



**NATIONAL SCIENCE FOUNDATION**

**NATIONAL ENDOWMENT FOR THE HUMANITIES**



---

## Documenting Endangered Languages

### SAMPLE APPLICATION NARRATIVE

This sample narrative conforms to a past set of guidelines.

Please consult the guidelines for the current competition on NSF's website at:  
<http://www.nsf.gov/pubs/2005/nsf05590/nsf05590.htm>

# Arapesh Grammar and Digital Language Archive

Lise Dobrin (PI)

David Golumbia (co-PI)

Daniel Pitti (co-PI)

The Institute for Advanced Technology in the Humanities

## 0. Overview

This project will document the endangered Arapesh languages, which are traditionally spoken by people living along the New Guinea north coast in East Sepik and Sandaun provinces, Papua New Guinea (PNG). Building on the principal investigator's past fieldwork on Arapesh, the project will transform an already collected corpus of texts and elicited linguistic data into (1) a written reference grammar that provides a synthetic overview of the Arapesh language family in addition to comprehensively describing Cemaun Arapesh synchronic grammar, (2) a Digital Language Archive and research tool, and (3) a multilingual, multimedia public-facing website to serve as an educational resource on Arapesh. Arapesh is of special significance to linguistic theory for its typologically unusual system of noun classification that elaborates phonological, as opposed to semantic, principles of morphological class assignment and syntactic agreement. Like other Arapesh varieties, the Cemaun dialect of Mountain Arapesh which this project documents most extensively is seriously endangered. In the Cemaun villages, language shift to Tok Pisin is well advanced, to the point that there are fewer than 100 fluent speakers, none of whom are under 40, and none of whom are Arapesh monolingual.

This project description is organized as follows. After introducing the project investigators in section 1, some background information on Arapesh is presented. Section 2.1 briefly describes the Arapesh people; 2.2 discusses the relationships among languages within the Arapesh family; 2.3 characterizes the language's endangered status along with the sociohistorical and cultural factors that shape the current linguistic situation; and 2.4 surveys the existing documentation for Arapesh, showing how it forms an inadequate basis for language preservation and for drawing sound linguistic conclusions. Section 3 describes the source materials that will be analyzed, digitized, and represented in the project's concrete products, a written reference grammar of Arapesh (section 4), a Digital Language Archive consisting of a Grammatical Database and a Text and Audio Collection (section 5), and a public-facing website (section 6). Section 4 describes the main features of the Arapesh grammar and the contribution this grammar will make to our understanding of grammatical architecture and the comparative study of New Guinea languages. The technical aspects of the project, its adherence to existing and proposed standards, and the significance and broader impacts of the Arapesh Digital Language Archive are discussed in section 5. The project's communication with the public, embodied in a public-facing website and a metadata stream about the archive's contents, are outlined in section 6. Planned future projects that will bring some of the results of this work to the Arapesh people are mentioned in section 7. Finally, in section 8, the roles and responsibilities of the project collaborators are specified and the project timeline is laid out.

## 1. The Project Investigators

The project is an interdisciplinary collaboration that brings together several areas of research interest and expertise at the University of Virginia. It also draws on the expertise of the leading Arapesh scholar working in the field of missionary linguistics.

**Lise Dobrin** (PI) teaches linguistics in the Department of Anthropology at the University of Virginia. She brings to the project a background in language description and linguistic theory, a professional and personal commitment to Arapesh people, and the experience and material results of 15 months of field research on Arapesh. From December 1997 through March 1999, Dobrin lived and conducted linguistic research in Wautogik, one of two Cemaun villages, during which

time she also visited other Arapesh-speaking regions to collect data on variation in the Arapesh family. Upon finding that, even in this remote mountaintop community, villagers routinely spoke not Arapesh but the PNG lingua franca Tok Pisin as their main language of daily life, Dobrin recognized the urgency of documenting the speech of the remaining fluent speakers for the future, and even though it was not strictly required for her immediate project (dissertation research on Arapesh phonology and noun classification), she applied herself to collecting the wide range of data that would be necessary to produce a comprehensive Arapesh reference grammar. The Filemaker Pro database Dobrin constructed in preparation for this fieldwork proved to be an important resource for her analysis of Arapesh noun classification (Dobrin 1998, 1999a), and it was her experience developing and working with it that formed the inspiration for the Arapesh Digital Language Archive. The format Dobrin evolved in the FileMaker Pro database will serve as a model for the expanded grammatical markup of Arapesh data to be undertaken as part of the present project.

After more than a decade spent developing and managing the development of software applications in industry and the non-profit sector, **David Golumbia** (co-PI) turned to academic work in the fields of media studies and cultural studies. He first wrote on the cultural politics of linguistic theories, with a particular focus on the philosophy of language (Golumbia 1998, 1999a, 1999b), but has since become increasingly interested in the way the western tradition of language theorizing constructs nonwestern peoples and languages. This subject is now a major theme of his teaching and research at the University of Virginia (Golumbia 2001, 2004). He has also worked for several years on the linguistic biases inherent in modern communications technologies, especially the web (Golumbia 2003a, 2003b). He has worked directly with linguists and community members in Canada on the East Cree Language Web, one of the first projects to provide all information in a native writing system (the Cree Syllabary) as well as digitized audio data. Through his work on this project, Golumbia has become convinced of the multimedia web's potential value not only for raising awareness of the global threat to linguistic diversity, but also for giving voice to linguistically peripheral peoples, many of whom speak non-written languages. His work on the present project reflects his commitment to extending the web's capabilities so as to better represent the world's linguistic diversity.

**Daniel Pitti** (co-PI), Associate Director of IATH, has more than ten years of experience in working with markup (SGML then XML) and database technologies. Working with archivists and librarians, he is the principal architect of one established standard (Encoded Archival Description) and one emerging standard (Encoded Archival Context). He has extensive experience working with humanists in the application of markup and database technologies to literary, historical, and cultural studies. He played a central role in the design and development of web-based database projects involving Tibetan and Chinese (see below), in particular in the application and extension of TEI encoding. While he does not have direct experience in the application of markup and database technologies to support linguistic research, his extensive experience in working collaboratively with humanists and social scientists in the use of both technologies as well as his international contacts within the literary and linguistic computing communities makes him well qualified to serve the technical objectives of this project.

For over a decade, the **Institute for Advanced Technology in the Humanities (IATH)** has been working together with humanists and social scientists to apply advanced technology to their research. Many of IATH's projects are text-based, and two major projects are on languages not written in the Latin alphabet, Tibetan and Chinese. IATH has direct experience creating and maintaining a language archive following OLAC (Open Language Archive Community) standards, the Tibetan-Himalayan Digital Library. IATH's facilities and expertise are ideal for the long-term support of work on endangered language preservation, though the current project is the first to devote IATH resources expressly toward such aims. In particular, IATH (with the University of Virginia Library) serves as one of the four hosts for the Text Encoding Initiative (TEI) Consortium, the organization responsible for the development and maintenance of the

XML-based standard for representing text to support linguistic and literary research. Thus, in addition to the direct goals of documenting and preserving Arapesh, one of the present project's broad impacts will be to lay a foundation for future work on endangered languages within the IATH infrastructure (see the letter of commitment submitted by Bernard Frischer, Director of IATH).

**Bob Conrad** (outside consultant; see CV and letter of commitment) has training in linguistics from the University of Pennsylvania and elsewhere, and has spent over two decades working in PNG with the Summer Institute of Linguistics (SIL) translating the New Testament into Arapesh languages and promoting vernacular literacy. Over this period he has gained fluency in several varieties of Arapesh that complement the Cemaun variety mastered by Dobrin. Conrad has been an active supporter of local projects responding to Arapesh language endangerment, such as training teachers for village vernacular language preschools. He brings to the project unmatched understanding of Arapesh grammatical structure and dialect variation, and the kind of culturally informed encyclopedic knowledge that can be critical for interpreting vernacular discourse.

An **Advisory Committee** of scholars with expertise in the areas of nonwestern languages and digital archive development has been established to oversee work on this project (see letters of commitment):

1. Ellen Contini-Morava (Professor and Chair of Anthropology at the University of Virginia) is a linguist and linguistic anthropologist specializing in Swahili morphology, syntax, and discourse. She has a longstanding research interest in the theoretical issues raised by noun classification, particularly those relating to its discourse function. She also has experience collaborating with IATH on this topic in the past.
2. David Germano (Associate Professor of Religious Studies at the University of Virginia) is a scholar specializing in Tibetan, Buddhist and Tantric studies. He is the principal architect of the Tibetan-Himalayan Digital Library created and maintained by IATH. He has extensive experience with the representation of non-western linguistic data on the web.
3. David Sapis (Professor of Anthropology at the University of Virginia) is a senior anthropologist who has conducted ethnolinguistic fieldwork in Sénégal on Kujamaat Jóola and who is now working on a web site bringing together linguistic and folklore material from the language.

## **2. The Language**

### **2.1 Ethnographic Setting**

Arapesh languages are traditionally spoken across a broad region extending inland from the north coast of PNG in East Sepik and Sandaun Provinces. The region was colonized a century ago by Germany, followed by Australia, until 1975 when Papua New Guinea received independence. Due to the long history of Catholic mission activity in the region, most Arapesh villagers are Catholic, though the traditional form of Catholicism is now giving way to Pentecostal and charismatic Christianity. Arapesh villagers subsist mostly on sago and the varied produce of traditional swidden gardens, supplemented by store-bought commodity foods like rice and tinfish. Money is obtained by cash-cropping (chiefly copra), selling produce in markets, and by remittance from urban wage-earning kin. In the 1930s, the "Mountain Arapesh" people were the subject of a famous study by the anthropologist Margaret Mead (1935), who portrayed men and women alike as sharing a nurturing and maternal "feminine temperament," an oversimplification challenged in a little-known essay by the anthropologist and linguist Reo Fortune (1939), who had been her husband and partner during their profoundly troubled 8-month-long Arapesh fieldwork (Dobrin and Bashkow in prep, Bashkow and Dobrin in prep). It is Fortune's 1942 grammar of Rohwim Arapesh that remains the most widely cited linguistic resource on the languages of this family. Recent ethnographic studies of other Arapesh-speaking people include

Tuzin on the Ilahita Arapesh (1976, 1980, 1997) and Leavitt on the Bumbita (1989, 1995, 1997, 2000).

## **2.2 Genetic Affiliation**

On the basis of lexicostatistical evidence, the Papuan Arapesh language family belongs to the Kombio Stock of the Torricelli phylum (Laycock 1973, 1975; see also Foley 1986, 2000). The conventional classification of Arapesh includes three languages: Mountain Arapesh or Bukiyip (1998 pop. 14,000), Southern Arapesh or Muhiang (1998 pop. 11,000), and Bumbita Arapesh or Weri (1994 pop. 4,000; Grimes 2000). Nekitel (1985, 1986) has carved out from this classification a fourth language, Abu', though it is open to question whether Abu' constitutes a distinct language or a northern (Supari) dialect of Muhiang. Given the considerable chaining of phonological, lexical, and grammatical features as one moves from village to village (a distributional pattern not unfamiliar in New Guinea), Arapesh is perhaps best understood as one long dialect chain on which the language classification outlined above is superimposed (Conrad 1978). Cemaun is a hitherto undocumented dialect of Mountain Arapesh. It is spoken in Wautogik and Kotai villages, which are located at the northeastern border of the Arapesh dialect continuum in East Sepik Province, where it contacts the Ndu family Boikin language.

## **2.3. Endangered Status**

According to Dobrin's residential census in 1999, the two Cemaun villages Wautogik and Kotai have a combined population of approximately 800 (cf. the official figure of 258 arrived at in the 1990 PNG census; see PNG National Statistical Office 1994). But such population figures present an inaccurate picture of Arapesh vitality, since vernacular fluency is no longer predictable from village of residence. This is increasingly true throughout the entire "West Coast" and "Sepik Highway" areas where Arapesh is spoken. As with so many other languages in the Sepik provinces, which has the highest percentage of Tok Pisin users on mainland New Guinea (see Wurm, Mühlhäusler, and Tryon 1996; Aikhenvald forthcoming), Arapesh is giving way to Tok Pisin, and increasingly to English. The advancement of western values and commodities (and with them massive borrowing) has been unrelenting, and Tok Pisin is now the main language used in the Cemaun village communities, as well as by Cemaun people living in town. The youngest good Cemaun speakers are all above forty, putting Arapesh on the verge of "seriously endangered" according to the model of Wurm 1998; see also Crystal 2000. Many Cemaun speakers show clear deterioration in their grammar and fluency relative to their elders, and their speech in virtually all genres is marked by frequent, unsuppressable code-switching. Below middle age, there is a precipitous drop in vernacular competence in every respect. Young people are unable to produce extended stretches of vernacular speech, and although some young villagers have limited passive competence, most children cannot understand even the simplest everyday commands and greetings. There are fewer than 100 fluent speakers of Cemaun. No monolinguals remain.

In many ways, the Cemaun Arapesh situation is a textbook case of language shift leading toward extinction. Paradoxically, this process is being accelerated by the Cemaun people's unusually high rate of success in education and formal employment, which makes them reliant on Tok Pisin and English and draws the population, as well as prestige, away from the village and traditional way of life. Indeed, in the last generation, so many people have left Wautogik that more of its people now live in towns and cities than in the village itself. The Wautogik diaspora is extraordinarily well represented in the national elite, and includes several of PNG's top diplomats, politicians, businessmen, teachers, and professors, as well as several internationally trained PhDs. Fully aware that their language is dying, the people of Wautogik have raised funds for language revitalization activities such as building a vernacular language preschool in their home village. Village leaders have been intensely supportive of linguistic work on the language and have

repeatedly articulated their hope that the activities of western linguists will not only document the language for their descendants but also help reinvent its use as a spoken language.

Language shift among the Arapesh has proceeded in large part in response to culture-external factors associated with colonialism and the postcolonial national framework. These include the use of Tok Pisin as the official medium of Catholic missionization since 1931 and the use of English as the official language of national discourse and schooling since the 1950s (Mihalic 1996, Romaine 1992, Nidue 1990). Perhaps more importantly, however, Arapesh people's eager embrace of Tok Pisin has been facilitated by the culturally particular ways in which foreign linguistic codes have come to be associated ideologically with the prestige of western modernity, Christianity, and development. While Tok Pisin is by all means a local language, Arapesh people nevertheless strongly associate its use with the west (see also Kulick 1992). Arapesh people as young as 50 resolutely believe the Tok Pisin they speak to be the native language of Germany, an earlier colonial power (the glossolalia that occurs in new charismatic forms of Catholic worship is also believed to channel actual European languages through people's voices). Traditional Arapesh society is culturally predisposed to be open to diffusion and to value most highly those cultural forms, including languages, that can be "pulled in" from outside (hence Mead's 1938 characterization of Arapesh as an "importing culture"). While one hesitates to predict the most dire scenario for Arapesh's linguistic future, the cultural factors that have guided the language's history to this point make the outlook discouragingly dim.

#### **2.4. Previous Documentation**

Cemaun has not been previously documented. Published grammar-length descriptions exist for the neighboring Rohwim dialect of Mountain Arapesh (Fortune 1942), the inland Bukiyip variety (Conrad and Wogiga 1991), and the western coastal But dialect (Gerstner 1963). There are also a few unpublished sources, such as Nekitel's dissertation on Abu' (1985), which provides a linguistic overview of the variety of Arapesh spoken in the author's native village, Womsis, some manuscripts by Conrad (Conrad 1996 is one example), and a handful of published articles such as Alungum, Conrad, and Lukas 1978.

Nevertheless, none of these is an adequate comprehensive general reference source on Arapesh. Fortune's grammar succinctly describes the Rohwim dialect's typologically unusual phonologically-based noun classification system, which has attracted recent theoretical interest (Aronoff 1992, 1994, Fraser and Corbett 1995, Dobrin 1995, 1998, 1999a; see section 4 below). But it fails to indicate mood, a ubiquitous and criterial inflectional category on the verb. Fortune's texts are also highly problematic in that they underrepresent phonological information, and represent it erratically; they also fail to distinguish morphemes from words. Using a tagmemic model, Conrad and Wogiga's Bukiyip grammar analyzes syntactic and discourse-level features, especially those that are relevant to translation. But it gives little attention to lower levels of grammar; the phonology and remarkably complex morphophonemics are together covered in a mere 6 pages. Gerstner's grammar is highly inaccessible (it is a typewritten manuscript with handwritten marginalia, available only on microfilm), and it suffers from severe inadequacies in transcription, e.g., conflating voiced and voiceless segments in the word-final position crucial for understanding the morphology. It is also couched in terms of traditional grammar, inappropriately applying notions such as "future" and "dative" to a language that has neither tense nor case. In short, the published documentary record on Arapesh is uneven and incomplete. It provides an insecure basis for linguistic generalization, and it is inadequate for purposes of preservation, both for the community of linguists and for the Arapesh people. Moreover, as Dobrin found when studying the available materials in preparation for fieldwork, the fragmentary literature on Arapesh does not support a coherent picture of the family. It comprises a dizzying patchwork of language and village names, transcription conventions, grammatical categories, and time periods, which limit the ability of typologists and theoreticians to make linguistic assertions about Arapesh with confidence.

This project aims to overcome these limitations through the preparation of a comprehensive grammar that accurately describes Arapesh's structural particularities and that balances a thorough description of a single variety with systematic reference to others, thus representing both the remarkable unity and revealing diversity of the Arapesh family as a whole. Clarifying the interrelations among the Arapesh varieties requires a broad range of data that has been framed in comparable terms and that can ideally be interpreted against a background understanding of the settings from which the data was obtained. Thus, one important contribution of the Arapesh grammar to be produced within the bounds of this project is to present data on diverse Arapesh varieties that was collected by a single researcher (with a single set of biases) at a single point in time, in addition to interpreting data from earlier sources within a calibrating descriptive framework. Such integration is necessary if we are to develop a coherent picture of the genetic, areal, and geographic relationships that hold within the Arapesh family.

### **3. Source Data**

The primary source of data for the project will be the Arapesh field materials collected by Dobrin during a 15 month research period in PNG in 1997-1999. Fieldwork was supported by an NSF Dissertation Improvement Grant (Award No. SBR-9707681), a Fulbright-Hays Training Grant for Doctoral Dissertation Research Abroad (Award No. P022A70043), and a Wenner-Gren Foundation for Anthropological Research Predoctoral Grant (Award No. 6156). The materials include the following:

1. 320 handwritten pages of Cemaun Arapesh language texts, which have been transcribed phonemically, annotated and glossed in Tok Pisin, and checked with native speakers while in the field. These texts represent origin/land charter stories, both traditional and modern legends, spontaneous narratives, casual conversations with multiple participants, and "recipes" for constructing objects of cultural value. Another 40-plus pages of texts documenting other Arapesh varieties were collected and have been similarly processed; these include samples of the Arapesh varieties spoken at Woginara (the Rohwim dialect of Mountain Arapesh described by Fortune), Dogur and Balam (two distinct western coastal dialects of Mountain Arapesh), Balif (central Muhiang), Hwamsuk (southern Abu'), and Bumbita village (Bumbita Arapesh, the morphologically most divergent Arapesh language which has not previously been documented).
2. Cassette tape recordings of most of the texts. To ensure the long-term security of the data, the recordings were copied and the originals archived at the University of Chicago Language Laboratories and Archives.
3. Extensive handwritten notes on grammar and phonology, much of it recopied and reordered to reflect preliminary analyses arrived at in the field. This grammatical data occupies more than a dozen field notebooks and is reasonably well organized. Vocabulary was recorded according to semantic domain (e.g., species of vines, yams, birds; color terms; winds and seasons; kinship terms) or according to formal properties (e.g., plural ending; reduplicative sound structure; verb valency). Many notebook sections are devoted to demonstrating specific grammatical generalizations or justifying particular conclusions (e.g., the syntactic incompatibility of object suffixes and lexical object NPs; lists of minimal pairs; evidence for the borrowed status of phonotactically anomalous forms.) Dobrin 1999b, 2001 are based on some of these notes.
4. A FileMaker Pro database of 1495 nouns compiled by Dobrin prior to fieldwork that gathers together information from the main published sources on Arapesh and includes not only nouns and glosses, but also notes on their morphological irregularities and attested morphosyntactic behavior along with indexing to the published sources. Similar but less extensive databases of Cemaun nouns and verbs have also been compiled.
5. In addition to the texts described under (1) above, the project will make use of the 60 pages of legends that accompany Reo Fortune's 1942 grammar of Rohwim Arapesh.

Dobrin's Arapesh language field materials were collected with IRB approval, and may be incorporated in the Digital Language Archive and published in the written grammar. However, since web publication was not specifically addressed in the protocol for that research project, additional IRB review specifically addressing web distribution is being sought in order that selected field data may be included for illustrative purposes in the public-facing website. Until such time as this approval is granted, only a general description will be posted on the public-facing website, and the digital form of all field data will be closed to the wider public. There are no copyright restrictions limiting the use of Fortune's previously published materials, making it possible to edit, digitize, and electronically distribute them through a link on the public-facing website (see note from Ann McLean, Fortune's literary executor).

#### **4. Written Reference Grammar**

The written grammar will be a theoretically informed but ecumenical reference grammar that comprehensively describes the Cemaun dialect of Mountain Arapesh while devoting systematic attention to the variation that exists within the Arapesh family. It will begin with a chapter contextualizing the language by describing its sociocultural and recent historical setting, and it will be illustrated by maps showing the distribution of the languages, the villages of sampling, and important dimensions of variation such as the west-to-east *l* to *r* transition along the coast, the switch in the internal order of possessive NPs which seems to follow a north-south path from Abu' to Bumbita, and the dramatic differences in the structure and use of kinship vocabularies throughout the family. Although such dimensions of linguistic variation are well known to Conrad and Dobrin, nowhere in the literature is this kind of information available. Whenever possible, generalizations presented in the grammar will be exemplified with data from the digitized texts.

The topics addressed by the grammar will be arranged "bottom up" roughly according to traditional grammatical domains. The section on the phonology will cover the distribution of phones, distinctive features, syllable structure and phonotactics, vowel harmony, and word-level prominence. Because the morphophonemics are rather extensive, alternations involving contrasting sounds and the factors that condition them may warrant a short chapter of their own. Realizational morphology will be conceptually distinguished from morphosyntax, suggesting a division of labor in the grammar outline. The morphology section will deal with the types of formal operations used to construct words, derivation and root compounding, inflectional categories and the organization of paradigms, and the two major lexical form classes, nouns and verbs (it is not yet clear whether whether adjectives constitute a proper form class or a subclass of verbs). The section on morphosyntax will be a large and important one. It is here that the grammar will describe the principles of noun classification and agreement which constitute the most central organizing subsystem of the language. By providing numerous examples of how noun class marking works over extended stretches of discourse, the grammar will contextualize the interpenetration of phonology, morphology, and syntax which was elucidated in Dobrin 1999 within an analysis of the overall grammatical and communicative system. Within the rubric of morphosyntax such matters as valency-changing operations and the behavior of clitics will also be treated. The syntax section will cover topics including phrase and clause types and structuring, word order, pronominal affixes, interclausal relations, and the function of the ubiquitous particles. Given the Arapesh tendency for short sentences with shallow embedding, it is possible that the discussion of syntax will merge with the discussion of discourse structuring and genres.

Information about Arapesh variation will appear throughout the grammar, with the major dimensions of variation described and exemplified as an integral part of the presentation of each construction or phenomenon. Thus, for example, the geographic and lexical distribution of *l* and *r* will be discussed in the section on phonology, making it clear that no Arapesh phonological system uses the two sounds contrastively; it will also be treated in the chapter on morphosyntax insofar as differences in sounds in noun-final position are associated with differences in noun

classification and agreement. The detailed descriptions presented in earlier chapters will also be summarized in a final chapter (or possibly set of chapters) that provides a synthetic overview of the Arapesh family as a whole.

A grammar of Arapesh constructed in this way will provide a crucial foundation on which to base hypotheses about the historical relationships that hold among Arapesh varieties; this in turn may lead to progress on the larger question of affiliations among the languages grouped together within the Torricelli phylum. In laying this foundation, the project will contribute not only to Melanesian historical linguistics, but also to a cross-disciplinary conversation regarding prehistoric human migration patterns in the Sepik area. Anthropological studies integrating evidence from archaeology, oral history, linguistics, and biology have produced a reconstruction of Sepik prehistory in its very broad outlines (see e.g., Swadling 1984; Roscoe 1989, 1994). While Torricelli populations in the Sepik appear to be ancient, recent settlement histories indicate that the migrations of the Arapesh and neighboring Ndu-speaking peoples have moved through the region following a northward trajectory. This finds support in the impression, which can be tested and elaborated on the basis of the data analyzed in this study, that the Arapesh language family constitutes a miniature “spread zone” embedded in what is otherwise one of the most “residual” regions of New Guinea (see Nichols 1992), and that the linguistic variation within Arapesh is rather more pronounced in inland areas than it is along the coast.

Perhaps most immediately, the grammar of Arapesh will contribute to research on the typology of noun classification devices (Corbett 1991, Aikhenvald 1999, Grinevald 2000, Senft 2000a). Complementing frequently found systems in which noun classes have a semantic basis, the noun class systems of northern Arapesh languages such as Cemaun are unusual in the extent to which they are organized around phonological features. In most cases a noun’s class, and hence its plural form, is determined by its final phonological element, usually a consonant phoneme, irrespective of the noun’s meaning; thus, *nîmbat* ‘dog’ belongs to the *t*-class, the borrowing *kap* ‘cup’ to the *p*-class, etc. Moreover, these class-determining sounds are copied surprisingly faithfully in agreement on syntactically related units, yielding a pattern of “literal” alliterative concord across words. Phonological classification and agreement highlights the organizing role phonological form can play in synchronic linguistic categorization, countersupporting the “Saussurean dogma” of modern linguistics (Nichols 1986, 1989; Aronoff 1992, 1994; Dobrin 1995, 1998, 1999a). Arapesh noun classification is also relevant to understanding the path by which noun classes arise diachronically. It is known, for example, that classifier systems may derive from “repeater” constructions in which a copy of the noun itself is used as its classifier (Aikhenvald 2000:361-2; see also Senft 2000b:40). Arapesh noun class agreement bears an obvious resemblance to repeater classification; Arapesh also has a limited number of classifier constructions (Fortune 1942, Dobrin 1999). If it can be established that these are remnants of an earlier classifier system, it would suggest the grammaticalization of repeaters as a possible source for closed grammatical noun classes, a development which has so far not been attested.

The typologically unusual phonologically-based noun class assignment and agreement systems of Arapesh also present the community of linguists with a phenomenon that requires us to rethink fundamental theoretical assumptions about the way the core components of grammar interact. In the lexicalist spirit that has guided much research in the generative paradigm, syntax has been claimed to be “phonology-free” (Zwicky 1987; Zwicky and Pullum 1986; Pullum and Zwicky 1988; Miller, Pullum, and Zwicky 1997; Corbett 2001 provides a clear restatement). If true, this would place significant constraints on grammatical architecture. However, agreement evidence from the northern Arapesh varieties strongly suggests that a phonological system may directly influence syntactic operations, since it is a noun’s phonological composition that is the determining factor not only in selecting among agreement morphemes, but in determining their concrete realization.

## **5. Digital Language Archive**

As a tool for research on the grammar and as a means of preserving Arapesh linguistic data for the benefit of linguistic science and future generations of Arapesh people, the project will create an Arapesh Digital Language Archive. The Archive will consist of two major software components, a Grammatical Database and a Text and Audio Collection, which will be linked together in such a way as to enable the isolation and investigation of specific, grammatically relevant features in the corpus of materials in both parts of the Archive. This structure will allow us to implement software functions including sorts and searches for lexical entries and grammatical features, alone and in combination; “keyword in context” displays of search results; the organization of Database contents in multiple dictionary formats; and fine grained coordination of written texts with parallel audio files. The result will be a flexible research tool of use not only to linguists interested in Arapesh grammar, but also to scholars working in folklore studies and anthropology. In addition to its utility as a research tool, the Arapesh Digital Language Archive will provide a robust and flexible, enduring record of this endangered language.

The Digital Language Archive will not only be useful as a research tool once it has been completed. The process of creating it, and the many analytical decisions this will entail, will feed directly into Dobrin’s work on the written Arapesh grammar. Indeed, these two parts of the overall project will overlap, most dramatically in the second year. As with any formalization of grammatical hypotheses, the creation of the Digital Language Archive will require many aspects of the grammar (e.g., the identification of grammatical categories and word boundaries) to be made precise and explicit, making the input and coding of data a direct test of the linguist’s analysis.

The Archive will be constructed in accordance with current standards and best practices recommendations for linguistic documentation and archiving (see 5.7 below) and will actively contribute to their further development. There are several aspects of Dobrin’s texts that are likely to pose technical challenges as archiving proceeds. These include language and dialect variation across the Text Collection as a whole, a high frequency of code-switching within most of the contemporary texts, and the presence of simultaneous multiple voices in the recorded conversations. We will interact with the larger markup and database community to develop techniques for productively integrating these textual features into the Digital Language Archive, and we anticipate sharing our solutions with other linguists through bodies such as E-MELD and OLAC as a broader impact of the work.

### **5.1 Grammatical Database**

The Grammatical Database will provide a powerful technological tool for accessing Arapesh linguistic data according to precisely delimited criteria. It will allow the retrieval of lexical entries by grammatical features, phonological shape, gloss, and language variety. This capability will be instrumental for Dobrin’s work on the grammar, as it will facilitate the formation and testing of hypotheses regarding Arapesh language structures by allowing systematic and comprehensive access to examples in the Text and Audio Collection. The Grammatical Database will be constructed in such a way as to be inherently open and extensible, allowing new search and report functions to be created as new linguistic features relevant to the language are identified (see 5.1.2 below).

#### **5.1.1 Legacy Database**

The Grammatical Database component of the Digital Language Archive is a significant extension of a legacy Filemaker Pro application created by Dobrin (see section 3 above). This custom tool was used to organize Arapesh data extracted from the documentary sources, and it was specifically designed to represent the morphophonological and morphosyntactic properties of

nouns that were relevant to her work on Arapesh noun classification. The following table illustrates the kinds of information contained in a single noun record.

<b>Sample database entry for the noun <i>mbul</i> 'pig'</b>			
English	<i>pig</i>	This near spkr: sg	<i>whaguhuda'</i>
Tok Pisin	<i>pik</i>	This near spkr: pl	
Singular	<i>mbul</i>	This near hr: sg	
Plural	<i>mbuluguH</i>	This near hr: pl	
Sex		Distal pro: sg	
Adjective suf: sg	<i>ri</i>	Distal pro: pl	
Adjective suf: pl		Proximal pro: sg	
Verb pre: sg	<i>ra</i>	Proximal pro: pl	<i>agwaguH</i>
Verb pre: pl	<i>wu, wha</i>	Determiner	<i>biaguH 2, anar/unar 1, baiguhatuH 3, anagaguH</i> 'some'
Verb suf: sg	<i>r, ar, ur</i>	Possessor	<i>ir, iguh, iguH, er</i>
Verb suf: pl	<i>aguH, guwh</i>	Interrogative	
Data Source	Fortune 1942		
Page cites, comments	34 (noun class 10 in table), 25 (borrowed or adapted from neighboring dialects?), 54, 78, 87, 96, 97, 104, 108 ( <i>warabai-gwihi</i> 'many of them'), 110:Sa (default agreement when humanness unclear), 116, 118, 126, 132, 134, 140, 144, 148, 154, 156, 162, 172, 180 (many examples), 184, 196, 200		
Rightmost 1: sg	<i>l</i>	Rightmost 1: pl	<i>H</i>
Rightmost 2: sg	<i>ul</i>	Rightmost 2: pl	<i>uH</i>
Rightmost 3: sg	<i>bul</i>	Rightmost 3: pl	<i>guH</i>
Rightmost 4: sg	<i>mbul</i>	Rightmost 4: pl	<i>uguH</i>
Rightmost 5: sg	<i>mbul</i>	Rightmost 5: pl	<i>luguH</i>

In addition to glosses in Tok Pisin and English, the record contains phonological, morphological, and semantic information about the noun itself. It also includes information about the attested forms of pronouns and other syntactically associated elements. Of particular interest are the ten calculated fields labeled “Rightmost 1” through “Rightmost 5” for both singular and plural forms of the noun. These fields enable searches on the basis of string-final sounds (letters were used because FileMaker Pro would not accommodate non-Roman fonts). This innovation was necessary in order to directly target the ends of words, where the key morphosyntactic information relevant for noun classification resides. These functions will be replicated and greatly expanded in the Grammatical Database created as part of the present project.

### 5.1.2 New Grammatical Database

The existing FileMaker Pro database will be migrated to a PostgreSQL database. The PostgreSQL database will conform to standards promulgated under E-MELD (Electronic Metastructure for Endangered Languages Data) and the input interface will conform to the emerging data input protocols of FIELD (see E-MELD no date-b). The field structure of the new database will be greatly enriched to accommodate verbs and other parts of speech. Some limited coding for syntactic position may also be incorporated. The search and report features of the legacy database will be maintained and enhanced to include the expanded field structure; it will also be enhanced by relating the grammatical data with textual data and associated audio files in the Text and Audio Collection. Such interrelations will allow users to navigate easily between grammatical

analysis, text, and audio, with each type of data complementing and enhancing the understanding of the others.

While both FileMaker Pro and PostgreSQL are based on Structured Query Language (SQL), PostgreSQL is open source software with an extensive international maintenance and development community committed to compliance with public, open standards. SQL is an ISO standard first codified in 1986 and substantially revised in 1992 and again in 1999. SQL will ensure a lossless migration of data and functionality from FileMaker Pro to PostgreSQL. As open source software, PostgreSQL provides additional assurance that the data will endure ongoing changes in hardware and software for the foreseeable future. PostgreSQL is ported to Unix, including Linux and OSX. Release 8.0, scheduled for late 2004, will be ported to Windows operating systems as well. Extensible Markup Language (XML) has emerged as both an alternative method for representing data and as a data transport or communication syntax. PostgreSQL has integrated XML support. The incorporation of XML support into PostgreSQL facilitates the interrelation of the text and audio data, as XML will be used for representing the text, and for interrelating text with audio data.

### **5.2 Text and Audio Collection**

The other key component of the digital language archive will be a set of electronic texts encoded using the Text Encoding Initiative (TEI) Guidelines for use of XML, and a digitized audio collection captured according to the latest digital library standards and best practices. TEI is the international standard for representing texts to facilitate literary and linguistic analysis and research. Embedded TEI tags will be used to mark the language variety of stretches of text of any length; this feature is critical because of the multiple Arapesh varieties represented in the Archive, as well as because of the frequent code-switching that occurs within individual texts. If the given TEI suite of tags and attributes proves to be inadequate for this purpose, TEI-conforming extensions will be written to accommodate such tagging. The main body of texts will be those recorded and transcribed by Dobrin in the field; the archive will also include a reanalyzed corpus of Arapesh legends published by Fortune in 1942 (see section 3 above for details on these materials). Audio versions of the texts will be digitized and correlated with their written transcriptions through time-offset addressing. The Metadata Encoding and Transmission Standard (METS) will be used to correlate audio and text, as well as to associate descriptive and administrative metadata with each audio file. METS is an international XML standard maintained by the Library of Congress. Using TEI and METS will enable navigation between text and audio passages. Dobrin's field tapes will be digitized in conformance with E-MELD's and the Library of Congress' best practices with respect to both the format (WAVE) and associated descriptive and administrative data.

### **5.3 Transcription and Character Encoding**

The Arapesh texts and grammatical data will be transcribed using IPA (International Phonetic Alphabet) and the IPA characters will be represented using Unicode UTF-8. Both XML, by default, and PostgreSQL, optionally, support UTF-8 (for a description of the Unicode standard, see Unicode Consortium 2003a; for the IPA set, see Unicode Consortium 2003b). Both TEI and METS use English-based names for tags. Grammatical field labels as well as descriptive and administrative data will be in English. Tok Pisin, which occurs frequently in the texts, will be used as one of the glossing language and written using the standard Latin alphabet encoded in UTF-8.

### **5.4 Linking Between Grammatical Database and Text and Audio Collection**

A major structural component of Digital Language Archive is the linking of the Grammatical Database with the Text and Audio Collection. Though the exact methods need to be determined, it is anticipated that integration of the querying of the Database and Audio and Text Collection

will be accomplished using Cocoon 2.0, an open source XML publishing platform that includes SQL and XQuery (XML Query) modules, and XSLT for transforming and synthesizing query results. Such an approach will optimize use of the underlying database and markup technologies, exploiting the strengths of the representational architecture of each. PostgreSQL querying functions provide a wide array of search capabilities across field labels and contents, and by interrelating it to the XML text and through the text to the audio data, the rich analytical detail thus accessible in the Grammatical Database will be expanded by the ability to see contextually situated examples in the Text and Audio Collection (for Cocoon 2.0, see Apache Cocoon Project n.d.; for XML and its compatibility with databases, see Bourret 2004). The development of this linking software may serve as a model for other similar tools for language archiving, and constitutes another broader impact of this project.

### **5.5 User Interface**

The Digital Language Archive will be accessible to users via web page interfaces that display archived information in a functional and user-friendly way. These web page interfaces will be designed for maximum efficiency and functionality, using layouts and formats established jointly with project linguists to accommodate their research needs. A suite of tools for creating web page interfaces will be provided by IATH as part of its standard server setup. The main user interface functions will be inputting, manipulating, searching, and reporting, with pages to be created for tasks such as inputting data into the Grammatical Database, inputting data into the Text Collection and applying markup tags for language variety, searching and browsing the Grammatical Database contents (with both list and individual record views), and searching and browsing the Text and Audio Collection. Both components of the Digital Language Archive will be accessible through a single web page interface, with search result displays utilizing an integrated report structure.

### **5.6 Server Location and Support**

IATH has extensive experience with secure long-term data archiving and is institutionally committed to maintaining the availability of its holdings over the long term. System administration for IATH's remote and resident servers (including regular backups, security, and 24x7x365 support) is provided by the University of Virginia's Department of Information Technology and Communication (ITC). The project includes the purchase of server hardware according to IATH's specifications, to be housed at IATH, with system administrative support provided by ITC and IATH's resident systems administrator.

### **5.7 Standards Implementation and Compliance**

The project investigators are committed to working within the recommendations and standards proposed by the multi-university E-MELD (Electronic Metastructure for Endangered Languages Data) consortium, which was formed to maximize the interoperability and long-term robustness of electronically archived endangered languages data. The design of this project has been developed in coordination with E-MELD best practice outlines from the outset, specifically its proposed GOLD standard (General Ontology for Language Description; see E-MELD 2003) for linguistic category markup. E-MELD offers several tools that may be useful to ensuring the standards compliance of this project. These include ORE (the OLAC Repository Editor), which enables the direct creation of OLAC-compliant metadata; CharWrite<sup>®</sup>, a tool that enables the direct input of IPA markup into web-based forms; and FIELD (Field Input Environment for Linguistic Data; E-MELD no date), which was designed to facilitate the digital markup and storage of endangered languages data. E-MELD's FIELD tool has already been used successfully to create a GOLD compliant archive of Potawatomi linguistic information by importing a legacy database set up in FileMaker Pro.

Project investigators will provide ongoing feedback to E-MELD regarding its proposed GOLD standard in light of the Arapesh language data, and will work with E-MELD on other issues that arise in the creation of the Arapesh Digital Language Archive, including the sharing of software tools (Bird and Simons 2003, E-MELD no date-a, no date-c). The project's potential to significantly extend current standards and best practice outlines for the digital archiving of endangered languages is one of its important broader impacts.

In addition to E-MELD standards and best practices, the project investigators are also committed, under the direction of IATH faculty and staff, to adhere to digital library standards. These standards include but are not limited to XML and the appropriate standards within the XML family, such as XSLT, XPATH, and XQuery, and community-based derived standards such as TEI and METS and descriptive and administrative schemas used by TEI and METS.

## **6. Public-Facing Website, Metadata Broadcasting, and Access**

In accordance with E-MELD and OLAC standards, metadata about the Arapesh Digital Language Archive will be broadcast to OLAC (Open Language Archive Community) servers. The broadcast data will include a description of the Arapesh Digital Language Archive and will explain how scholars and others interested in the language can apply for access. This information will also appear on the public-facing website.

The public face of the project will be a small, open-access website, designed and hosted by IATH, to provide culturally and sociolinguistically contextualized "lessons" about the Arapesh language. In addition to including selected photographic images, the public-facing website will present representative linguistic data (with proper rights certification), a small demonstration version of a dictionary, and examples of interesting grammatical features, particularly the phonologically-based Arapesh noun classification and agreement patterns. It will also have an audio component, allowing visitors to hear as well as visualize the Arapesh utterances presented. It is anticipated that the website will provide useful teaching materials for Dobrin (in her courses on morphological theory, language and culture in Melanesia, and languages of the world) as well as for other scholars teaching about human languages.

The inclusion of audio will make the project's public-facing website not only more informative for western students of language; it will also make it more accessible and appealing to those Arapesh people who visit the site on computers at their workplaces or while visiting relatives in town. Despite efforts by the SIL and others to promote vernacular literacy, Arapesh remains an essentially spoken language. Many of the people who speak and understand Arapesh would be unable to comprehend the information on the website without the audio dimension, and even for those Arapesh who do not speak the language, spoken forms of the vernacular are more natural, valued, and interpretable than written ones.

Because the website will be viewable by Arapesh and other Papua New Guinean people, the website text will be written not only in English, but in Arapesh and Tok Pisin as well. This is particularly important if we are to keep the technological medium from passively reinforcing the association between monolingual English and the high-tech web's cosmopolitan prestige. By designing the site in such a way as to instead harness the web's prestige to lend value to Arapesh and multilingualism, this project will model one important way in which the web can be a resource that serves endangered language speaking communities, rather than something that, because of its format, merely furthers their alienation. By highlighting the cultural setting of Arapesh speakers and addressing the community in their own languages, the project website hopes to benefit all endangered language communities. This too may be considered a broader impact of the project.

## **7. Plans for Future Projects Stemming from this Work**

At present, Cemaun speakers have virtually no access to a written version of their language. At the initiative of village leaders, Dobrin helped native speakers of Cemaun compose vernacular

versions of The Lord's Prayer and Hail Mary which are regularly recited at Catholic church services in the village. Cemaun speakers can also read Christian pamphlets that are written in the neighboring Rohwim dialect, which is mostly comprehensible to Cemaun speakers. But these cannot be the sole local instrument of language revitalization. In part this is because they can only be used by people who are both fluent in the vernacular and literate, and for historical reasons such people are few. Moreover, since the orthography in which these are written is only partially phonemic, people experience interference from English as well as a feeling of estrangement from their own language when they try to read them, which they typically do haltingly aloud. This feeling is intensified by the ubiquitous presence of features (most saliently "sh" instead of "ch") that mark the texts clearly as "Rohwim." This carries infelicitous, counter-motivating social meaning, as the Cemaun and Rohwim people were traditional enemies and remain antagonistic today.

In response to the problem of the near mutual exclusivity of literacy and vernacular competence, Dobrin will use the Arapesh Digital Language Archive as a resource for a subsequent project to create printed pedagogical materials oriented toward children, for use in the Cemaun villages' vernacular preschools and local primary school. Funding for this future project will be sought elsewhere. These materials will draw on the Arapesh Digital Language Archive to produce a reverse nominal dictionary of Cemaun that incorporates morphosyntactic information (in addition to glosses in Tok Pisin and English) in a way that graphically demonstrates the language's phonologically based noun classification patterns. The unusual reverse organization of the dictionary is suggested by the linguistic structure of Arapesh itself, and, as Dobrin's experience eliciting nominal vocabulary in the field suggests, it more closely parallels the formal word-searching mechanisms used spontaneously by Arapesh speakers than does conventional alphabetical order. The reverse nominal dictionary, along with other planned educational materials written in an essentially phonemic orthography, will be particularly beneficial for Arapesh revitalization, because the logic of the pervasive noun classification system is becoming increasingly opaque as key phonological contrasts on which it depends are being lost from young people's speech and linguistic perception. This opacity contributes to further deterioration of the grammar while increasing people's sense that the native language is "too hard" to learn (see also Nekitel 1984, 1992). Laying the groundwork for such pedagogical materials is yet another of the current project's broader impacts.

## **8. Work Plan**

### **8.1 Roles and Responsibilities**

Lise Dobrin, PI, is the chief researcher for the project and the investigator responsible for all linguistic matters, including work with Conrad. She will work closely with other project personnel to design software and write specifications. Her main tasks will include analyzing the linguistic data in order to define and refine Grammatical Database fields and markup tags for the Text and Audio Collection; entering her fieldnotes and handwritten texts into the Text Collection and Grammatical Database; digitizing audio tapes and editing audio files to coordinate with the texts; and training and supervising an undergraduate assistant to aid with data entry. Dobrin will act as general editor for all linguistic data. She will carry out grammatical analysis for the reference grammar and write the manuscript. As this work proceeds, she will also write research papers for presentation at conferences and publication in journals. As Principal Investigator, Dobrin will see to it that the project is subject to appropriate ethical oversight. She will also oversee the budget and write additional funding documents.

David Golumbia, co-PI, will serve as the primary contact between linguistic and engineering personnel on the project team, overseeing all software design and development. He will lead in the writing of technical and functional specifications and perform direct technical tasks in database development and markup. As the work proceeds, he will also write research papers for presentation at conferences and publication in journals concerning the representation of

endangered and non-western languages in digital technology. He will serve as project liaison to outside technical bodies such as E-MELD and OLAC, and assist Dobrin in writing additional funding documents.

Daniel Pitti, co-PI, will serve as the chief engineer for the project and chief liaison with the rest of IATH. In addition to contributing to specification writing, reviewing technical requirements, and performing direct technical tasks in database development and markup, Pitti will supervise the work of IATH engineers, programmers, and designers, and also interact with the larger markup and archiving community to solve technical challenges raised by the project and disseminate information about successful approaches taken within it.

IATH is the project's home and chief technical resource. IATH will provide programming and scripting resources, web design, web hosting, digitization support, development support, grantwriting support, and ongoing project support for the Digital Language Archive.

## **8.2 Project Timeline**

This project will take three years to complete. The work plan for each year is as follows:

### First Year (September 2005 to August 2006)

Purchase equipment and configure server; hire undergraduate assistant

Plan specifications and develop prototypes for the main Digital Language Archive components:

Grammatical Database, Text and Audio Collection, linking software layer, and user interfaces

Input sample data into prototypes; revise and debug until applications are functional based on initial testing

Digitize audio tapes

Begin data entry into Grammatical Database and Text and Audio Collection

Begin focused work on written reference grammar

### Second Year (September 2006 to August 2007)

Complete data entry into Grammatical Database and Text and Audio Collection

Edit digital audio files

Revise and debug design of Digital Language Archive components as needed

Plan the content of the public-facing website

Continue work on grammar and solicit publishers

### Third Year (September 2007 to August 2008)

Complete data entry into Grammatical Database and Text and Audio Collection

Maintain server and Digital Archive components

Design and launch public-facing website

Work intensively on reference grammar and submit completed draft to publisher for review

Solicit funding for follow-on project to create Arapesh pedagogical materials

In addition to being in regular contact with Dobrin by email and phone, project consultant Bob Conrad will make two four-day visits to Virginia in each of the three years to consult on questions of Arapesh grammatical structure, the interpretation of texts, and issues regarding variation within the Arapesh family. He will also aid in the Arapesh translation for the public facing website.

Project investigators will be in ongoing contact with members of the Advisory Board on an individual, as-needed basis for the duration of the project. The Advisory Board and project investigators will convene in the spring of each year to assess completed work, monitor progress in relation to project goals, and discuss the project's future direction.

## References

- Aikhenvald, Alexandra Y. 2000. *Classifiers: A Typology of Noun Categorization Devices*. Oxford: Oxford University Press.
- Aikhenvald, Alexandra Y. Forthcoming. Language Endangerment in the Sepik Area of Papua New Guinea. *Materials of The Fourth International Conference on Endangered Languages of the Pacific Rim*. Kyoto, 23-25 November 2002.
- Alungum, John., Robert J. Conrad, and Joshua Lukas. 1978. Some Muhiang Grammatical Notes. *Workpapers in Papua New Guinea Languages 25: Miscellaneous Papers on Dobu and Arapesh*. Ukarumpa, Papua New Guinea: Summer Institute of Linguistics.
- The Apache Cocoon Project. n.d. Cocoon 2.1. Software framework and description. Online at <http://cocoon.apache.org/2.1/>.
- Aronoff, Mark. 1992. Noun Classes in Arapesh. In Geert Booij and Jaap van Marle (eds.), *Yearbook of Morphology 1991*, 21-32. Dordrecht: Kluwer.
- Aronoff, Mark. 1994. *Morphology by Itself*. Linguistic Inquiry Monograph 22. Cambridge: MIT Press.
- Bashkow, Ira, and Lise M. Dobrin. In prep. The Great Arc of Human Possibilities and a Small Circle of Friends: The Social Microcosm of Mead's *Sex and Temperament*. For publication in *History of Anthropology*.
- Bird, Steven and Gary Simons. 2003. Seven Dimensions of Portability for Language Documentation and Description. *Language* 79/3:557-582.
- Bourret, Ronald. 2004. XML and Databases. Whitepaper. Online at <http://www.rpbourret.com/xml/XMLAndDatabases.htm>
- Bray, Tim, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler, and François Yergeau. 2004. Extensible Markup Language (XML) 1.0. Third edition. W3C (World Wide Web Consortium) Recommendation, February 04, 2004. Online at <http://www.w3.org/TR/REC-xml/>.
- Clark, James and Steve deRose. 1999. XPath: XML Path Language Version 1.0. W3C Recommendation, November 16, 1999. Online at <http://www.w3.org/TR/xpath>.
- Conrad, Robert J. 1978. A Survey of the Arapesh Language Family of Papua New Guinea. *Workpapers in Papua New Guinea Languages 25: Miscellaneous Papers on Dobu and Arapesh*. Ukarumpa, Papua New Guinea: Summer Institute of Linguistics.
- Conrad, Robert J. 1987. Kinds of Information in Bukiyip Oral Narrative Discourse. *Language and Linguistics in Melanesia* 16:23-40.
- Conrad, Robert J. 1996. More on Arapesh Noun Classes. Unpublished ms.

- Conrad, Robert J., and Kepas Wogiga. 1991. An Outline of Bukiyip Grammar. *Pacific Linguistics C 113*.
- Corbett, Greville. 1991. *Gender*. Cambridge: Cambridge University Press.
- Corbett, Greville. 2001. Why Linguists Need Languages. In Luisa Maffi (ed.), *On Biocultural Diversity: Linking Language, Knowledge, and the Environment*. Washington and London: Smithsonian Institution Press.
- Crystal, David. 2000. *Language Death*. Cambridge: Cambridge University Press.
- Dobrin, Lise M. 1995. Theoretical Consequences of Literal Alliterative Concord. In . In Dainora et al. (eds.), *The Proceedings of the 31st Regional Meeting of the Chicago Linguistic Society, Volume 1: The Main Session*, 127-142. Chicago: Chicago Linguistic Society
- Dobrin, Lise M. 1998. The Morphosyntactic Reality of Phonological Form. In Geert Booij and Jaap van Marle (eds.), *Yearbook of Morphology 1997*, 59-81. Dordrecht: Kluwer.
- Dobrin, Lise M. 1999a. *Phonological Form, Morphological Class, and Syntactic Gender: The Noun Class Systems of Papua New Guinea Arapeshan*. Unpublished doctoral dissertation, University of Chicago.
- Dobrin, Lise M. 1999b. Lexical Splitting in the Kinship Vocabulary of the Buki (Arapesh) Languages. *Language and Linguistics in Melanesia* 28:99-119.
- Dobrin, Lise M. 2001. Arapesh. In *Facts About the World's Languages: An Encyclopedia of the World's Languages, Past and Present*. Jane Garry and Carl Rubino (eds.), 33-38. New York: H. W. Wilson.
- Dobrin, Lise M. and Ira Bashkow. In prep. Arapesh Warfare: Mead and Fortune's Clash of Ethnographic Temperament. For publication in *History of Anthropology*.
- Dyson, Tom. 2003. PostgreSQL and XML. Whitepaper, March 12, 2003.  
[http://www.throwingbeans.org/tech/postgresql\\_and\\_xml.html](http://www.throwingbeans.org/tech/postgresql_and_xml.html).
- EMELD. 2003. General Ontology for Linguistic Description (GOLD) Ontology Tree. Online at <http://cf.linguistlist.org/cfdocs/emeld/school/ontology-tree.cfm>.
- EMELD. n.d.-a. Archival Formats for Audio Digitization. Best Practices recommendations. Online at <http://cf.linguistlist.org/cfdocs/emeld/school/classroom/audio/archive.html>.
- EMELD. n.d.-b. FIELD: Field Input Environment for Linguistic Data. Software program. Description at <http://www.emeld.org/tools/fieldinput.cfm>.
- EMELD. n.d.-c. What Are Best Practices? Best Practices recommendations. Online at <http://cf.linguistlist.org/cfdocs/emeld/school/what.html>.
- Foley, William A. 1986. *The Papuan Languages of New Guinea*. Cambridge: Cambridge University Press.
- Foley, William A. 2000. The Languages of New Guinea. *Annual Review of Anthropology* 29: 357-404.

- Fortune, Reo F. 1939. Arapesh Warfare. *American Anthropologist* 41:22-41.
- Fortune, Reo F. 1942. *Arapesh*. Publications of the American Ethnological Society 19. New York: J. J. Augustin.
- Fraser, Norman M. and Greville G. Corbett. 1995 ms. *Defaults in Arapesh*. University of Surrey.
- Gerstner, P. Andreas. 1963. *Grammatik der Alubänsprache*. Micro-Bibliotheca Anthropos 37. Bonn: St. Augustin.
- Golumbia, David. 1998. Quine's Ambivalence. *Cultural Critique* 38:5-38.
- Golumbia, David. 1999a. Feminism and Mental Representation: Analytic Philosophy, Cultural Studies, and Narrow Content. In Emmanuela Bianchi (ed.), *Is Feminist Philosophy Philosophy?* 202-211. Evanston: Northwestern University Press.
- Golumbia, David. 1999b. Quine, Derrida, and the Question of Philosophy. *The Philosophical Forum* 30/3:163-186.
- Golumbia, David. 2001. Hierarchies, Evidentials, and the History of Constituency. In Sunyoung Oh and Naomi Sawai, eds., *Proceedings of WSCLA 6: The Workshop on Structure and Constituency in Languages of the Americas*. Vancouver, BC: University of British Columbia Working Papers in Linguistics. 51-64.
- Golumbia, David. 2003a. Computation, Gender, and Human Thinking. *differences: A Journal of Feminist Cultural Studies* 14/2:27-48.
- Golumbia, David. 2003b. Metadiversity: On the Unavailability of Alternatives to Information. *Electronic Book Review* 3.0. Online at [http://www.electronicbookreview.com/v3/servlet/ebr?essay\\_id=golumbiaaltx&command=view\\_essay](http://www.electronicbookreview.com/v3/servlet/ebr?essay_id=golumbiaaltx&command=view_essay).
- Golumbia, David. 2004. The Interpretation of Nonconfigurationality. *Language & Communication: An Interdisciplinary Journal* 24:1.
- Grimes, Barbara, ed. 2000. *Ethnologue: Languages of the World*. 14<sup>th</sup> Edition. Website and CD-ROM. Dallas: SIL International. Online at <http://www.ethnologue.com/>.
- Grinevald, Colette. 2000. A Morphosyntactic Typology of Classifiers. In Gunter Senft (ed.), *Nominal Classification*, 50-92. Cambridge: Cambridge University Press.
- Kulick, Don. 1992. *Language Shift and Cultural Reproduction: Socialization, Self, and Sycretism in a Papua New Guinean Village*. Studies in the Social and Cultural Foundations of Language 14. Cambridge: Cambridge University Press.
- Laycock, Donald. 1973. Sepik Languages: Checklist and Preliminary Classification. *Pacific Linguistics* B25.
- Laycock, Donald. 1975. The Torricelli Phylum. In S. Wurm (ed.), *New Guinea Area Languages and Language Study, 1: Papuan Languages and the New Guinea Linguistic Scene*. Pacific Linguistics C38:767-80.

- Leavitt, Stephen C. 1989. *Cargo, Christ, and Nostalgia for the Dead: Themes of Intimacy and Abandonment in Bumbita Arapesh Social Experience*. Doctoral dissertation, University of California, San Diego.
- Leavitt, Stephen C. 1995. Political Domination and the Absent Oppressor: Images of Europeans in Bumbita Arapesh Narratives. *Ethnology* 34/3:177-189.
- Leavitt, Stephen C. 1997. Cargo Beliefs and Religious Experience. In J. P. Spradley and D. McCurdy (eds.), *Conformity and Conflict*. New York: Longman.
- Leavitt, Stephen C. 2000. The Apotheosis of White Men? A Reexamination of Beliefs about Europeans as Ancestral Spirits. *Oceania* 70:304-323.
- Library of Congress. n. d. METS: An Overview and Tutorial. Whitepaper. Online at <http://www.loc.gov/standards/mets/METSOverview.v2.html>.
- Mead, Margaret. 1935. *Sex and Temperament in Three Primitive Societies*. New York: Morrow.
- Mead, Margaret. 1938. *The Mountain Arapesh. I. An Importing Culture*. New York: American Museum of Natural History.
- Mihalic, Rev. Francis. 1996. The Medium of the Message. *Verbum SVD* 37:167-176.
- Miller, Philip H., Geoffrey K. Pullum, and Arnold M. Zwicky. 1997. The Principle of Phonology-Free Syntax: Four Apparent Counterexamples in French. *Journal of Linguistics* 33:67-90.
- Nekitel, Otto. 1984. What Is Happening to Our Vernaculars? *Bikmaus* 2:89-98.
- Nekitel, Otto. 1985. Sociolinguistic Aspects of Abu', A Papuan Language of the Sepik Area, Papua New Guinea. Unpublished doctoral dissertation, Australian National University.
- Nekitel, Otto. 1986. A Sketch of Nominal Concord in Abu' (An Arapesh Language). *Papers in New Guinea Linguistics No. 24, Pacific Linguistics A70*.
- Nekitel, Otto. 1992. Culture Change, Language Change: The Case of Abu' Arapesh, Sandaun Province, Papua New Guinea. In Tom Dutton, ed., *Culture Change, Language Change—Case Studies from Melanesia. Pacific Linguistics C-120*, 49-58.
- Nichols, Johanna. 1986. On Form and Content in Typology. In W. P. Lehmann (ed.), *Language Typology 1985*. (Current Issues in Linguistic Theory 47), 141-162. Amsterdam: John Benjamins.
- Nichols, Johanna. 1989. The Nakh Evidence for the History of Gender in Nakh-Daghestanian. In Howard Aronson (ed.), *The Non-Slavic Languages of the USSR: Linguistic Studies*, 158-175. Chicago: Chicago Linguistic Society.
- Nichols, Johanna. 1992. *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press.
- Nidue, Joseph. A. 1990. Language Use in a New Guinea Village: A Triglossic Profile of Makopin I. *Language and Linguistics in Melanesia* 21:47-69.

- Papua New Guinea National Statistical Office. 1994. *1994 Report on the 1990 National Population and Housing Census in East Sepik Province*. Port Moresby: National Statistical Office.
- Pullum, Geoffrey K. and Arnold Zwicky. 1988. The Syntax-Phonology Interface. In Frederick J. Newmeyer, ed., *Linguistics: The Cambridge Survey. I: Linguistics Theory: Foundations*, Cambridge: Cambridge University Press.
- Romaine, Suzanne. 1992. *Language, Education, and Development: Urban and Rural Tok Pisin in Papua New Guinea*. Oxford: Clarendon Press.
- Roscoe, Paul B. 1989. The Flight from the Fen: The Prehistoric Migrations of the Boiken of the East Sepik Province, Papua New Guinea. *Oceania* 60/2:139-154.
- Roscoe, Paul B. 1994. Who Are the Ndu? Ecology, Migration, and Linguistic and Cultural Change in the Sepik Basin. In Andrew J. Strathern and Gabriele Stürzenhofecker (eds.), *Migration and Transformations: Regional Perspectives on New Guinea*, 49-84. Pittsburgh: University of Pittsburgh Press.
- Senft, Gunter. 2000a. *Systems of Nominal Classification*. Cambridge: Cambridge University of Press.
- Senft, Gunter. 2000b. What Do We Really Know about Nominal Classification Systems? In G. Senft (ed.), *Systems of Nominal Classification*, 11-49. Cambridge: Cambridge University of Press.
- Swadling, Pamela. 1984. Sepik Prehistory. In *Sepik Heritage: Tradition and Change in Papua New Guinea*. Nancy Lutkehaus, Christian Kaufmann, William E. Mitchell, Douglas Newton, Lita Osmundson, and Meinhard Schuster (eds.), 71-86. Durham, N.C.: Carolina Academic Press.
- The Tibetan & Himalayan Digital Library. The THDL Tibetan Collaborative Dictionaries. Digital language archive. Online at <http://iris.lib.virginia.edu/tibet/reference/dictionary.html> and <http://www.thdl.org>.
- Tuzin, Donald F. 1976. *The Ilahita Arapesh: Dimensions of Unity*. Berkeley: University of California Press.
- Tuzin, Donald F. 1980. *The Voice of the Tambaran: Truth and Illusion in Ilahita Arapesh Religion*. Berkeley: University of California Press.
- Tuzin, Donald F. 1997. *The Cassowary's Revenge*. Chicago: University of Chicago Press.
- The Unicode Consortium. 2003a. The Unicode Standard, Version 4.0.1, defined by: *The Unicode Standard, Version 4.0*. Reading, MA, Addison-Wesley. As amended by Unicode 4.0.1, online at <http://www.unicode.org/versions/Unicode4.0.1/>.
- The Unicode Consortium. 2003b. IPA Character Extensions. Part of Unicode 4.0.1. Online at <http://www.unicode.org/charts/PDF/U0250.pdf>.