The Documentation of Endangered Linguistic, Lexical and Cultural Knowledge of the Marquesan and Tuamotuan Languages of French Polynesia

Gabriele Cablitza, Fasan Chongb and Edgar Tetahiotupac
aSeminar für Allgemeine und Vergleichende Sprachwissenschaft der CAU zu Kiel, Olshausenstrasse 40, 24098 Kiel, Germany
bCultural Association Te Reo o te Tuamotu, B.P. 4025, Tahiti, 98714 Papeete, French Polynesia
cLaboratoire de Recherche en Sciences Humaines de Polynésie française (LARSH), Servitude Pékin, Taunoa, 98714 Papeete, French Polynesia
gcablitz@googlemail.com

ABSTRACT
This paper reports on an interdisciplinary project within the DoBeS-program of endangered languages (Volkswagen foundation) in which digital multimedia encyclopaedic lexica are created for the endangered Marquesan and Tuamotuan languages of French Polynesia with the web-based lexicon tool LEXUS. LEXUS has a flexible scheme of linking multimedia documents to lexical entries as well as the possibility of creating relational links which allow to make associations between words and their concepts in and across languages and cultures visible via Information Technology. One major objective is to motivate the speech communities to actively participate in the process of creating these multimedia lexica. Over the course of several years, linguistic, lexical and cultural data have been gathered, analysed and annotated by researchers (linguistics, anthropology, botany) in close cooperation with the Marquesan and Tuamotuan speech communities in order to build up digital multimedia language archives housed by the Max Planck Institute for Psycholinguistics (Nijmegen, Netherlands). Apart from giving details about the specifics of the LEXUS project mentioned above, this paper also talks about the language documentation procedure and methodology, the major purposes of such documentations (e.g. for pedagogical exploitation, anthropological studies) and how other speech communities of the Pacific can become major agents of the documentation, maintenance and revival of their endangered languages and cultures.

Keywords
Endangered languages; language documentation; multimedia lexicography

1. INTRODUCTION
This paper reports on an interdisciplinary project within the DoBeS-programme of endangered languages1 in which we aim at creating a digital multimedia encyclopaedic lexicon for the endangered Marquesan and Tuamotuan languages of French Polynesia with the help of the newly developed multimedia lexicon tool LEXUS which is currently being developed at the Max Planck Institute for Psycholinguistics in Nijmegen (Netherlands) [1]. It is a co-joint project between linguists, technicians and speech communities in which the functionalities of the LEXUS tool as well as its user interface for different user groups are further developed. LEXUS is a web-based tool which has a flexible scheme of linking multimedia documents to lexical entries as well as the possibility of creating relational links which allow to make associations between words and their concepts in and across languages and cultures visible via IT technology. The development of a multimedia encyclopedic lexicon (=MEL) is an important step towards language documentation as a means of language maintenance and preservation of endangered linguistic, lexical and cultural knowledge. One major objective in this project is to motivate the speech communities to actively participate in the process of creating these multimedia lexica by a) learning basic methodology of lexicography and linguistic software, b) working out a system of collaborative workspaces for future online participation by the speech community, c) writing monolingual definitions and encyclopaedic articles of vernacular (i.e. Marquesan and Tuamotuan) words for language maintenance purposes, and d) creating their own “ethno-ontologies” and the organisation of relations between words which are based on indigenous categorisation alone.

This indigenous or emic perspective is complemented by the outsider’s or researcher’s perspective (=etic perspective)2. The etic perspective documents words according to scientific categories often drawing on classification according to standard linguistic criteria or other disciplines such as anthropology, botany, ornitology or marine biology and their standardised scientific classifications (e.g. scientific names for fish, plants, birds, and their taxonomies, etc.). The LEXUS tool with its new specific features of knowledge representation – also called relational linking – can create a dense network of lexical as well as cultural data in ways which are meaningful for different kinds of user groups, allowing a multilayered organisation of lexical and cultural knowledge.

For documentary linguistics, the multimedia lexicon creation as envisaged in our MEL-project with LEXUS is in fact a new form of language archiving within the structural frame of a lexicon as it

1 Funded by the Volkswagen foundation in Germany.
2 The emic/etic-distinction has been introduced by the cultural anthropologist Kenneth Pike [2].
combines data from a language documentation archive with a structured lexical database. It also goes beyond conventional dictionary making as our multimedia lexicon will consist of a dense network of lexical entries with all sorts of media files, and thus presenting the meaning of words in a new way.

2. MOTIVATIONS FOR THE PROJECT
In previous development and research on multimedia lexica, user studies have revealed that conventional paper dictionaries are limited in its usefulness in language maintenance and language revival of endangered languages [3]. Due to the often quite academic set-up, design and internal organisation of conventional paper dictionaries (alphabetical order, conventions, abbreviations, etc.), they only apply to a limited group of users, but not to a wide range of different user groups with differing demands and interests. [3] have found that members of endangered language communities are much more interested in navigating through a lexicon by following semantic links such as mango > fruit > tree > breadfruit tree > material culture > canoe-building, etc. rather than browsing through an alphabetically ordered lexicon. These semantic links can be created in so-called knowledge spaces through the relational linking mechanism which – unlike the one-to-one hyperlinking mechanism – can create chains of links between words (e.g. lexemes) and multimedia files or any other kind of data. With this new technology, words and their associations to cultural concepts can be organised in such a way that it represents indigenous categorisations and understandings, thus creating so-called ethno-ontologies. The navigation through a lexicon via knowledge spaces offers a more natural entry point into the lexicon for indigenous communities. How these cultural concepts (and ethno-ontologies) are labelled is a matter of careful semantic analysis and methodology.

The present MEL-project which has evolved from a language documentation archive by following semantic links such as mango > fruit > tree > breadfruit tree > material culture > canoe-building, etc. rather than browsing through an alphabetically ordered lexicon. These semantic links can be created in so-called knowledge spaces through the relational linking mechanism which – unlike the one-to-one hyperlinking mechanism – can create chains of links between words (e.g. lexemes) and multimedia files or any other kind of data. With this new technology, words and their associations to cultural concepts can be organised in such a way that it represents indigenous categorisations and understandings, thus creating so-called ethno-ontologies. The navigation through a lexicon via knowledge spaces offers a more natural entry point into the lexicon for indigenous communities. How these cultural concepts (and ethno-ontologies) are labelled is a matter of careful semantic analysis and methodology.

The present MEL-project which has evolved from a language documentation archive by following semantic links such as mango > fruit > tree > breadfruit tree > material culture > canoe-building, etc. rather than browsing through an alphabetically ordered lexicon. These semantic links can be created in so-called knowledge spaces through the relational linking mechanism which – unlike the one-to-one hyperlinking mechanism – can create chains of links between words (e.g. lexemes) and multimedia files or any other kind of data. With this new technology, words and their associations to cultural concepts can be organised in such a way that it represents indigenous categorisations and understandings, thus creating so-called ethno-ontologies. The navigation through a lexicon via knowledge spaces offers a more natural entry point into the lexicon for indigenous communities. How these cultural concepts (and ethno-ontologies) are labelled is a matter of careful semantic analysis and methodology.

Another new aspect and motivation for the creation of multimedia lexica is also the visualisation and contextualisation of the meaning of a word. Full contextualisation of word meaning is possible through linking all kinds of different annotated media files (video- and audio-files, still images, drawings) with words of a lexical entry (mostly headwords) from a digital language archive or a local file system. This presents the meaning of a word in its full and natural (i.e. socio-cultural) contexts without losing the basic information structure of a lexical entry.

3. LEXICOGRAPHIC WORK IN PREVIOUS DOCUMENTATION PROJECT
The gathered lexical database for the present MEL-project has been collected in the DoBeS-language documentation project Documentation of the Marquesan languages and culture in French Polynesia (funded by the Volkswagen foundation between 2003-2006). It consists of a trilingual general lexical database (Marquesan dialects, French, English) as well as thematic glossaries of topics which have become important foci of the documentation (food preparation, plant medicine, fish and fishing, breadfruit varieties). The lexical entries in our lexical databases are mostly corpus-based, but have been added by field notes and elicited wordlists collected during fieldwork periods. Our lexical databases will not turn into a complete dictionary as there are well-known time constraints in short term documentation projects [4]. That is also one of the reasons why we have chosen a thematic approach for the built-up of our MEL-lexicon. The thematic approach has the advantage to complete a sub-domain of the culture in a short period of time which often has motivating effects on the speech community [4].

The lexical data has been prepared with the MDF 4.0 Shoebox/Toolbox (lexical) database type which can be imported into LEXUS by representing the same hierarchical structure as in the corresponding Toolbox MDF database type. For the Tuamotuan languages we are currently working on several glossaries (fish/fishing, shellfish, plaiting, food preparation, plants, birds, material culture of coconut tree) which are also prepared with the MDF Toolbox database type before being imported into LEXUS. Lexical data linked with media files in LEXUS will look like the following figure:

Figure 1. Linking of lexical data (Marquesan fish glossary) with media files in LEXUS.

---


4 Credits for drawing (left): K. Tomita/B.W. Halstead, World Life Research Institute; credits for photo (right): John A. Long ARPS.
4. DOCUMENTATION OF WORD MEANING DURING FIELDWORK

Doing lexicography in and for endangered speech communities requires some practical orientation in the working procedure. Apart from finding adequate translations in the target language for (vernacular) words of the source language, which is a general problem of bi- or multilingual lexicography [5], there is also the problem that the headword of many lexical entries denote complex phenomena, procedures or concepts which are specific to the source language and culture and do not have a translation equivalent in the target languages at all. Contrary to conventional bi- or multilingual dictionaries, we often provide - next to translation equivalents (if possible) - detailed definitions of the headwords in our lexical entries, thus giving more encyclopaedic information than a conventional bi- or multilingual dictionary would normally include.

When first starting to document lexical knowledge during fieldwork, the initial approach to word meaning of every field lexicographer is extensional and referential [6]. Local language consultants tend to explain and exemplify word meaning in relation to concrete things and situations of the world. For the field lexicographer it is often a hard task to find out about the extensional meaning of words, what precise meaning distinctions they encode and how they are semantically related to other words of the source language [6]. In this case it is not sufficient to rely solely on a good corpus of the language to tackle this task. There are basically two methods of documenting indigenous word meaning: the writing of vernacular (i.e. monolingual) definitions and elicitations.

An important step towards the understanding of indigenous word meaning is the writing of vernacular definitions of words by native speakers themselves because it documents word meaning from the native speaker’s understanding of a word. This methodological approach can clarify possible misinterpretations between researcher and local language consultant because vernacular definitions can serve as a source of reference of authentic indigenous word meaning and metalinguistic knowledge of speakers. Misinterpretations are not uncommon because the fieldworker as well as the local language consultant often communicate via a contact language which is not their native language. This can result in inadequate and inexact documentation of word meaning [6].

Even with this methodology, the documentation of lexical knowledge of speakers is not an easy task. Field lexicographers like [7] have observed that lexical knowledge is not necessarily well-structured and commonly shared knowledge across the speech community, in particular in endangered speech communities which undergo rapid linguistic change caused by a dominant contact language [6].

Careful, but systematic elicitation is another essential method of accessing word meaning. It should be avoided to let the local language consultants directly translate words from the contact language into their native language [8]. One way of avoiding the usage of a contact language is to start off by showing stimuli (photos, drawings, films) or modelling actions to consultants, depending on which semantic domain is investigated. Consultants then should describe the action they perceive in their native language. After these first basic techniques other elicitations such as devising interactive tasks between consultants or careful questioning can be applied to further refine the semantic analysis.

Another way of approaching indigenous word meaning - in particular for motion verbals - can be achieved by filming local language consultants when they act out the meaning of verbs in their natural contexts without verbalising or commenting of what they are doing. As LEXUS can include multimedia documents, the vernacular definitions and definitions in the target languages are accompanied by non-verbal video clips which depict the interpretation of the meaning of a headword by native speakers. This method does not only have the advantage of contrasting the different senses of a headword via visual media (i.e. having video clips for each sense of a headword), but it can also visually contrast and compare a whole semantic field of verbals such as CUT, BREAK, CARRY, MOVE, PUT, etc. in a created knowledge space with the relational linking mechanism.

5. REPRESENTING LEXICAL AND CULTURAL KNOWLEDGE IN LEXUS

This new technology of relational linking in LEXUS does not only open up possibilities to compare and view a semantic field in one and the same PC window and therefore serve as a research tool to understand and analyse the semantic typology of verbs, but it allows to integrate all kinds of information (lexical and multimedia) in whatever way defined by the user. Speech community members can define and organise how words and other data are grouped together and what (semantic) relations they hold to each other, thus creating spaces which are meaningful to native speakers alone (=emic perspective). For the Marquesans whose culture was traditionally based on plants, they decided that much of their lexical and cultural data should be presented via an ethno-technological perspective by focusing on the plant as a point of departure. A digital drawing of a plant - which often shows the characteristics of the plant much better than a photo does - can be uploaded into the knowledge space and the particular parts of the plant are named and linked with the respective lexical entries in the lexicon as shown in figure 1:

![Figure 1. Cultural knowledge space of breadfruit plant in LEXUS.](image)

Information of how the plant parts are used in the material culture exist in the Marquesan language archive which can then be linked with the archive. Even comparisons and links between the Marquesan and Tuamotuan lexica can be made which are useful for cross-linguistic and cross-cultural comparisons.
6. SUMMARY
Multimedia lexicon tools such as LEXUS are excellent resource and research tools for scientific as well as endangered speech communities interested in lexical, cultural and ethnographic documentation. LEXUS does not only allow to visualise lexical as well as cultural knowledge representations in ways defined by the users, but it can also contextualise the meaning of words in new ways. In particular for speech community members a tool such as LEXUS represents a more natural entry point into a lexicon than conventional paper dictionaries do and it can be an important tool for language maintenance and language revival. For documentary linguistics, the creation of multimedia lexica with LEXUS is a new form of language archiving which goes beyond conventional dictionary making as well as so far practised language archiving as it will consist of a dense network of lexical entries with all sorts of media files which are part of a language documentation archive and thus presenting a culture’s network of meanings in a new way.

7. ACKNOWLEDGMENTS
We like to thank the Volkswagen foundation for the generous support of our DoBeS-grants. We are grateful for the enthusiastic support of members of the cultural associations Te U‘i hou o te Fenua ‘Enata (Marquesas islands) and Te Reo o te Tuamotu (Tuamotu islands). We like to thank in particular Mathias (Teaiki) Tohetiaatua, Grégoire (Taiea) Epetuhui, Marie-Madeleine Bruneau, Lucien Mataiki, Pascal Pati, Mimio Puhetini (†) and Tahia Tuohe for their contributions.

8. REFERENCES