

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

**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

ABSTRACT.....	ii
ACKNOWLEDGEMENTS.....	iii
TABLE OF CONTENTS .....	iv
LIST OF FIGURES .....	vii
LIST OF TABLES .....	viii
1. INTRODUCTION.....	1
1.1 Minority and endangered languages.....	1
1.2 Documentation.....	3
1.3 Language Documentation.....	4
1.4 Multimedia Language Documentation.....	6
1.5 Stakeholders and ethics.....	8
1.6 The scope of this study .....	14
2. THE NEED FOR LANGUAGE DOCUMENTATION.....	16
2.1 The current state of the languages of the world.....	16
2.2 Language death .....	18
2.3 The Value of Languages.....	23
2.3.1 Languages as objects of beauty .....	23
2.3.2 Language as an essential part of culture.....	25
2.3.3 Language as a reflection of the mind.....	27
2.3.4 Language as a vehicle to express spirituality .....	28
2.3.5 Linguistic diversity as a key to knowledge about ourselves .....	29
2.3.6 Language as a part of our heritage .....	31
2.4 Responses to Language Extinction.....	31
2.4.1 Conservation .....	32
2.4.2 Documentation .....	33
3. THE SCOPE OF LANGUAGE DOCUMENTATION - TYPES OF DATA.....	35
3.1 The nature of data .....	35
3.2 The traditional view of documentation.....	35
3.3 SIL language documentation.....	36
3.4 Simon's suggestion .....	37
3.5 Himmelmann's proposal .....	40
3.6 Data sampling approaches.....	42
3.6.1 Anthropological approach.....	42
3.6.2 Linguistic approach.....	42
3.6.3 Functional approach .....	43

3.7 A user-oriented documentation .....	44
4.    THE MEDIA AND FORMAT OF DATA .....	49
4.1 Media for the data .....	49
4.2 Analogue media.....	50
4.2.1 Recording.....	50
4.2.2 Data manipulation.....	50
4.2.3 Integration of media .....	50
4.2.4 Reproduction .....	51
4.2.5 Access.....	51
4.2.6 Archiving .....	51
4.3 Digital media .....	54
4.3.1 Recording.....	54
4.3.2 Data manipulation.....	54
4.3.3 Integration of media .....	54
4.3.4 Reproduction .....	55
4.3.5 Access.....	55
4.3.6 Archiving .....	57
4.4 Summary - preferred media for documentation.....	60
4.5 The format of the data.....	61
4.5.1 Images.....	61
4.5.2 Sound .....	63
4.5.3 Video .....	64
4.5.4 Text.....	65
4.5.4.1 ASCII.....	65
4.5.4.2 ASCII Extensions .....	66
4.5.4.3 ANSI.....	66
4.5.4.4 IPA.....	67
4.5.4.5 Unicode.....	68
4.5.4.6 Text on the internet.....	70
5.    THE LOGICAL STRUCTURING OF THE DATA .....	72
5.1 Content structure .....	72
5.1.1 Filing structure .....	74
5.2 Marking up content structure.....	76
5.2.1 HTML.....	77
5.2.2 SGML.....	80
5.2.3 TEI.....	82
5.2.4 XML.....	83
5.2.5 XHTML.....	85
5.2.6 SF.....	85
5.2.7 RSF.....	86
5.3 Conclusion .....	86

6.	THE INSTRUMENT .....	88
6.1	Organisation of the documentation .....	88
6.2	General formatting recommendations for each file type .....	90
6.2.1	Markup.....	90
6.2.2	Original documents .....	90
6.2.3	Text.....	91
6.2.4	Sound, picture and video .....	92
6.3	Recommendations for specific sections.....	92
6.3.1	General Introduction.....	92
6.3.2	Historical Materials .....	93
6.3.3	Cultural Notes.....	93
6.3.4	Readers, literature and translated material.....	93
6.3.5	Instructional materials .....	95
6.3.6	Language data .....	95
6.3.7	Lists .....	96
6.3.8	Analyses .....	97
7.	SAMPLE DOCUMENTATION.....	98
7.1	Introduction .....	98
7.2	Folders and Files .....	98
7.3	Using the documentation - the <i>ReadMe</i> file .....	99
7.4	Sample views .....	101
8.	CONCLUSION.....	127
	APPENDIX A. ORGANISATION .....	130
	APPENDIX B. EXAMPLES OF MULTIMEDIA LANGUAGE DATA FROM THE WORLD WIDE WEB ....	132
	REFERENCES.....	138
	GLOSSARY AND INDEX .....	147

## LIST OF FIGURES

1.1.	Tensions affecting the linguist .....	11
4.1	A Rosetta disk.....	52
4.2.	Online Language Populations .....	57
5.1.	Linear structure .....	72
5.2.	Hierarchical structure.....	73
5.3.	Hypermedia structure .....	74
5.4.	Information with a relational structure .....	76
5.5.	HTML and web browser display of the document fragment .....	78
5.6.	A dictionary entry marked up using HTML.....	79
5.7.	A comparison of fragments of SGML and XML.....	84
5.8.	Relationships between selected members of the SGML family .....	85
6.1.	The user's view of the documentation .....	89
6.2.	Organisation of folder and files .....	90
7.1.	The opening window of the Rumu Documentation sample.....	98
7.2.	The database folder.....	99
7.3.	Title page .....	102
7.4.	Main table of contents.....	103
7.5.	A section contents page .....	104
7.6.	A bibliography .....	105
7.7.	Who's Who with thumbnails.....	106
7.8.	Cultural notes illustrated with photos .....	107
7.9.	Blowup of a small illustration.....	108
7.10.	Video.....	109
7.11.	Tables and graphs .....	110
7.12.	Terms and definitions in columns .....	111
7.13.	A scanned chart .....	112
7.14.	A hand-drawn map .....	113
7.15.	Scanned coloured map.....	114
7.16.	Educational book pages as thumbnails.....	115
7.17.	A scanned printed page image.....	116
7.18.	A commentary.....	117
7.19.	Dictionary.....	118
7.20.	Straight Rumu text showing use of non-ASCII characters .....	119
7.21.	Bilingual text .....	120
7.22.	Simple interlinear text and access via multiple paths .....	121
7.23.	Interlinear glossed text with sound .....	122



7.24.	Raw interlinear text data .....	123
7.25.	IPA phonetic symbols.....	124
7.26.	Interlinear phonetic data display with sound .....	125
7.27.	Browser view of dictionary database with XML tags.....	126
B1.	Whole text with translation (and sound) .....	132
B2.	Interlinear text.....	132
B3.	Interlinear text with special characters.....	133
B4.	Text in non-Roman orthography.....	133
B5.	Pronunciation guide with program-controlled sound.....	134
B6.	Pronunciation guide with hyperlinked sound .....	134
B7.	Paradigm with sound.....	135
B8.	Phonetic text example with normal and slow speech sound.....	135
B9.	Transcribed field notes.....	136
B10.	Anthropological notes.....	136
B11.	Browsable dictionary with links to encyclopedic information.....	137
B12.	Dictionary with finderlist and thesaurus .....	137

## LIST OF TABLES

2.1	Rumu possessive adjectives.....	23
3.1.	A framework for the repository.....	36
3.2.	Comparison of Simon's and SIL's categories with Himmelmann's .....	42
4.1.	Numbers of people (millions) using the internet by region of the world .....	53
4.2.	Analogue vs. digital media.....	57
4.3.	The upper 128 characters of some common 8-bit ASCII_based codes .....	63
4.4.	Examples of ASCII and Unicode.....	65
A1.	A selection of organisations concerned for minority and endangered languages worldwide .....	130
A2.	A selection of online organisations concerned with developing tools and standards for multimedia documentation of languages.....	131