The Arapesh "suitcase miracle": The interpretive value of reproducible research

LISE M. DOBRIN

_____

_____

The Arapesh "suitcase miracle": The interpretive value of reproducible research

Lise M. Dobrin
University of Virginia

Abstract

Recent discussions about reproducibility in linguistics emphasize that access to the original documentation allows analyses to be checked against the data that underlie them. But every annotated recording is itself always an interpretation, with some information and perspectives highlighted and some obscured. For this reason, access to the original documentation offers something better than scientific accountability: it offers a view into the underlying acts of interpretation made by other researchers and the consultants and assistants they worked with. These acts of interpretation – even problematic or "incorrect" ones – can generate insight into both linguistic structure and the social relations developed and mobilized in the context of research. This point is illustrated through analysis of legacy documentation of the Bukiyip variety of Arapesh (Papua New Guinea), which was collected fifty years ago by SIL linguist Robert Conrad and transcribed by a native speaker. Analyzing the transcribed texts against the original recordings shows the transcriptions to be problematic, with significant stretches of speech elided and "corrections" introduced by the transcriber. Yet these discrepancies do something more valuable than disconfirm the texts' validity: they reveal linguistic patterns that would otherwise go undetected, and in some cases are only detectable now because of the discrepancies.

1. Introduction: The call for reproducible research

When linguists collect recordings of naturally occurring and elicited speech, they are expected to archive the materials so they will be accessible to others (Himmelman 1998; Linguistic Society of America 2010; Woodbury 2011). This has not always been the case. It was not very long ago that "primary recordings […] were regarded as not having any intrinsic value [as] can be seen by the way in which they were treated – left in filing cabinets, in boxes in

garages, in deceased estates, with no catalogs of their contents" (Thieberger 2014: 156). Even when researchers saw their materials as valuable, the prospect of sharing them often made them uncomfortable; as one senior linguist put it, "fifty years ago, most field researchers that I know considered the data collected in the field to be the exclusive property of the field researcher […]. Asking someone to see their primary data was like asking to see them in their underwear – you were transgressing on very private territory" (Daniel Everett, personal communication, 2020-07-12). But within the contemporary documentary linguistic paradigm it is almost unthinkable that a linguist would carry out fieldwork on a language without plans to archive the resulting documentation. We have returned to the ethos of an earlier era guided by the Boasian aim to document and preserve linguistic practices in all their diversity, so that the "relationship between language documentation and the archive" has once again become nearly "inseparable" (Henke & Berez-Kroeker 2016: 416).

The imperative to create, preserve, and share primary language data is supported by many new forms of infrastructure in linguistics: language archives, data management tools, and journals like this one that are devoted to publishing about language documentation; funding streams that support the documentation of endangered languages; training programs and conferences that are focused on linguistic fieldwork and documentary research methods; a rich literature on the ethical dimensions of linguistic research in indigenous and minority language communities; guidelines and models for the professional evaluation of documentary outputs; even an archivists' award for outstanding language corpora.[1] These major disciplinary investments are justified by the understanding, widely shared since the publication of Hale et al. (1992), that primary language data is not just an incidental byproduct of documentary linguistic research but indeed one of its most important outcomes. The thoughtful collection and preservation of primary language data helps put linguistics on solid footing for the future given that so many languages are now or are soon likely to be endangered, making opportunities for later documentation and description uncertain. It also enables the production of language learning and reference materials to support language maintenance and renewal.

But primary data is still relatively backgrounded in linguistic publications, the key locus where academic value is expressed and assessed. How important can documentary sources be to the knowledge production process if linguists do not need to directly build upon them in their written outputs? To bring publications into line with a renewed emphasis on primary data as central for

---

[1] See https://www.delaman.org/delaman-award/.

linguistic analysis and theorizing, researchers are encouraged to help "foster a culture shift" in linguistics by increasing the transparency of the links between their claims and data sources in their publications (Gawne et al. 2017: 178). They are encouraged to provide more elaborate discussions of "the methodologies employed in data collection, management, and analysis", and to cite their sources in a way "that will allow the reader to not only locate the larger data set of recordings and/or fieldnotes in an archive, but also to resolve back to the particular datum within the set" (Gawne et al. 2017: 176–177).

The argument for more transparent data citation practices is grounded in the idea that linguistic research should be reproducible: it should allow for independent assessment of the original observations on which analyses are based. This is proposed as a desideratum for researchers in all subfields of linguistics – not just documentary linguists but even theoretical syntacticians: they should cite their data sources and make them accessible. Access to the original data is argued to "play a key role in increasing verification and accountability in linguistic research" by allowing interpretations to be checked against the underlying observations that are claimed to substantiate them (Berez-Kroeker et al. 2018: 1–2; see also Gawne et al. 2017: 158). This call for reproducibility is justified by an appeal to the requirements of science. The idea is borrowed from disciplines like field biology where claims cannot be confirmed by the replication of experiments, but where data originally collected in uncontrolled conditions can at least be independently scrutinized.

In what follows, I argue that access to underlying documentary sources such as recordings and fieldnotes contributes to the knowledge production process in linguistics not by giving us the power to scientifically verify or disconfirm analyses, but by providing additional interpretive angles from which to view and understand them. The argument is based on work with linguistic legacy materials, records of language that were previously produced by others. Reproducibility would seem to encourage work with legacy materials, since it suggests that linguists should be scrutinizing others' analyses and reading them against their data. The experience I have had doing something like this provides an alternative framework for appreciating what references to primary data can contribute to linguistics. Even though documenters may aspire to produce a maximally transparent record of what was said and what it meant, every annotated recording is itself always an interpretation, with some information and perspectives highlighted and some obscured (see Dobrin & Schwartz, this volume). For this reason, access to the original recordings offers something other, and perhaps better, than accountability: it offers a view into the underlying acts of interpretation made by other researchers and participants involved in the materials' production. Making visible these acts of interpretation – even problematic ones leading to "incorrect" transcriptions and analyses – can generate important new insights into both linguistic structure and the social circumstances that gave shape to the materials.

2. The Arapesh "suitcase miracle"

The focus of my argument is a legacy corpus of Bukiyip (ISO 639-3 ape), an Arapesh language of the New Guinea Sepik plains that is now in an advanced state of shift to Tok Pisin, the lingua franca of Papua New Guinea (PNG). The Bukiyip corpus includes 75 typed texts – some 1300 pages of computer printouts such as the one shown in Figure 1 – that were collected in the early 1970s by SIL linguist Robert J. Conrad as part of his preparation for translating the New Testament into Bukiyip. By the time I began working on

Figure 1. Scanned excerpt from Conrad's Bukiyip texts

the Bukiyip corpus Conrad and I were already well acquainted. His publications on Arapesh constituted some of my most treasured sources, including a survey of the language family (Conrad 1978), a grammatical sketch of the Muhiang (also called Mufian) variety (Alungum et al. 1978), and a full-length grammar of Bukiyip, *An Outline of Bukiyip Grammar* (Conrad & Wogiga 1991). Conrad and his wife Jo Ann had facilitated my entry into Arapesh country when I first arrived in PNG in 1997 as a graduate student preparing to conduct dissertation fieldwork on the role of phonology in Arapesh noun classification. Conrad used his extensive networks to help me establish a village home, and we saw each other frequently over the fifteen months I spent there. We traveled together throughout the region, sharing notes and talking excitedly whenever we met about Arapesh language structures and dialectology, Melanesian culture, common acquaintances in PNG and in the

US, and developments in linguistic theory. Over that time our relationship evolved into one of mutual care and admiration, and we remained dear friends and colleagues until his death in 2021 at the age of 89 while this paper was in preparation.

In 2007, Conrad carried the Bukiyip texts with him in a suitcase on a visit to my home in Charlottesville, VA, and he offered them to me to preserve alongside my own field materials on other varieties of Arapesh. I had not known about the texts' existence until that time. My research assistant Amanda Glass and I quickly realized we were bearing witness to what we began calling a "suitcase miracle". The texts, which Conrad told us had been transcribed by a native speaker shortly after they were recorded, represented the unpublished database for his published grammar. The suitcase also held some minimally labeled moldy cassette and reel-to-reel tapes that we arranged to have deposited in the Tuzin Archive of Melanesian Anthropology at UC San Diego along with the original manuscript materials.[2] At the time we had little confidence that the information on the recordings would be recoverable or useful, though as I discuss below that turned out not to be the case. The manuscript materials included not only the computer printouts of the Bukiyip texts but hand-sketched maps of the May River region where Conrad had worked with Sepik Iwam speakers prior to switching his focus to Arapesh languages in the Torricelli foothills, computer printouts of Mufian legends, Mufian phonology materials, and a number of other loose texts. As we were processing the materials, Conrad explained that the original recordings the Bukiyip texts were based on had unfortunately been lost.

Bukiyip is similar in structure and lexicon to Cemaun, the variety of Coastal Arapesh I have studied most closely, but it is spoken inland across the Prince Alexander mountains, so I had been relying on Conrad & Wogiga's grammar to guide my understanding of the language (see Figure 2).

Since the Bukiyip grammar had been published without the supporting texts, and since an absence of glossing limited the texts' utility for future work of any kind, we began taking steps to translate and preserve them. In collaboration with the University of Virginia's Institute for Advanced Technology in the Humanities we had the texts scanned, keyboarded, encoded in XML, and preserved in a secure digital environment. Conrad and I began meeting through video calls and in person in Charlottesville and at the SIL retirement center in Dallas, TX, where he lived at the time to review and

---

[2] Conrad, Robert (ca. 1960–1990). Robert Conrad Papers, MSS 732, Special Collections and Archives, University of California San Diego Library. https://library.ucsd.edu/speccoll/findingaids/mss0732.html.
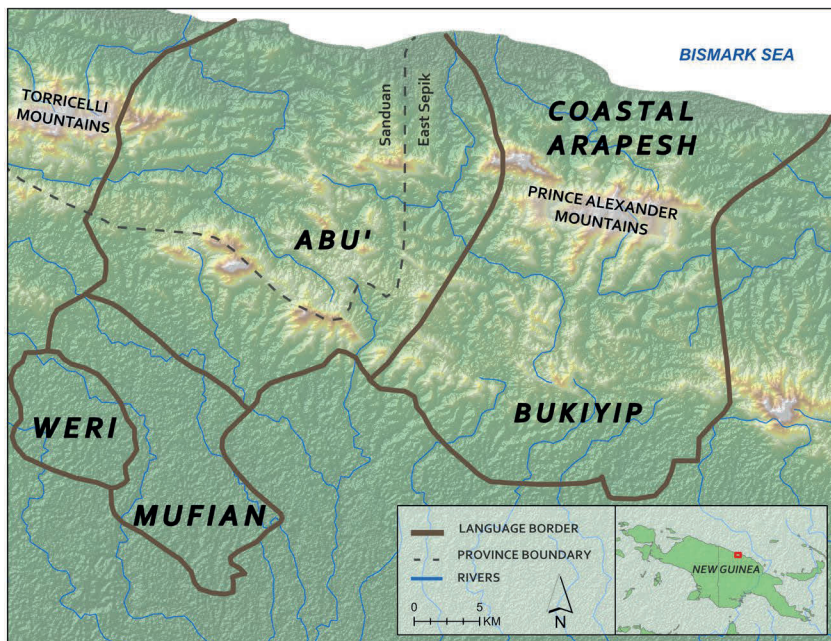
Figure 2. The Arapesh language area

translate them. We worked on the texts together like this more or less weekly over a period of five years, from 2011 to 2016. Our work was partially supported by an NEH Chairman's grant as a kind of addendum to the Arapesh Grammar and Digital Language Archive that I was then developing with NSF-NEH support.[3]

In many respects it was exhilarating work. Through our shared investment in the materials my relationship with my senior colleague deepened. We paid visits to one another's homes and families, pored over the texts together in order to check the transcribed forms and provide them with line-by-line glosses, and created new ethnographic records as I typed up his reminiscences of life in the Torricelli foothills during the early 1970s. Our discussions led Conrad to revisit some of his firsthand experiences with the magico-religious Mount Hurun Cult and the somewhat more practically focused Peli Association, dynamic social movements that rocked the region at the time he

---

[3] Dobrin, Lise M., collector and editor. Arapesh Grammar and Digital Language Archive. Institute for Advanced Technology in the Humanities, University of Virginia. http://www.arapesh.org/.

began working there, attracting thousands who were deeply dissatisfied with their colonial situation and eager for development in the lead-up to PNG's independence from Australia (see, e.g., May 1982; Camp 1983; Roscoe 2004). Conrad spoke to me about the personal qualities and biographies of the individual speakers who produced the texts, enriching the materials with precious metadata. With time I found myself able to understand more of the Bukiyip, which I now appreciate is quite a close relative of Coastal Arapesh despite originally appearing to me to be more distant because of how it was rendered in writing by Conrad and his SIL team.

But our joint glossing project could also be a frustrating one. The texts were riddled with inconsistencies, and Conrad's memory was fading. Guesses about uninterpretable stretches of text could be checked against a small Bukiyip dictionary that Conrad and his SIL collaborators had made (Juagu & Conrad 2008), but the dictionary was itself problematic: full of gaps, written in an insufficient orthography, and inconsistent in its use of citation forms, forcing you to guess and hunt around when you wanted to look something up. Should you go for the root or an inflected form? If the latter, which one? I occasionally posted a query based on our work with the texts to an Arapesh Facebook group I belong to, but only rarely did these elicit informative answers. I left many of our text glossing sessions wondering if this was really a worthwhile use of our time.

Just as we were nearing the end of this painstaking work, my colleagues and I received a package in the mail: it was the audio materials we had submitted to the Tuzin Archive, now digitized and returned to us on DVDs. What we discovered when we listened to them was that the source recordings corresponding to the Bukiyip texts had not all been lost as Conrad had believed. To the contrary, we were now holding about a third of them in our hands. We were thus given the chance to check many of the texts, and our work interpreting them, against the original audio – a kind of a natural reproducibility test. Doing this revealed the texts to be highly problematic, with stretches of speech elided and systematic "corrections" introduced by the transcriber. But these discrepancies did something far more interesting than confirm or disconfirm analyses: they revealed linguistic patterns that otherwise went undetected, and in some cases were only detectable now *because of* the gaps between the audio and the transcript. They also provided perspective on the interactions that gave rise to the collection.

Conrad could unfortunately tell us little about who the transcriber was or what he understood his task to be. He only recalled that the transcription was done by a Bukiyip man who knew how to write, and that having him do this work saved Conrad a great deal of time. In contrast to the recorded speakers, whom Conrad could almost always name and characterize, the transcriber seemed not to be a key participant in the documentation process from Conrad's point of view. But from our perspective as users of the materials fifty years later, this unnamed individual could not have been more influential.

In making decisions as he transcribed, he acted as a kind of linguistic and cultural filter who left his imprint on the materials as he added, removed, or adjusted information, and started or stopped his transcriptions at the points that made sense to him. There is no neutral basis for judging the decisions he made, even when they lead the texts to depart significantly from what was recorded. Fidelity to the audio recording is just one of the ways that a transcript can be "right"; as we will see below, a transcript can also be right because it is consistent in the code the speech is expressed in, or because it makes use of standard spellings. These evaluations highlight the profound difference between orality and literacy as modes of language production that the task of transcription makes it necessary to navigate and resolve. The intellectual investments of native speaker transcribers and editors are a valuable source of insight into how people think about their language in both its oral and (in this case only newly) written forms (Mosel 2012, 2015).

3. Discrepancies between recordings and texts

Conrad's Bukiyip recordings from the early 1970s are to my knowledge the earliest extant examples of audiorecorded Arapesh speech. They thus provide a window into the historical trajectory of language shift to Tok Pisin, the creole lingua franca of PNG that is now rapidly overtaking local languages across the Sepik region and elsewhere in the country. Prior to receiving the recordings, my research assistant Amanda Glass, whose ear had been trained on recordings of Cemaun Arapesh that I had collected in the late 1990s, had noticed that there was much less code switching in the Bukiyip texts than she was used to encountering in my materials. Conrad's documentation was carried out nearly thirty years earlier, we reasoned; it would only make sense that people spoke a more monolingual form of the vernacular then. But the digitized recordings taught us that our speculative reasoning was wrong. The Bukiyip recordings abound in Tok Pisin code switching and borrowings. What this discrepancy shows is that first, Tok Pisin was already well entrenched in the speech habits of Bukiyip speakers at that time, and second, the apparent purity of the Bukiyip texts was a creative restoration by the transcriber.

The following example shows how significantly, and how systematically, the original transcripts differ from what is heard on the recordings. In the original transcript of text RG, which represents a genre that Conrad labels "mythic" or "folkloric", there are 41 tokens of the discourse marker *ali* 'and, and so, then, well'.[4] Most of these appear utterance-initially, where they

---

[4] For continuity I maintain the system of text genre labels established by Conrad whether or not I agree with how the texts are categorized. The first of the two letters in each of Conrad's text labels indexes the genre: H – "hortatory", N – "narrative", R –

introduce new steps in a narrative sequence. The marker is a variant of *alia* or *əriə*, which is used with a similar range of functions across the New Guinea north coast and which itself may be an earlier borrowing into Arapesh from a neighboring language (Ross 2001: 150–51). But as we learned by listening to the recording, what the speaker, a Bukiyip man named Peilug, actually uses in order to fulfill this function throughout the recorded discourse is not *ali* but its Tok Pisin equivalent *orait*. An example utterance from the text illustrating this discrepancy is given in (1).[5] Here and throughout the exposition, transcriptions from the Conrad collection are presented in all caps on the top line. This follows the convention used in the original printed texts we inherited (see Figure 1). Our own transcriptions are presented in lower case. Indeed, there

(1) LISTEN (RG_14.39–14.45)

ALI	APAK	DOUMUN	MW-E-YAGULEH	BUKIYɨP
orait	apak	doumun	mw-e-yaguleh	bukiyip
and	1.PL.PRO	now	1.PL-R-speak	Bukiyip.language

'And so now we speak Bukiyip.'

are only two instances of the *ali* marker actually attested on the recording; everywhere else Peilug says *orait*. We infer from this that *ali* was supplied by the native-speaker transcriber, presumably with the intention of correcting or improving what he heard as he converted the audio-recorded speech to written form.[6]

"myths, folktales", etc. The second letter indexes the text's position in the sequence of that genre. So NA is the first narrative text in the collection, NB is the second, NC is the third, etc. Time-stamped audio links with these labels reference segmented recordings corresponding to the texts in the author's active research files. The recordings may be accessed on a single page at http://www.arapesh.org/suitcase_miracle/ as well as through the clickable "LISTEN" links throughout this article.

[5] The following abbreviations are used in examples: 1, 2 – first person, second person; I, II, III, etc. – noun class I, noun class II, etc.; BEN – benefactive; CTPT – centripetal (movement toward speaker or topic); COMPL – completive; COND – conditional; DEM – demonstrative; DU – dual; F – feminine; FUT – future; IR – irrealis; LOC – locative; M – masculine; MOD – modifier; NEG – negative; PFV – perfective; PL – plural; POSS – possessive; PRO – independent pronoun; R – realis; SG – singular; TP – Tok Pisin.

[6] We cannot know whether decisions such as these were made on the transcriber's own initiative or at Conrad's urging. However, they would at least have been made with Conrad's approval, since I know from being with him in the field that his standard text production method involved independent work by a native transcriber or Bible translator followed by a joint review of the text in order to check the person's work and make any corrections.

We also find that Tok Pisin predicates are systematically transcribed as their Bukiyip equivalents. In Arapesh languages, predicates are often borrowed from Tok Pisin by incorporating them into a light verb construction. The examples I will use to illustrate this are taken from text NC, an instance of Conrad's default "narrative" genre, which was provided by a speaker named Joni. In the original transcript, the light verb + Tok Pisin predicate *ɲ-i-ne stret* 'VIII.SG-IR-do/make correct(TP)' was nativized to the Bukiyip stative verb *yopwi-ɲ* 'be.good-VIII.SG', as shown in the ELAN screen shot in Figure 3.[7] Figure 4 shows how the light verb + Tok Pisin borrowing *y-e-ne hamamas* '1.SG-R-do/make happy(TP)' in the same text was similarly nativized in the transcript by substituting for it the Bukiyip verb *y-ə-nəhilau* '1.SG-R-be.happy/fly'. The native expression for 'be happy' is one that even the eldest Arapesh speakers can no longer confidently provide, as I learned in 2013 when I overheard an unprompted debate about it among people in an Arapesh community who were trying to decide what vernacular form with this meaning to put on a poster. The Bukiyip recordings provide a clue as to why that might be: already for many decades the Tok Pisin variant has been the preferred form.
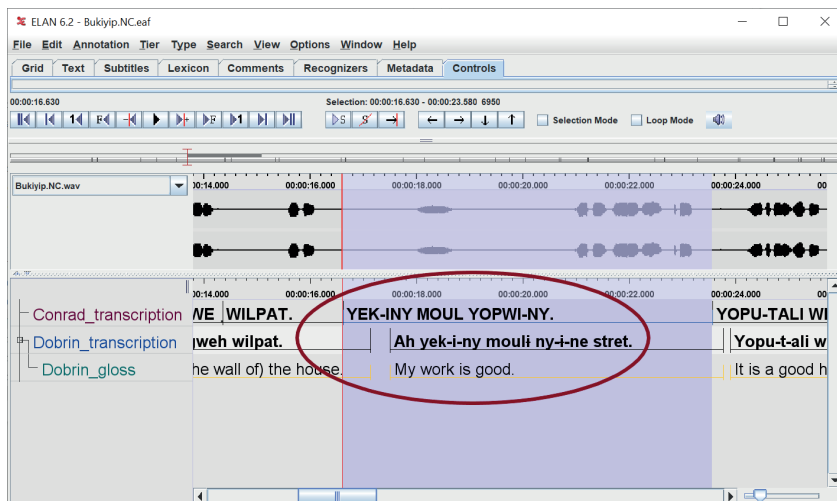
Figure 3. LISTEN (NC_0.18–0.24). *Nativization of Tok Pisin predicate ɲ-i-ne stret 'VIII.SG-IR-do/make correct' in text NC.*

---

[7] A description of the ELAN media annotation tool can be found at https://archive.mpi.nl/tla/elan.
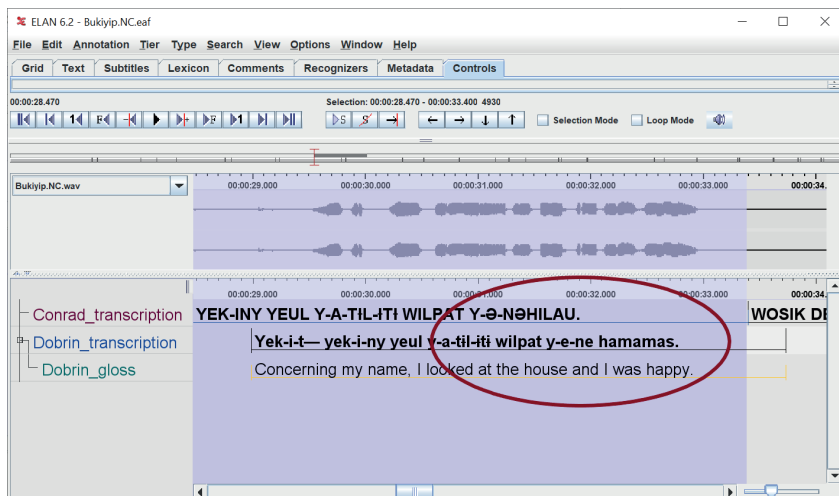
Figure 4. LISTEN (NC_0.29–0.34). *Nativization of Tok Pisin predicate* y-e-ne hamamas *'1.SG-R-do/make happy' in text NC.*

Another systematic intervention made by the transcriber was to remove repetition. Doing this has the benefit of moving the text along more efficiently, but it distorts the flow of the discourse as it does so. As in other Papuan languages, an important device used to create coherence in Arapesh is tail-head linkage, where repetition of a clause (or just its verbal head) marks a discourse relation between two events. According to de Vries (2005: 365) it is found most commonly in narratives and procedural texts, genres which account for the majority of the texts in the Bukiyip collection. Tail-head linked chains have a distinctive intonational contour, with a fall on the first instance of the verbal element and a rise, often followed by pause, on the repeated element marking the shift to a subsequent clause. Example (2) is an instance of tail-head linked clauses from a Cemaun Arapesh text spoken by Scola Sonin that I recorded. In it she explains how to make the traditional Arapesh feast food *saməhəs*, taro or banana that has been boiled, mashed, and rolled in grated coconut. The symbol <#> indicates a falling intonational contour; </> indicates a rising contour followed by pause.

(2) LISTEN (Samahas_2.24–2.46)

Oke	h-a-dik-ec	pinis
okay(TP)	M.PL-R-shoot-VIII.PL	PFV(TP)

'So the men pound them

ei	y-e-rih –	y-o-wah
1.SG.PRO	1.SG-R-XII.PL	1.SG-R-remove

and I [remove] them – I remove the taro roots

ga   marɨp   y-a-rɨh-ir
LOC   saucepan   1.SG-R-XII.PL-be.inside
[that are] in the saucepan

dəbei-hʷ-i   yaurəhʷ   oke   h-ə-hir
big-XII.SG-MOD   plate   okay(TP)   M.PL-R-carry
and put them on a big plate, and then they take

əɲin-də   rowog apokwam   orait   teibog
VIII.SG-DEM   tree   dish   alright(TP)   pounder
this wooden dish along with a pounder,

[TAIL   ]   [HEAD   ]
h-a-dik-irih #   H-a-dik-irih   pinis /
M.PL-R-shoot-XII.PL   M.PL-R-shoot-XII.PL   PFV(TP)
and they pound them. Then, when they are done pounding them,

[HEAD (REPEAT)]
na   pinis /   na   h-a-rupw-ec #
and(TP)   PFV(TP)   and(TP)   M.PL-R-cut-VIII.PL
all done, they cut them up.'

In (2) the inflected verb *hadikirih* '(the men) shoot (i.e., pound) them' is pronounced with falling intonation; this is the tail of the first clause. The verb is then repeated at the start of the next clause, where it is followed by the Tok Pisin perfective marker *pinis* and a rise, indicating that the action was completely wrapped up before the next one began: *harupec* '(the men) cut them'. In this example the completion of the prior action is further emphasized by a second, partial, repetition of the head: *na pinis*.

Such discourse-structuring repetition is frequently elided from Conrad's Bukiyip texts. In (3), for example, we have an extract of a narrative taken from an audio-recorded text in Conrad's "hortatory" category spoken by Peilug in which a tail-head linked structure is used.

(3) LISTEN (HP_0.26–0.37)

[TAIL   ]
ch-a-wok   on-ogw   yabigw #
VIII.PL-R-consume   some-XI.PL coconut.soup
'They drank some coconut soup.

[HEAD   ]
ch-a-wok   on-ogw   yabigw /
VIII.PL-R-consume   some-XI.PL coconut.soup
Having drunk some coconut soup,

əlia	ch-a-tɨk	solich /
and.then	VIII.PL-R-remove	dishes

they took out the dishes,

ch-o-kule	alia	əgɨnɨ	yopu-gɨn-umu #
VIII.PL-R-remove	and.then	here	good-LOC-MOD

and then they put them here out in the open.'

But you would never know this by looking at the text because it only includes one instance of the clause *chawok onogw yabigw* 'they drank some coconut soup'. The repetition is not transcribed.

4. Discourse problems caused by the discrepancies

The divergence of native-speaker transcripts from their spoken originals is by no means an unprecedented phenomenon, and it is worth studying because it contributes to what Nikolaus Himmelmann (2018: 35) has called "the transcription challenge": understanding "what […] speakers and researchers actually do when they transcribe". Two related cases from the Americas are discussed in the volume *Natural histories of discourse*, which explores the processes by which texts are distilled or "entextualized" out of recorded speech (Silverstein & Urban 1996; see also Jung & Himmelmann 2011; Marten & Petzell 2016). In his contribution to the volume, Greg Urban analyzes transcripts of Indigenous discourse from southern Brazil that were produced by two local assistants whose relationships to the speaker and the outside researcher differed in ways that affected the transcribers' decisions and thus the resulting texts. John Haviland analyzes the rendering of Tzotzil recordings by native speaker transcribers from the Zinacantec community of Mexico where they were collected. What he concludes is that even individuals "whose experience with reading and writing in any language (let alone their own) [was] next to nil" were "able implicitly to indicate *what a text should be like*" (Haviland 1996: 48–49). (Any linguist transcribing documentary corpora would benefit from reading these fascinating studies.) The authors of these articles find systematic patterns in what happens when spoken discourse undergoes "transduction" to writing. One is that elements whose presence contributes primarily to coherence in discourse, rather than to the referential meaning of the talk, tend to get edited out (Haviland 1996: 53 et seq., 69). Another is that transcribers "routinely and consciously purged overlaps and repetitions" in order to "make things come out more neatly" (Haviland 1996:

60–61). The compression of tail-head linked structures we find in the transcribed Bukiyip texts simultaneously instantiates both of these patterns.[8]

The fact that so many aspects of the Bukiyip texts do not faithfully represent what was originally uttered but are creative interventions made by the transcriber – and I should emphasize that some of the texts diverge from the recordings dramatically – makes them a tenuous basis for language description, since they do not in fact capture "how the people actually communicate with each other" (Himmelmann 2006: 7).[9] But the kinds of transcriber adjustments we find in these texts are problematic for reasons that go beyond the limits they place on the texts' utility for grammatical analysis. They also distort the speakers' voice, "the particular [culturally shared] ways […] in which subjects produce meanings" through implicit structure, or *how* they say what they say (Blommaert 2006: 10). The ethical imperative for students of language to recognize and validate marginalized ways of speaking through the textual products of their research was one of the driving forces behind the movement for ethnopoetic presentation of Native American narratives pioneered by Dell Hymes in the 1980s and 1990s (Hymes 1981, 1996; Kroskrity & Webster 2015). The goals of this movement should resonate with any documentary linguist who understands their charge as not only to build a database of oral texts that will support linguistic description, but also to preserve forms of expression that meaningfully and accurately represent the cultural world from which they are drawn.

For an example illustrating how the speaker's voice can be muted by a seemingly minor choice made by the transcriber, consider another selection, (4), from the "hortatory" text HP discussed above. These two lines occur about midway through the recorded event, where Peilug initiates the topic of how he objected to serving guests food in western-style dishes, or *peletog* (the Tok Pisin noun *pelet* 'plate' marked with the formally appropriate Bukiyip plural), instead of traditional coconut shell bowls. (The reason for his

---

[8] In a study of Teop legends that had been transcribed by native speakers on the Pacific island of Bougainville, Mosel (2015) finds that tail-head linked structures are *overused*: they are introduced where paratactic constructions were actually uttered. Mosel hypothesizes that transcribers do this in an effort to more concretely specify the nature of the relationship between two clauses.

[9] The following are some excerpts from the notes my research assistant Amanda Glass began keeping on the Bukiyip text-audio correspondences as their divergences were becoming apparent. Regarding text NB: "[T]wo lines transcribed at the end that aren't on the audio file." Regarding text NN: "[Lines] 123–124, 242, 413 are not in the recording." Regarding text NO: "The audio *really* doesn't align with the transcription." Regarding text NW: "Recording continues beyond the transcription. Transcription just cuts off." Regarding text RL: "Transcription ends at a song. The story picks back up after the song but the transcript does not."

objection was clear to neither Conrad nor me.) The top line in all-caps is how Peilug's speech is presented in the original transcript. The lower line of transcript is the one we produced based on listening to the audio recording. Separate interlinear glosses are provided where the transcriptions diverge. The elements to focus on are in bold.

(4) LISTEN (HP_1.16–1.21)

PELETOG	**Y-A-KLI**	**WAK**
dishes	**1.SG-R-say**	**NEG**
peletog	**mi(TP)**	**tambu(TP)**
dishes	**1.SG**	**taboo**

'(Concerning making up and serving people) dishes (of food), I forbade it.

Y-A-KLI	**WAK**-ɨMU	P-I-CH-AHU	PELETOG
1.SG-R-say	**NEG**-BEN	2.PL-IR-VIII.PL-serve	dishes
Y-a-kli	**wak**-imu	p-i-ch-ahu	peletog

I said no to you serving (food in) dishes

P-Ɨ-NAK	P-Ɨ-K-ECH	**WAK**
2.PL-IR-go	2.PL-IR-give-VIII.PL	**NEG**
p-i-nak	p-i-k-ech	**wak**

and going and giving (them) to them, no!'

    As can be seen in (4), the transcriber "corrected" the Tok Pisin *mi tambu* 'I forbade' to its Bukiyip equivalent *yakli wak*. This might seem to be of little consequence, since the exact words substituted in fact appear in the very next line. But Peilug's use of Tok Pisin here was not just a slip; it was a strategic attempt to create one of "the great variety of rhetorical effects that can be achieved through code-switching", a common feature in Papua New Guineans' speech (Kulick 1992: 77). In this case Peilug was availing himself of "self-repetition, where a speaker says one thing in one language and then repeats the same thing in another language […] to emphasize a command or a warning" (Kulick 1992: 77). The emphatic intent of Peilug's speech – "I said no!" – is suggested by the high pitch and amplitude on the final *wak* 'no' in example (4). It becomes even more apparent as Peilug continues (5):

(5) LISTEN (HP_1.22–1.31)

Wak. Em bai yakli wak. Yakli wak stret. Yek, yakli. Yek Peilɨk yakli wak ɨmu pichahu peletogw pɨkech.

'No. About that I said no. I said no way. I, I said it. I, Peilug, said no to you serving [food in] dishes and giving it to them.'

So while the transcriber's nativization of the Tok Pisin in (4) may improve the text in one respect, by making the code more consistent, it also undoes the emphatic voice that was effected by Peilug's code switch. This erasure of communicative effect is particularly poignant given Peilug's insistence in what follows in (5) that his words should have force because he himself is the one speaking them.

The expression of voice is likewise disrupted by the transcriber's removal of direct repetition. Some background on the poetic structure of Arapesh narrative is required in order to understand what I mean by this. Bukiyip narratives – and, more generally, extended stretches of speech produced by people in PNG's Sepik region regardless of which code they are using – often make use of a discursive rhythm organized around units of four. In ethnopoetic terms this often takes the form of four-verse stanzas, sequences of verses made up of lines, typically with one verb per line (Dobrin 2012b). (6) is a simple example of a four-verse stanza from Bukiyip text NN. The recorded event, spoken by a man named Jeliga, consists of a series of verbal snapshots of imagined happenings in ordinary village life. Here he describes the activity of cutting a new garden, which is idealized as taking place over the course of one day.[10]

(6) LISTEN (NN_0.53–1.07)

1	*Komon muna mulib yawihas*	1	Tomorrow we'll go cut gardens.
2	*Mɨyatəgənik*	2	We'll finish the area completely,
	wabigən mɨnaki		then in the afternoon we'll come back.
3	*Munek waligən*	3	We will make food
	mugɨnah		and eat it.
4	*Michuh*	4	We will sleep.

Utterances grouped in this way feel coherent and complete to listeners who are accustomed to this speech pattern, just as sequences of threes feel rhythmically complete to those with an English-language narrative sensibility. (Hence the satisfaction English speakers get from a basic story frame centered around three objects or attempts at an action like "Goldilocks and the Three Bears", memorable triads like "Life, liberty, and the pursuit of happiness", and five-paragraph essays consisting of three main points bracketed by an introduction and a conclusion.) Part of the evidence that these verses constitute a unit is that the line that follows them begins with *biyebih* 'the day

---

[10] The sound on this recording is unfortunately marred by flutter from the irregular drive of the tape. The music that is audible in the background seems to have already been on the tape when Conrad recorded over it to capture Jeliga's story.

after tomorrow', a time adverb parallel to the initial element in (6), *komon* 'tomorrow', suggesting the start of an equivalent unit, i.e., a new stanza.

Returning to example (3), copied in again here as (3'), we are now in a position to recognize this sequence of utterances as another instance of a four-verse stanza whose rhetorical force was diminished by the removal of repetition, in this case through simplification of a tail-head linked structure in the transcript.

(3') LISTEN (HP_0.26–0.37)

[TAIL   ]
ch-a-wok   **on-ogw**   **yabigw #**
VIII.PL-R-consume   some-XI.PL   coconut.soup
'They drank some coconut soup.

[HEAD   ]
ch-a-wok   **on-ogw**   **yabigw /**
VIII.PL-R-consume   some-XI.PL   coconut.soup
Having drunk some coconut soup,

əlia   ch-a-tɨk   solich /
and.then   VIII.PL-R-remove   dishes
they took out the dishes,

ch-o-kule   alia   əgɨnɨ   yopu-gɨn-umu #
VIII.PL-R-remove   and.then   here   good-LOC-MOD
and then they put them here out in the open.'

Paul Kroskrity (2015) calls such erasure of indigenous verbal artistry in the entextualization of oral speech "discursive discrimination". Although Kroskrity was looking at Native American texts that had been written down by Europeans, the decisions discussed here have the effect of muting locally meaningful Bukiyip rhetorical features even though the conversion of speech to writing was carried out by a native transcriber.

5. Linguistic insights facilitated by the discrepancies

Any time transcription is done by a human being, future users will have to do the interpretive work of trying to understand the traces that person left on the resulting texts. This need not be seen as a drawback. While one might think that the ideal transcriber would be a machine, the fact that transcribers are themselves interpreters exercising linguistic judgment can also generate insight, as their judgment adds a potential source of information for those who come after them. So far we have seen how analyzing the discrepancies between text and audio that resulted from the Bukiyip transcriber's choices

allows us to correct the timeline of language shift and more fully appreciate what the speaker was trying to express. In what follows I will show that the transcriber's approach to entextualization also provides insights into structural features of Arapesh at a time of rapid language change.

The first case I will discuss involves the grammaticalization of the particle *ta ~ tə ~ ətə* that appears in both Cemaun (Coastal) Arapesh and Bukiyip, often but not always in future contexts. The meaning of this particle is difficult to pin down. It is mentioned neither in Fortune (1942) nor in Conrad & Wogiga (1991), the two published Arapesh grammars, though *ta* does receive treatment in Conrad & Simatab's (1998) unpublished manuscript *Arapesh and Bukiyip (Mountain Arapesh) comparative grammar notes*, where it is said to indicate 'strong future certainty'. Example (7) illustrates the use of *ətə* in a Cemaun Arapesh narrative that I recorded in 1998.

(7) <u>LISTEN</u> (Dog.that.became.a.girl_1.42–1.48)

ei   uwe   i-nak
1.SG.PRO   NEG   1.SG.IR-go
'Me, no way [am I staying], I'm going;

ətə   i-pwe-um   ipə   p-i-nə   p-i-gək
?   1.SG.IR-stay-COND   2.PL.PRO   2.PL.IR-go   2.PL.IR-die
if I stay you guys will go off and die

bai   ino   inap   i-tik
FUT(TP)   NEG(TP)   be.able(TP)   1.SG.IR-see
and I won't be able to see [what happened to you].'

The particle never appears in the Bukiyip texts, which is undoubtedly why it is not discussed in Conrad & Wogiga's grammar; after all, the grammar is based on the texts. But reading the texts against the recordings it becomes clear that the particle does in fact occur, indeed it occurs frequently, but it is systematically transcribed wherever it occurs as the temporal adverb *wotak*, glossed 'more, not yet, still' in Conrad & Wogiga (1991: 41). Such completion by native speaker transcribers of utterances that are felt to be "partial", like an adverb undergoing grammaticalization to a particle, is another well-documented feature of entextualized speech. As Urban (1996: 29) describes it, "the spoken image [is reshaped] into something resembling the spoken form produced in the relatively more formal, self-conscious context of elicitation", thus making "semantically explicit" the transcribers' pragmatic inferences. In other words, the transcriber's "improvement" here provides insight into a native speaker's metalinguistic knowledge about how forms are related, something one might assume is only activated during elicition of grammaticality judgments or in other interview-like settings.

Example (8) below is from Bukiyip text RF, an instance of Conrad's genre "myths and folktales". The first line, in all caps, shows how the utterance is originally transcribed. The line below it is our transcription based on listening to the recording. Wherever the speaker said *ətə*, the transcriber wrote *wotak*.

(8) LISTEN (RF_3.48–3.52)

N-A-KLI	ƏN-DAK	**WOTAK**	N-A-NAK-I
n-a-kli	ən-dak	**ətə**	n-a-nak-i
M.SG-R-say	M.SG.1-DEM	**?**	M.SG-R-go-CTPT

'He said, 'this one' and then he came.'

I have often heard contemporary Arapesh speakers use *wotak* with the meaning 'still, not yet', or 'a little while later'. But never have I had anyone associate the particle with the adverb. I infer that this is because the particle has by now become so semantically bleached, or conversely because the meaning of *wotak* has so narrowed, that the connection is no longer salient. Either way, the language has changed since the transcriber produced the text, and the particle now seems to be fully grammaticalized. In this it contrasts with some other particles that fluent speakers I worked with in the late 1990s would still readily associate with a full lexical form. An example is *bo* 'already, before', which speakers understood to be a variant of *nimbo*, itself a reduced form of *nimbotik* 'yesterday' (p.113).[11] In any event, reading the transcript against the recordings strongly suggests that *wotak* (whatever its precise meaning was at the time) is the lexical source of the contemporary grammatical element *ta ~ tə ~ ətə*. This is not something I ever would have hypothesized – much less been able to provide evidence for – based on interviewing Arapesh speakers, reading only the texts, or listening only to the audio. But because of the discrepancy, analyzing the transcriber-interpreted texts in conjunction with the recordings provides access to knowledge that has since been lost, enabling us to make the connection.

Another area where the discrepancy between text and audio reveals a linguistic pattern that neither source could on its own is agreement. Arapesh languages have a typologically remarkable tendency to categorize nouns according to their final phonological elements, prototypically consonant phonemes, and to express syntactic agreement using markers that alliterate

---

[11] Here and elsewhere data citations in the format (letter.number) refer to my field notebooks by (notebook.page); see http://www.arapesh.org/field_notebooks.php.

with those final phonological elements (Dobrin 2012a).[12] As illustrated in
(9a–c) with examples from a Cemaun text spoken by Arnold Watiem, the
default noun class agreement marker echoes the controller's final consonant in
a highly concrete way. The agreement markers and noun-final sounds
determining agreement are both marked in bold in the examples; the
controllers are also underlined.

(9a) ə**b**-idə   yowei-**b**-i   wa**b**
      I.SG-DEM   bad-**I.SG**-MOD   night(**I.SG**)
      'this bad night' (g.1)

(9b) bora**n**   e**ɲ**-idək   eti-**ɲ**
      talk(**VIII.SG**)   **VIII.SG**-DEM   only-**VIII.SG**
      'This is all [I have] to say.' (= 'The end.') (g.2)

(9c) mare-**g**   api**g**   **g**-a-pwe
      what-**III.SG**   leafy.greens(**III.SG**)   **III.SG**-R-exist

      ga   nimbarig   ba   u-rək-ə**g**
      LOC   garden   FUT   F.PL.IR-get-**III.SG**
      'The women will get whatever greens are there in their gardens.' (g.5)

Arapesh has an optional alternation between the segments *k* and *p* in word-
or morpheme-final position. Consider for example the Cemaun verb root
yabik 'to show'. Its default realization is with final [k], as in *c-e-yabik-əp*
VIII.PL-R-show-1.PL 'they showed us' (p.175). But it can alternatively be
realized with final [p]: *h-i-yabipw-ec* M.PL-IR-show-VIII.PL '[when] they(M)
show them' (g.110).[13] The alternation is bidirectional, with final underlying *p*
also realized sometimes as [k]. There are two sources of evidence that the
Bukiyip noun *bolup* 'mountain' underlyingly ends with *p*. It takes an
allomorph of the *s* plural, which is standard across the Arapesh family for *p*-
final (but not *k*-final) singulars, and it occurs with final *p* in the two other
varieties for which it is attested: Cemaun *borup* ~ *borugwis*, Balam *balup* ~
balugwis. In Conrad's mythic text RC, spoken by Lumumbuli of Bonohwitam
village, we find *bolup* transcribed with *k* and pronounced with final glottal

---

[12] There is one exception: in Weri, agreement with non-human nouns is sensitive not to
class but only to singular/plural number. As in all Arapesh languages, agreement with
human nouns in Weri follows the controller noun's semantics.

[13] In Cemaun the consonant *p* is typically pronounced with a labial offglide when a
front vowel follows. That the underlying segment in *yabik* is *k* rather than *p* is further
suggested by the verb's partial identity with non-alternating *bik*, which also has the
meaning 'show'.

stop [boluʔ], the most common reflex of *k* in this speaker's dialect (the segment also sometimes deletes, as can be seen in forms like *yek* [ye] 'I' and *nənak* [nəna] 'he went' in example (12) below). Because a phonological contrast between noun-final *p* and *k* implies a difference in noun classification, the selection of the *k*-alternant might be expected to trigger a shift in agreement. And this is in fact what happens: the noun takes agreement with *k*, realized phonetically in this dialect as glottal stop (10).

(10) LISTEN (RC_0.29–0.34)

Underlying *bolup* realized as [boluʔ] 'mountain, island':

KW-Ə-NAK   OKWOK-I-**K**   BOLU**K**
kw-ə-naʔ   oʔwoʔ-i-**ʔ**   bolu**ʔ**
F.SG-R-go   F.SG.PRO-POSS-**ʔ** mountain
'she went to her mountain'

But at what level of representation does the *k ~ p* alternation occur? Is *bolup* reclassified as a *k*-final noun in the lexicon as a result of having an alternate listed form belonging to a different noun class, or is the shift just a matter of surface realization? The question is worth asking because it has implications for the nature of agreement. If the noun's shift of final sound is a superficial phonetic phenomenon rather than a lexical one, it means that the agreement marker is responding not just to an abstract lexical class feature but to the noun's concrete phonetic form. The *bolup* example does not provide enough evidence for us to answer. But another example where the transcript departs from the recording does.

In line with the *k ~ p* alternation, the Tok Pisin borrowing *motobaik* 'motorcycle' can be realized in Bukiyp as *motobaip*, with final [p]. Although transcribed in the texts as *MOTO BAIK*, the word is pronounced with final [p] throughout narrative text NB, spoken by Kepas Wogiga. In this text, Kepas tells about a day trip he and Conrad took together on a motorcycle to the nearby town of Yangoru. Although in Kepas's dialect *k* is most often realized as [ʔ] or deleted, he does have the phone [k] in his repertoire, even in word-final position, as evidenced by his articulation of the place name *Kumərik* [kumərik] and at least one instance of the highly frequent verb 'to go' *nak* ([cenak] 'they went'). A final [k] can be clearly heard in the Tok Pisin borrowing *trak* [trak] 'truck' at multiple points in the same text (11a–c).

(11a) LISTEN (NB_10.43–10.45)

*yelech monak tra**k***
'I brought them and we went to the truck.'

(11b) LISTEN (NB_10.24–10.30)

*nakli nye nyiyil Wanguenomwi pinak tra**k***
'He said, 'you come along [lit. 'hang on'] with Wangiwen and those folks and go to the truck.''

(11c) LISTEN (NB_8.26–8.32)

*ənən okwok chənaki tra**k***
'The man and the woman came to the truck.'

When we listen to the audio recording of NB, we hear that most tokens of *k* are realized either as [ʔ] or Ø. But *motobaik* is pronounced with final [p] in (12), and then in the next clause we find the verb *bik* 'put' expressing agreement with the noun by prefixing with *p*.[14]

(12) LISTEN (NB_15.32–15.42)

DOU	YEK	Y-E-YATU	ƏNAN	N-Ə-LAU
dou	ye	y-e-yatu	ənan	n-ə-lau
now	1.SG.PRO	1.SG-R-stand	M.SG.PRO	M.SG-R-get

MOTO BAI**K**	N-Ə-NAK	WILARU
motobai**[p]**	n-ə-na	Wilaru
motorcycle	M.SG-R-go	Wilaru

'So I waited while he got the motorcycle and went to Wilaru [mission].'

N-Ə-NAK	N-A-**Pɪ**-BɪK-ɪK	N-Ə-NAK-I
n-ə-na	n-a-**pi**-biʔ-iʔ	n-ə-naʔ-i
M.SG-R-go	M.SG-R-**ʔ**-put-COMPL	M.SG-R-go-CTPT

He went and parked it and came,

Y-Ə-HƏL	ECHƏ-DAK	Y-A-HW-ECH	W-Ə-NAK-I
y-ə-həl	echə-daʔ	y-a-suh	w-ə-naʔ-i
1.SG-R-lift	VIII.PL-DEM	1.SG-R-hold	1.DU-R-go-CTPT

and I picked up these things [we got] and held [them] and we two came.'

The fact that the Bukiyip transcriber systematically wrote *motobaik* with a final *k* throughout this text suggests that he recognized the [p] sound he was

---

[14] The glossing in (12) follows the transcription we made based on listening to the recording. It departs slightly from Conrad's text, though not in ways that bear on the present argument.

hearing to be a low-level surface phenomenon, not part of the noun's underlying lexical form. If the role of the noun-final consonant in Bukiyip is to assign nouns to classes in the lexicon, mere vagaries of pronunciation such as this should have no bearing on a noun's syntactic behavior. And yet here, precisely because of the discrepancy between transcript and recording, we can show that the vagaries of pronunciation *do* affect syntactic behavior: when pronounced with the final phone [p], *motobaik* is treated in agreement as a *p*-final noun. In other words, the transcriber's decision to represent what he heard by the noun's lexically underlying sound *k* rather than its surface manifestation [p] makes it possible to demonstrate a typologically unusual syntactic pattern that can no longer be reliably documented given the current advanced stage of language shift: agreement not just with a word-final phonological element, but with its concrete phonetic form.

6. Insights into the social production of the materials

The last point I want to make about reading Conrad's Bukiyip texts against the original audio recordings is that doing so provides insights into the interactional context surrounding the materials' production. When we document a language, we transform human events (elicitation interviews, recording sessions) into archivable objects (audio and video files, transcripts, fieldnotes). In that process some information is always lost: even when recordings are (unlike the Bukiyip recordings) associated with good metadata, not every observation about the context that might be relevant to their interpretation can possibly be included with the materials. At the same time, as we have just seen, new and potentially useful information is introduced in the act of transcription. So considering the two sources, text and recording, in relation to one another gives secondary analysts a helpful triangulation point from which to view what the participants were doing – and how they were thinking about what they were doing – as the original documentary events were undertaken.

I will illustrate this with Bukiyip text RD, a narrative of Conrad's genre "myths and folktales". It is a difficult story with thematic components that appear in variant forms throughout the Arapesh region. The story is about Yeldə, a *masalai* or bush spirit who is associated with the speaker's clan lands and who is the hero of their men's cult. In my discussions with Conrad over video call in ca. 2013 as we were collaboratively glossing the transcribed text, Conrad described Yeldə as a "lusty, magical character". Lusty is an understatement. In the story, Yeldə targets and repeatedly rapes a young woman before drowning and dismembering her in the river pool where he lives. He then takes up residence in a spirit house in the villagers' midst and terrorizes them by grabbing their children and eating them. When the villagers try in vain to destroy him by cutting him up and setting him on fire, he climbs

up the spirit house's central post, removes it, and carries it off to the ocean. He is a violent, malevolent character and exceedingly powerful.[15]

As is typical in Melanesian storytelling (and Melanesian speech more generally), the speaker offers few clues about what the story means or why it is being told. Instead, it is incumbent on listeners to make sense of what they hear. To a person accustomed to this, it is therefore unsurprising that the transcript begins simply with, "I am Ibara. I am going to tell a story from long ago. Four women went to wash (in a river)." The action continues from there without commentary until the text ends with, "He came down to the ocean. The talk that I, Ibara, have spoken is finished. As for this (talk), I have finished it."

Conrad provides some metadata in a lead at the start of the recording, but it is not transcribed, and listening to it raises more questions about the meaning of the text than it answers. Conrad introduces what is to follow by saying in English, "this is a story about *yahalok*, Tok Pisin *ton*, told by Ibara of Bubuamo [village], December 8th, 1970". A *yahalok* is a wild lychee tree, and it does feature in the story insofar as Yeldə throws *yahalok* fruits at the young woman to get her attention at the river, getting the narrative action underway. But Ibara's story could hardly be said to be "about a *yahalok*". Why would Conrad have described the story in this strange oblique way? Perhaps he did not fully realize what Ibara was about to say for his recorder. Or perhaps he was avoiding naming – or was even actively deemphasizing – the violence and sexuality that he knew would be imminent in Ibara's telling. He was, after all, a Christian missionary whose overriding purpose in being there was to spread the gospel among Arapesh-speaking people. Celebrating Yeldə as a hero was not exactly consistent with his perspective.

Evidence from the recording supports the second interpretation. The speech on this six-minute-long narration sounds bizarre: each word constitutes its own intonational phrase. Some examples are given in (13a–b).

(13a) LISTEN (RD_0.21–0.28)

yek	Ibara	y-a-kli	i-yaguleh	bolaɲ	eiwak-i-ɲ
1.SG.PRO	Ibara	1.SG-R-want	1.SG.IR-say	speech	long.ago-POSS-VIII.SG

'I am Ibara. I want to tell a story from long ago.'

---

[15] In summarizing this text Conrad suggested that Yeldə's last act in the story was to create the ocean. I recognize this motif from stories told by neighboring Arapesh people, but nowhere on the recordings does Ibara actually say this.

(13b) <u>LISTEN</u> (RD_1.44–1.51)

wab	Yəldɨ	n-a-kih-i	n-a-wich
night	Yeldə	M.SG-R-go.up-CTPT	F.SG-R-enter

'At night Yeldə came up and entered (her house).'

It sounds as if Ibara is reading a written text aloud one word at a time. This suggests that he had become literate enough in Bukiyip to at least read, and perhaps even to have himself written down, this ancestral story. This is consistent with the standard SIL literacy training practice of having learners write down their own well-known stories (see, e.g., Weber et al. 2007). If that is the case, the "oral discourse" documented here is of a very unusual kind: text RD would not have been spoken, recorded, and then transcribed, but rather written down, read aloud, recorded, and then re-inscribed.

It would also mean that by the time the recording was made, the narrative was already an object of focus and reflection for Conrad in his work with Ibara; indeed, the two would most likely have collaborated on producing the written version of the text just as they were now collaborating on recording Ibara's reading of it. When Conrad and I were glossing the transcript of RD, even before we had recovered the audio, he described Ibara as a good friend and senior "white magic leader" who "humbled himself" by attending the Conrads' literacy school and who later sent his son to learn from Conrad and his wife how to read, write, and understand "the Good News". Given this, it is hard to imagine that Conrad had been shielded from the story's unsavory content until the moment Ibara began speaking.

There is another intriguing feature of the recording that is revealed by comparison with the transcript. When Ibara's Bukiyip narration ends, the recording continues with a Tok Pisin translation provided by Ibara that was not transcribed. Conrad was assembling a collection of Bukiyip texts as a database for grammatical analysis, so the Tok Pisin translation would have been produced not as data but as an informal resource for helping him understand the Bukiyip. Ibara's speech in the Tok Pisin translation sounds natural, i.e., not like he is reading but rather like he is speaking (14).

(14) <u>LISTEN</u> (RD_TokPisin_8.11–8.24)

Orait, ol ipilai istap moa, kam daun, kisim wanpela.
'And while they were playing, (Yeldə) came down (from the spirit house) and took (another child).

Go kaikai moa, na em ilukim ol nau, ino istap.
He ate it, and when (the grandmother looking after the children) looked for them, she saw they weren't there.

Em tok, "Tupela man we?" Em itok, "Mipela ino save."
She said, "Where are the two boys?" and (someone) said, "We don't know".'

I have three interrelated observations to make about this segment of the recording. The first is that in his Tok Pisin telling Ibara includes a number of details that are missing from the vernacular version. For example, he mentions that Yeldə takes the form of a python. He describes how the rapes take place in graphic detail,[16] and he notes that Yeldə went to a village called Dogur when he ran away to the coast. (Creating a context for speakers to perform their authoritative knowledge of proper names like this is one important function of "traditional" Melanesian storytelling.) The existence of these discrepancies between the two versions in itself calls for explanation. When Arapesh speakers provide a Tok Pisin translation for a story they just told in the vernacular, I have often been impressed by how closely it tracks with their original speech, to the extent that the speaker may even stumble at the same point in each version. I have found this to be true whether the story is canonical or a spontaneous original, and even when the speaker was not expecting to provide a translation. The fact that this translation includes many details that are absent from the original again makes more sense if the Bukiyip version of Yeldə was read aloud rather than told, because in that case the content would not have been in Ibara's active memory when he was telling it in Tok Pisin.

Second, Ibara's Tok Pisin translation reveals that an audience was present when the recording event took place. One or more other voices are detectable in the background on the recording of the Bukiyip version, but once Ibara commences his Tok Pisin telling these voices come alive. They actively participate in the negotiation that takes place when Conrad asks Ibara to translate, and they continue to chime in throughout the storytelling that follows, shushing children, discussing among themselves what the next step in the action should be, and prompting Ibara to fill in details that he forgot to include. From this we learn that reading as opposed to speaking resulted not only in elision of important information, it also transformed what would otherwise have been a dialogic interaction into a monologic one, i.e., a public reading event in which others could not or would not participate.[17]

Finally, the Tok Pisin version concludes with a surprising coda that brings the story out of narrative time and into the moment of speaking. In this coda,

---

[16] Here is an example: *Ikam long arere long haus. Meri silip indai pinis, em igo troimwe kok go insait long kan bilong em.* '(Yeldə) came up alongside the house. When the woman was fast asleep, he went and shoved his cock into her cunt.'

[17] The erasure of such "conversational back-channel" was similarly observed by Haviland (1996: 63) in the conversion of Tzotzil oral stories to written text.

Ibara avers that as descendants of Yeldə, he and his clans-people – and he is at pains to say he means all of them, including *susa, meri, kandere* 'sisters, women, aunties/nieces' – are *man bilong puspus nabaut* 'people who go around fucking'. There is no mistaking what he says, and one can even hear a woman in the background trying to contain nervous laughter as he speaks. Maybe Ibara was attempting to say how much he and his clans-people were in need of the moral transformation Conrad had come to provide? Maybe he was posturing in the presence of his fellow clansmen? Maybe he was making an advance apology to the missionaries that their efforts were unlikely to have their intended effects? We can only speculate. But these comments leave no doubt that in animating the Yeldə story for Conrad's recorder and then elaborating on it as he did when he retold it in Tok Pisin, Ibara was doing something other than humbly assisting Conrad with Bible translation tasks by offering him an anodyne example of vernacular discourse. Although it is not easy to interpret what he was doing or why, he was clearly making some kind of move that, from the point of view of those listening, was a meaningful – and likely challenging or subversive – comment in relation to Conrad and his project.

Having access to the original audio recording of text RD gives us insight into the work practices and social relations that gave rise to Conrad's Bukiyip texts. From the recording of the story itself we infer that the text was written down and read aloud rather than spontaneously narrated. From Conrad's fifteen-second lead before Ibara begins to speak we detect him sidestepping the challenge Ibara's story presented to the American Christian values he was aiming to impart through his missionary work. From the ten minutes of talk that follow Ibara's reading we discover that multiple Arapesh people were present when the recording was made, and we learn details that were left out of the original narrative but that other participants in the storytelling event considered important. At the same time, studying text RD in relation to the recording reveals how a seemingly straightforward monologic text of the genre "myths and folktales" was distilled from an at times awkward encounter between Conrad, with the Bible translation and literacy promotion agendas that motivated his text collecting project, and Arapesh people like Ibara, with their own interests, goals, and sources of power. Listening to the recording against the text expands our understanding of what kinds of "context" might be relevant to the interpretation of the text, including what took place before the transcribed portion of talk started, and what took place after it stopped.

7. Conclusion: The interpretive value of reproducible research

The Arapesh suitcase miracle makes it possible for us to "check" Conrad's Bukiyip texts against the original recordings. But what we learn from doing this is so much more complex and interesting than that the native transcriber

did or did not accurately capture what was said, or that Conrad's claims about Bukiyip grammar are or are not correct. We learn that the reach of Tok Pisin was much further advanced in the inland Arapesh villages in the early 1970s than we otherwise would have imagined. We learn that although the collected texts were meant to represent "Bukiyip oral narrative discourse" (Conrad 1987), they depart from Bukiyip oral discourse patterns in important respects, reshaping their rhythms and thereby partially obscuring the speakers' expressive intent. It is true that some elements of the recorded speech are inaccurately transcribed, but even these inaccurate transcriptions add value to the recordings because they reflect native speaker judgments. The inaccuracies make linguistic features discernable that otherwise would not be from either text or recording alone, such as the lexical sources of now grammaticalized particles and a remarkable syntactic agreement rule that refers not to a noun's morphosyntactic specification but to its concrete phonetic form. Finally, by reading the texts against the recordings we come away with a richer understanding of the interpersonal setting in which the texts were produced. We see some of the research methods Conrad adopted, such as having speakers dictate or write a story and then read it aloud, or having speakers follow up a Bukiyip narration with a translation in Tok Pisin. And we gain insight into the ways that Conrad and his Bukiyip collaborators interacted across a major cultural divide.

This exercise in studying the transcripts against the original documentation shows how the work of engaging with archival materials depends upon a series of interpretations, with no step further back in the chain being definitive. Grammatical claims are interpretations. Accessing the texts that underlie them tells us more about how those interpretations were made, and the audio more still, but only through further acts of interpretation. If we had video of Ibara speaking we might be able to infer something from his facial expression when he offered his lascivious coda. If we had a 360-degree camera or multiple cameras we might infer something by noticing who his gaze shifted to at the moment he announced that he and his clanspeople were *man bilong puspus nabaut* 'people who go around fucking'. But there is no point at which the material becomes self-explanatory, so that by "checking" it we can now know that our interpretations are correct and complete. The same limitation besets metadata. It is by all means helpful to have some basic information about who Ibara was, e.g., "a male village leader in his 50s who knew white magic", or where the recording was made, e.g., "Bubuamo village". But that is only the tip of an iceberg. *Who Ibara was* in sharing the Yeldə story with Conrad on December 8th, 1970, is something we begin to understand more deeply by listening to this recording, and the perspective we gain from it on *who Conrad was* bears on how we interpret all the rest of the material in the collection. Being able to talk through the text (and later the recordings) with Conrad was enormously helpful, but what he told me was no more definitive than the material itself, as his comments too require interpre-

tation (e.g., how exactly does he understand the concept of "white magic"?). And this is not only because his memory was fading, or because he was a missionary linguist, or even because he was someone other than myself: I have Arapesh texts I recorded myself that I could still not begin to make sense of until many years later, when I had gained enough perspective to interpret what the speaker was trying to say (Dobrin 2012b).

The call for reproducibility implies that documentation-plus-metadata will make archival materials transparent, so that future users will be able to check a linguistic description against them and judge whether or not that description was right. I sympathize with this perspective, up to a point. After all, I have taken pains throughout this exposition to provide links to the audio for key examples so that readers can verify my statements for themselves and become convinced of my analysis. But having links to the audio does not straight-forwardly confirm that my analyses are right. It lets readers join me in the unending project of interpreting my predecessors' voiced and inscribed interpretations, just as it allows readers to interpret mine. As I hope to have shown, studying the Bukiyip transcriber's interpretations against the original documentation raises questions that go far beyond the realm of (dis)confirmation. Working with archival materials in the human sciences is more akin to historiography than it is to positivist science. The best we can aim for is insightful syntheses that are veridical to our sources: that conform to the materials, that do not contradict them, and that "represent reality in some way which is generally accepted as satisfying, coherent, and meaning-ful" (Humphreys 1980: 3). Despite what we might wish, there is no depth of capture, no level of annotation detail, no amount or type of metadata that will allow documentary linguistic materials to speak for themselves. We are engaged in an irreducibly interpretive process, but one that can nevertheless produce valuable insight into language structure and communicative context alike.

References

Alungum, John, Robert J. Conrad, & Joshua Lukas. 1978. Some Muhiang grammatical notes. *Workpapers in Papua New Guinea Languages 25: Miscellaneous papers on Dobu and Arapesh*, 89–130. Ukarumpa: Summer Institute of Linguistics.

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice, & Anthony C. Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1). 1–18.

Blommaert, Jan. 2006. Ethnopoetics as functional reconstruction: Dell Hymes' narrative view of the world (Working Papers in Urban Language & Literacies Paper 32). London: The Centre for Language, Discourse, and Communication, King's College London.

Camp, Cheryl, R.S.M. 1983. The Peli Association and the New Apostolic Church. In Wendy Flannery (ed.), *Religious movements in Melanesia today* vol. 1, 78–93 (Point Series No. 2). Goroka: The Melanesian Institute.

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*, 57–55. Ukarumpa: Summer Institute of Linguistics.

Conrad, R[obert]. J. 1987. Kinds of information in Bukiyip oral narrative discourse. *Language and Linguistics in Melanesia* 16. 23–40.

Conrad, Bob [Robert J.] & Aluis Simatab. 1998. Arapesh and Bukiyip (Mountain Arapesh) comparative grammar notes. SIL [Summer Institute of Linguistics] and BTA [Bible Translators Association]. (Unpublished manuscript.)

Conrad, Robert J. & Kepas Wogiga. 1991. *An outline of Bukiyip grammar* (Pacific Linguistics C-113). Canberra: Australian National University.

de Vries, Lourens. 2005. Toward a typology of tail-head linkage in Papuan languages. *Studies in Language* 29(2). 363–384.

Dobrin, Lise M. 2012a. *Concreteness in grammar: The noun class systems of the Arapesh languages.* Stanford: CSLI Publications.

Dobrin, Lise M. 2012b. Ethnopoetic analysis as a methodological resource for endangered language linguistics: The social production of an Arapesh text. *Anthropological Linguistics* 54(1). 1–32.

Fortune, R. F. 1942. *Arapesh* (Publications of the American Ethnological Society, Vol. 19). New York: J. J. Augustin.

Gawne, Lauren, Barbara F. Kelly, Andrea L. Berez-Kroeker, & Tyler Heston. 2017. Putting practice into words: The state of data and methods transparency in grammatical descriptions. *Language Documentation & Conservation* 11. 157–189.

Gawne, Lauren & Andrea L. Berez-Kroeker. 2018. Reflections on reproducible research. In Bradley McDonnell, Andrea L. Berez-Kroeker, & Gary Holton (eds.), *Reflections on language documentation 20 years after Himmelmann 1998*, 22–32 (Language Documentation & Conservation Special Publication 15). Honolulu: University of Hawai'i Press.

Hale, Kenneth, Colette Craig, Nora England, Laverne Jeanne, Michael Krauss, Lucille Watamahogie, & Akira Yamamoto. 1992. Endangered languages. *Language* 68. 1–42.

Haviland, John B. 1996. Text from talk in Tzotzil. In Michael Silverstein & Greg Urban (eds.), *Natural histories of discourse*, 45–78. Chicago: University of Chicago Press.

Henke, Ryan & Andrea L. Berez-Kroeker. 2016. A brief history of archiving in language documentation, with an annotated bibliography. *Language Documentation & Conservation* 10. 411–457.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36. 161–195.

Himmelmann, Nikolaus P. 2018. Meeting the transcription challenge. In Bradley McDonnell, Andrea L. Berez-Kroeker, & Gary Holton (eds.), *Reflections on language documentation 20 years after Himmelmann 1998*, 33–40 (Language Documentation & Conservation Special Publication 15). Honolulu: University of Hawai'i Press.

Humphreys, R. Stephen. 1980. The historian, his documents, and the elementary modes of historical thought. *History and Theory* 19(1). 1–20.

Hymes, Dell. 1981. *"In vain I tried to tell you": Essays in Native American ethnopoetics.* Philadelphia: University of Pennsylvania Press.

Hymes, Dell. 1996. *Ethnography, linguistics, narrative inequality: Toward an understanding of voice*. London: Taylor and Francis.

Juagu, Junny & Bob [Robert J.] Conrad. 2008. *Bukiyip dictionary*. Ukarumpa: Summer Institute of Linguistics.

Jung, Dagmar & Nikolaus P. Himmelmann. 2011. Retelling data: Working on transcription. In Geoffrey Haig, Nicole Nau, Stefan Schnell, & Claudia Wegener (eds.), *Documenting endangered languages: Achievements and perspectives,* 202–220. Berlin: Mouton De Gruyter.

Kroskrity, Paul V. 2015. Discursive discriminations in the representation of Western Mono and Yokuts stories: Confronting narrative inequality and listening to Indigenous voices in Central California, 135–163. In Paul V. Kroskrity & Anthony K. Webster (eds.), *The legacy of Dell Hymes: Ethnopoetics, narrative inequality, and voice.* Bloomington: Indiana University Press.

Kulick, Don. 1992. *Language shift and cultural reproduction.* Cambridge: Cambridge University Press.

Linguistic Society of America. 2010. Resolution recognizing the scholarly merit of language documentation. https://www.linguisticsociety.org/resource/resolution-recognizing-scholarly-merit-language-documentation (accessed 2021-12-30).

Marten, Lutz & Malin Petzell. 2016. Linguistic variation and the dynamics of language documentation: Editing in 'pure' Kagulu. In Mandana Seyfeddinipur (ed.), *African language documentation: New data, methods and approaches*, 105–129 (Language Documentation & Conservation Special Publication 10). Honolulu: University of Hawai'i Press.

May, R. J. 1982. The view from Hurun: The Peli Association of the East Sepik district. In May, R. J. (ed.), *Micronationalist movements in Papua New Guinea*, 31–62. Canberra: Australian National University.

Mosel, Ulrike. 2012. Creating educational materials in language documentation projects: Creating innovative resources for linguistic research. In Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts, & Paul Trilsbeek (eds.), *Potentials of language documentation: Methods, analyses, and utilization*, 111–117 (Language Documentation & Conservation Special Publication 3). Honolulu: University of Hawai'i Press.

Mosel, Ulrike. 2015. Putting oral narratives into writing: Experiences from a language documentation project in Bougainville, Papua New Guinea. In Bernard Comrie & Lucía Golluscio (eds.), *Language contact & documentation (Contacto lingüístico y documentación)*, 321–342. Berlin: De Gruyter Mouton.

Roscoe, Paul B. 2004. The evolution of revitalization movements among the Boiken, Papua New Guinea. In Michael E. Harkin (ed.), *Reassessing revitalization movements: Perspectives from North America and the Pacific Islands*, 176–180. Lincoln: University of Nebraska Press.

Ross, Malcolm. 2001. Contact-induced change in Oceanic languages in north-west Melanesia. In Alexandra Aikenvald & R.M.W. Dixon (eds.), *Areal diffusion and genetic inheritance: Problems in comparative linguistics*, 134–166. Oxford: Oxford University Press.

Silverstein, Michael & Greg Urban. 1996. The natural history of discourse. In Michael Silverstein & Greg Urban (eds.), *Natural histories of discourse*, 1–17. Chicago: University of Chicago Press.

Thieberger, Nicholas. 2014. Digital humanities and language documentation. In Lauren Gawne & Jill Vaughan (eds.), *Selected papers from the 44th conference of the Australian Linguistic Society*, *2013,* 144–159. Melbourne: University of Melbourne.

Urban, Greg. 1996. Entextualization, replication, and power. In Michael Silverstein & Greg Urban (eds.), *Natural histories of discourse*, 21–44. Chicago: University of Chicago Press.

Weber, Diana Dahlin, Diane Wroge, & Joan Bomberger Yoder. 2007. Writers' workshops: A strategy for developing indigenous writers. *Language Documentation & Conservation* 1(1). 77–93.

Woodbury, Anthony C. 2011. Language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 159–211. Cambridge: Cambridge University Press.