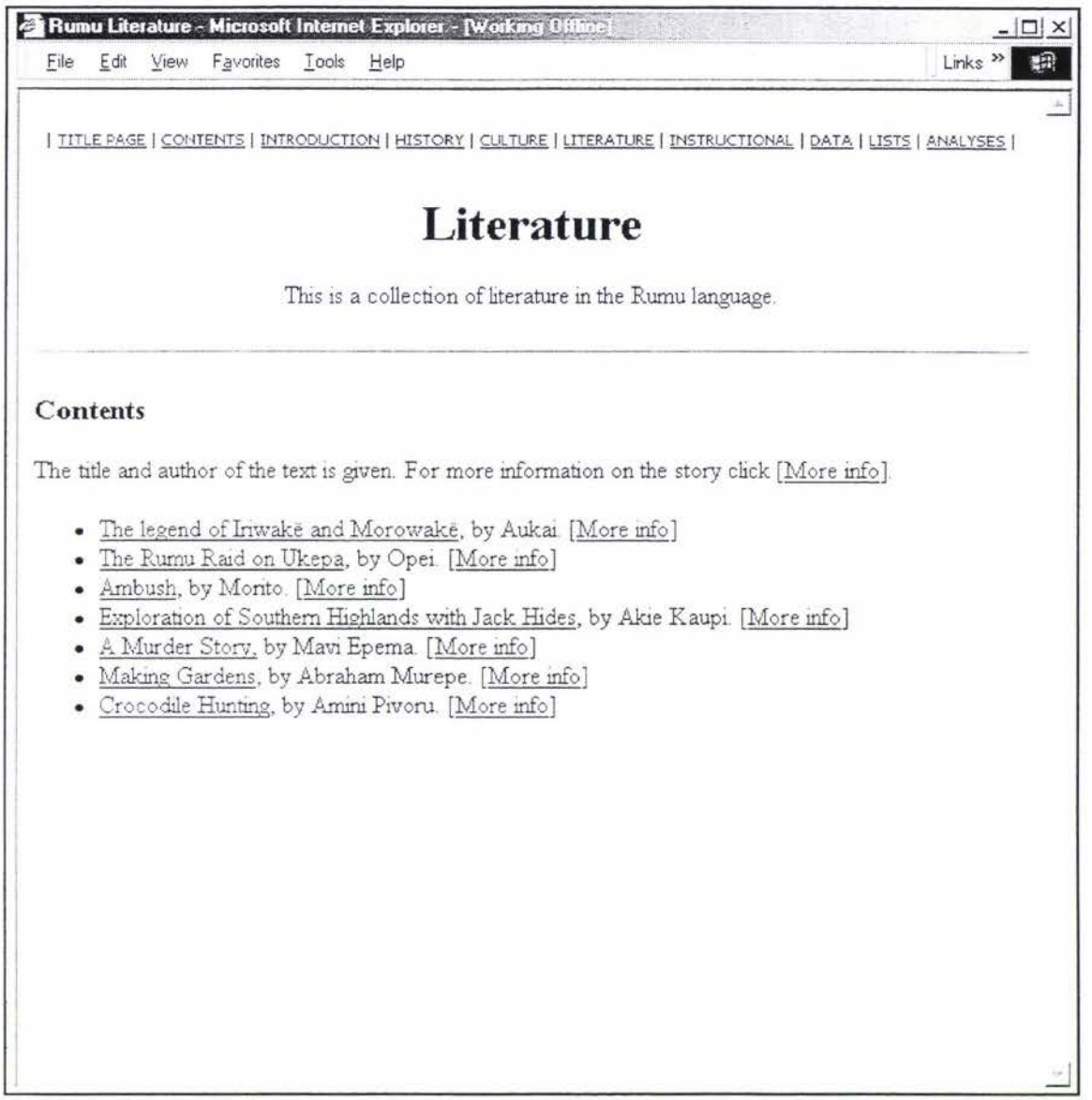
Comment

The main table of contents has one link for every major section of the documentation together with information on the highlights of each section. The sections especially likely to be of interest to the language community come first; those for the linguist come later. A link to tools for the linguist is out of view, as is also a link back to the title page.

Access path

Title Page > Table of Contents

Figure 7.5. A section contents page



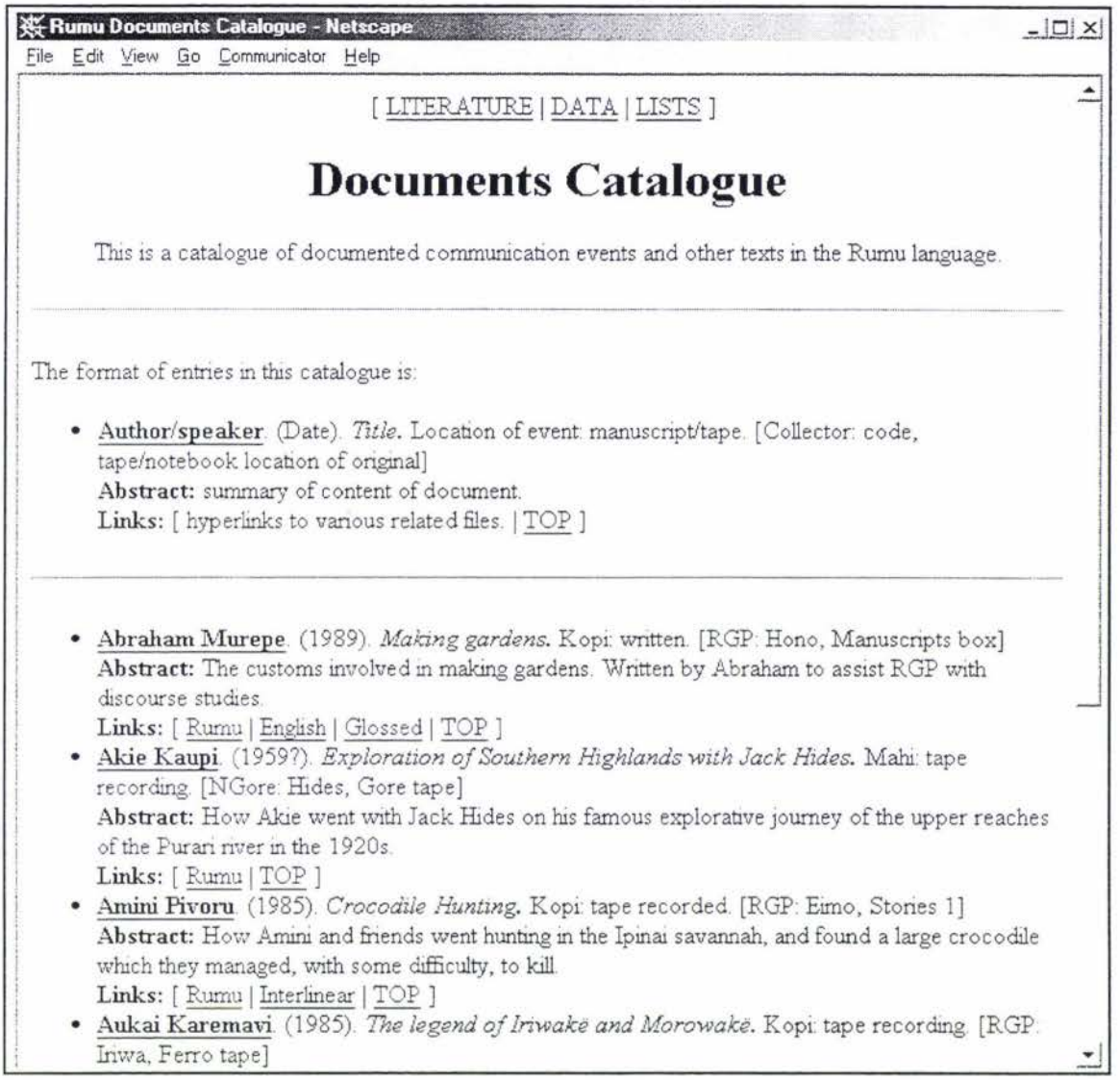
Comment

A click on a section name in the main table of contents takes the user to a section contents table. Section contents tables all have the same format, with a coloured navigation bar at the top, section title and brief explanation, and then links to the various documents pertaining to the section. The navigation bar takes the user to any other section, or to the main contents page or even the title page. Each document in the section may be related to a number of other files. A click on the [More info] link takes the user to a catalog where all files related to a particular document are listed.

Access Path

Title Page > Table of contents > Literature

Figure 7.6. A bibliography



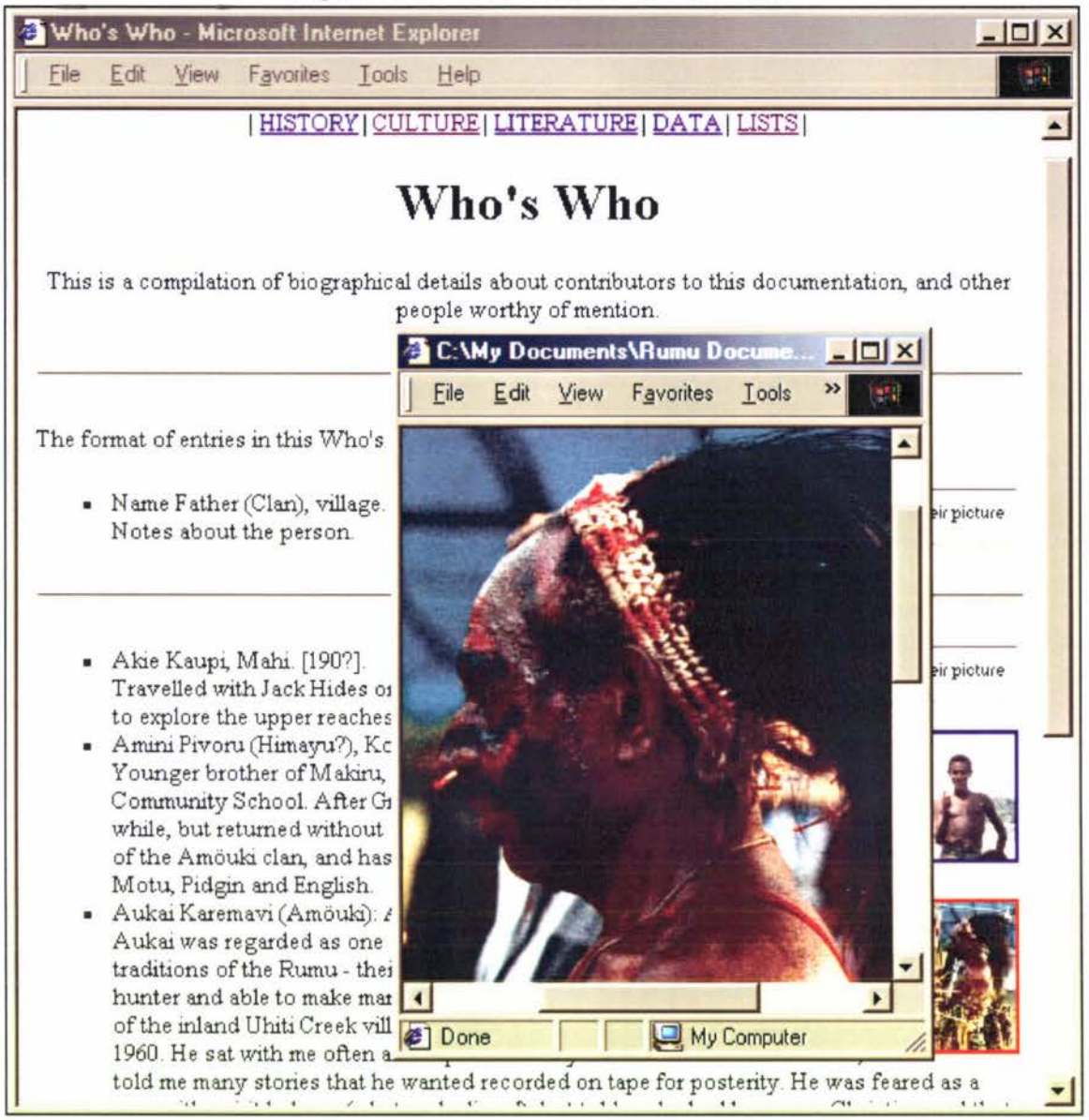
Comment

This catalogue uses a bibliographic format, which is explained near the top of the page. (The explanation can be used as a template for new entries.) Each entry contains not only bibliographical information, but also an abstract (or summary) and links to various versions of the document. In fact, this catalogue is *the* place where links are maintained whenever a new version, commentary or analysis of a text is produced.

Access Path

Title Page > Table of Contents > Lists > Catalogue

Figure 7.7. Who's Who with Thumbnails



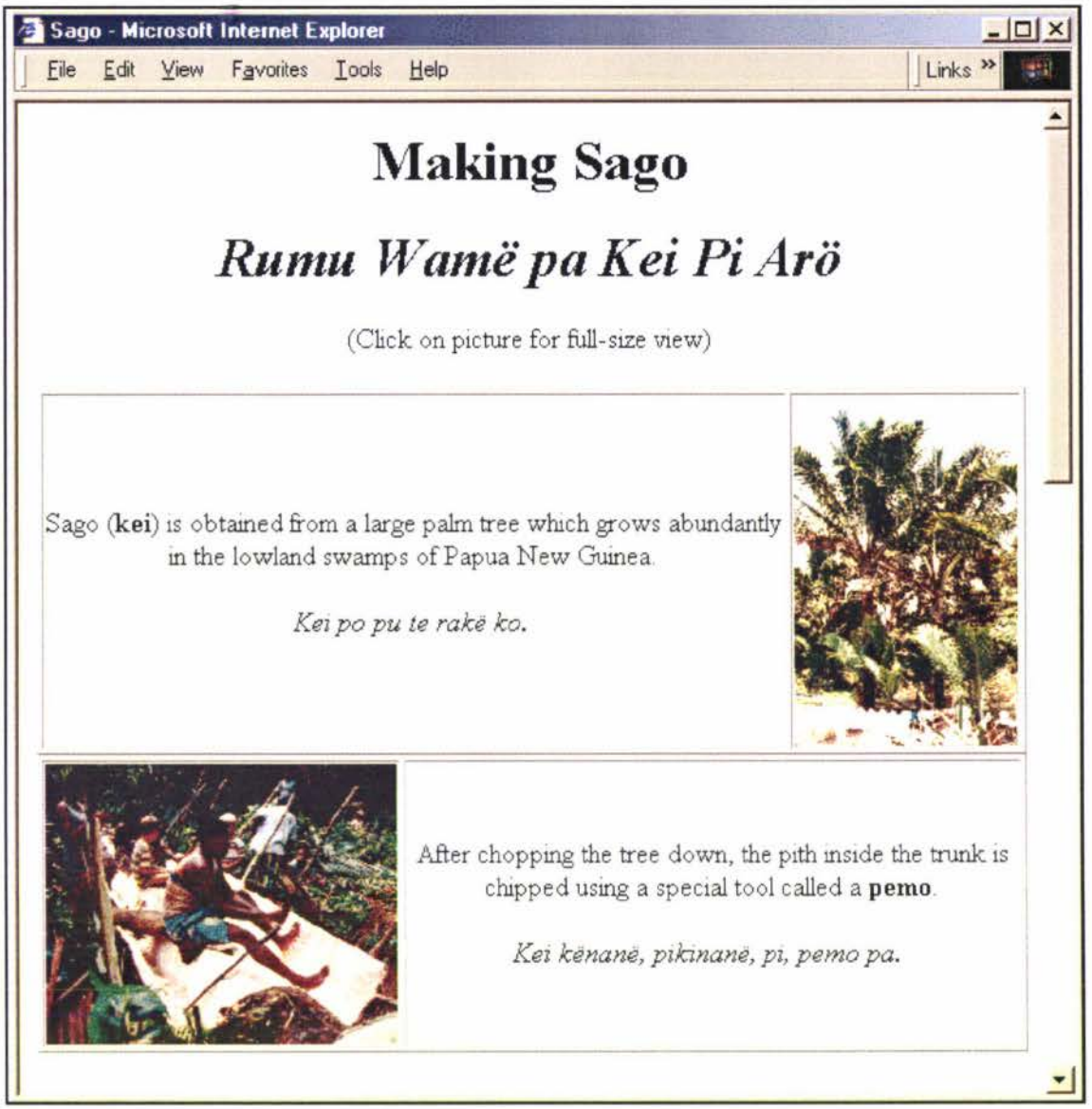
Comment

The Who's Who contains biographical notes on writers, story-tellers and other people who are important to the language documentation. The tag structure of the format explanation near the top of the page can also be used as a template for new entries. Each person may, if they desire, have a thumbnail portrait (e.g. a JPEG of 1 or 2kb) of themselves included, and a click on this opens up a larger view (e.g. a JPEG file of 10-100kb) in a separate window.

Access Path

Title Page > Table of contents > Lists > Who's Who

Figure 7.8. Cultural notes illustrated with photos



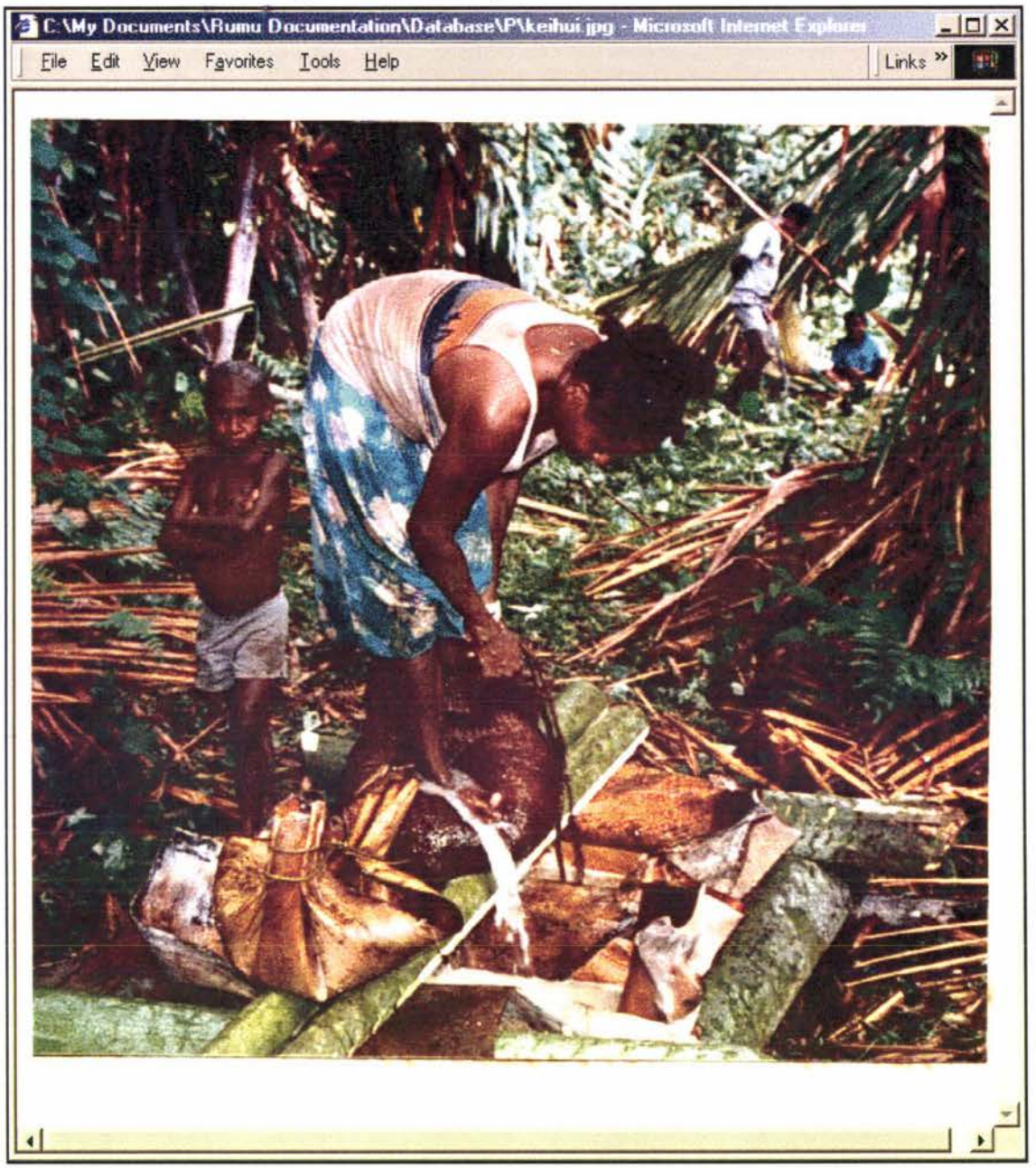
Comment

Observations of material culture illustrated with scanned coloured photographs of reasonable resolution (JPEG around 70kb). In this document the photos are reduced in size to fit into the presentation, but can be viewed full size with a click (see Figure 7.9). The captions are in both English and Rumu, and introduce lexical items that are hard to define out of context.

Access Path

Title Page > Table of contents > Culture > Material Culture > Making Sago

Figure 7.9. Blowup of a small illustration



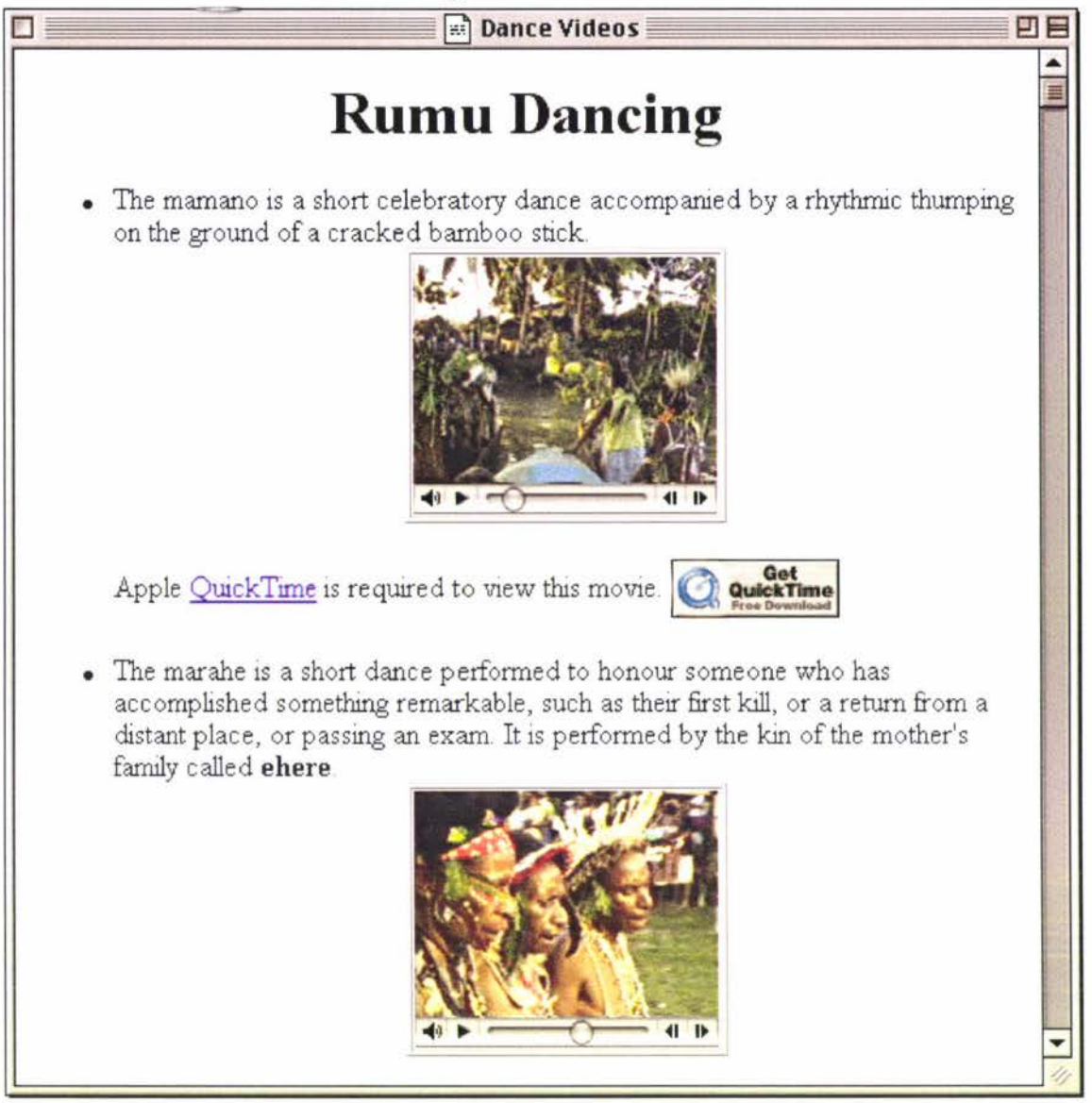
Comment

When the user clicks on one of the small illustrations in the page, part of which is shown in Figure 7.8, a full-size view of the picture is displayed in a separate window. These pictures are in JPEG format and are 70-140 kbyte in size.

Access paths (this file has several)

Title Page > Table of Contents > Culture > Material Culture > Sago making > keihui.jpg

Figure 7.10. Videos



Comment

This presentation of Rumu dances has embedded video clips that can be played on a click of the "play" button on a player bar.

The video clips are QuickTime video files taken from analog video tapes using Apple Video Player (version Z2-1.7.3) on an AV Macintosh computer, and compressed using QuickTime Player Pro (version 5.0.2). The compression process reduced the size of the first video clip (which lasts 27 seconds) from 29.3 to 7 Mbytes.

The player bar seen here is that of Apple QuickTime on a Macintosh computer.

Another version of the second video in AVI format is also included.

Access Path

Title Page > Table of contents > Culture > Videos of Rumu Dances

Figure 7.11. Tables and graphs

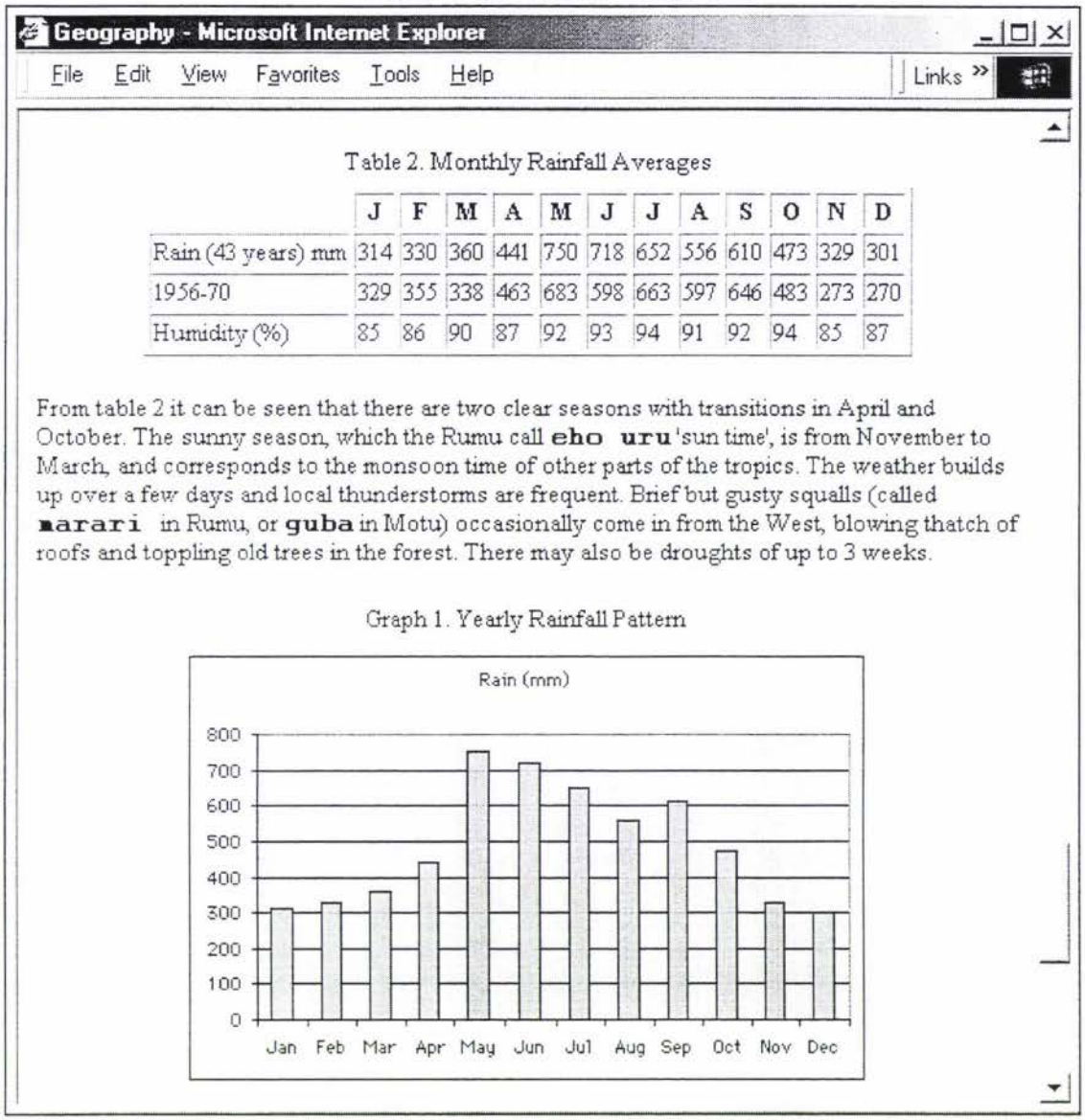


Table 2. Monthly Rainfall Averages

	J	F	M	A	M	J	J	A	S	O	N	D
Rain (43 years) mm	314	330	360	441	750	718	652	556	610	473	329	301
1956-70	329	355	338	463	683	598	663	597	646	483	273	270
Humidity (%)	85	86	90	87	92	93	94	91	92	94	85	87

From table 2 it can be seen that there are two clear seasons with transitions in April and October. The sunny season, which the Rumu call **eho uru** 'sun time', is from November to March, and corresponds to the monsoon time of other parts of the tropics. The weather builds up over a few days and local thunderstorms are frequent. Brief but gusty squalls (called **marari** in Rumu, or **guba** in Motu) occasionally come in from the West, blowing thatch of roofs and toppling old trees in the forest. There may also be droughts of up to 3 weeks.

Graph 1. Yearly Rainfall Pattern

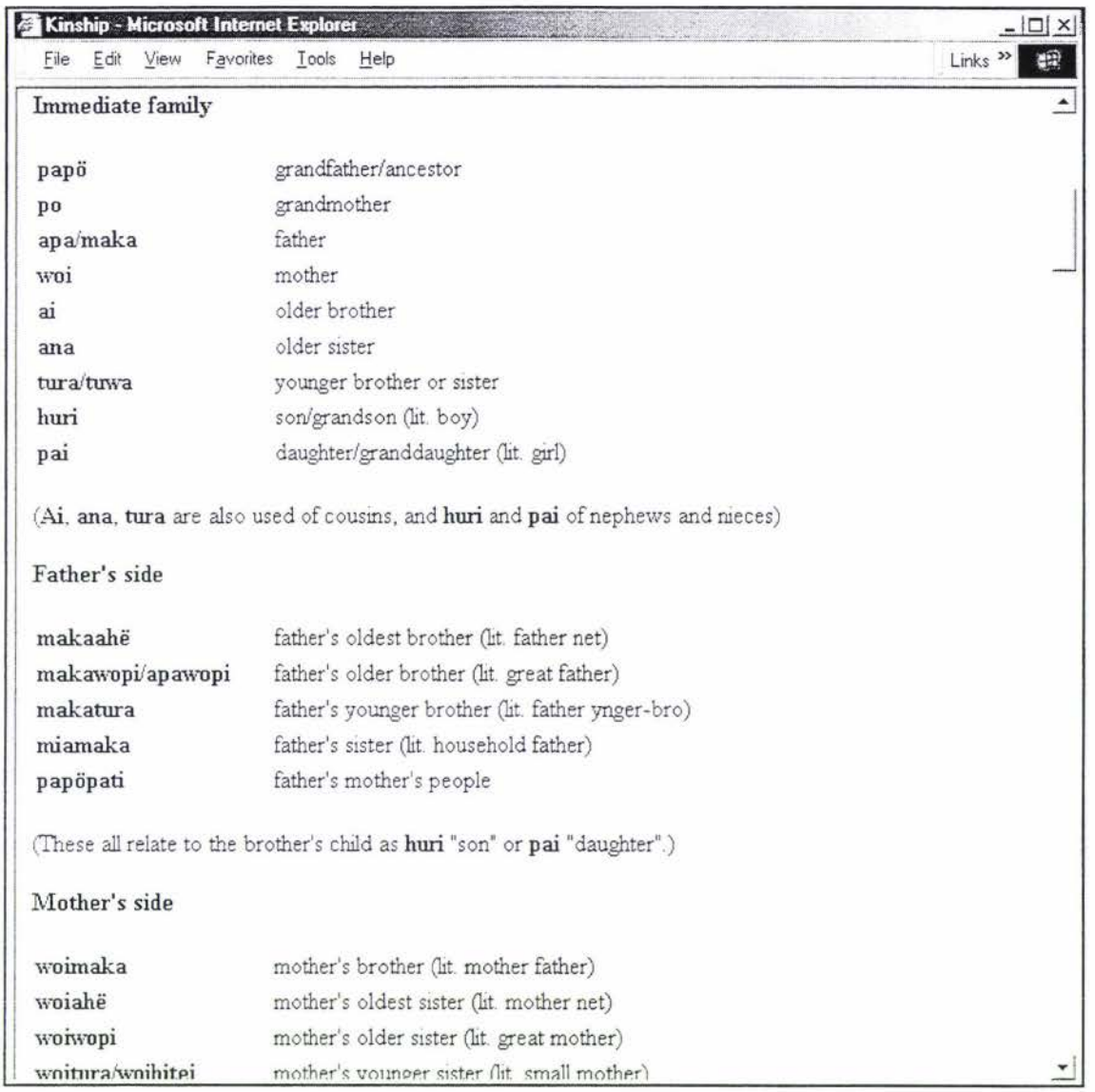
Comment

Notes on climate introduce vernacular lexical items for seasons, illustrated with real climate data presented in tables and graphs. The table was generated using HTML tags, and the graph was generated in a spreadsheet program (Microsoft Excel) and converted to a GIF image file (4.8kb).

Access Path

Title Page > Table of contents > Culture > Geography

Figure 7.12. Terms and definitions in columns



Comment

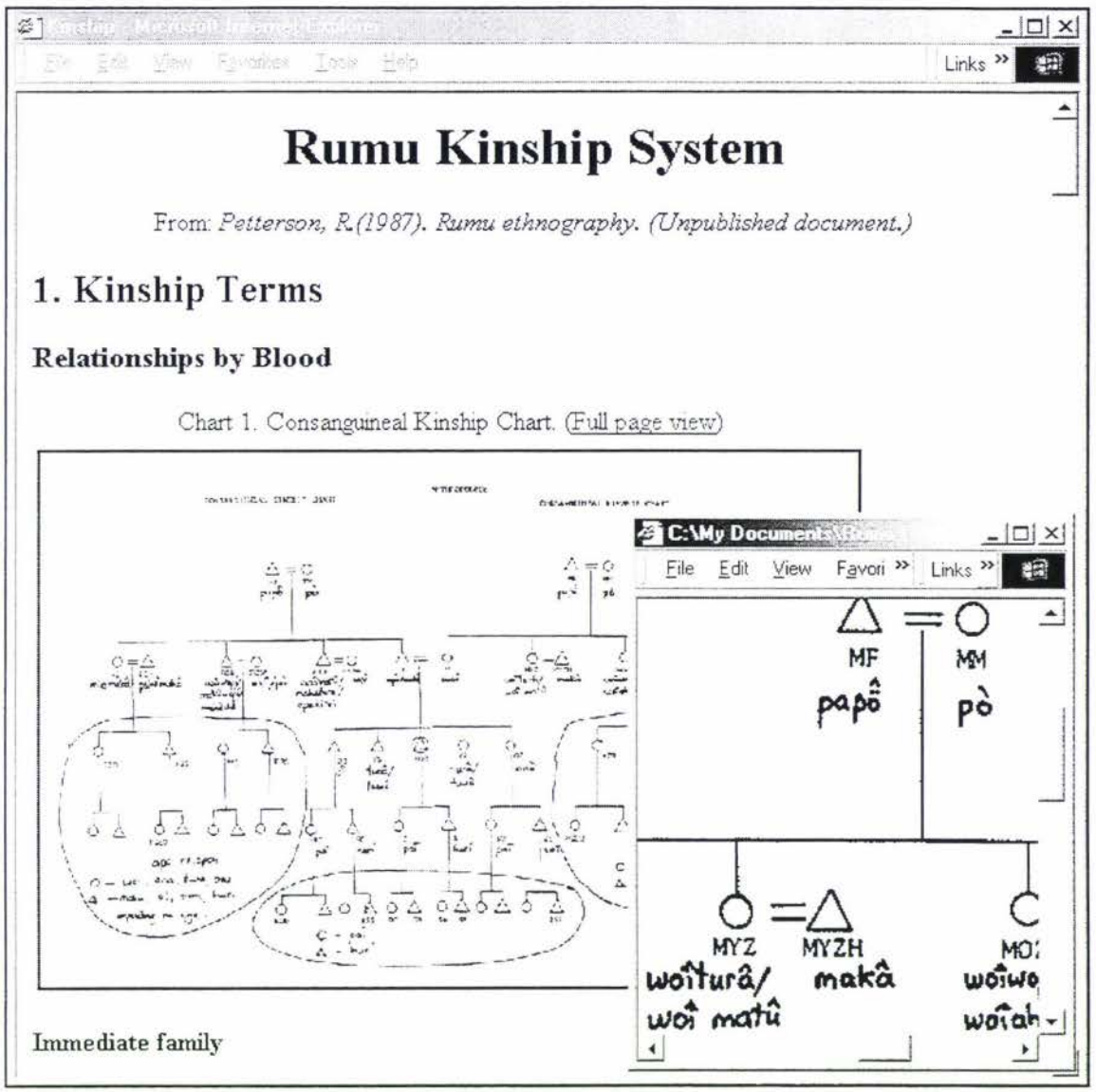
Rumu kinship terminology and their explanations in English are presented in two columns.

The data is formatted in XHTML using borderless tables.

Access Path

Title Page > Table of contents > Culture > Kinship

Figure 7.13. A scanned chart



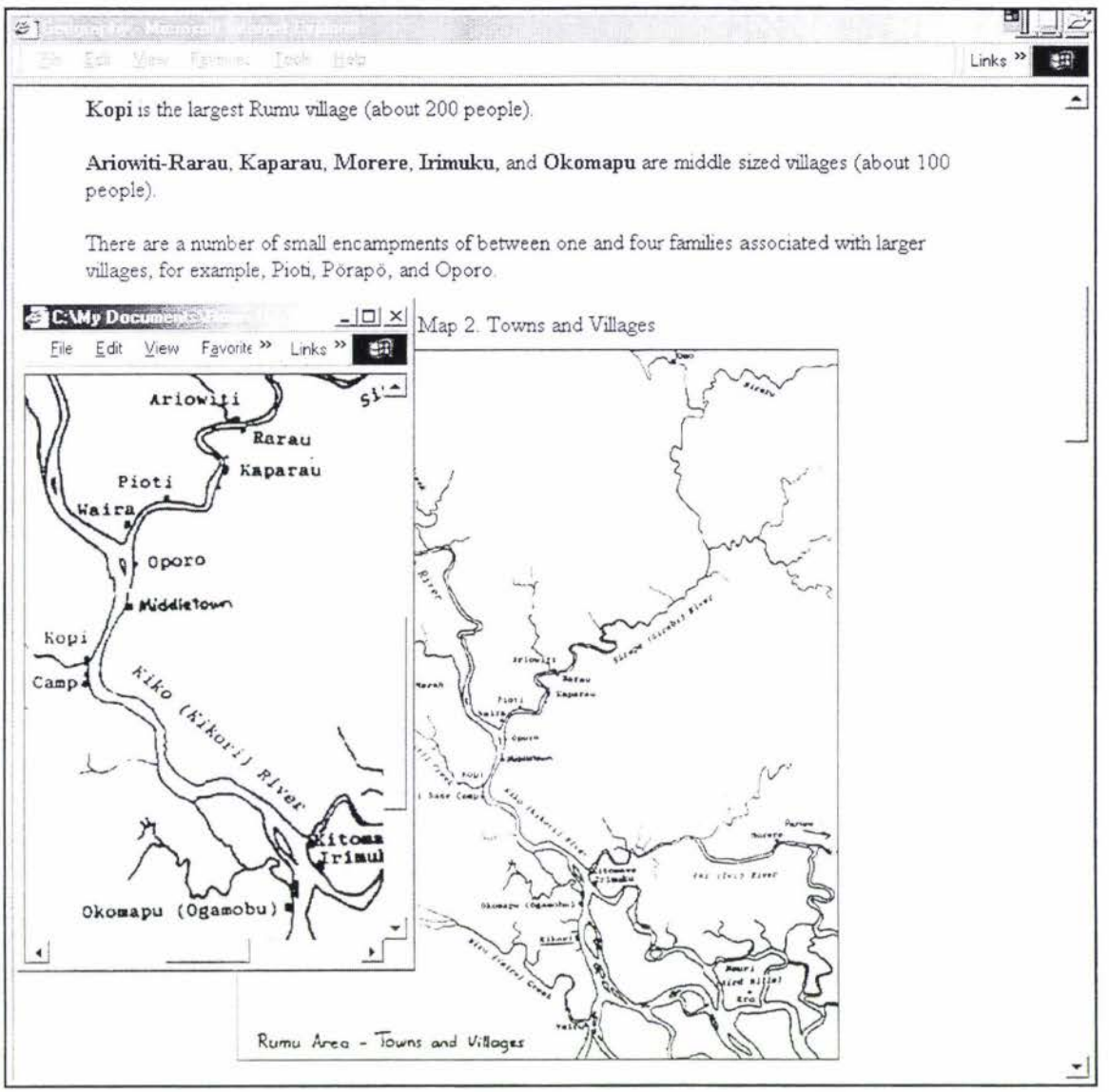
Comment

The section on kinship has a hand-drawn chart with a landscape aspect which does not fit inside the document window with a resolution such that the handwriting can be read. A click on the "Full page view" link opens a larger copy of the image in another window with a much better resolution. (The chart was scanned in as a "greyscale" image and saved in JPEG and PNG formats. The enlargement is the PNG formatted image.)

Access Path

Title Page > Table of contents > Culture > Kinship

Figure 7.14. A hand-drawn map



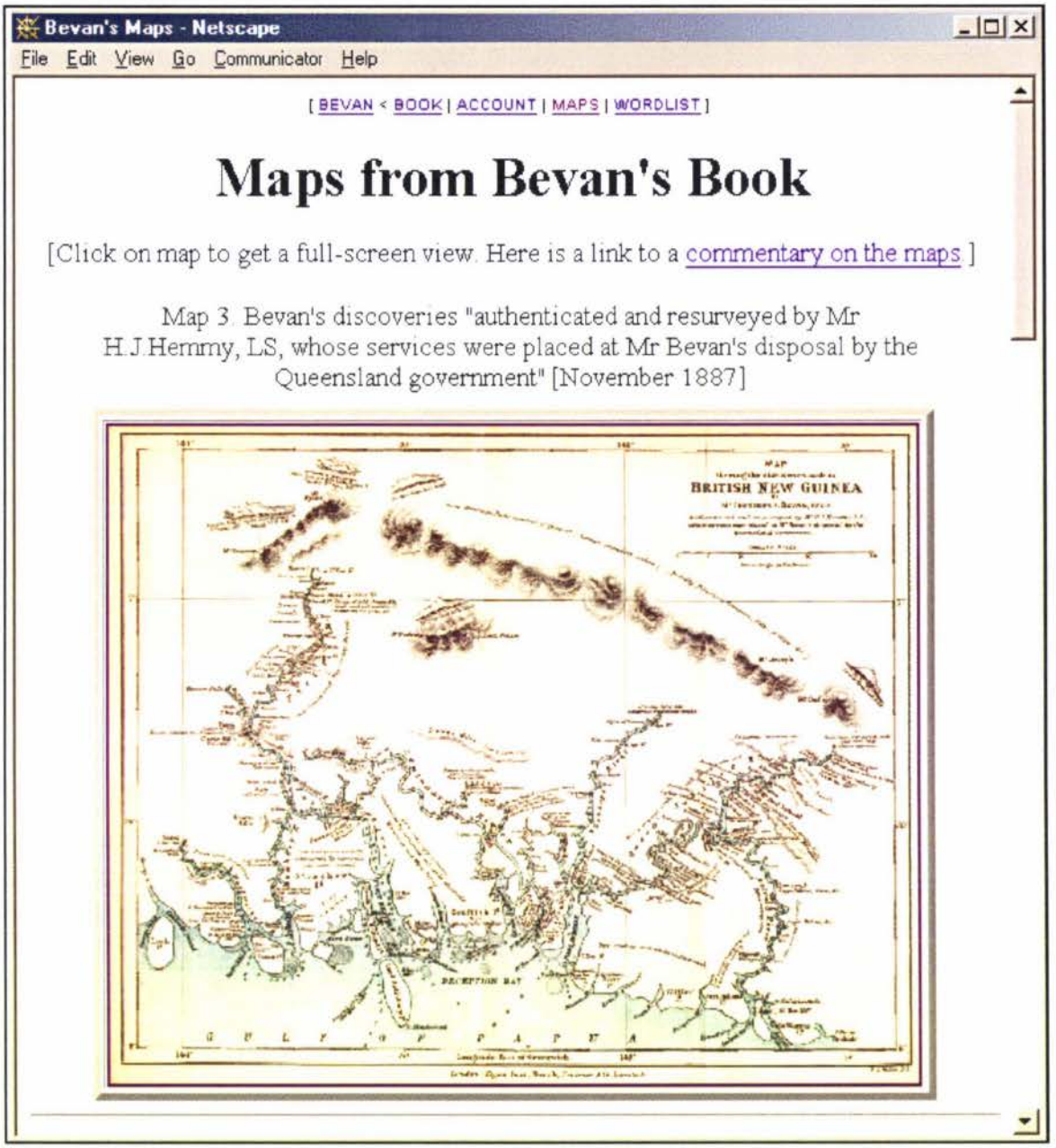
Comment

This map had been sketched by hand and the printed labels cut out and pasted on. It was scanned in and saved in both high resolution and low resolution formats. The low resolution file fits comfortably into a window, but the labels on the map are difficult to read. The map image is hyperlinked to the image file with greater resolution which opens in a separate window, and whose labels are now legible.

Access Path

Title Page > Table of contents > Culture > Geography

Figure 7.15. Scanned coloured map



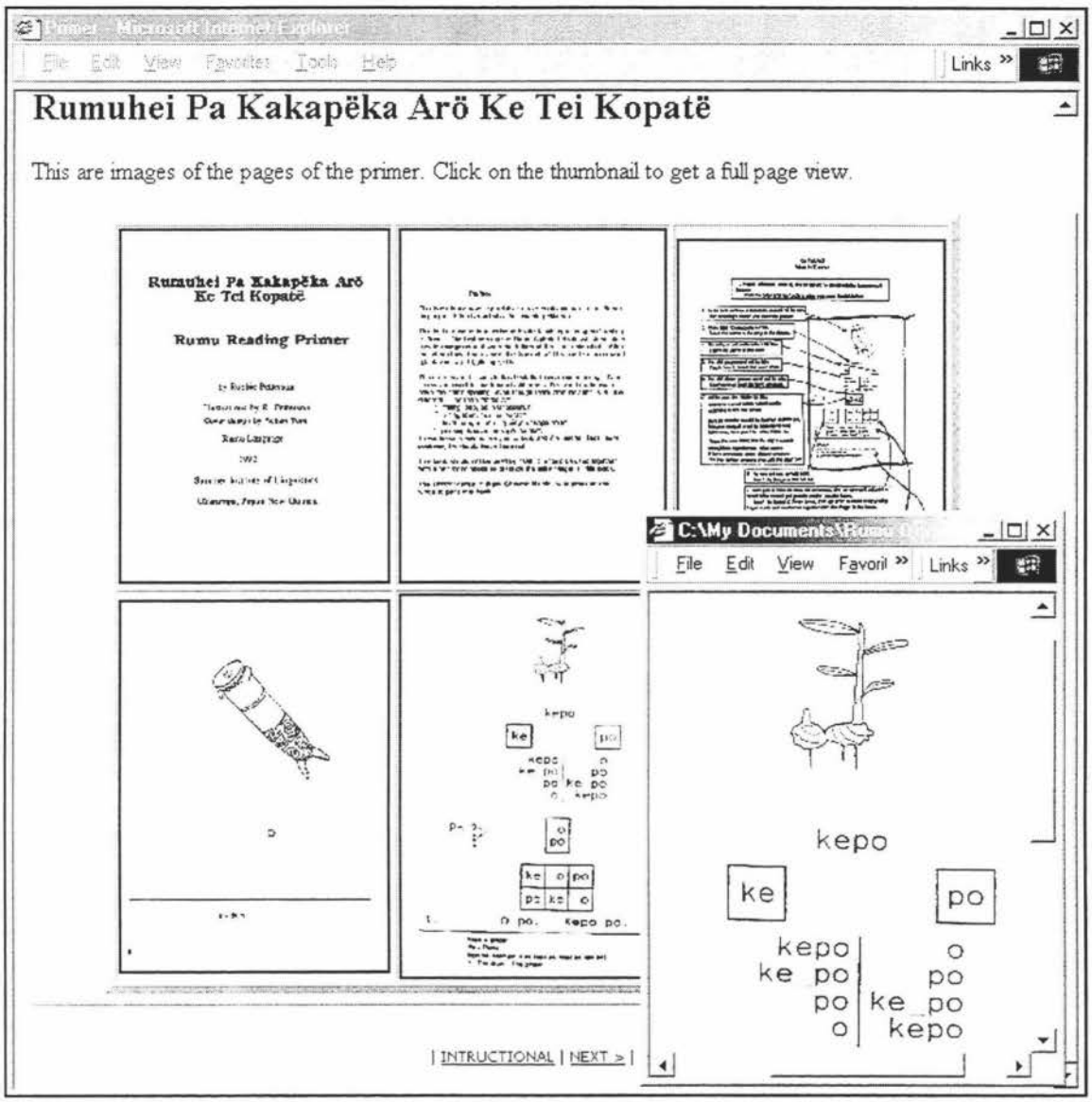
Comment

This map is one of three scanned from a rare book. High resolution (300dpi) images were obtained for printing, and these were then optimised for display on the screen by reducing to low resolution (72dpi) image. The low resolution images were displayed in a reduced size (width of 450-500 pixels) that fit the window neatly, but a click on an image opens up a full-size view of it in a separate window in order to allow closer inspection.

Access Path

History > Theodore Bevan's Account > Maps from Bevan's Book

Figure 7.16. Educational book pages as thumbnails



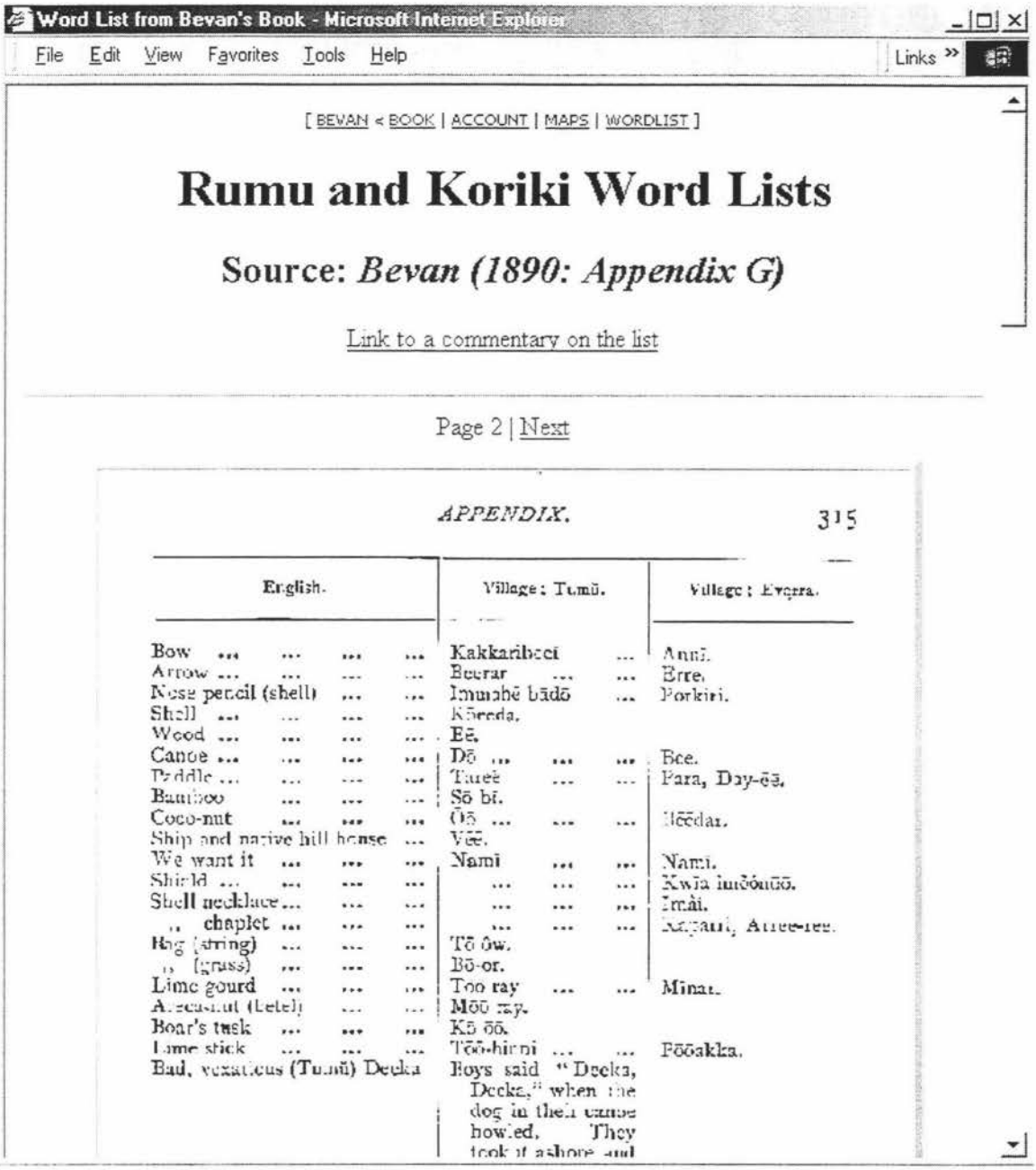
Comment

This primer was created in Word 5 for DOS using the extended character set, and pictures were pasted in on the printouts. It was easier to scan the whole page in than to combine computerised text and images. A thumbnail for each page is given, which can be expanded with a click. (Thumbnail and full page images are both GIF format.)

Access Path

Title Page > Table of contents > Instructional > Primer

Figure 7.17. A scanned printed page image



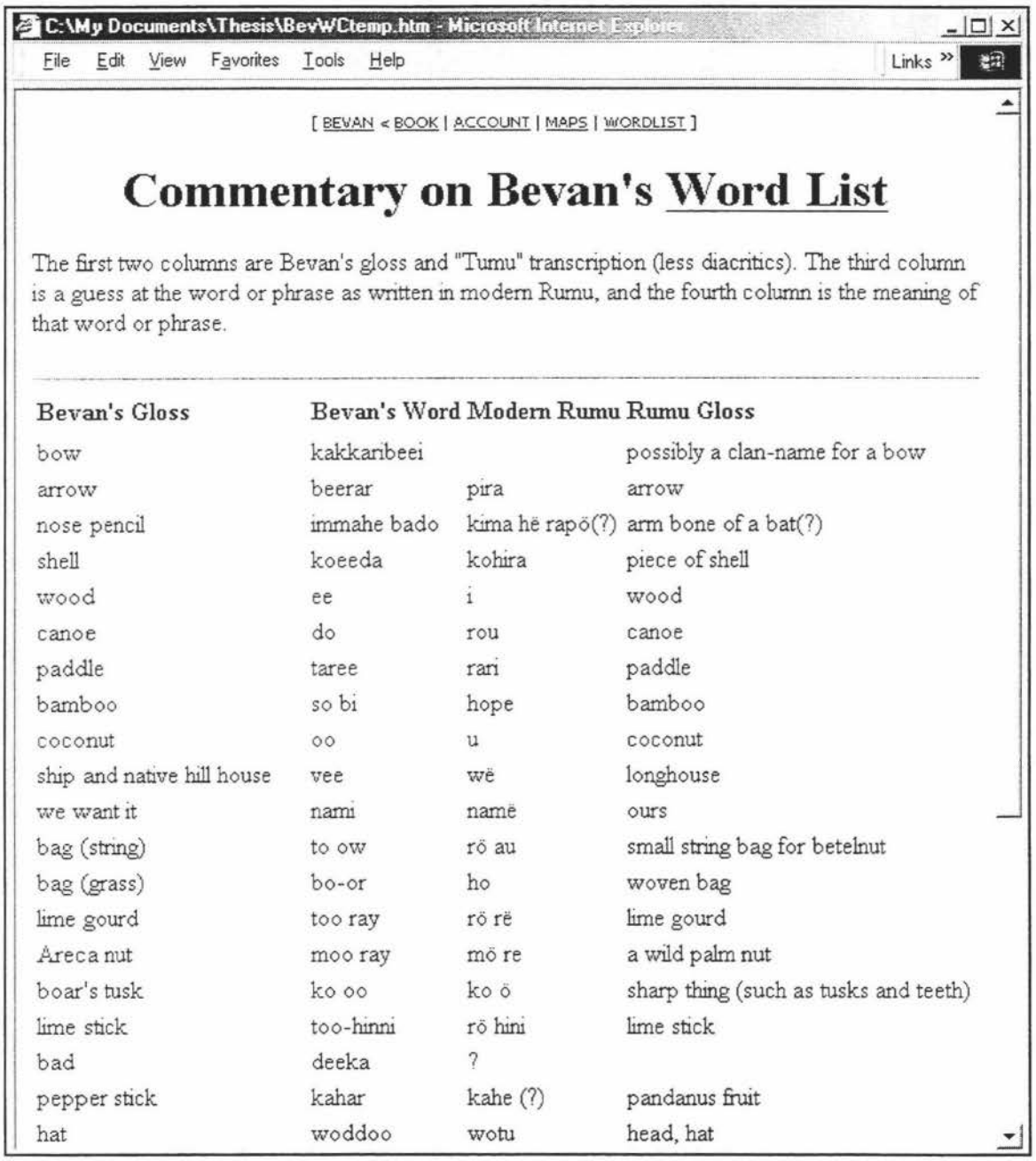
Comment

This shows a wordlist from an old book which has been scanned as a page image (GIF). The text of the page contains many diacritics, and so I decided it was better not to try to convert this to a text file at once. There is a text version of the word list in the commentary, which has a link to it in this page.

Access path

Title Page > Table of Contents > History > Bevan's Account > Word List

Figure 7.18. A commentary



Comment

This is an example of a commentary on another document, as recommended by Himmelmann. In this case the table shows an attempt to interpret the words elicited by Bevan (see Figure 7.17) in terms of my knowledge of modern Rumu. The data is set out in XHTML using a table. The transcription in column two omits the macrons used by Bevan.

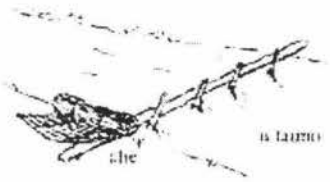
Access path

Title Page > Table of Contents > History > Bevan's Account > Word List > Commentary

Figure 7.19. Dictionary

Rumu Dictionary

ahë [ahə̀]
n. reke; fishing net or weir
- Ahë po u tai ko. *A fish weir is a thing for blocking a stream. /*
- Imo ahë maimo, hope pa. *I'm binding together a net of bamboo.*



Note: traditionally the :ahë is a curtain of bamboo strips which is put across a stream. The fishermen use derris root to stun the fish, which are trapped behind the net by the ebbing tide.

Ahëmakëau [Ahëmakëau]
n name of a bay and former village where Kopi is today
From: ahë *net* + maakë *sitting* + aū *bay*

ahi [ahī]
n. kumi; bundle
- mō ahi: *a bundle of bows and arrows /*
- ì ahi: *a bundle of firewood.*

ahi rëka [ahī rëkâ]
phr-v [T] to tie up
- Ahi rënane ahutu. *She tied up the bundle and slung it on her back. /*
- Hei po ahi rënō po, kooto te patë hekō wa. *If the dispute gets tied up in knots, it will be taken down to court.*

ahiko [ahikô]
n. See: **ehiko** plate.

ahini [ahinì]
n. hisisi, siahu; pain (of a sting)

Comment

The dictionary has headwords in regular form in bold and the full phonemic form (with tone diacritics as GIF images) in square brackets following. Senses are a numbered (none shown in this sample page) and examples are bullet-pointed. Cross-references have links to other entries, and illustrations have links to blowups.

The original dictionary was produced in Standard Format with Shoebox and formatted for publishing using Multi-Dictionary Formatteror and Microsoft Word. This XHTML sample was converted roughly from the Standard Format version by a Perl program, and then tidied up by hand.

Access path

Title Page > Table of Contents > Lists > Rumu-Motu-English Dictionary > In HTML format

Figure 7.20. Straight Rumu text showing use of non-ASCII characters

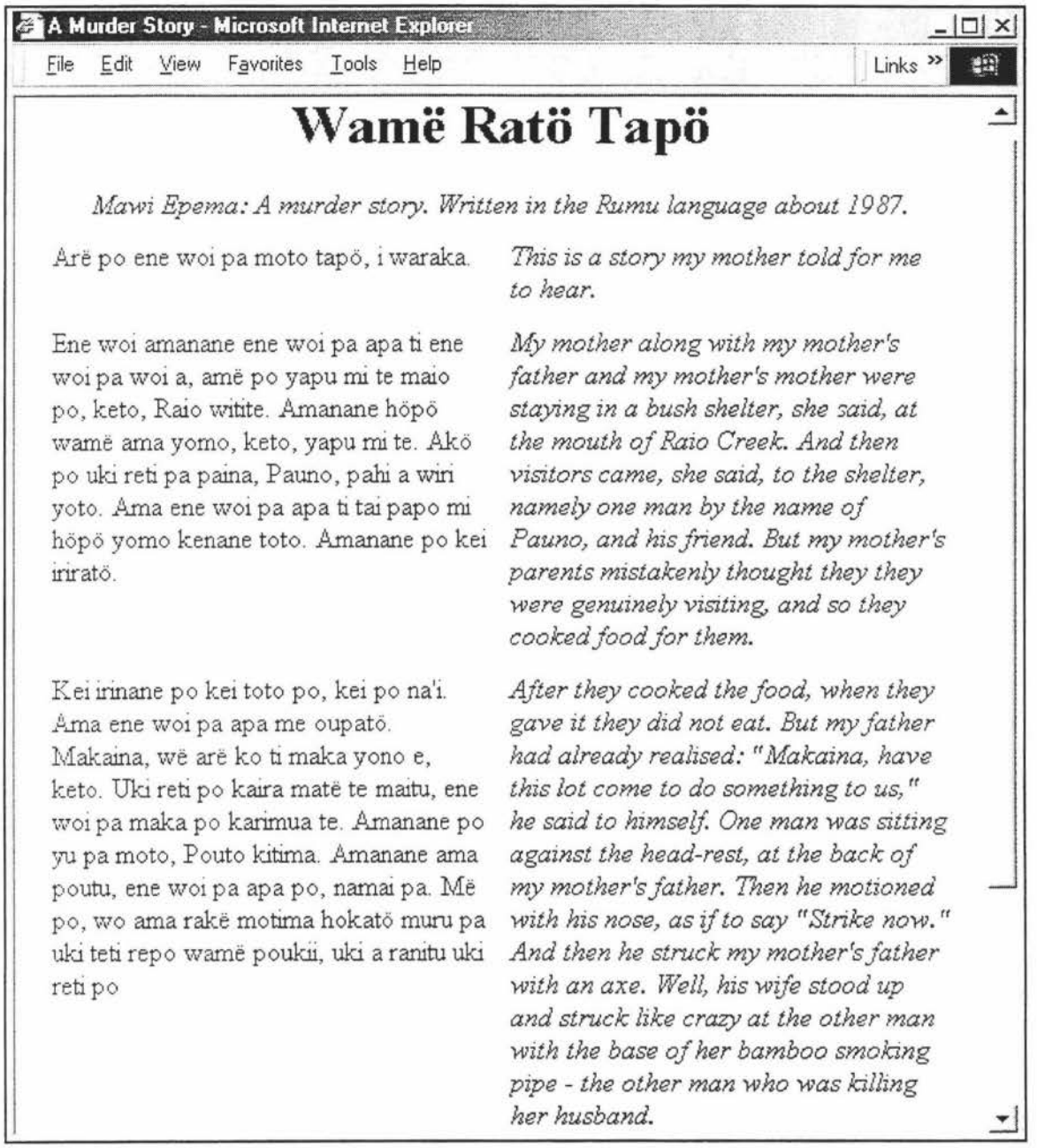
Comment

This is a story formatted for inclusion in a Rumu reader. The Rumu orthography uses some non-ASCII characters which are encoded in (X)HTML using mnemonic character entity references: *ë* for **ë** and *ö* for **ö**. (See CER and NCR in the Glossary.) To ensure that these are visible in the browser, the Unicode encoding UTF-8 should be selected.

Access path

Title Page > Table of Contents > Literature > A Murder Story

Figure 7.21. Bilingual text



Comment

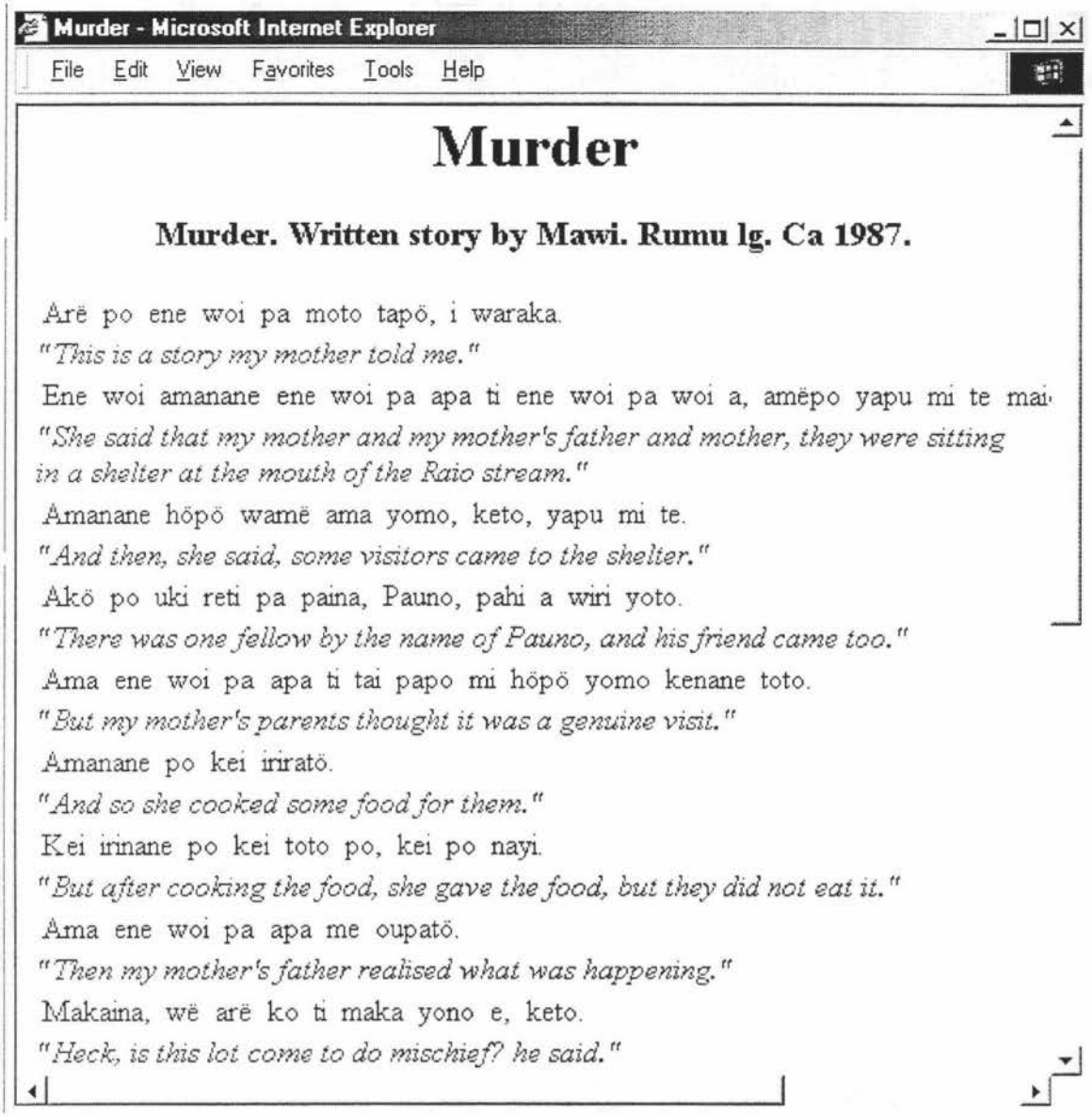
This bilingual version of the text seen in the previous figure has the English on the right in italics, matching paragraph for paragraph. It is common practice amongst linguists to put examples of the language under study in italics. In this case the English is in italics since this document is designed primarily for study by the Rumu where it is English that is the language under study!

The paragraph matching is achieved by putting each paragraph in a table cell.

Access path

Title Page > Table of Contents > Literature > A Murder Story > Rumu-English

Figure 7.22. Simple interlinear text and access via multiple paths



Comment

This is a simple sentence-by-sentence interlinear version of a text made when there is no time or no necessity to make a fully glossed interlinear version. To help the eye, the English is given in a different colour and style to the Rumu.

Because of the network of hyperlinks in the documentation, there are often several access paths to any particular document.

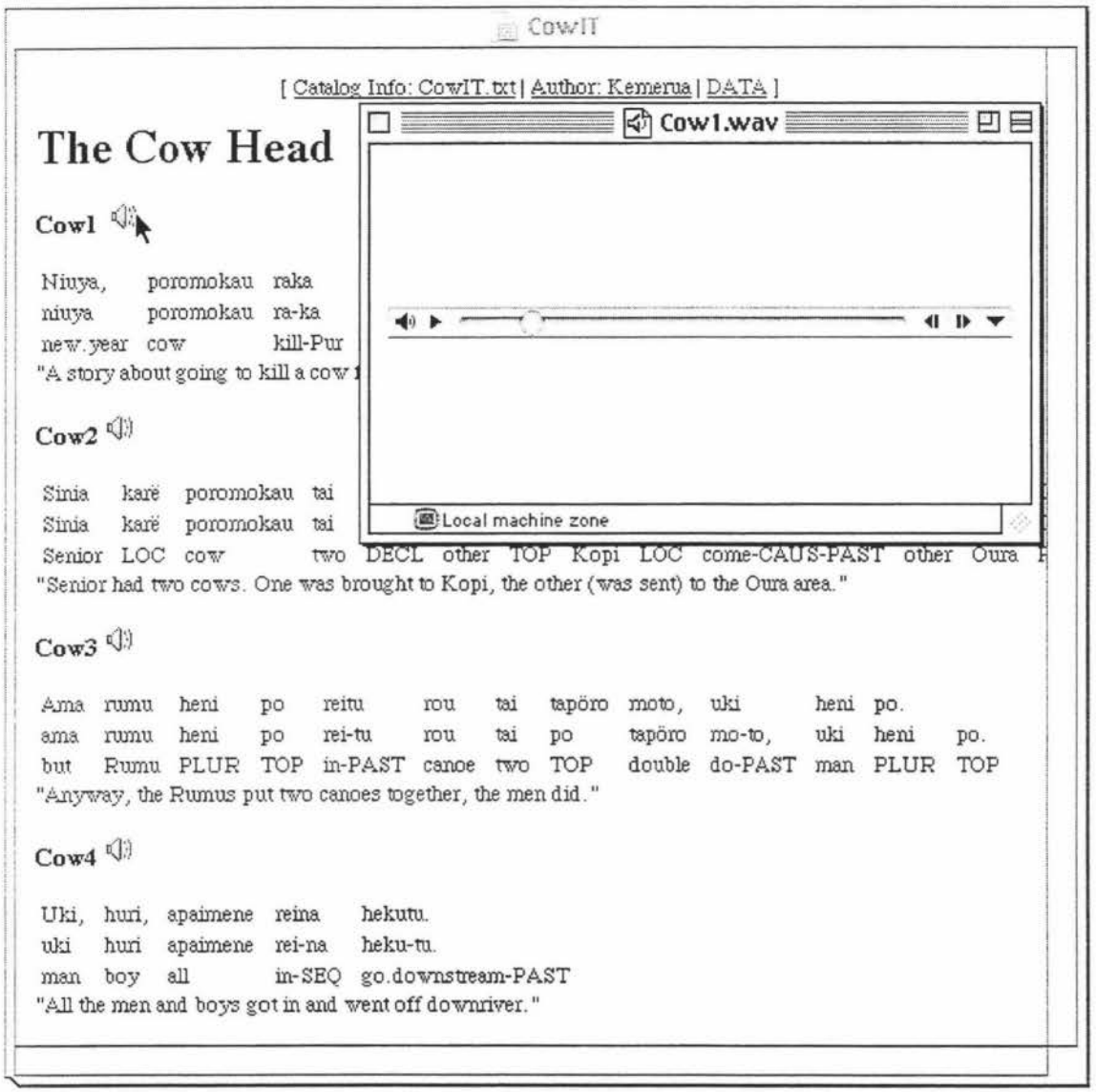
Access paths

Title Page > Table of Contents > Lists > Catalogue> Mavi Epema (1989) > Interlinear

Title Page > Table of Contents > Data > Mavi's Murder Story

Title Page > Table of Contents > Literature > A Murder Story > Interlinear version

Figure 7.23. Interlinear glossed text with sound



Comment

This interlinear text has morphemic analysis, English gloss and free translation under each line. Sound for each line can be played with a click on the loudspeaker icon. A sound player bar appears in a separate window.

The original file was glossed using Shoebox, and converted from SF to XHTML using a Perl program.

The sound files are WAV format.

Access Path

Title Page > Table of Contents > Data > Cow Head

Figure 7.24. Raw interlinear text data

file:///...terson/Rumu Doc2/Database/T/CowIT.txt

```
\id CowIT.txt. Kemerua. Tape RS7B
\name The Cow Head

\ref Cow1
\t Niuya, poromokau raka utu tapsˇ.
\m niuya poromokau ra-ka u-tu tapsˇ.
\g new.year cow kill-Pur go-Past story
\f A story about going to kill a cow for New Year

\ref Cow2
\t Sinia karʻ poromokau tai po. Ti po Kopi te yaratsˇ, ti po
\m Sinia karʻ poromokau tai po ti po Kopi te ya-ra-tsˇ, ti po
\g Senior LOC cow two DECL other TOP Kopi LOC come-CAUS-PAST
\f Senior had two cows. One was brought to Kopi, the other (w

\ref Cow3
\t Ama rumu heni po reitu rou tai tapsˇro moto, uki heni po.
\m ama rumu heni po rei-tu rou tai po tapsˇro mo-to, uki heni
\g but Rumu PLUR TOP in-PAST canoe two TOP double do-PAST man
\f Anyway, the Rumus put two canoes together, the men did.

\ref Cow4
\t Uki, huri, apaimene reina hekutu.
\m uki huri apaimene rei-na heku-tu.
\g man boy all in-SEQ go.downstream-PAST
\f All the men and boys got in and went off downriver.

\ref Cow5
\t Pereke morinanʻ hekutu. Mamansˇ pakari hekutu pa, hekutu p
\m pereke mo-ri-nanʻ heku-tu mamansˇ pakari heku-tu pa heku-t
\g fasten do-CAUS-SEQ go.down-PAST dance with go.down-PAST g
\f They went off downriver with the the canoes fastened toget

\ref Cow6
\t Mʻ, Sinia ne poromokau po ama toto.
\m mʻ Sinia ne poromokau po ama to-to.
\g done Senior ERG cow TOP then give-PAST
\f Well, Senior then gave them the cow.
```

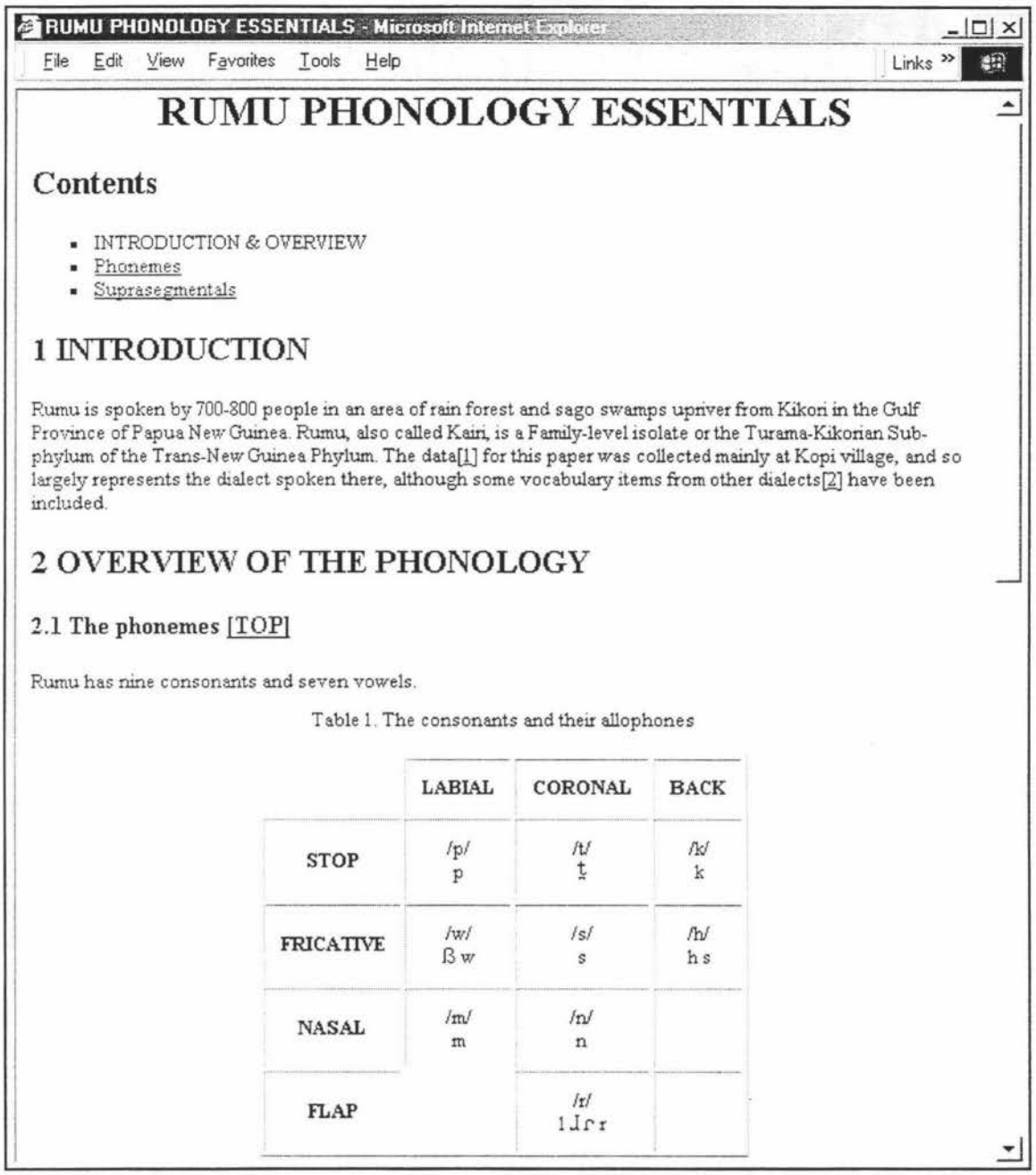
Comment

This is a view of the raw glossed text from which the display in Figure 7.23 was produced. Texts like this in their original formats are kept in the documentation so that they can be used directly in the original software. In this case the markup is Standard Format and the interlinearising was carried out using Shoebox. The strange sˇ and ʻ are a result of inconsistent character coding between computer operating systems (see section 4.5.4.2).

Access Path

Title Page > Table of Contents > Data > The Cow Head > CowIT.txt

Figure 7.25. IPA phonetic symbols



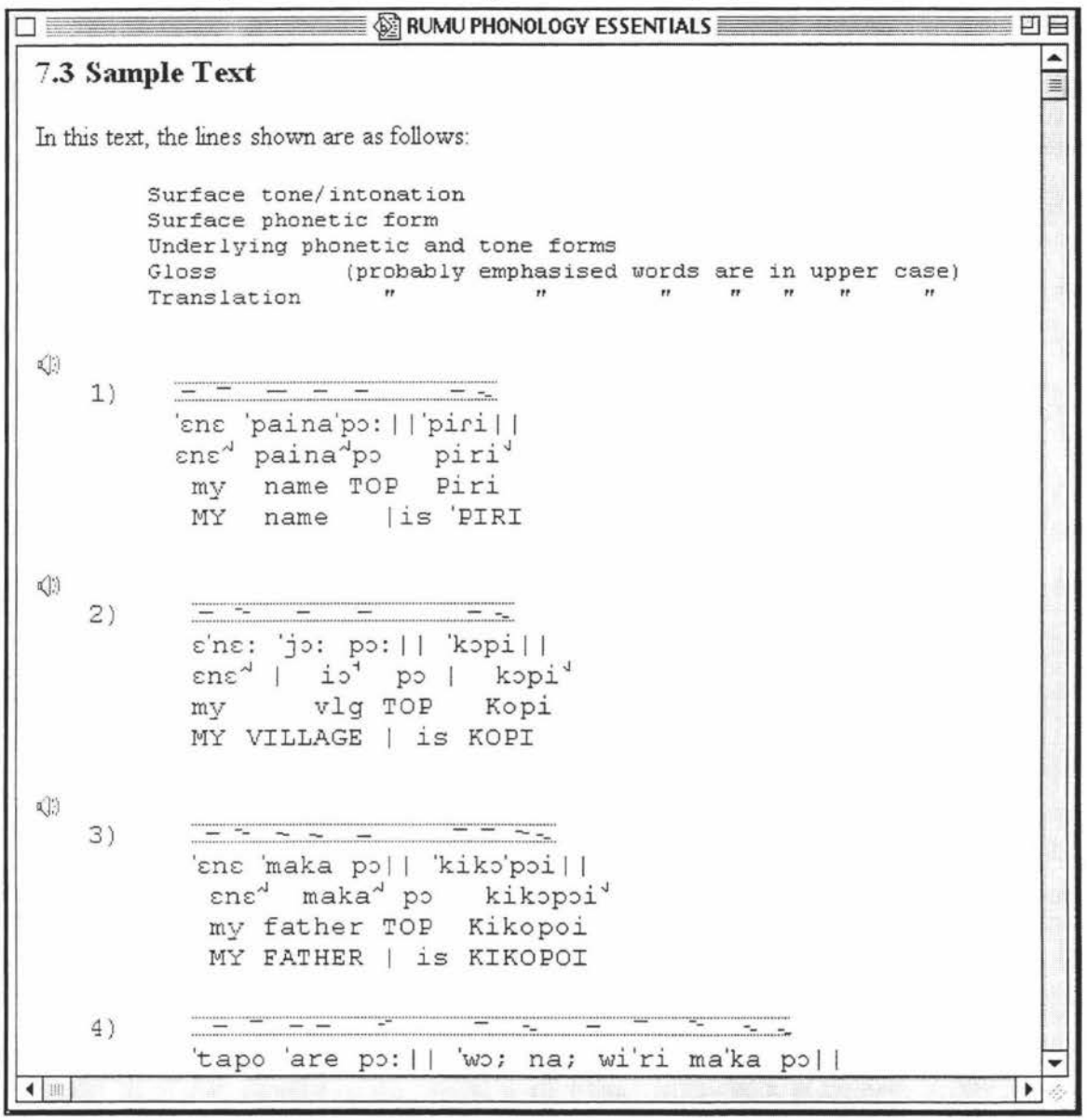
Comment

The beginning of the phonology paper contains some tables with phonetic symbols. The consonant table is visible here. Because Unicode IPA characters are not yet available for all computer platforms, the symbols are displayed as small GIF images.

Access path

Title Page > Table of Contents > Analyses> Phonology > Introduction and Overview

Figure 7.26. Interlinear phonetic data display with sound



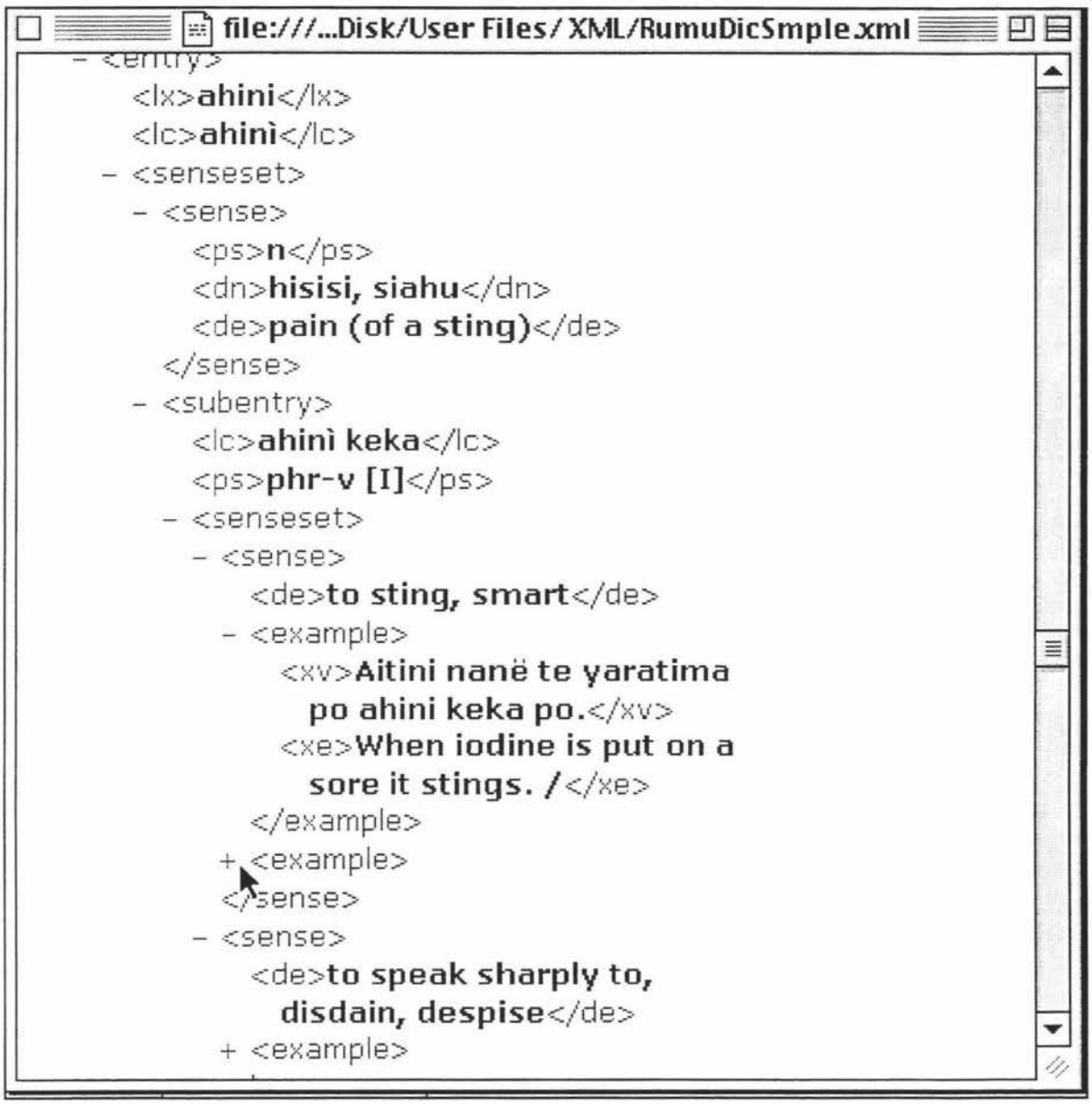
Comment

This sample of phonetic data was produced in a study on tone and intonation, and shows intonation "tram-tracks", segmental data in IPA, and phonemic data. A click on a loudspeaker icon will play the sound in a separate window. The original file for this sample was a text file marked up using SF codes for printing with Manuscripter. The intonation tramtracks used fonts developed for the FindPhone phonetic analysis software. In the XHTML file the tramtrack components and IPA symbols are GIF images, and the sound files are AIFC format.

Access Path

Title Page > Table of Contents > Analysis > Phonology > Sample

Figure 7.27. Browser view of dictionary database with XML tags



Comment

This is a view of a dictionary entry marked up in XML as it is displayed in Internet Explorer version 5, which was the only browser at the time of writing that could be found that could handle XML. The - and + signs indicate the beginning of a tag containing embedded tags. The + sign indicates that the embedded material has been collapsed. When the user clicks on the + sign, the embedded material is immediately expanded.

Access Path

Title Page > Table of Contents > Lists > Rumu-Motu-English Dictionary > In XML format

8. Conclusion

In writing this thesis I found that linguists commonly use metaphors that compare language to life. This seems reasonable, for language is a part of life, indeed a defining part of human life, for no other life form has a communication system that could be properly called language. One such metaphor is *a language as a living being* which can live, weaken, die and even be revived. This seems a little inadequate as a metaphor in that a language can "weaken" and "die" in some places, but remain "vital" in another. The other, more apt metaphor is *a language as a living species* which evolves and has an ecology and a potential for survival or extinction.

Using these metaphors I have discussed the alarming status of the world's languages, noting the parallel between the language situation in Australia, where 90% of the languages have become extinct or moribund in the last 200 years, with that of the world, where, according to some estimates, 90% of the world's languages could be in the same situation 200 years from now. I have argued that language documentation is one valid response to the impending extinction crisis, a response that not only preserves maximal language data for posterity beyond language extinction, but, with the proper methodology and attitude, a response that could help to bring about a reversal in language decline.

While working to develop a suitable methodology for documentation in this thesis, I have found a number of tensions between different interests that need to be resolved. One tension is sophistication versus simplicity. On the one hand there is a desire to develop sophisticated software that will allow cooperative research programs, manage projects, and enable powerful analyses. A number of useful tools now exist in workable form. In particular, I have some personal experience with Shoebox, and have experimented with LinguaLinks in the course of this study. On the other hand the kind of documentation that is required to meet the present crisis does not need to wait for more and more powerful software tools that demand much effort in developing and testing and training in their use. It may be better for a language documentalist to learn to

use a few simple tools that are good for the accumulation and organisation of data, and then to get on with the work in hand.

For the methodology I have proposed, as a minimum,[28] the field worker should know how to (a) create and edit text documents that cope with the writing system of the target language, (b) create and edit web pages, and (c) capture sounds and images as digital files and put them into an optimum format. It would also be very useful to know how to (d) use a linguistic database program (such as Shoebox or LinguaLinks or Kura) that can cope with the writing system and assist the researcher with interlinearisation of texts and dictionary work, and (e) work with XML. A team approach may be necessary, because, as Himmelmann (1998, p.171) notes, "the compilation of a high-quality language documentation generally requires interdisciplinary cooperation as well as close cooperation with members of the speech community."

Another tension related to this first one is desirable tools versus available tools. It is desirable, for example, to use Unicode for all text character needs, but Unicode fonts are not universally available on all computer platforms just yet. Because of that GIF "image fonts" may need to be used in the meantime. Another example is the use of XML as a markup language. Tools (especially linguistic tools) that can develop and interpret XML files are not yet universally available, and in the meantime RSF or some other format will need to be used which can be converted easily to HTML or XHTML, and, in the future, XML.

A third tension is the desires of the academic community versus the desires of the researched language community. I have argued for a methodology of documentation that involves the language community as much as possible, seeking to meet their perceived short-term language-related needs as a priority. It can be expected that large amounts of data suitable for analysis and support of theories developed in the academic community will be a by-product, as it were, of this kind of documentation.

A fourth tension is the accessibility of digital data by the academic community versus the accessibility of analogue data by the language community. This thesis has

[28] Given, of course, proper and adequate training in linguistic and anthropological field methods.

argued for a documentation that is primarily formatted in the digital medium, using international standards, and structured as an internet-ready hypermedia "Web site". This would serve as an easily upgradable repository for valuable linguistic and cultural *taonga*. It would also be accessible by both the academic community and the ethnic community, provided they have ready access to computer technology. If they do not, it would be easy to obtain analogue versions of the digital data for them. (These analogue versions should also be made for archives, because, as has been illustrated, analogue data tends to be more easily recoverable than digital in the long term.)

A fifth tension concerns the rights of access to and control over the data by various stakeholders. The rights of the language community must be given every consideration, especially when some of the material is regarded as sensitive or "tabu". The implications of "ownership" of the data need to be fully explained as clearly as possible, as also the implications of making data available over the internet. For people not used to the concepts of cooperative research and the sharing of data with the world at large, the field worker has a responsibility to educate and to try to avoid imparting unrealistic expectations.

Finally, the results of a pilot project have been illustrated, a pilot project which explored the incorporation of all the main forms of document to be included in the proposed language documentation, and which tested many of the structures and formats proposed for it. Although it can be expected that the technology (both hardware and software) that have been used will be superseded in the near future, the formats and structures employed will still be usefully accessible for a long time to come. Feedback about this pilot project from the academic and ethnic community should result in an improved model, on the basis of which a full documentation of the language can be pursued. Linguistic fieldwork can be a long term occupation, and so various versions of the documentation should be published on CD and on the World Wide Web from time to time so that the expectations of the stakeholders can be met in good time.

APPENDIX A. ORGANISATIONS

Table A1. A selection of organisations concerned for minority and endangered languages worldwide

Acronym	Name and details	URL
ELF	**The Endangered Language Fund**	http://www.ling.yale.edu:16080/~elf/
	The ELF aims to support the study and maintenance of endangered languages and the dissemination of the results of these efforts to both the academic world and the language communities.	
FEL	**Foundation for Endangered Languages**	http://www.ogmios.org/
	The FEL has aims similar to those of the ELF, and maintains a useful set of links to like-minded organisations.	
ICHEL	**The International Clearing House for Endangered Languages**	http://www.tooyoo.l.u-tokyo.ac.jp/ichel/ichel.html
	ICGEL maintains the UNESCO Red Book of Endangered Languages (Wurm et al., 2002) and also an extensive bibliography on endangered languages (Tsunoda, 2000).	
LDC	**Linguistic Data Consortium**	http://www.ldc.upenn.edu/
	The LDC maintains a directory of language Web-based archiving projects (Bird, n.d.)	
LSA	**Linguistic Society of America**	http://www.linguistlist.org/el-page/
	The LSA works with LinguistList to maintain the Endangered Languages Homepage.	
SIL	**SIL International** (formerly **Summer Institute of Linguistics**)	http://www.sil.org/
	SIL has a large membership of linguists working in minority and endangered language groups.	
Terra-lingua	**Partnerships for Biological and Linguistic Diversity**	http://www.terralingua.org/
	Terralingua aims to foster linguistic and biological diversity in the world through research, information and advocacy.	

Table A2. A selection of online organisations concerned with developing tools and standards for multimedia documentation of languages

Acronym	Name and details	URL
ELSNET	**European Network of Excellence in Human Language Technologies** ELSNET's main objective is "to advance human language technologies in a broad sense by bringing together Europe's key players in research, development, integration or deployment in the field of language and speech technology and neighbouring areas."	http://www.elsnet.org/
LDC	**Linguistic Data Consortium** The LDC maintains directories of organisations which have useful resources for linguistic exploration (Bird, 2000) and annotation (Bird & Liberman, 2001).	http://www.ldc.upenn.edu/
OLAC	**Open Language Archives Community** One of the aims of OLAC is to "develop consensus on best current practice for the digital archiving of language resources".	http://www.language-archives.org/
SALTMIL	**Special Interest Group on Speech and Language Technology for Minority Languages** SALTMIL has as one of its aims to provide "a channel of communication between minority language researchers and those active in speech and language technology in general."	http://isl.ntftex.uni-lj.si/SALTMIL/
SIL	**SIL International** (formerly **Summer Institute of Linguistics**) SIL has developed some excellent software tools for linguistic analysis.	http://www.sil.org/

APPENDIX B. EXAMPLES OF MULTIMEDIA LANGUAGE DATA FROM THE WORLD WIDE WEB

Text

Figure B1. Whole text with translation (and sound)

Jiwarli	**English**
Kapakurta mantharta mikalyaji paja yananyja manthartawu. Yini pipijunkurru. Maatha ngunha manthartanyjarriyi pipijunkurru. Warri nhukuparnti ngunha paja yananyja. Ngunhakayi kajiriwari kamparninyjalu kajiriyi kamparninyja ngunhipa yirrara. Ngunha wirlkamanta	The nightjar and bat were angry with a man. His name was *Pipijunkurru*. That *Pipijunkurru* was the boss of the people. They didn't go along angry from nearby. After they first heated (straightened) spears at Mt Florrie, they heated them there at the top. There is

This shows part of a text with side-by-side translation that was formatted in HTML using a two-column table. There are hyperlinks (not shown) to sound files (AU format) for the whole text in both Jiwarli and English. [Jiwarli language, from Austin (1998) at http://www.linguistics.unimelb.edu.au/research/projects/jiwarli/story.html]

Figure B2. Interlinear text

4. *Ita*	*bele*	*fahe*	*Ema*	*Timor*	*nia*	*moris*	*iha*	*fatin*	*rua.*
we.INCL	can	divide	person	Timor	POSS	life	in	place	two

'We can categorise East Timorese and their lives into two.'

This is a glossed and translated sentence taken from an analysed text of several paragraphs. Alignment is achieved in HTML using tables with three rows, where each word in the first two lines is in a separate table cell, and the free translation is contained in just one cell of the third row. [Tetun language, from Newman (2000) at http://www.massey.ac.nz/~wwlingui/Tetun/gloss.html]

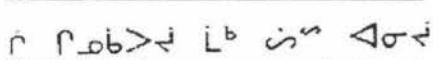
Figure B3. Interlinear text with special characters

```
ok0008
Nɛ́ɛ̀tâỳ          apá       tíɔ́pa       ɔltʊŋánì          apá  ojî      Inkimís,
n -ɛ-ata -ɪ       apá       tè -ɔpá      ɔl -tʊŋánì        apá  oj-ɪ      Inkimís
CN-3-have-PASS    long.ago  OBL-former   MSG-person:ACC    ago  say-PASS  Inkimís
'A long time ago there was a person named Inkimis.'
```

This is interlinear text prepared using Shoebox 4.0. The special characters are GIF images. Alignment is achieved in Shoebox by the use of spacing, and this has been transferred to HTML by converting the space characters to non-breaking space characters (in HTML:). [Maasai language, from Payne (2000) at http://darkwing.uoregon.edu/~dlpayne/maasai/matext.htm]

Figure B4. Text in non-Roman orthography

Line 8



Chii chinukaapuyuu maak shaash aniyuu ukusa.
He was tall already his son.

This sentence is written in a special syllabary with a non-Roman font. For display on the Web it is has been converted to a single GIF image. The sentence is from a larger text which is given in full on another page, where the whole text is a single image. Each sentence also has a transliteration in a Roman-based orthography and a very literal translation in English. The sound recording is as an embedded Quicktime MOV file. Each line is formatted in HTML as a row in a table. [Eastern Cree language, from Edwards (2000) at http://uiuuii.com/eastcreegrammar/eng/texts/ht/ht_lines.html]

Sound

Figure B5. Pronunciation guide with program-controlled sound

The [ɛ] sound is identical to the vowel sound of *pet*. The letters ê, è and e are normally pronounced [ɛ].

bête	*beast*	mère	*mother*
belle	*beutiful*	telle	*such*
mène	*bring*		

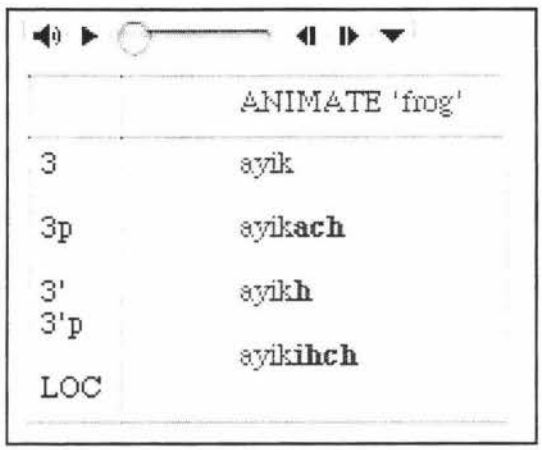
When the user clicks on the speaker icon, a recording of someone reading through the example is played with sound controls appearing in a separate small window. The sound file format is MP3, and is started up by a JavaScript function. The 'open e' IPA symbol is a GIF image (see section 6.2.3). [French language, from Blackmon (2001) at http://www.frenchlesson.com/grammar/pronun/vowels_e.html]

Figure B6. Pronunciation guide with hyperlinked sound

Swahili	*as in English*	*Examples*	
a	ah!	baba	"father"
e	say	wewe	"you"
i	be	kiti	"chair"
o	ho!	moto	"fire"
u	too	tu	"only, just"

The pronunciation of a word can be heard by clicking on the word itself. The sound file format is **WAV**, and (unfortunately) the sound player takes over the browser window, so that one has to push the Back button to see the text again. [Swahili language, from Hinnebusch & Mirza (1995) at http://www.cis.yale.edu/swahili/sound/pronunce.htm]

Figure B7. Paradigm with sound



	ANIMATE 'frog'
3	ayik
3p	ayik**ach**
3'	ayik**h**
3'p	ayik**ihch**
LOC	

The paradigm is displayed as a simple table. The sound recording (MOV format) is of someone reading through the paradigm. [Eastern Cree language, from Junker (2001) at http://www.carleton.ca/~mojunker/eastcreegrammar/pages/english/nouns/paradigms.html]

Figure B8. Phonetic text example with normal and slow speech sound

ʔɪskà: də ra:na:

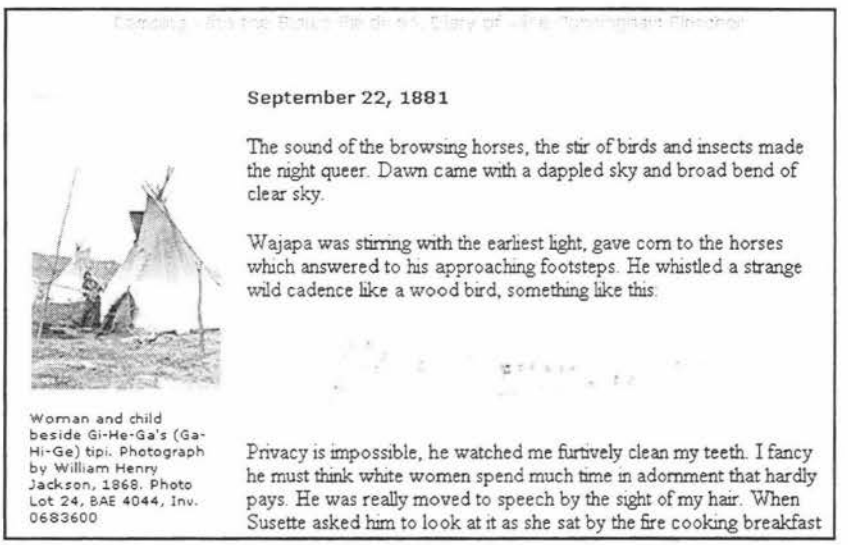
wətə ra:na:, də ʔɪskər hùntu:rù:
tə ʔərè:wəʔ də ra:na: sukə ji
gərdəmà: ʔə kən ko:wà:tʃe:tʃə:
dəgə tʃikɪnsù tə ɸi k'ərɸi:

Hausa passage spoken normally (412K) | Hausa passage spoken slowly (616K)

This phonetic example can be played at normal speed and as if spoken slowly. The slowing down of the speech while maintaining pitch was achieved using software called Praat. The phonetic text is a single GIF image. [Hausa language, from Newman (2001) at http://vcsymposium.massey.ac.nz/presenters/presentations/Newman/www/John%27s%20WebCT/Hausa/Hausa.html]

Images

Figure B9. Transcribed field notes



September 22, 1881

The sound of the browsing horses, the stir of birds and insects made the night queer. Dawn came with a dappled sky and broad bend of clear sky.

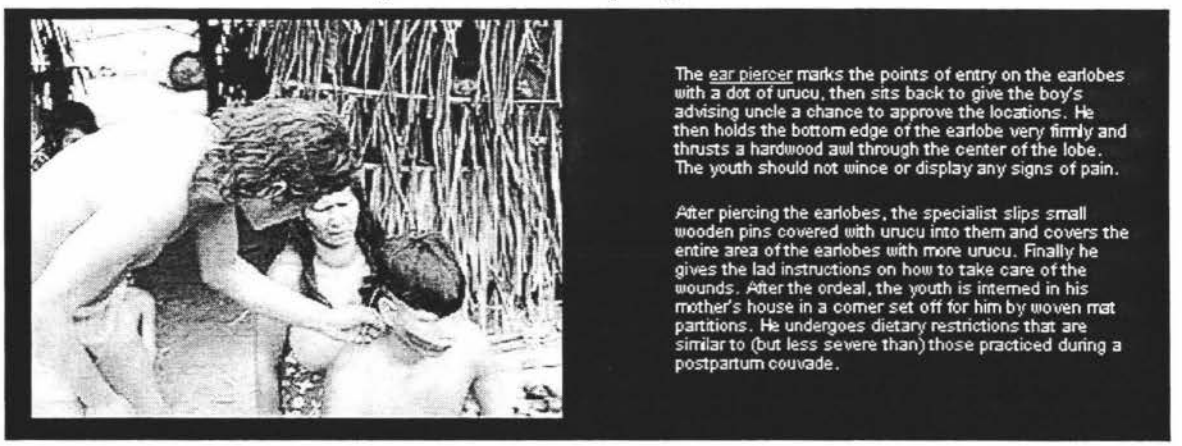
Wajapa was stirring with the earliest light, gave corn to the horses which answered to his approaching footsteps. He whistled a strange wild cadence like a wood bird, something like this.

Woman and child beside Gi-He-Ga's (Ga-Hi-Ge) tipi. Photograph by William Henry Jackson, 1868. Photo Lot 24, BAE 4044, Inv. 0683600

Privacy is impossible, he watched me furtively clean my teeth. I fancy he must think white women spend much time in adornment that hardly pays. He was really moved to speech by the sight of my hair. When Susette asked him to look at it as she sat by the fire cooking breakfast

This is a neat example of a transcription of an originally hand-written field diary, a typewritten version of which was processed using OCR. Accompanying sketches and photographs have been digitised in JPEG format. [Sioux people, from Leopold (2001) at http://www.nmnh.si.edu/naa/fletcher/acf_sept_22.htm]

Figure B10. Anthropological notes



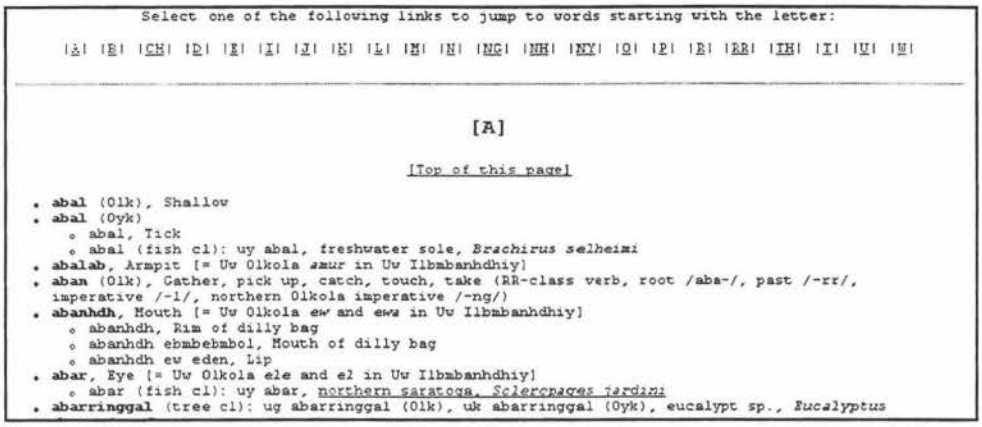
The ear piercer marks the points of entry on the earlobes with a dot of urucu, then sits back to give the boy's advising uncle a chance to approve the locations. He then holds the bottom edge of the earlobe very firmly and thrusts a hardwood awl through the center of the lobe. The youth should not wince or display any signs of pain.

After piercing the earlobes, the specialist slips small wooden pins covered with urucu into them and covers the entire area of the earlobes with more urucu. Finally he gives the lad instructions on how to take care of the wounds. After the ordeal, the youth is interned in his mother's house in a corner set off for him by woven mat partitions. He undergoes dietary restrictions that are similar to (but less severe than) those practiced during a postpartum couvade.

An illustrated account of an event important to a culture, it contains a series of photographs (JPEG) and text, formatted in HTML using tables. [Canela people of Brazil, from Crocker & Leopold (n.d.) at http://www.nmnh.si.edu/naa/canela/canela1.htm]

Hyperlinks

Figure B11. Browsable dictionary with links to encyclopedic information



Select one of the following links to jump to words starting with the letter:

|A| |B| |CH| |D| |E| |I| |J| |K| |L| |M| |N| |NG| |NH| |NY| |O| |P| |R| |RR| |TH| |T| |U| |W|

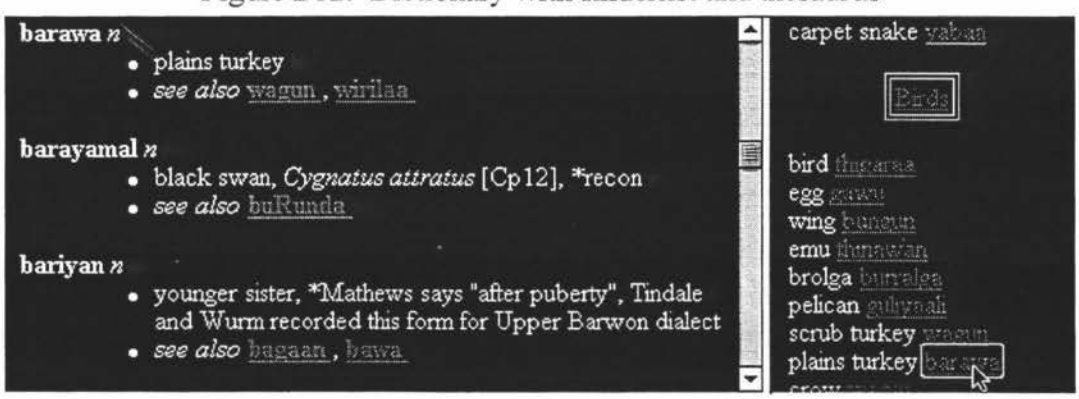
[A]

[Top of this page]

- **abal** (Olk), Shallow
- **abal** (Oyk)
  - abal, Tick
  - abal (fish cl): uy abal, freshwater sole, *Brachirus selheimi*
- **abalab**, Armpit [= Uw Olkola *amur* in Uw Ilbmbanhdhiy]
- **aban** (Olk), Gather, pick up, catch, touch, take (RR-class verb, root /aba-/, past /-rr/, imperative /-l/, northern Olkola imperative /-ng/)
- **abanhdh**, Mouth [= Uw Olkola *ew* and *ewa* in Uw Ilbmbanhdhiy]
  - abanhdh, Rim of dilly bag
  - abanhdh ebmbebmbol, Mouth of dilly bag
  - abanhdh ew eden, Lip
- **abar**, Eye [= Uw Olkola *ele* and *el* in Uw Ilbmbanhdhiy]
  - abar (fish cl): uy abar, northern saratoga, *Scleropages jardini*
- **abarringgal** (tree cl): ug abarringgal (Olk), uk abarringgal (Oyk), eucalypt sp., *Eucalyptus*

This dictionary is in one browsable page, with links at the top to allow the user to jump to the start of a letter section. There are links in some of the entries to encyclopedic information, including photographs. [Uw Olkola and Uw Oykangand, from Hamilton (1996) at http://www.geocities.com/Athens/Delphi/2970/olkola.htm]

Figure B12. Dictionary with finderlist and thesaurus



barawa *n*
- plains turkey
- *see also* wagun , wirilaa

barayamal *n*
- black swan, *Cygnatus atratus* [Cp12], *recon
- *see also* buRunda

bariyan *n*
- younger sister, *Mathews says "after puberty", Tindale and Wurm recorded this form for Upper Barwon dialect
- *see also* bagaan , bawa

carpet snake yabaa

Birds

bird thugaraa
egg gunu
wing buneun
emu thunawan
brolga burralea
pelican gulyuah
scrub turkey
plains turkey baragga

This dictionary has extensive cross-referencing using hyperlinks. An English finder list or thesaurus (right-hand frame) can be used to find an entry in the main dictionary (left-hand frame). [Kamilaroi language, from Austin & Nathan (1996) at http://coombs.anu.edu.au/WWWVLPages/AborigPages/LANG/GAMDICT/GAMDICTF.HTM]

REFERENCES

Altova. (2002). XML Spy. Retrieved 18 Mar 2002 from http://new.xmlspy.com

Antworth, E. L., & Valentine, J. R. (1998). Software for doing field linguistics. In J. Lawler & H. A. Dry (Eds.), *Using computers in linguistics: A practical guide* (pp. 170–196). London & New York: Routledge.

Antworth, E. L., & Valentine, J. R. (January 11, 1999). Software for doing field linguistics: Online appendix. Retrieved 14 Mar 2002 from http://www.sil.org/computing/routledge/antworth-valentine/

Apple. (2001). Writing systems and script systems. Retrieved 31 Jul 2001 from http://www.devworld.apple.com/techpubs/mac/Text/Text-28.html

Associated Press. (June 18, 2001). Half of the word's 6,800 languages could die by 2100. Retrieved 20 Jun 2001 from http://asia.cnn.com/2001/US/06/18/disappearing.languages.ap/index.html

Austin, P. (1998). Jiwarli: A language of Western Australia. Retrieved 24 Jan 2001 from http://www.linguistics.unimelb.edu.au/research/jiwarli/index.html

Austin, P., & Nathan, D. (1998). Gamilaraay (Kamilaroi) dictionary and knowledge base. Retrieved 24 Jan 2001 from http://www3.aa.tufs.ac.jp/~austin/GAMIL.HTML

Baker, P. (2001). Foundation for endangered languages: Manifesto. Retrieved 4 Jan 2002 from http://www.ogmios.org/manifesto.htm

Bell, J., & Bird, S. (2000). A preliminary study of the structure of lexicon entries. Retrieved 4 Dec 2001 from http://morph.ldc.upenn/exploration/expl2000/papers/bell/bell.html

Berners-Lee, T. (2001). Information management: A proposal. In R. Packer & K. Jordan (Eds.), *Multimedia: From Wagner to virtual reality* (pp. 189–205). New York & London: W. W. Norton & Company. (Original article: Berners-Lee, T. (1989) *Information management: A proposal*. Geneva, Switzerland: CERN.)

Biggs, B. (1973). *Let's learn Maori: A guide to the study of the Maori language*. Wellington: Reed Education.

Bird, S. (September 25, 2000). Linguistic exploration. *Linguistic Data Consortium*. Retrieved 23 Jan 2001 from http://www.ldc.upenn.edu/exploration/

Bird, S. (n.d.). Archives for language documentation and description. Retrieved 29 Jan 2001 from http://www.ldc.upenn.edu/exploration/archives.html

Bird, S., & Liberman, M. (January 1, 2001). Linguistic annotation. *Linguistic Data Consortium*. Retrieved 21 Mar 2001 from http://morph.ldc.upenn.edu/annotation/

Bird, S., & Simons, G. (2000). Linguistic exploration. Workshop on web-based language documentation and description. Retrieved 23 Jan 2001 from http://www.ldc.upenn.edu/exploration/expl2000/

Blackmon, T. (2001). Frenchlesson.org. *Language Revolution.* Retrieved 26 Feb 2002 from http://www.frenchlesson.com/index.htm

Boersma, P. (2001). Praat program. Retrieved 19 Feb 2002 from http://fonsg3.let.uva.nl/praat/manual/Praat_program.html

Bray, T., & Sperberg-McQueen, C. M. (November 14, 1996). Extensible Markup Language (XML). Retrieved 19 Mar 2002 from http://www.w3.org/TR/WD-xml-961114

Burnard, L., & Sperberg-McQueen, C. M. (1993). Living with the guidelines: An introduction to TEI tagging. Retrieved 3 Jul 2001 from http://www.oasis-open.org/cover/teiu5-uva.html

Buszard-Welcher, L. (2001). Can the Web help save my language? In L. Hinton & K. Hale (Eds.), *The green book of language revitalisation in practice* (pp. 331–352). San Diego: Academic Press.

Cairncross, F. (1997). *The death of distance: How the communications revolution will change our lives*. Boston, Mass.: Havard Business School Press.

Cameron, D., Frazer, E., Harvey, P., Rampton, B., & Richardson, K. (1993). Ethics, advocacy and empowerment: Issues of method in researching language. *Language & Communication*, 13, 81–94.

CIEMEN. (1996). Universal declaration of linguistic rights. Retrieved 20 Apr 2000 from http://www.egt.ie/udhr/udlr-en.html

Clark, J. (n.d.). DSSSL. Retrieved 11 Mar 2002 from http://www.jclark.com/dsssl/

Craig, C. G. (1992). Fieldwork on endangered languages: A forward look at ethical issues. Paper pre-circulated at the plenary session on endangered languages presented at XVth International Congress of Linguists, Quebec, August 10 1992.

Crocker, W. H., & Leopold, R. (n.d.). Canela body adornment. *Smithsonian Institute*. Retrieved 29 Jan 2001 from http://www.nmnh.si.edu/naa/canela/canela1.htm

Crowley, T. (1999a). Review of the book *The rise and fall of languages*. *Australian Journal of Linguistics*, 19, 109–115.

Crowley, T. (1999b). The socially responsible lexicographer in Oceania. *Journal of Multilingual and Multicultural Development*, 20, 1–12.

Crystal, D. (1997). *The Cambridge encyclopedia of language* (2nd ed.). Cambridge: Cambridge University Press.

Crystal, D. (2000). *Language death*. Cambridge: Cambridge University Press.

Dixon, R. M. W. (1991). The endangered languages of Australia, Indonesia and Oceania. In R. H. Robins & E. M. Uhlenbeck (Eds.), *Endangered languages* (pp. 229-256). Oxford & New York: Berg.

Dixon, R. M. W. (1997). *The rise and fall of languages*. Cambridge: Cambridge University Press.

Dorian, N. C. (1998). Western language ideologies and small-language prospects. In L. A. Grenoble & L. J. Whaley (Eds.), *Endangered languages: Language loss and community response* (pp.3-21). Cambridge: Cambridge University Press.

Edmondson, J. A., & Burquest, D. A. (1994). *A survey of linguistic theory* (2nd ed.). Dallas, TX: Summer Institute of Linguistics.

Edwards, L. (June 19, 2000). East Cree texts: Hard times (line-by-line version). Retrieved 26 Feb 2002 from http://uiuuii.com/eastcreegrammar/eng/texts/ht/ht_lines.html

Global Reach. (2001a). Global Internet Statistics (by language). Retrieved 17 Jul 2001 from http://www.glreach.com/globstats/index.php3

Global Reach. (2001b). Global internet statistics: sources & references. Retrieved 18 Jul 2001 from http://www.glreach.com/globstats/refs.php3

Goldfarb, C. F. (1996). The roots of SGML: A personal recollection. Retrieved 20 Feb 2002 from http://www.sgmlsource.com/history/roots.htm

Grenoble, L. A., & Whaley, L. J. (Eds.). (1998). *Endangered languages:Language loss and community response*. Cambridge: Cambridge University Press.

Grenoble, L. A., & Whaley, L. J. (2002a). What does digital technology have to do with Yaghan? *Linguistic Discovery*, 1, http://linguistic–discovery.dartmouth.edu/WebObjects/Linguistics.woa/1/page/article/101.htmlOnce

Grenoble, L. A., & Whaley, L. J. (2002b). Stylesheet. Retrieved 20 Feb 2002 from http://linguistic-discovery.dartmouth.edu/stylesheet.html

Grimes, B. F. (2000). Ethnologue: Languages of the world. *SIL International*. Retrieved 13 Dec 2001 from http://www.ethnologue.com

Grimes, J. E. (1975). *The thread of discourse*. The Hague: Mouton.

Hale, K. (1971). A note on a Walbiri tradition of antonymy. In D. D. Steinberg & L. A. Jakobovits (Eds.), *Semantics: An interdiscplinary reader in philosophy, linguistics and psychology* (pp. 472–482). Cambridge: Cambridge University Press.

Hale, K. (1998). On endangered languages and the importance of linguistic diversity. In L. A. Grenoble & L. J. Whaley (Eds.), *Endangered languages: Language loss and community response* (pp. 192–216). Cambridge: Cambridge University Press.

Hale, K., Craig, C., England, N., Jeanne, L. M., Krauss, M. E., Watahomigie, L., & Yamamoto, A. (1992). Endangered languages. *Language*, 68, 1–42.

Hamilton, P. (1996). Uw Oykangard and Uw Olkola multimedia dictionary. Retrieved 25 Jan 2001 from http://www.geocities.com/Athens/Delphi/2970/

Himmelmann, N. P. (1998). Documentary and descriptive linguistics. *Linguistics*, 36, 161–195.

Hinnebusch, T., & Mirza, S. (1995). Swahili pronunciation guide. *Yale Program in African Languages*. Retrieved 27 Feb 2002 from http://www.cis.yale.edu/swahili/sound/pronunce.htm

Hinton, L. (2001). Audio-video documentation. In L. Hinton & K. Hale (Eds.), *The green book of language revitalisation in practice* (pp. 265–272). San Diego: Academic Press.

Hinton, L. (2001). Sleeping languages: Can they be awakened? In L. Hinton & K. Hale (Eds.), *The green book of language revitalization in practice* (pp. 413–417). San Diego: Academic Press.

Hinton, L., & Hale, K. (Eds.). (2001). *The green book of language revitalisation in practice*. San Diego: Academic Press.

Hiscock, P. (October 28, 1997). Vanuatu cultural research policy. Retrieved 4 Jan 2002 from http://arts.anu.edu.au/arcworld/vks/contre.htm

Hockey, S. (2000). Towards a model for Web-baased language documentation and description: Some contributions from digital libraries and humanities computing research. Retrieved 15 Mar 2001 from http://morph.ldc.upenn.edu/exploration/expl2000/papers/hockey/hockey.htm

Hockey, S. (n.d.). Describing electronic texts: The Text Encoding Initiative and SGML. Retrieved 4 Dec 2001 from http://lcweb.loc.gov/catdir/semdigdocs/hockey.html

Ide, N., & Romary, L. (2000). XML support for annotated language resources. Paper presented at Workshop on Web-based Language Documentation and Description, Philadelphia, 12/12/00-15/12/00.

Information science. (1989). In *The new Encyclopedia Britannica* (Vol. 6, p. 312). Chicago: Encyclopedia Britannica.

Information theory. (1989). In *The new Encyclopedia Britannica* (Vol. 6, p. 312). Chicago: Encyclopedia Britannica.

Itkonen, L. (1998). Endangered North American Indian languages. Retrieved 19 May 2001 from http://www.helsinki.fi/lehdet/uh/298j.html (Also available from Quarterly of the University of Helsinki 1998(2))

IUCN. (September 28, 2000). Confirming the global extinction crisis. Retrieved 21 Feb 2002 from http://www.iucn.org/redlist/2000/news.html

Jacobson, M., & Michailovsky, B. (2000). A linguistic archive on the Web. Retrieved 15 Mar 2001 from http://morph.ldc.upenn/exploration/expl2000/papers/michilovsky/index.htm

Jahnke, H. (1998). Ethnographic research: A Maori-centred approach. Lecture presented at International Pacific College, Palmerston North, New Zealand, 27/5/98.

Jennings, T. (1999). ASCII: American Standard Code for Information Infiltration. Retrieved 30 Jul 2001 from http://www.wps.com/texts/codes/index.html

Junker, M.-O. (2001). Eastcree.org. *Carleton University*. Retrieved 26 Feb 2002 from http://www.carleton.ca/~mojunker/eastcreegrammar/pages/english/index_menus/index_about.html

King, J. (2001). Te kohanga reo: Maori language revitalisation. In L. Hinton & K. Hale (Eds.), *The green book of language revitalisation in practice* (pp. 118–131). San Diego: Academic Press.

Krauss, M. E. (1992). The world's languages in crisis. *Language*, 68, 4–10.

Landow, G., & Delany, P. (2001). Hypertext, hypermedia and literary studies: The state of the art. In R. Packer & K. Jordan (Eds.), *Multimedia: From Wagner to virtual reality* (pp. 206–217). New York & London: W. W. Norton & Company. (Original article: Landow, G. & Delany, P. (1991). Hypertext, hypermedia and literary studies: The state of the art. In Landow & Delany (Eds.) (1991). *Hypermedia nad literary studies*. Massachusetts: Massachusetts Institute of Technology.)

Laudon, K. C., & Laudon, J. P. (1998). *Information systems and the internet* (4th ed.). Fort Worth: The Dryden Press.

LeMay, L. (2001). *Teach yourself Web publishing with HTML and XHTML in 21 days*. Indianapolis: Sams Publishing.

Leopold, R. (Ed.). (2001). Camping with the Sioux: Fieldwork diary of Alice Cunningham Fletcher. *Smithsonian Institute*. Retrieved 29 Jan 2001 from http://www.nmnh.si.edu/naa/fletcher/fletcher.htm

Linguistic Society of America. (1991, March). 1990-1991 Annual Meeting resolution. *LSA Bulletin*, 131.

Longacre, R. E. (1972). *Hierarchy and universality of discourse constituents in New Guinea languages: Discussion*. Washington, D.C.: Georgetwon University Press.

Loving, R. (Ed.). (1983). *Technical studies handbook* (3rd ed.). Ukarumpa, PNG: Summer Institute of Linguistics.

Lyovin, A. V. (1997). *An introduction to the languages of the world*. New York & Oxford: Oxford University Press.

Media Design in-Progress. (1999). Emile. Retrieved 15 Mar 2002 from http://www.in-progress.com/emile

Mason, J. (n.d.). The Rosetta Project. Retrieved 20 Feb 2002 from http://www.rosettaproject.org:8080/live

Mithun, M. (1998). The significance of diversity in language endangerment and preservation. In Grenobe & L. J. Whaley (Eds.), *Endangered languages: Language loss and community response* (pp. 163–191). Cambridge: Cambridge University Press.

Morgan, G. (2001). Welsh: A European case of language maintenance. In L. Hinton & K. Hale (Eds.), *The green book of language revitalisation in practice* (pp. 106–113). San Diego: Academic Press.

Mühlhäusler, P. (1996). *Linguistic ecology: Language change and linguistic imperialism in the Pacific region*. London & New York: Routledge.

Mühlhäusler, P. (1998). A rejoinder to Siegel's review of Linguistic Ecology. *Australian Journal of Linguistics*, 18, 219–225.

Newman, J. (2000). Tetun. *Massey University*. Retrieved 9 Mar 2002 from http://www.massey.ac.nz/~wwlingui/Tetun/

Newman, J. (2001). Creating multidimensional extensions of the classroom using online resources. Retrieved 9 Mar 2002 from http://vcsymposium.massey.ac.nz/presenters/presentations/Newman/www/home.html

Nua Internet Surveys. (2001). How many online? Retrieved 18 Jul 2001 from http://www.nua.net/surveys/how_many_online/index.html

OED. (1989). *Oxford English Dictionary* (2nd ed.). Oxford: Oxford University Press.

Payne, D., L. (June 2000). Maasai Language Project. Retrieved 25 Jan 2001 from http://www.uoregon.edu/~dlpayne/maasai/madict.htm

Quinn, E. M. (2001). Can this language be saved? Retrieved 11 Mar 2003 from http://www.cs.org/publications/CSQ/252/ (Also available from *Cultural Survival Quarterly* 25(2))

Rempt, B. (1999). Kura: an open-source multi-language multi-user linguistics database application. Retrieved 1 Jul 2001 from http://www.xs4all.nl/~bsarempt/linguistics/proposal.html

Renkema, J. (1993). *Discourse studies: An introductory textbook*. Amsterdam & Philadelphia: John Benjamin.

Reuters. (July 13, 2001). Taliban bans net from Afghanistan. Retrieved 17 Jul 2001 from http://www.msnbc.com/news/599931.asp?cp1=1

Robins, R. H., & Uhlenbeck, E. M. (Eds.). (1991). *Endangered languages*. Oxford & New York: Berg.

SBE Science Nuggets. (n.d.). A language at its genesis. Retrieved 20 Feb 2002 from http://www.nsf.gov/sbe/nuggets/028/nugget.htm

Searle, S. J. (1999). A brief history of character codes in North America, Europe, and East Asia. Retrieved 2 Jul 2001 from http://tronweb.super-nova.co.jp/characcodehist.html

SGML Users' Group. (June 11, 1990). A brief history of the development of SGML. Retrieved 20 Feb 2002 from http://www.sgmlsource.com/history/sgmlhist.htm

Sherzer, J. (n.d.). The Archive of the Indigenous Languages of Latin America. Retrieved 17 Jul 2001 from http://uts.cc.utexas.edu/~ailla/htdocs/mainindex.html

Siegel, J. (1997). Review of Mühlhäusler (1996). *Australian Journal of Linguistics*, 17, 219–237.

SIL. (2001). SIL Home Page. Retrieved 18 June 2001 from http://www.sil.org/

Simons, G. (June 12, 2000). Developing and infrastructure for online linguistic archives. Retrieved 23 Jan 2001 from http://www.talkbank.org/resources/simons.html

Skutnabb-Kangas, T. (1997). Terralingua definition: Minority. Retrieved 4 Jan 2002 from http://www.terralingua.org/Definitions/DMinority.html

The Jesus Film Project. (n.d.) The Jesus Film Project: Languages. *Campus Crusade for Christ*. Retrieved 18 Mar 2002 from http://www.jesusfilm.org/languages/index.html

TSD. (1985). *Translator's handbook*. Ukarumpa, PNG: SIL Technical Studies Department.

TSD. (n.d.). *Linguistic Section: Detailed requirements*. Ukarumpa, PNG: SIL Technical Studies Department.

Tsunoda, T. (May 22, 2000). Bibliography on language endangerment. *Department of Asian and Pacific Linguistics, University of Tokyo*. Retrieved 25 Feb 2002 from http://www.tooyoo.l.u-tokyo.ac.jp/~tsunoda/dlg_lst.html

Unicode. (1999). A standard compression scheme for Unicode. http://www.unicode.org/unicode/reports/tr6/tr6-3.3.html

Unicode. (2002a). Unicode 3.0 chapter1: Introduction. Retrieved 11 Mar 2002 from http://www.unicode.org/unicode/uni2book/uc20ch1.html

Unicode. (2002b). Code charts (PDF version). Retrieved 17 Mar 2002 from http://www.unicode.org/charts/

Unicode. (n.d.). What is Unicode? Retrieved 20 Feb 2002 from http://www.unicode.org/unicode/standard/WhatIsUnicode.html

von Bieberstein Koch-Weser, M. R. (December 22, 1999). And yet there is hope: From a century of loss to a century of environmental gains. *IUCN*. Retrieved 21 Feb 2002 from http://www.iucn.org/info_and_news/press/dg2000.html

W3C. (March 7, 2002). Extensible Markup Language (XML) activity statement. Retrieved 14 Mar 2002 from http://www.w3.org/XML/Activity

Wells, J. (2000). SAMPA computer readable phonetic alphabet. Retrieved 20 Feb 2002 from http://www.phon.ucl.ac.uk/home/sampa/home.htm

Wells, J. (December 6, 2001). The International Phonetic Alphabet in Unicode. Retrieved 8 Mar 2002 from http://www.phon.ucl.ac.uk/home/wells/ipa-unicode.htm

Werlich, E. (1982). *A text grammar of English*. Heidelberg: Quelle & Meyer.

Wired Digital. (February 1, 1999). No Morse Code. Retrieved 15 May 2001 from http://www.wired.com/news/print/0,1294,17641,00.html

Woodbury, A. C. (1993). A defense of the proposition, 'when a language dies, a culture dies'. Retrieved 29 Jan 2001 from http://uts.cc.utexas.edu/~ailla/woodburysalsa.html (Also available from *Texas Linguistic Forum*, 33, pp. 101-129.)

Wurm, S. A. (1996). *Atlas of the world's languages in danger of disappearing*. Paris & Canberra: Unesco Publishing/Pacific Linguistics.

Wurm, S. A., Tsuchida, S., Heine, B., Brenzinger, M., Crevels, M., Adelaar, W., Salminen, T., & Janhunen, J. (February 8, 2002). UNESCO red book of endangered languages. Retrieved 25 Feb 2002 from http://www.tooyoo.l.u-tokyo.ac.jp/Redbook/index.html

Yamane, L. (2001). New life for a lost language. In L. Hinton & K. Hale (Eds.), *The green book of language revitalisation in practice* (pp. 429-432). San Diego: Academic Press.

Zepeda, O., & Hill, J. H. (1990). *The condition of native American languages in the United States*. Unpublished manuscript.

GLOSSARY AND INDEX

Term	Explanation	Section
AIFF	(Audio Interchange File Format) a sound compression format useful for the World Wide Web	4.5.2
analogue data	the representation of one physical quantity (e.g. sound wave amplitude) by another (e.g. magnetic field strength in an audio tape, or the amount of deviation of a groove in a vinyl record)	4.2
ANSI	(American National Standards Institute) an extension to ASCII used by Windows (ISO-8859)	4.5.4.3
ASCII	(American Standard Code for Information Interchange) the standard character coding system that works across all computer systems (ISO-646)	4.5.4.1
AU	the "mu-law" sound compression format giving files of small size but low quality	4.5.2
CER	(Character Entity Reference or Entity Name) a way of including a Unicode character in HTML text by giving them a name, e.g. ligature ae (æ) has an entity name æ . Cf NCR.	7.4 (Fig. 7.20)
codec	Digital video compression/decompression algorithm	4.5.3
CSS	(Cascading Style Sheets) stylesheets used with HTML	
digital data	the representation of discrete elements of data by numbers, especially binary numbers in a computer	4.3
documentation	the accumulation, classification and dissemination of information	1.2
DPI	(dots per inch) a way of measuring the resolution of image files	4.5.1
DSSSL	(Document Style Semantics and Specification Language) a language for writing SGML style sheets	5.2.2
DTD	(Document Type Definition) a formal definition of the tagging structure used in a TEI, XML or other SGML-based document	5.2.2

Term	Explanation	Section
EBCDIC	(Extended Binary Coded Decimal Interchange Code) an 8-bit character code set used by IBM, but not recommended for language documentation today	
ECS	(Extended Character Set) an outmoded 8-bit extension to ASCII used by IBM PCs running DOS	4.5.4.2
endangered language	any language of a community which is not learned any more by children, or at least by a large part of the children of that community	1.1, 2.2
ethnic community	a language community or the community of their descendents if the language has become extinct	1.1
FindPhone	a phonetic data analysis program developed by SIL	7.4 (Fig 7.26)
GIF	(Graphic Interface Format) an image format suitable for displaying symbols and drawings in Web pages	4.5.1
GML	(Generalised Markup Language) the precursor to SGML	4.4.2
hypermedia	text, sound and image files linked together in a network which allows access of one file from another	4.3.1
HTML	(Hypertext Markup Language) an application of SGML used to markup Web pages	4.4.1
hypertext	text documents linked together in a network which allows access of one document from another	4.3.1
information block	a unit of documentation structure	4.3.1
information science	the theory and practice of the management of information on computers	4
information theory	theory dealing with the problems of information transmission	4
IPA	International Phonetic Alphabet	6.2.3
JPEG	(Joint Photographic Experts Group) an image format suitable for displaying photographs in Web pages	4.5.1
Kura	Linguistic software with planned XML capability (see Rempt, 2001)	5.2.4

Term	Explanation	Section
language community	the community of speakers of a minority language	1.1
language community	the community of speakers of a minority language	1.1
language description	the record of a language as a system of abstract elements, constructions, and rules that constitute the invariant underlying structure of the utterances observable in a speech community	1.3
language documentation	a comprehensive record of the linguistic practices characteristic of a given speech community	1.3
language shift	where one language is replaced by another as the common language of the community	2.2
Latin-1	a 256-character code set compatible with Unicode (ISO-8859-1)	
LinguaLinks	a linguistic data management program developed by SIL	3.3, 4.1.2, 5.2.7
MacOS	A computer operating system (Macintosh)	
Manuscripter	a computer printing program that interprets SF markers to format manuscripts. Formerly used by SIL.	7.4 (Fig 7.26)
minority language	a language spoken by a minority group in a country	1.1
moribund language	a language on the verge of extinction, having only a few elderly speakers	2.1
MOV	(Movie) Quicktime audio/video format	4.5.3
MPEG	(Motion Picture Experts Group) a digital audio/video format	4.2.3
multimedia	the integration of text, sound, and still or moving images, especially using digital technology	1.4
NCR	(Numeric Character Reference) a way of including a Unicode character in HTML text by referring to its number, e.g. ligature ae (æ) has a decimal Unicode number 230, or hexadecimal 00E6. if inserted in HTML text as æ or æ it will display correctly on all computer platforms running modern browsers. Cf CER.	7.4 (Fig. 7.20)

Term	Explanation	Section
PDF	(Portable Document Format) a cross-platform document format based on PostScript printer language that preserves the look and feel of the original	4.5.4.6
platform	a computer running a certain operating system, e.g. Windows, MacOS, Unix, etc	
PNG	(Portable Network Graphics) an image format suitable for archiving photographic images and displaying them in Web pages	4.5.1
Praat	Speech analysis software	4.5.2
RSF	Revised Standard Format (Cf SF)	5.2.7
Rosetta Project	a project for longterm archiving of language data	4.2.6
SAMPA	(Speech Assessment Methods Phonetic Alphabet) a method of representing IPA in ASCII	4.5.4.4, 6.2.3
SGML	(Standard Generalised Markup Language) a markup language upon which HTML, TEI and XML are based (ISO-8879)	5.2.2
Shoebox	a linguistic data management program developed by SIL	5.2.6
SF / SFM	Standard Format markup system used in SIL software such as Shoebox	5.2.6
SpeechAnalyzer	Speech analysis software	4.5.2
TEI	(Text Encoding Initiative) an application of SGML used to markup literary texts	4.3.1
UCS	(Universal Character Set) The Unicode Consortium's character set	
Unicode	a character coding system (also called UCS) developed by the Unicode Consortium incorporating ASCII but with enough coding positions for all the alphabets of the world. Adopted for use with HTML and XML (ISO-10646)	4.5.4.5
Unix	A computer operating system	
WAV	(WAVeform) a sound compression format useful for the World Wide Web	4.5.2

Term	Explanation	Section
Windows	A computer operating system (Microsoft)	
XHTML	(Extensible Hypertext Markup Language) a version of HTML that conforms to the XML standard	4.4.5
XML	(Extensible Markup Language) a modern streamlined variety of SGML	4.4.4
XSL	(Extensible Stylesheet Language) a way of formatting XML documents (derived from DSSSL)	5.2.4