Approaches to digitization and annotation: A survey of language documentation materials in the Alaska Native Language Center Archive

Gary Holton

University of Alaska Fairbanks

1. Introduction

The design of best practices for digitizing and annotating texts and field recordings should be informed by knowledge of the types of texts and recordings which are produced by documentary linguists. One way to gain insight into this typology is to examine materials in existing language archives. This paper describes the structure of existing documentary texts and recordings, based on an informal survey of items in the Alaska Native Language Center (ANLC) archive. Knowledge of data in existing language archives is important in at least two respects. First, any markup or encoding schemes will need to be robust enough to handle existing data, so prior knowledge of some of the legacy data will assist with the eventual encoding of those data. Second, and perhaps more important, knowledge of existing data provides indirect but empirically valid insight into the way in which linguists approach language documentation. Theorizing about the structure of language documentation materials may unwittingly lead us to examine idealized models of language documentation. An example of such a model is the oft-cited three-line gloss, a text-encoding format which is in practice much more diverse than might at first appear. Examining existing data permits the development of best practice to be grounded in what field linguists actually do, rather than what we think they do.

The ANLC archive contains approximately 10,000 paper documents and 5,000 recordings comprising nearly everything written in or about Alaska Native languages (cf. Krauss & McGary 1980). The archive also contains substantial holdings of materials on related languages spoken outside Alaska. Admittedly, the archive still lacks geographic breadth, in the sense that it does not represent a typologically broad sample of the world's languages. The majority of Alaska's Native languages fall into one of two families: Eskimo-Aleut and Athabaskan-Eyak-Tlingit. However, the time depth of the materials and the comprehensive nature of the coverage ensure that the archive is representative of a broad range of linguistic traditions and field worker styles. Materials in the archive span the entire period of the development of modern linguistics, including scholarly traditions from Europe, Asia and the Americas.

The description presented here is in no way intended to be statistically representative. Rather, materials have been chosen in an ad-hoc manner in order convey an admittedly subjective impression of the diversity of text and audio recordings contained in the ANLC archive. That said, the examples presented here are also not outliers; that is, these data are typical of the types of materials encountered in the

archive. It is perhaps not too surprising that the unpublished documentary materials produced by field linguists are less standard and uniform than the descriptive products of linguistic analysis. The range of existing documentary materials argues for flexibility in the development of best practice.

In the next section I describe the annotated text resources in the ANLC archive, providing some examples of annotations. In section 3 I very briefly describe a constraint on digitizing audio recordings. Finally, section 4 concludes with some recommendations for the development of best practice.

2. Texts

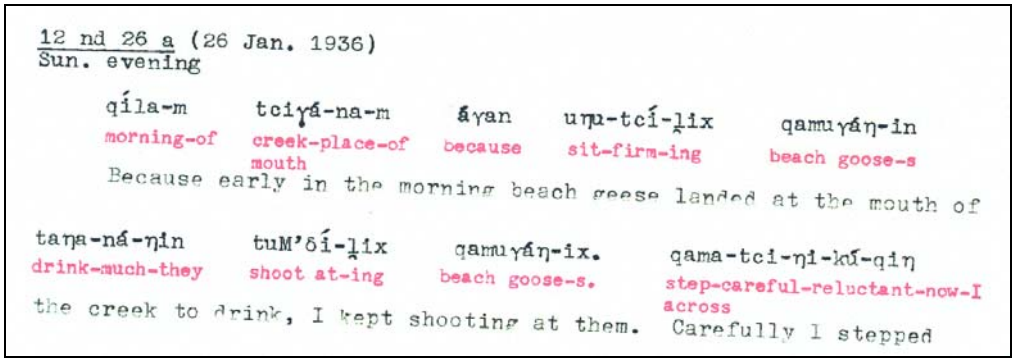
Annotated texts are found in abundance in the ANLC archive. Most represent transcriptions from original audio recordings. In some cases the physical medium on which the recording was made can be identified via references on the transcription, though this is often not the case. Annotated texts are archived in paper format, either handwritten, typed or printed. While some texts were evidently originally transcribed electronically, only the paper printouts of these texts are archived at ANLC.

2.1 Typology of annotation

Texts in the ANLC archive exhibit many types of annotation, which may be broadly classed as glosses, comments, uncertainty, and editorial annotations. These types are distinguished on an ad hoc basis and thus do not form a true typology; however, they do provide a useful entry point to the archive data. Examples of each of these four types are discussed below.

2.1.1 Glosses

The prototypical annotated text consists of the “interlinear” or “three-line” gloss representing annotation at the morpheme, word and sentence/paragraph level. This format is the staple of documentary linguistics, and most linguists are familiar with this theme. The purpose of the gloss is to provide source language equivalents for target language morphemes, words, or phrases. This type of annotation is also well-represented in the ANLC archive; an example is given below.


12 nd 26 a (26 Jan. 1936)
Sun. evening

qíla-m — morning-of
tciɣá-na-m — creek-place-of mouth
ǎɣan — because
uɱu-tcí-ɮix — sit-firm-ing
qamuɣáŋ-in — beach goose-s

Because early in the morning beach geese landed at the mouth of

taŋa-ná-ŋin — drink-much-they
tuM'δí-ɮix — shoot at-ing
qamuɣáŋ-ix. — beach goose-s.
qama-tci-ŋi-kʊ́-qiŋ — step-careful-reluctant-now-I across

the creek to drink, I kept shooting at them. Carefully I stepped

Figure 1: Typical three-line annotation (Ermeloff 1937)

The first line represents the original text, broken into morphemes using hyphens. The second line (in red ink) represents an English gloss of each morpheme. The third line contains an English free translation of the entire line of the original text. In this case there are two free translations, one corresponding to each of the two lines of original text in this text fragment. There are of course many variations on this theme, corresponding to different levels of representation.

2.1.2 Comments

A surprisingly common form of annotation found in the ANLC archive is the general comment. Comments may occur at any level (morpheme, word, phrase, etc.) and provide information on a range of subjects ancillary to the actual transcription, including semantics, ethnographic information, and transcription notes. Sometimes comments are written directly onto the manuscript as marginal notes, as in the following example.

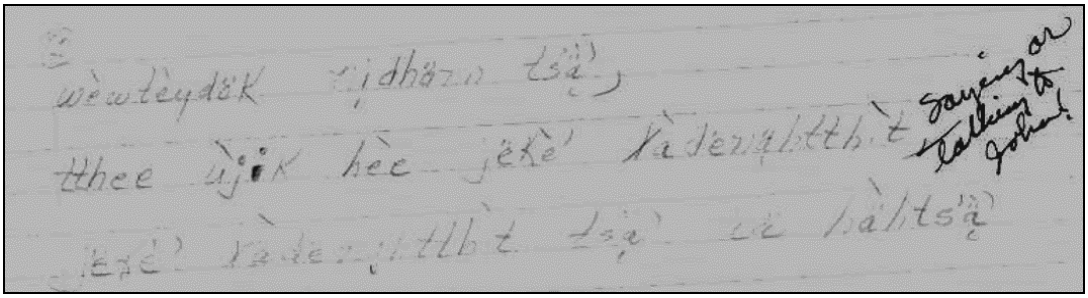


Figure 2: Annotation showing comments as marginal notes (Paul 1977)

Comments may also be written inter-linearly, as in Figure 3. Here the comment “stutter” indicates reference to paralinguistic information.

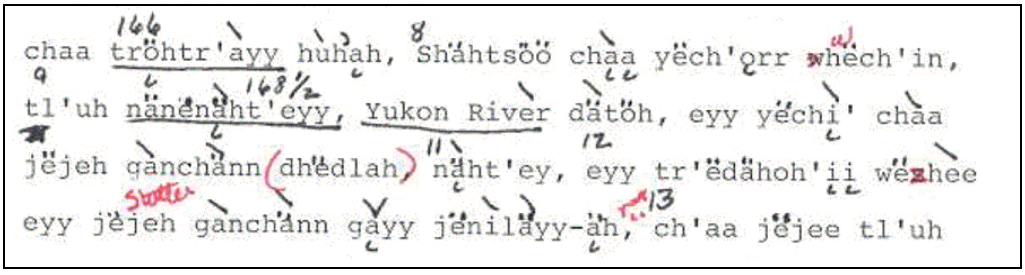
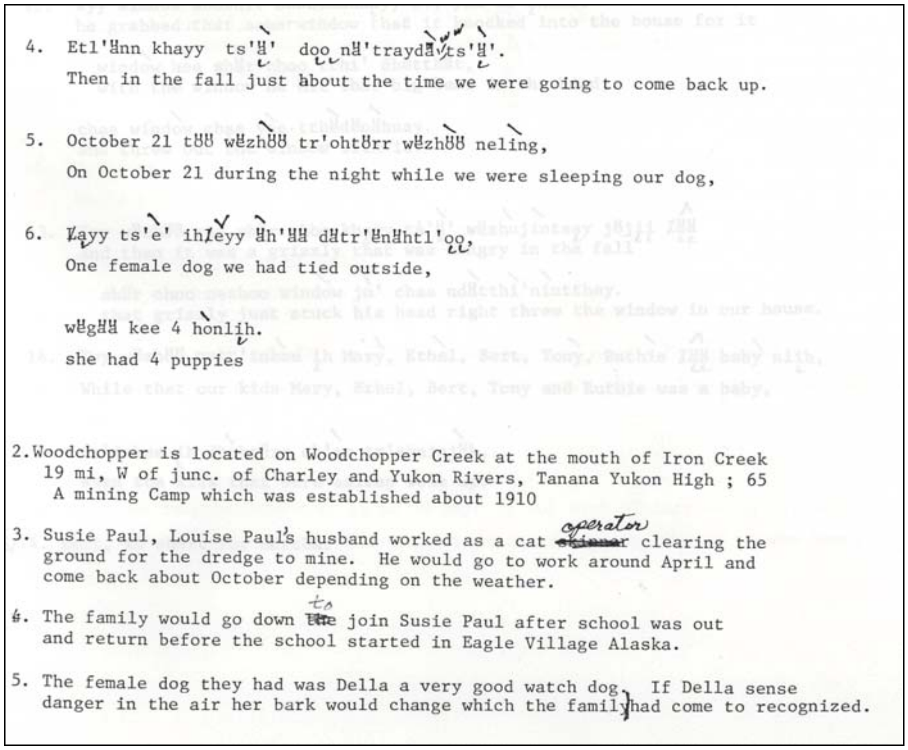


Figure 3: Annotation showing comments as marginal notes (Paul 1977)

In other cases comments are entered as footnotes or endnotes. The example in Figure 5 shows ethnographic comments entered as footnotes corresponding to numbered lines of transcription.



4. Etl'änn khayy ts'ä' doo nä'traydä ts'ä'.
Then in the fall just about the time we were going to come back up.

5. October 21 töö wëzhöö tr'ohtörr wëzhöö neling,
On October 21 during the night while we were sleeping our dog,

6. Łayy ts'e' ihłeyy äh'ää dätr'änähtl'oo,
One female dog we had tied outside,

wëgää kee 4 honlih.
she had 4 puppies

2. Woodchopper is located on Woodchopper Creek at the mouth of Iron Creek 19 mi, W of junc. of Charley and Yukon Rivers, Tanana Yukon High ; 65 A mining Camp which was established about 1910

3. Susie Paul, Louise Paul's husband worked as a cat ~~skinner~~ operator clearing the ground for the dredge to mine. He would go to work around April and come back about October depending on the weather.

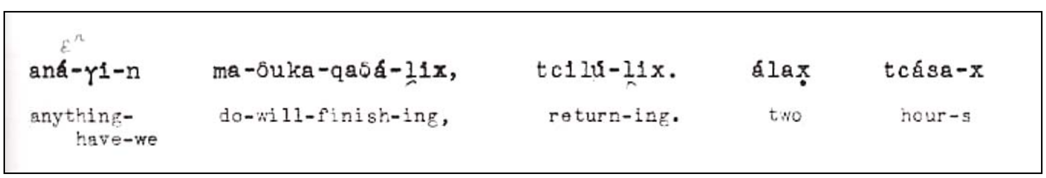
4. The family would go down ~~the~~ to join Susie Paul after school was out and return before the school started in Eagle Village Alaska.

5. The female dog they had was Della a very good watch dog. If Della sense danger in the air her bark would change which the family had come to recognized.

Figure 4: Annotation showing comments as footnotes (Paul 1977)

2.1.3 Uncertainty or elaboration

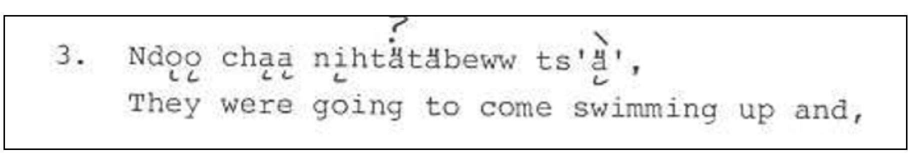
Annotations often indicate uncertainty on the part of the field worker. This may be represented in the form of comments followed by a question mark or simply by multiple annotations at a single level. An example of such uncertainty is the transcription of alternate levels of phonetic detail. Often this is indicated as a marginal or interlinear marking, as in the first morpheme of the following example.



aná-γi-n	ma-ðuka-qaðá-ḻix,	tcilú-ḻix.	álax̣	tcása-x
anything-have-we	do-will-finish-ing,	return-ing.	two	hour-s

Figure 5: Annotation showing greater phonetic detail (Ermeloff 1937)

The annotation above may represent either uncertainty of phonetic transcription or greater elaboration of phonetic detail. The following example clearly indicates uncertainty with a question mark above the second vowel of the third word.



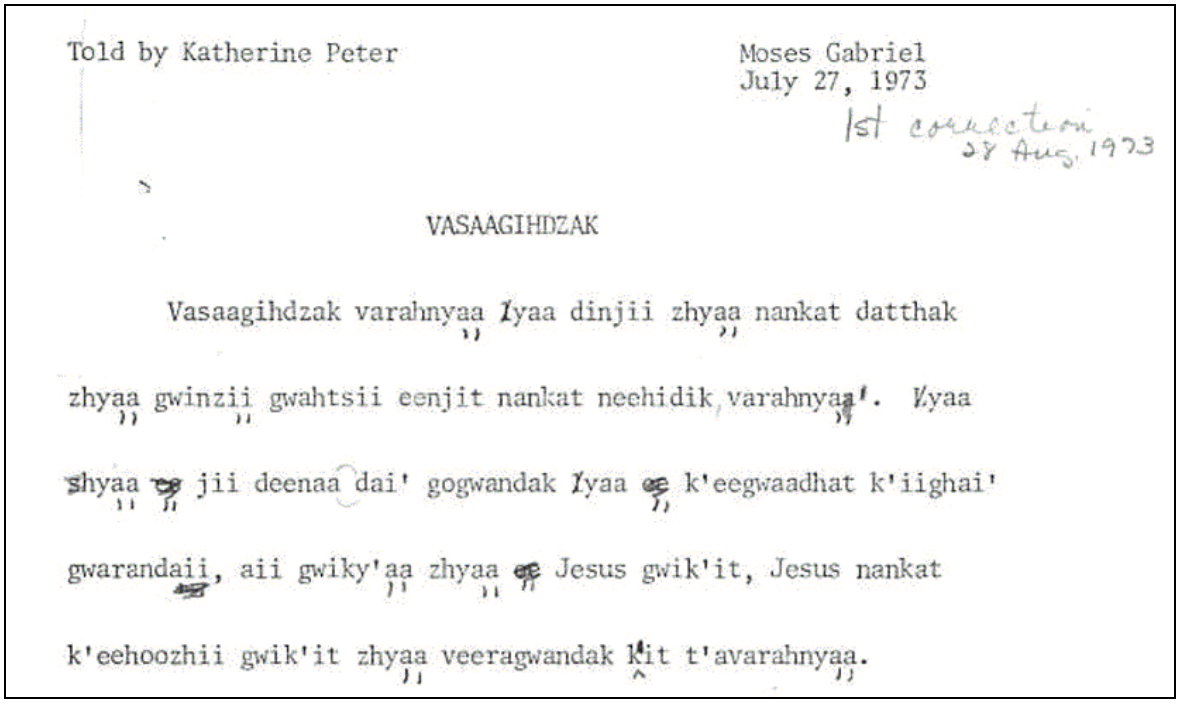
3. Ndoo chaa nihtätäbeww ts'ä',
They were going to come swimming up and,

Figure 6: Annotation showing uncertainty (Paul 1977)

2.1.4 Editorial annotations

Linguistic documentation is commonly a collaborative product of more than one language worker. The ANLC archive contains many examples of texts which have been annotated editorially by multiple authors. Subsequent authors or editors may enter hand editorial corrections or comments or additions to the original annotation. These usually take the form of secondary pen or pencil markings on the original document. Sometimes the author or editor is explicitly identified; other times she is not.

Editorial annotations may take several forms. Most commonly such annotations involve corrections to the transcription, for example, the addition or revision of tone diacritics. Sometimes the nature of the annotation can be directly identified from the manuscript, as in the following example. Here the manuscript directly indicates that a "correction" has been applied to the original transcript. Dates are indicated for both the original and the correction.



Told by Katherine Peter

Moses Gabriel
July 27, 1973
1st correction
28 Aug. 1973

VASAAGIHDZAK

Vasaagihdzak varahnyaa Ɂyaa dinjii zhyaa nankat datthak

zhyaa gwinzii gwahtsii eenjit nankat neehidik varahnyaa'. Ɂyaa

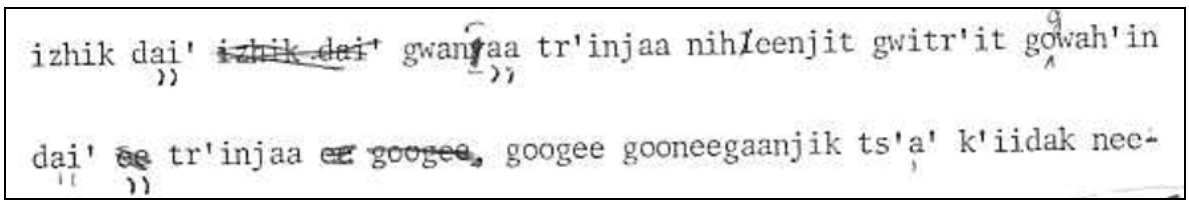
shyaa ~~ee~~ jii deenaa dai' gogwandak Ɂyaa ~~ee~~ k'eegwaadhat k'iighai'

gwarandaii, aii gwiky'aa zhyaa ~~ee~~ Jesus gwik'it, Jesus nankat

k'eehoozhii gwik'it zhyaa veeragwandak Ɂit t'avarahnyaa.

Figure 7: Editorial annotation by multiple authors (Peter 1973)

Editorial corrections may involve the simple addition or deletion of a diacritic or graph, but they may also include addition or deletion of words and phrases, as Figure 8 below.



izhik dai' ~~izhik dai'~~ gwanɁaa tr'injaa nihɁeenjit gwitr'it gowah'in

dai' ~~ee~~ tr'injaa ~~ee~~ ~~googee~~, googee gooneegaanjik ts'a' k'iidak nee-

Figure 8: Editorial deletions (Peter 1973)

Sometimes such annotations represent alternate transcriptions for a particular word or section of text, as in the following example.

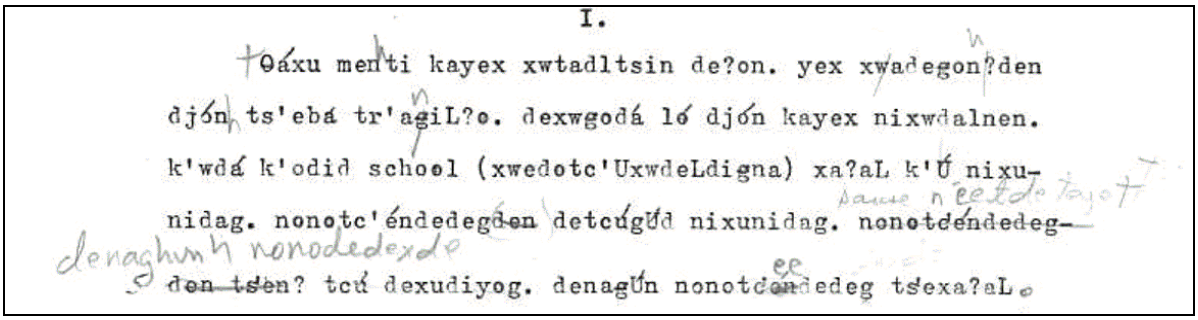


Figure 9: Editorial additions (Charlie 1961)

In other cases the nature of the annotation can be difficult to determine. For example, the following example contains editorial annotations marked in red ink. While these appear to be corrections, we have no way of determining which version is more "correct".

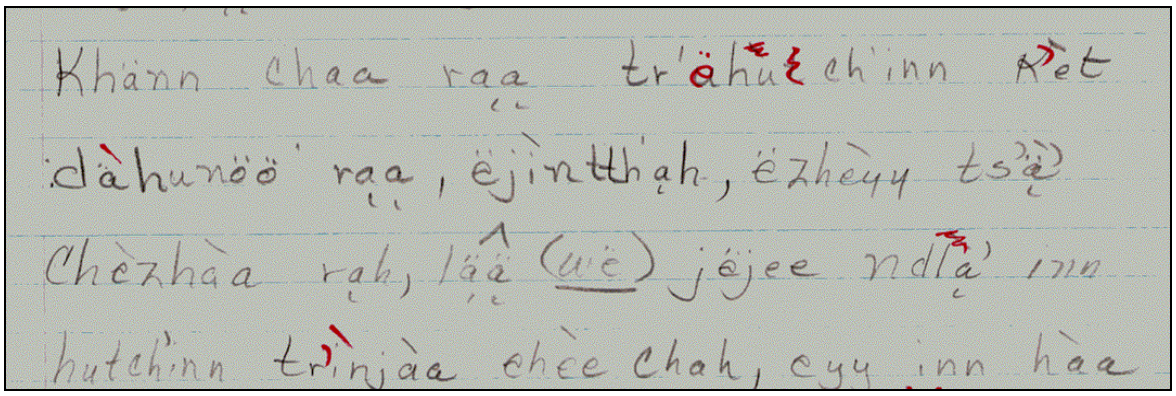


Figure 10: Editorial corrections (Paul 1977)

In such cases both the original and the edited or "corrected" form need to be preserved as separate layers of annotation. It may well be that the original turns out to be more "correct", or at the very least provide greater insight into the transcription.

2.2 The evolutionary nature of text annotation

Many of the examples of text annotation cited above exhibit an important feature of linguistic documentation which is often overlooked in approaches to preservation. This is the observation that language documentation is an inherently ongoing process. Primary linguistic data, including especially texts, evolve with time as linguists gain better understanding of the material. While we may tend to idealize the documentation as a mere snapshot of a language, most field workers readily acknowledge the evolutionary reality of field linguistics. Data may be "corrected" or annotated multiple times by one or multiple authors. Examples of such evolutionary annotation are well-attested in the ANLC archive.

Like other empirical sciences, documentary linguistics has sought to preserve both the original and the annotation or "corrected" forms of primary data. The paper trail left behind by such corrections often provides unanticipated insight into our understanding of language. Electronic approaches to creating and preserving linguistic documentation risk obliterating this paper trail. Best practices for electronic text annotation should not only readily accept legacy annotations but should also permit such annotations to be preserved in electronically-created data.

3. Recordings

At first glance it would appear that little needs to be said about standards for digitizing field recordings outside reference to technical parameters. However, an examination of existing recordings reveals interesting facts about the recording process which may need to be addressed by digitization standards. The primary observation is that there is a disjunction between the physical media and the recording session. A single medium, say a cassette tape, may contain several recording sessions, and a single session may span several media. Thus, these two approaches to representing recordings—the media and the session—represent cross-cutting and overlapping views of the same data.

For the archivist the physical medium may be the most important artifact. This is the object which can be handled and shelved. However, for the field linguist the medium may be largely irrelevant. From a linguistic point of view it is the recording session which is primary. And linguists have managed to record sessions in a variety of bizarre manners.

Mismatches between recording sessions and media are of two basic types. The first is the recording session which spans more than one physical medium. This happens when a recording session is continued onto another tape due to physical storage limitations. Such tapes are often labeled as a series, e.g., "tape 1 of 2". The second type of mismatch between sessions and media occurs when a single medium contains more than one session. Often this occurs due to a field worker's (perhaps misguided) desire to conserve physical media by beginning a recording session in the middle of a tape. Many recordings in the ANLC archive belong to both types, that is, they contain multiple sessions and portions of sessions. In all cases, the cataloged archival artifact is the physical medium.

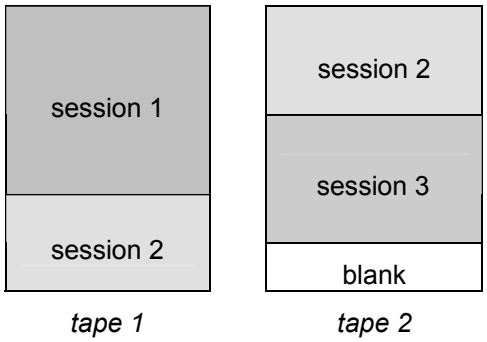


Figure 11: Mismatch between sessions and media

With the advent of digital technology we have the opportunity to place greater emphasis on the recording session, as proposed for example by IMDI (Broeder et al 2000). Sessions can be essentially reconstructed into single coherent units from several media. While the linguistic merits of such an approach are quite strong, this approach may pose difficulties for language archives which emphasize the physical media. For the foreseeable future language archives, including ANLC, will continue to archive physical media even after digital copies of those media are created. Thus, what is needed is a standard which allows for two views of the data, both from the point of view of the physical medium and the recording session.

As technologies such as direct digital recording become more popular, the distinction between sessions and media will continue to blur. While this convergence may simplify some aspects of recording preservation, it will clearly present new challenges to language archivists, as even the concept of an “original” recording becomes more difficult to maintain.

4. Recommendations

While this survey is by its nature of a very limited scope, it has clear implications for the development of best practices for digitizing and annotating texts and field recordings. The diversity of text formats and recording practices found in the ANLC archive alludes to an even great variety of approaches to textual documentation across the world. The practices of field linguists and the artifacts they create are as diverse as the languages which they document. Thus, any markup or encoding schema will need to be flexible enough to handle these diverse approaches. This point has already been recognized more generally, for example by Simons in his paper for this workshop:

> “A single markup schema that sanctions all common practices in structuring the content of a particular kind of resource will be too permissive to constrain any single resource to the specific plan of its creator.... [Thus] there will be multiple markup schemas, even in the context of best practice.”

It may well be that the propagation of best practice(s) will lead to the development of more standardized approaches to creating digital language resources. In this way of thinking the diversity of existing text and audio materials may be a relic of non-digital approaches to language documentation. However, it seems more likely that this diversity results directly from linguistic and methodological diversity among languages and language workers, respectively. In any case, even in a future digital world linguists will continue to rely on such diverse types of annotations. Successful annotation standards and tools will need to permit and facilitate such diversity.

References

Broeder, Daan, Pirkko Suihkonen & Peter Wittenburg. 2000. Developing a standard for media-descriptions of multi-media language resources. Papers from the workshop on Web-Based Language Documentation and Description, Philadelphia, December 12-15. Philadelphia: Institute for Research in Cognitive Science.

Charlie, Moses. 1961. Tanana Texts, edited by Michael Krauss. Manuscript, ANLC.

Ermeloff, Afenogin. 1937[2001]. Wreck of the Umnak Native, translated by Jay Ellis Ransom. The Dalles, Oregon: Western America Institute for Exploration.

Krauss, Michael E. & Mary Jane McGary. 1980. Alaska Native Languages: A Bibliographical Catalogue: Part One: Indian Languages. (Alaska Native Language Center Research Paper 3.). Fairbanks: ANLC.

Paul, Louise. 1977. Han Texts, transcribed Ruth Ridley. Manuscript, ANLC.

Peter, Katherine. 1973 Vasaagihdzak, transcribed by Moses Gabriel. Manuscript, ANLC.

Simons, Gary. 2003. The electronic encoding of text resources: A roadmap to best practice. Workshop on Digitizing and Annotating Texts and Field Recordings. Lansing, Michigan, July 11-13.