LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 4 (2010), pp. 34-59
http://nflrc.hawaii.edu/ldc/
http://hdl.handle.net/10125/4466

Principles and Practicalities of Corpus Design in Language Retrieval: Issues in the Digitization of the Beynon Corpus of Early Twentieth-Century Sm'algyax Materials

Tonya N. Stebbins
La Trobe University

and

Birgit Hellwig
University of Erfurt and La Trobe University

This paper describes a pilot project to develop a machine-readable corpus of early twentieth-century Sm'algyax texts from a large collection of handwritten manuscripts collected by the Tsimshian ethnographer and chief William Beynon. The project seeks to ensure that the materials produced are maximally accessible to the Tsimshian community. It relates established principles for corpus design to practical issues in language retrieval, recognizing that the corpus will likely function as an intermediate stage between the original manuscripts and any language materials developed by the community. The paper is addressed primarily to linguists working on language retrieval projects but may also be of use to communities who are working with linguists, as it provides insight into the concerns and preoccupations that linguists bring to such tasks.

1. INTRODUCTION. This paper describes a pilot project exploring the practicalities of retrieving and converting a large body of Sm'algyax manuscript texts into an electronic machine-readable format to use in language-related activities within the community. We identify a range of practical issues in the interaction between linguists and community members that will need to be addressed if the project is to expand and continue.[1]

Sm'algyax is the ancestral language of the Tsimshian Nation, whose territory is on the north coast of British Columbia, Canada. The texts discussed here were collected by the

[1] This paper is jointly authored. Each of us brings particular areas of knowledge to the paper. Birgit Hellwig brings expertise in language documentation and corpus design. Tonya Stebbins is responsible for the discussion of the potential relationship between the corpus and the Tsimshian community. Her views are based on her experiences working on the Sm'algyax Dictionary Project with the community between 1995 and 2000, and irregular contact in subsequent years.

Thanks to Catherine Easton, Christina Eira, and Mark Planigale for discussing various aspects of this paper with us; and also to two anonymous reviewers for their suggestions. The pilot project described here was funded by a Canada-Asia Pacific Award grant from the International Council for Canadian Studies, and we also thank them for their support.

E-ISSN 1934-5275

Tsimshian ethnographer and chief William Beynon under the sponsorship of Franz Boas in the first half of the twentieth century. The texts were transcribed in a phonetic orthography and are today generally available only through archives and microfilm copies. The goal of the pilot project described here is to test a system designed to make these texts more readily accessible to members of the Tsimshian community.

For a long time, linguists have implicitly or explicitly assumed that any work we do on a language that is no longer widely spoken is inherently useful to the community concerned. However, anyone working closely with communities may have had experiences that indicate that their work is sometimes regarded with ambivalence or even hostility. In many cases the results of the linguists' research are not taken up by the community in the ways that we might have anticipated. Our increasing awareness about the endangerment of the world's languages has started a long-overdue debate about the relationship between linguists and speech communities. In their seminal article, Hale et al. (1992) posed the central question of what "responsible linguistics" would be. And today, the linguistic community recognizes not only its responsibility towards supporting endangered languages but also the central role—as well as the agency—of speech communities in all matters related to their language (Cameron et al. 1992, 1993; Grenoble and Whaley 2006; Grinevald 2003; Hinton and Hale 2001; Nettle and Romaine 2000; Wilkins 2000). In fact, linguists involved in language documentation projects increasingly see their goal to be empowering speech communities to conduct their own documentation work and to support their own efforts of revitalization.

One central aspect to which linguists can contribute is in constructing and making available a corpus that benefits the community. This paper is an account of one specific case study in the field of corpus development: making old language manuscripts available for the benefit of a speech community for their revitalization efforts. This paper can be seen as a complement to Henderson 2008, which discusses a comparable project: the digital rendering of old handwritten manuscripts for the Noongar community. Henderson acknowledges the community-related aspects of this project (and refers the interested reader back to their earlier manuscripts outlining the adopted protocols in more detail: Henderson 2008:216) but focuses on the technical implementation of the corpus. Our paper takes the opposite approach: while acknowledging the importance of technical issues, it is concerned with those community-related aspects that are relevant to the design of the corpus, focusing particularly on the relationship among the corpus, its source materials, products, developers and potential users.

This paper begins with an introduction to the context of the project in section 2. The design of the corpus is discussed in section 3, while the conduct of the project is considered in section 4.

2. BACKGROUND TO THE PROJECT. This section provides the reader with background information about the context of the project. In section 2.1, the socio-political setting is described and in section 2.2 more information about the Beynon manuscripts is provided.

2.1 SM'ALGYAX: THE LANGUAGE OF THE TSIMSHIAN NATION. The main town in the traditional homelands of the Tsimshian Nation in British Columbia is Prince Rupert. In the mid-1990s the small handful of younger fluent speakers of the language were between

30 and 50 years of age, and the bulk of speakers were in the generation above. The really fluent speakers were nearly all well over 70 years of age by this time. The Tsimshian community began working on language revitalization in the 1970s. By that time, the language had gone through two generations during which language loss and obsolescence was taking place before revitalization efforts began.

Linguistics and linguists have had ambivalent status in the process of language revitalization in the Tsimshian community. The community has periodically made use of the expertise offered by linguists and the records linguists have left, but they are also troubled by many aspects of linguistic work and worldview. (See Stebbins 2003a for a discussion of her experiences working on the Sm'algyax Dictionary Project.)

The conflict runs right through relations between linguists and the community and must be recognized and addressed in order to maintain effective working relationships. For example, even something as basic as the name of the language reflects competing perspectives. *Sm'algyax* is the name used within the community, while *Coast Tsimshian* is the name used by outsiders, including linguists. The label *Sm'algyax* is preferred by the community partly because in Boas's writings, the name *Tsimshian* was used with reference to Nisga'a, Gitksan, and Sm'algyax collectively. Linguists use *Coast Tsimshian* in order to draw a finer distinction between the language and the language family.[2] Perhaps partly out of respect for Boas, there is a strong feeling in some sections of the linguistics community that using the word *Sm'algyax* to refer to the language is misleading as well as inconvenient. After all, the argument goes, we don't call German *Deutsch*![3]

Given these competing preferences and the exercise of power reflected in their resolution, using the word *Sm'algyax* in the title of this paper is a political statement regarding the relationship of the linguist to the language. Some other issues associated with the distinctive views of the language community in comparison with linguistics are discussed in Eira and Stebbins 2008.

In order to assert ownership over the language and provide an interface for linguists working in the area, the Tsimshian Nation convened the Ts'msyeen Sm'algyax Authority to oversee its interests in relation to the language. Any decisions about moving forward with this project beyond the pilot stage will be made by the Ts'msyeen Sm'algyax Authority in consultation with the community and, of course, the Ts'msyeen Sm'algyax Authority is likely to be a significant stakeholder in the project if it continues.

As shown in other revitalization projects (Amery 1995; Cameron et al. 1992, 1993; Grenoble and Whaley 2006; Grinevald 2003; Hinton and Hale 2001; Nettle and Romaine 2000; Warner et al. 2007; Wilkins 2000), the agency of the community over the project is essential for success. At the same time, cooperation with linguists has been fruitful, provided that agreement was reached about the respective roles and responsibilities of participants; in addition it has been important to recognize that linguists and community

---

[2] A final member of the family, Sgüüxs or Southern Tsimshian was not recognized as a distinct additional member of the family until Dunn (1976).

[3] This confusion has not had the same impact on the names of the communities concerned. The Sm'algyax speaking community refer to themselves as members of the Tsimshian nation, while their neighbours are referred to as the Nisga'a and Gitksan nations.

members may have different priorities that need to be reconciled (see section 4 for details). The Sm'algyax speaking community has made a strong commitment to re-establishing conversational competence in all age groups within the community, and there are already many sufficient resources in place to facilitate this. In such a situation, it is important to recognize that a corpus-construction project such as this one may not be a priority for the community concerned. Given the recognised importance of conversational competence to communities and most especially given the importance of community-level leadership in this type of approach (see, e.g., Hallett et al. 2007), a decision not to pursue this type of project must also be respected.

2.2 THE BEYNON MANUSCRIPTS. William Beynon's working life as an ethnographer is summarized in Halpin 1978. The Beynon texts referred to here were collected by Beynon and sent to Boas from 1932 to 1939. These materials, totaling 252 mostly handwritten narratives and ethnographic reports, are currently unpublished, though they are available on microfilm. A review of the materials was conducted by Winter (1984). The quality of the microfilm is poor (though perhaps not poorer than most), and the orthography used by Beynon differs in a number of respects from the modern writing system, making these materials difficult and often inaccessible for community members and researchers alike. The texts include varying amounts of word-level glossing as well as free translation. They also include a limited amount of metadata such as the name of the speaker and the month and year of the recording. Often one or two biographical details, such as the speaker's age or chiefly status, are also included. Additional metadata about many of the speakers could be gathered by referring to the work of Garfield (1939), as she discussed the careers of many of the people Beynon worked with. Figures 1 and 2 show the opening pages of a sample text. These pages are unusually clear and give an idea of the type of information available to current users of the corpus.

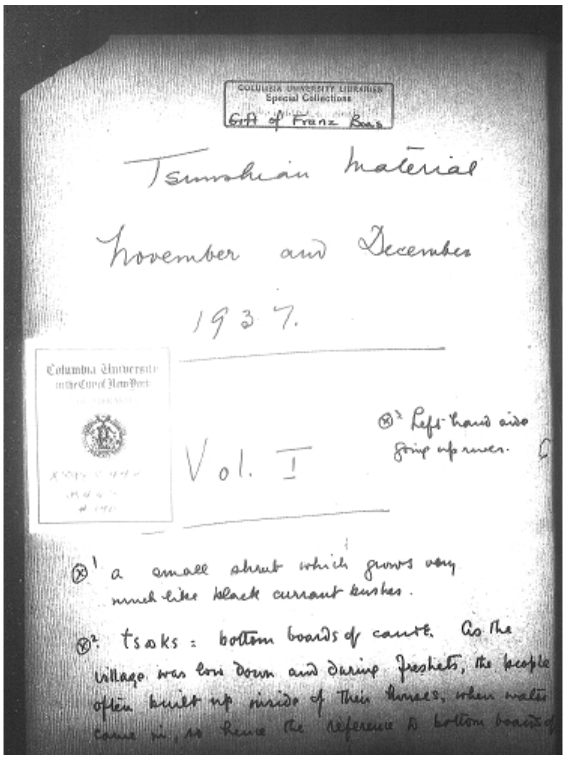
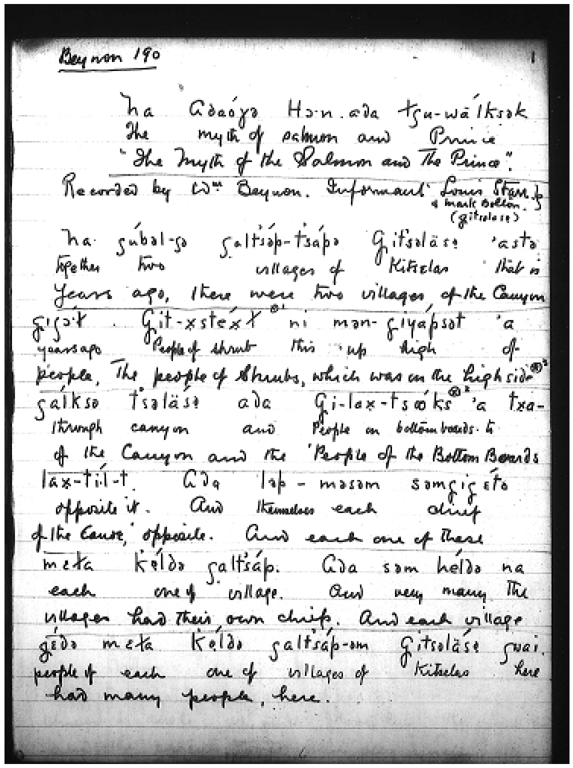


FIGURE 1: Reel 3, Notebook 13, Text 190, facing first page (reproduced with permission of Columbia University)


Beynon 190

"The Myth of the Salmon and The Prince".

Recorded by Wm Beynon.

Years ago, there were two villages of the Canyon people. The people of Shrubs, which was on the high side of the Canyon and the 'People of the Bottom Boards opposite. And each one of these villages had their own chief. And each village had many people, here.

FIGURE 2: Reel 1, Notebook 13, Text 190, page 1
(reproduced with permission of Columbia University)

This material represents a potentially rich linguistic resource for the community. However, past experience has shown that each of the few texts that has been retrieved and published has involved a great deal of work within the community: people have struggled with the poor condition of the copies of the manuscripts to which they have access, as well as with a range of linguistic issues posed by the spelling and language used within the texts. Beynon collected a great deal of other ethnographic material over the years (see Halpin 1978 for an overview), and while the majority is yet to be made more widely available, his account of a Gitksan potlatch was published as an edited volume by Anderson and Halpin (Beynon 2000).

The unedited Beynon manuscripts are not widely used in the Tsimshian community, although people are generally aware that they exist. Copies of the manuscript in microfilm and PDF formats are accessible to members of the Tsimshian community, but these formats impose a barrier for community members seeking to use these manuscripts. Microfilm is expensive to buy and requires special machinery to read; in addition, poor image quality makes it difficult to read the texts. A small selection of these texts has been published by Hutchingson et al. (1992), and these are available to the community as high quality, beautifully illustrated publications with parallel texts in Sm'algyax and English. A few other texts

have been incorporated into other publications on an occasional basis. Other researchers who have worked in the community, particularly Margaret Anderson and John Dunn, have re-transcribed selected texts from the collection for various purposes including teaching. These re-transcribed and edited texts have proven a useful resource in on-going language revitalization projects—a fact that lies behind the current attempts to make the entire Beynon corpus accessible.

Sm'algyax has two varieties: an everyday register used in general conversation and an elevated register used in formal public speaking (e.g., in recounting traditional law and other types of narratives). Consistent with the genre of the texts in the corpus, the elevated style predominates. It is readily identified through the choice of grammatical markers used by the speaker. The contrasts available in the elevated style involve five parameters: common vs. proper nouns, perfective vs. non-perfective aspect, ergative vs. absolutive case marking, indefinite location vs. definite location, and present vs. absent entity. In contrast, the everyday register is sensitive only to the marking of case and aspect.

Most of the Beynon texts are in the elevated register, but the corpus also includes some fragments of conversations. Clearly, even if the language of the texts reflects a rather specialized register—one that is not usually appropriate for learners to emulate—the texts nevertheless constitute a valuable resource of linguistic and non-linguistic information. In particular, they include important information on the culture and indigenous knowledge of the Tsimshian ancestors. This has been used in the development of materials and activities and other curriculum resources in the language programs. For more advanced learners of the language there is also the possibility of using the texts themselves as objects of study. That the texts, with their rich cultural heritage, exist at all is powerful in itself. A number of other studies also note the importance that such old sources have for the descendents, both in terms of their emotional significance and of their practical use in language revitalization and language teaching (Amery 1995; Warner et al. 2007; Yamada 2007). These studies also identify a number of accessibility issues that arise when trying to make available such resources (discussed in more detail in section 3).

Possible benefits for the community in making Beynon's texts available in a machine-readable format include the potential for the corpus to be used in higher-level language-teaching related activities (e.g., in senior high school language projects or projects for student language teachers) and to be more readily adapted for use in the community. The corpus would need to be glossed and annotated to allow for narrowly directed searches and analyses. With some training in using the glossed corpus, adult learners of the language could engage more fully with the texts.

3 CORPUS CONTENT DESIGN. The corpus is built around Beynon's original manuscripts and incorporates additional information in the form of re-transcriptions and linguistic analyses. Figures 1 and 2 illustrate two typical pages from Beynon's notebooks, and Table 4 conveys some of the additional information to be incorporated. It is clear from the differences in these representations that a number of issues arise in the course of corpus design. Henderson (2008) identifies some such issues in his discussion of the Laves Digitization Project, and these are also relevant to the Beynon project. This section outlines preliminary decisions in some of these key areas: issues in accounting for the histories of the texts are

discussed in section 3.1; issues of re-transcription are described in section 3.2; and the provision of metadata is explored in section 3.3.

3.1 ACCOUNTING FOR THE HISTORIES OF THE TEXTS. It is important to be aware of the way that the source manuscripts are perceived by the community. Linguists may assume that communities will value and trust manuscript materials on their languages, particularly if such material is scarce or has been difficult to obtain. And there are, indeed, studies that show the importance that such old sources have for the descendents of the original speakers (Amery 1995; Warner et al. 2007; Yamada 2007). However, this is not necessarily the case, since colonial records were often compiled by people who overall had profoundly damaging effects on the communities concerned. For example, the voices of earlier speakers of the language may have to be read from beneath the overlaid voices of colonial administrators or missionaries who often recorded these materials. In these circumstances communities may understandably have deeply ambivalent feelings towards the materials. For these reasons it is important to acknowledge that communities may be alienated from the manuscript materials themselves because of the identities of the compilers or the manner in which they were compiled (see especially the discussion in Amery 1995).

Tackling these issues may thus involve recognizing and acknowledging painful episodes in the history of the community, emphasizing the distinctions that can be made among voices in the manuscript, or seeking appropriate ways to distinguish information that either should or should not be shared. It may also involve reformulating materials so that they are in a different format. For example, if the difficulty related to a text was about restricted knowledge, and if the community were willing for the linguist to work with the text, then the product could be a wordlist with certain words omitted and a series of reformulated example sentences instead of the original text itself.

These issues are fortunately less significant in relation to the Beynon corpus than they might have been, since Beynon was himself a Tsimshian chief. His status in the community and his passion about his culture seem to have given him the confidence to record texts with great attention to detail,[4] and without concern for the ways they would be judged by outsiders. It is clear from his correspondence with Boas (Halpin 1978:148) that he was a sensitive and tactful researcher. Community members occasionally express concern about Beynon's Sm'algyax, since his mother was ethnically Nisga'a, though her father was certainly a prominent member of the Tsimshian community. Clah (Arthur Wellington), Beynon's grandfather, was translator for the first missionary to settle among the Tsimshian. In practice, when fluent speakers recheck the texts from the corpus they generally do not identify anything more than the very occasional problematic word (Margaret Anderson, pers comm.).

[4] This contrast is significant because an earlier Tsimshian consultant to Boas and Barbeau, Henry W. Tate, sometimes omitted sequences within texts that he felt were offensive to early twentieth-century Christian sensibilities. Also, whereas Beynon worked by taking dictation (Halpin 1978:143, citing Barbeau in a letter to Sapir, January 23 1915), Tate was apparently more secretive and did not make his activity public knowledge (this much discussed aspect of the narratives published as Boas 1916 is raised in Barbeau 1917:561).

For the practical design of the Beynon corpus, the considerations above make it necessary to put effort into accurately accounting for the histories of these texts. This involves identifying and acknowledging the originating author and any other voices that may be present. Authorship and the authority to recount these texts are central concerns within the community. For example, Roth (2008:167) reports that in the 1990s an *adawx*[5] was told by a member of the wrong lineage causing deep offense to the owners of the adawx. Not only must speakers who recount these texts be members of an appropriate lineage, they must also train for this role. Beynon usually attributed his texts to their narrators in metadata associated with each text, so for this corpus the issue of identifying authorship is essentially a matter of housekeeping rather than triggering larger problems about, for example, establishing where the text came from (see Section 3.3 for some issues arising in this context).

This identification has the advantage of reinforcing the primacy of authentic Sm'algyax voices in the text, partially addressing the concerns about language mixing between Sm'algyax and Nisga'a alluded to above. It also gives the individual voices within the corpus appropriate recognition and ensures that questions of intellectual property rights can be addressed. In addition, knowing the history of a source text and knowing the original narrator provides descendents within the language community with opportunities to reflect on how they want to make use of the text and make more informed editorial decisions if, for example, they intend to adapt the text to reflect current usage.

All these issues become important, since the digital re-transcriptions of source manuscripts can be used for various community-related and academic purposes during the course of language revitalization. Many of the source texts do not fall under copyright law proper, as they do not constitute original work (as succinctly summarized in Newman 2007). However, with the proper attribution comes an ethical obligation to learn about—and respect—the wishes of the original narrator, their descendants, and/or the language community with regard to the source materials. The specific details of the project therefore include the plan to assign the decision-making process to the Ts'msyeen Sm'algyax Authority or a similar body, so that the community itself can determine if and how the corpus or parts of it should be made available. This relates in particular to issues of access—i.e., who (within and outside the community) has access to which parts of the corpus and for what purposes, including the right to read material, to use it for community-related or academic projects, and to add or revise it (Craig 1993; Dwyer 2006; see also Warner et al. 2007:67–68 for a short summary of such issues in the North American context). All these decisions and restrictions then need to be implemented into the corpus design.[6]

In addition to an electronically readable version of the Beynon texts, there would be a link from each text back to a PDF of the original manuscript. This procedure ensures that

[5] An *adawx* is a general term for a traditional oral narrative. It is translated variously as: true telling, story, teaching narrative, myth, legend, or story.

[6] Depending on the community and the particular texts concerned, this potentially includes the rights of the original speakers (and their descendants), the rights of the speech community in these materials, and the rights of the people who initially recorded the texts (and their descendants). In the Tsimshian community, the practice is currently to identify the narrator and to have the rights of any publications be assigned to the Tsimshian Nation as a collective.

the corpus is reliable and trustworthy, as it allows users to go back to check the original source for themselves at the click of a button. One step that will be necessary is to tag every segment of text in the corpus to indicate its location in the original collection of manuscripts. The Beynon manuscript is organized into volumes that contain notebooks. Therefore, the reference would be to the volume number, the notebook number, the page number, and the line number where the relevant clause begins. To simplify checking the original manuscript, we will link each section of text directly back to the source location in a PDF of the original text.

Providing such links takes us back into the realm of copyright. The copyright for the Beynon manuscripts is held by Columbia University, since Boas worked for Columbia when he paid Beynon to collect the texts, which were at some stage incorporated into the Columbia archives. The texts were microfilmed circa 1980 by the Microfilm Corporation of America and are distributed by UMI. As a result of correspondence with staff at Columbia, we have established that, with permission from the Trustees of Columbia University in the City of New York, it would be possible to reproduce the original Beynon texts as part of this corpus. This permission would be granted on the basis that the materials would be published for educational purposes since the publication of these texts is allowed only on a cost-recovery basis.

3.2 RE-TRANSCRIPTION. The source manuscripts from Beynon form the basis for the corpus, and will be re-transcribed as faithfully as possible. However, to be useful to the community, the texts must incorporate a number of additions, including an orthographic representation and a revised translation. This decision is based on the following considerations.

A central issue for many speech communities is the language variety represented in the sources. This includes the inevitable fact that the language has changed over time. Because of these differences, speakers may find the old language inaccessible and hence of little practical use to their maintenance and revitalization efforts. Alternatively, they may dismiss their present-day language as a corruption, preferring to rely on the old language in their efforts. For the Tsimshian community, such issues are relatively easy to identify since most of the corpus is in a distinctive, widely recognized variety of the language.

It is not only the language variety itself that creates accessibility problems. Physical deterioration of manuscripts and illegibility of handwriting pose additional obstacles. Also, the use of phonetic symbols, linguistic terminology, abbreviations, and symbols is not always transparent to both linguists and native speakers (see also Henderson 2008, which traces comparable issues in the Laves Digitization Project). And although Beynon's transcription system is relatively straightforward in linguistic terms, it is different from the practical orthography used today and contains diacritics and symbols that are unfamiliar to community members. The remainder of this section addresses some of the practical barriers imposed by Beynon's orthography and his translation.

The texts are currently available in various versions of an orthography that Beynon developed over several years of working in collaboration with linguists such as Maurice Barbeau, Franz Boas, and Amelia Sussman. In order to allow users of the corpus to have a sense of the original manuscripts and to enable them to check the accuracy of our re-transcription, it is important to keep, or rather re-present, this orthography as accurately

as possible within the corpus. At the same time, this orthography is quite difficult to read, as it makes use of a range of symbols with which readers of the English orthography are not familiar. Furthermore, Beynon has subdivided words into morphemes and has adopted different principles for deciding on word boundaries in different places of his manuscripts, thus making it difficult for contemporary speakers to recognize words. The envisaged solution to this problem is not only to re-transcribe the source manuscripts in Beynon's orthography but also to add a transcription in a practical orthography.

In general terms, the key to ensuring community access to language retrieval materials is that the community, the ultimate users of the orthography, rather than the linguist, should also be its ultimate authors (or at least authorizers) (Coulmas 2003; Grenoble and Whaley 2006; Rogers 2005; Seifart 2006). To linguists, it seems rational to choose the "best" orthography from a technical standpoint (e.g., in relation to the consistency of representation of phonemes, pan-dialectal representation, or representation of tone or other relevant features). However, it is important to be aware that choices in this field are likely to be understood in other often more personal or political ways within the community (Eira 1998 and Easton 2007). It is likely to be better to choose a less ideal orthography if this is already established within the school system or if it is preferred by the people who are actively involved in the project.

In the case of the Tsimshian Nation, a practical orthography exists. Materials that are currently published in the language, particularly anything appearing under the auspices of the Ts'msyeen Sm'algyax Authority, are normally written in the Practical Orthography. The Practical Orthography is essentially a phonetic representation: i.e., it accurately reflects the spoken language, but does not represent phonemes. The challenge for any writer of the language is to keep track of the large number of graphemes entailed by an inventory of sixty-five phonemes and the many digraphs and diacritics required in order to modify the Roman alphabet to fit this purpose. Tables 1 and 2 show how the IPA representation of a selection of sounds in the practical orthography compares with Beynon's transcription system.

IPA	b	p	d	t	dz	ts	g	k	g^w	k^w	g^j	k^j	g	q	ʔ
Beynon	b	p	d	t	dz	ts	g	k	gw	kw	g·	-	g̣, q	q	‘
PO	b	p	d	t	dz	ts	g	k	gw	kw	gy	ky	g̲	k̲	‘

TABLE 1: Representation of Sm'algyax stops and affricates [7]

IPA	i	i:	e	e:	æ	æ:	a	a:	ɔ	ɔ:	u	u:	ɨ	ɨ:	ə
Beynon	i	i·	e, ɛ	-	a	ä	a	-	ɔ	ɔ·	u	u·	ə̥	ə̥	ə
PO	i	ii	e	ee	a	aa	a̲	aa̲	o	oo	u	uu	ü	üü	-

TABLE 2: Representation of Sm'algyax short and long vowels

[7] The contrast between the practical orthography <k> and <ky> was represented by the vowel alternation between <e> and <ɛ> in Beynon's system, particularly preceding a mid front vowel. The grapheme <ɛ> is associated with palatalization.

The symbols used in the Practical Orthography are largely accepted within the community; it has been in use in the community since the 1970s and is taught in schools. All the texts in the corpus would therefore be re-transcribed into this orthography. Three distinct grapheme sets that need to be handled in this project are thus the phonetic orthography used by Beynon, the Practical Orthography currently in use in the community, and the English orthography. Because of the generally transparent relationship of the practical orthography to IPA, no provision for rendering IPA is considered necessary.

Orthography design includes making decisions not only about individual symbols but also about representing words. This is not a straightforward issue, and we expect practical problems in this area, since Sm'algyax is a polysynthetic language in the sense that it makes use of incorporation, has large sets of bound morphemes, and uses derivational processes in the creation of clauses. As a result of these features, word boundaries in Sm'algyax pose interesting problems for both linguists and writers. The conventions around the representation of orthographic words in Sm'algyax are still developing. Stebbins (2003b:413–414) notes, however, that, "On the whole, Sm'algyax writers prefer not to create lengthy strings of graphemes (which they feel are more difficult to read) and tend not to treat lexical clitics as part of the stem orthographically. Occasionally this principle of writing meaningful units as separate orthographic words is extended to compounds and even derivational prefixes." Since Beynon varies in his own strategies of representing word boundaries, whatever conventions are decided upon for this project, there will be considerable regularization of the original text in this regard.

There is a dictionary available in the Practical Orthography that can assist in converting the spelling and boundaries of words in the original manuscript. Nevertheless, it is useful to be aware that individuals may have readily identifiable styles (see for example Stebbins 2003a:251, where she shows how the community is able to distinguish between orthographic styles of different linguists using what is ostensibly "the same orthography").

New words that have yet to be incorporated into the dictionary are also likely to occur in the manuscript, and a protocol for determining spellings of these words involving decision making within the Tsimshian community would be an important precursor to Tsimshian ownership of the final product. For example, though chiefly names were not collected in conjunction with the development of the dictionary, they continue to have a central place in Tsimshian public life and occur frequently in the corpus. In order to establish the preferred spelling for these names, it would be necessary for the linguist and/or community language workers to consult with the community. Until this is possible the names would be re-transcribed into the nearest equivalent in the current orthography without any attempt at regularization and marked so that they can be discussed with appropriate representatives from the community.

Another key issue that requires the addition of material is the translation of the texts. Beynon was certainly a competent speaker of English, but there is little information about his overall English proficiency. As regards his translations, he seemed to have favored literal translations (that accurately reflected the original Sm'algyax structure) over idiomatic English translations. Further research is certainly needed, and the project investigators will have to decide how to balance literal and idiomatic translations. Translating texts adequately is something of an art (see, for example, Evans and Sasse 2007; Foley 2007; Grenoble 2007; and Woodbury 2007 on issues of handling semantics and translations in language

documentation corpora). The goal is to protect the voices of the original manuscript while simultaneously taking steps to make the material readily accessible overall. The strategy we suggest here is to preserve the original translation for each line of text. In addition, a revised translation including changes motivated by the surrounding text would be prepared and included only if it seemed to represent a more accurate present-day translation than the original. For example, the word 'youths' was used by Beynon as a translation of the term *sum̲xsm 'yuut̲a* {*su=m̲xs=m 'yuut̲a* newly=grow.PL=DM man} 'young men'. The English term 'youth' can now carry negative connotations of potential criminality which Beynon was not likely to have intended. The principles applied in re-translation would be explicitly stated in the documentation accompanying the corpus, and these along with the original translations would allow the community control over the adaptation of materials for their own use.

3.3 THE AVAILABILITY OF METADATA AND ANNOTATION. It is widely recognized within the area of documentary linguistics that language materials are much more valuable when the circumstances of their recordings are available to subsequent users (see especially Austin 2006; Bird and Simons 2003). There is, of course, great variation in the amount and quality of metadata available within existing manuscript collections. At first glance it seems as if the sensible thing to do is to incorporate all available metadata about any text and its author into the corpus, but there are a range of contextual issues with this that deserve some consideration.

Keeping in mind that the corpus is destined for the community, and the communities are likely to have various different relationships to the speakers who were involved in the original texts, it is worth asking the following questions about the metadata:

- How reliable are the available metadata? Where are they from?
- Is it ethical to release (all of) the information contained in the metadata?

In cases where the documentary record is limited, these issues are unlikely, but they do crop up. This is true for the Tsimshian project, as a great deal of information about some of the consultants from whom Beynon recorded texts is available from other sources (particularly Garfield's 1939 dissertation "Tsimshian Clan and Society" which lays out in considerable detail the careers of various Tsimshian chiefs). From an academic perspective it seems worthwhile to incorporate any information from this source (and others like it) into the metadata since this additional information may well be of use in interpreting and contextualizing the texts. However, this material may also be read as being like gossip in some cases (especially when the information about the consultant is in some way unfavorable). For example, in her discussions of preparations for a number of funerals that took place around 1934–1935, Garfield went into some depth, specifically including comments on whether the amounts of compensation paid to various participants in the ceremonies were considered by others within the community to be satisfactory. This type of information is not likely to assist present-day readers in using the texts. In order to establish how much information from other sources (if any) should be added to the metadata, it will be necessary to consult with the descendents of the people featured in the text. Note that it would not be appropriate to render the author of the text anonymous (see section 3.1); and it is not feasible to prevent the tracing of unfavorable metadata back to a specific individu-

al. However, it would certainly be possible to provide links to these other sources without foregrounding any troublesome and irrelevant information in the corpus.

Based on Stebbins' experiences working in the community in the late 1990s, we expect that as the texts are used within the community, a range of issues are likely to emerge relating to the attitudes of various community members to the language of these texts. It will therefore become important to provide metadata information on the type of language used. In particular, there is a need to signal different registers and dialects.

First, texts need to be tagged for the register used. As we observed in section 2.2, differences in register are expressed through the choice of grammatical markers used by the speaker, with far more elaborate choices available in the elevated register. The connectives for each register are shown in table 3. Note that for common nouns in the everyday register the connective is always *=a* regardless of other factors, whereas there is a choice of five forms against three series of parameters in determining the form in the elevated register.

	ELEVATED REGISTER				EVERYDAY REGISTER		
	Common Nouns		Proper Nouns		Common nouns	Proper nouns	
	Non-Pfv	Pfv	Non-Pfv	Pfv		Non-Pfv	Pfv
A							
Indefinite	*=a*		*=t*	*?*	*=a*	*=t*	*=s*
Present	*=da*	*=sda*	*=dat*	*?*			
Absent	*=ga*	*=sga*	*=gat*	*=s*			
S/O							
Indefinite	*=a*		*=s*	*=t*	*=a*	*=s*	*=t*
Present	*=sda*	*=da*	*=das*	*=dat*			
Absent	*=sga*	*=ga*	*=s*	*=gat*			

TABLE 3: The full set of core argument dependency markers;
Non-Pfv = Non-perfective, Pfv = Perfective

Second, texts need to be tagged for the dialect used. Dialect differences tend to be associated with specific forms of lexemes. In a large number of cases, one dialect has a different word from the others. In the following pair of examples, the word ***meɬük*** 'pillow' is shown to be in the Kitkatla dialect but for speakers of other dialects the form is ***meɬüü***. The following examples from Stebbins 2003a illustrate these differences.[8]

[8] Abbreviations in these examples: 1 = first person, 3 = third person, A = subject of transitive clause, CN = connective (marker of grammatical dependency), FUT = future, O = subject of intransitive clause, PREP = preposition.

(1) Kitkatla dialect

Dm	txa'yaawkdu	meɬük	dziɬa	lisaym	ɬa'ati.
dm	txa'yaawk-t-u	meɬük	dziɬa	lisay-m	ɬa'ati
FUT	take.along-3O-1A	pillow	when	watch-CN	ball

'I'm going to take a cushion along when I watch the ball game.'

(2) Other dialects

P'lk'wa	hoyksit	da	meɬüü.
p'lk'wa	hoyks-t	da	meɬüü
feather.down	be.used-3S	PREP	pillow

'Feather down is used in pillows.'

(Stebbins 2003a:260)

In addition, the corpus would include metadata that refer to the original speaker as well as clear cross-referencing to the location of the text sequence in the original document (see also Section 3.1). Mark-up would also allow for overt discussion of editorial issues relating to the interpretation of the original document (e.g., where the writer's marks are ambiguous). Like most languages spoken on the Pacific coast of Canada, Sm'algyax has a very rich consonant inventory resulting in a complex orthography for expressing phonetic details. This was the case in Beynon's texts. In places, it is difficult to ascertain whether marks on the manuscript page are deliberate diacritics as opposed to incidental spots from discoloration and so on. Sometimes the manuscript quality is so poor that even larger marks can be unclear. An example is provided in figure 3.

In this example, the writing is generally very faint, particularly for the glossing line and the translation, which were no doubt added after the recording was made. The first word on the second line is particularly difficult to interpret, since both the Sm'algyax word and the gloss are unclear. Nor is there any indication in the free translation about what the word is likely to be. This word cannot be identified without assistance from a fluent speaker of the language, ideally someone familiar with this particular narrative from another context.

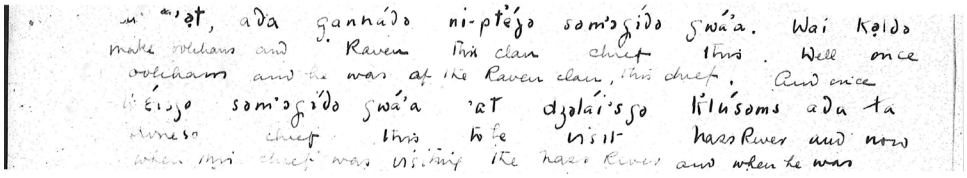


FIGURE 3: Excerpt from Reel 1, Notebook 6, Text 84, page 2 (reproduced with permission of Columbia University)

Finally, we propose to annotate the texts within the corpus for a range of linguistic features (e.g., parts of speech and grammatical functions) and to provide updated translations of the texts where this seems to be appropriate. In practical terms, these requirements must be incorporated into the design of the corpus while ensuring that it is flexible enough

to facilitate a range of uses. It must also be compiled in accordance with the principles of best practice in the area of corpus design. Annotation is argued to be a significant source of added value in a corpus (Leech 1997:2 and McEnery 2003:454–455) and could facilitate more sophisticated engagement with the texts by community members learning about Sm'algyax (e.g., in senior high school classes and beyond). A sample of the types of annotation to be included in the database is shown in table 4. Additional tiers will be necessary in order to accommodate metadata relating to issues in specific areas of the text (e.g., ambiguities in the manuscript).

1. Reference number (making the link to the location of the line in the source text).
2. Beynon original. This is the digital representation of Beynon's handwriting.
3. Practical orthography. This is the re-transcription of the Sm'algyax words in the original.
4. New analysis. This analysis of words in the clause and the morphemes will be based on the description of Sm'algyax presented in Stebbins (2003a, b). As new information about the language emerges in conjunction with the project the analysis is likely to gradually be refined.
5. Beynon gloss. This is a copy of Beynon's glossing.
6. New gloss. Based on the analysis, generally reflecting the glosses used by Beynon but also the translations available in the *Sm'algyax Learner's Dictionary* (see also comment regarding line 8, below).
7. New part of speech labels. Based initially on the categories presented in Stebbins 2003a, b and refined as new information emerges from the project.
8. New function labels. Based initially on the analysis presented in Stebbins 2003a, b and refined as new information emerges from the project. This line will be concerned mainly with labelling the heads of phrases according to the functions of the phrases.
9. New translation. This may involve adding information omitted in Beynon's original (as in the example in the table) as well as changes to the text to reflect modern usage as discussed in section 3.2. In the example in table 4, this has lead to differences in the syntactic relations represented in the translation where the syntactically more representative version 'Many were each of the Kitselas villages' people here' is non-idiomatic. The weight of different factors in determining the style of the translation will be determined in consultation with the community.
10. Beynon translation. This is a straight copy of Beynon's translation.

1. Reference number (to source text)										
2. Beynon original:	ɑdɑ	səm	héldə	nɑ	gédə	meɬɑ	kꞌə́ldə	gɑltsꞌáp-əm	gitsəlɑ̈́sə	gwɑi
3. Practical orthography:	Ada	sm	heelda	na	gyeda	meła	k'üülda	galts'abm	Kitselas	gwai
4. New analysis:	ada	sm	heelt-a	na	gyet-a	meła	k'üül-da	galts'ap-m	Kitselas	gwai
5. Beynon gloss:	and	very	many	the	people of	each	one of	villages of	Kitselas	here
6. New gloss:	and	really	many= CN	POSS	person = CN	each	one= CN	village = CN	Kitselas	here
7. New part of speech labels:	CONJ	ADV	QUANT= COM.N.CN	POSS	N= POSS.CN	MOD	NUM= HEAD.N.CN	N= COMP.N.CN	PN	DEM
8. New syntactic function labels			PRED						HEAD OF S	LOC
8. New translation:	And there were many people in each of the villages here at Kistelas.									
9. Beynon translation:	And each village had many people, here.									

TABLE 4: Sample of text from the corpus with complete analysis.[9]

[9] This is analysed as a verbless clause in which the quantifier *heelt* 'many' acts as the predicate with the possessed NP *nagyeda meła k'üülda galts'abm Kitsela* 'each of the Kitselas village's people' functioning as the subject. Abbreviations: ADV = adverb, COM.N.CN = connective marking common noun headed NP to follow, COMP.N.CN = connective marking head of compound noun to follow, COMP.N.CN = connective marking head of NP to follow, CONJ = conjunction, DEM = demonstrative, HEAD OF S = head of subject NP, LOC = locative phrase, MOD = modifier, N = noun, NUM = number, PN = proper noun, POSS = possessive marker, POSS.CN = connective marking possessor NP to follow, PRED = predicate, QUANT=quantifier.

The corpus would allow users the option to view only the lines of representation and mark-up fields that they require. For example, if users only wanted to view the material from the original manuscript, these lines could easily be selected.

4 CONDUCT OF THE PROJECT. The preceding section has identified a number of issues that have arisen in the design of the Beynon corpus: accounting for the histories of the text as a prerequisite for making informed decisions on questions of access rights and editing materials (section 3.1); re-transcribing the texts in the Practical Orthography (and in this process expanding on existing conventions and dictionaries) and re-working translations (section 3.2); and finally incorporating metadata information, especially about narrators and types of language used, as well as providing additional annotations to make the texts more useful to contemporary speakers and researchers (section 3.3). In all cases, it has become clear that while the linguist is in a good position to help and advise, the community alone has the knowledge and the authority to address these issues properly. This conclusion is of course at the heart of a responsible linguistics, and a number of articles beautifully show various steps in the interaction between linguists and communities in designing a corpus (Cameron et al. 1992, 1993; Dobrin 2005; Grenoble and Whaley 2006; Grinevald 2003; Hale 2001; Hinton and Hale 2001; Mosel 2006; Nettle and Romaine 2000; Otsuka and Wong 2007; Warner et al. 2007; Wilkins 2000; Yamada 2007).

This section now looks in more detail at how practical cooperation is envisaged for designing the Beynon corpus, with a particular focus on the potential pitfalls. It explores the steps that would be involved in developing the Beynon manuscripts into a corpus for use in the community. In order to ensure that the project is sustainable, it is important to establish good working relationships based on clear roles and responsibilities and a shared understanding of the activities involved. This section explores a range of strategies for organizing work and personnel in relation to the project.

The basic flow of work in relation to the project would be as follows:

1. Data entry: type up Beynon's texts, glosses, and free translation so that they exist in electronic machine readable format;
2. Re-transcribe the texts into the practical orthography;
3. Re-analyse the texts to produce new (more complete and consistent) interlinear glosses and additional annotations; and
4. Write a modernized translation particularly where the words used by Beynon are obsolete or have changed their meaning.

Many of the questions associated with how this work is conducted boil down to the balance of involvement between the community and the linguist. The community clearly possesses the best knowledge of the language and the best sense of how the texts could be translated into English today, while the linguist provides expertise in language analysis and the more technical aspects of the project. The assignment of roles in any project beyond the pilot stage would be negotiated with the Ts'msyeen Sm'algyax Authority.

The following sections consider the costs and benefits of different approaches to distributing the work involved in the project. Section 4.1 outlines the resources and personnel potentially available to the project. Section 4.2 considers two broad approaches to sharing

the work: one in which the linguist does the bulk of the data entry with periodic consultations with the community, and another more integrated approach in which the linguist and community members work more closely together. Section 4.3 focuses on the necessity of ensuring useable and timely results.

4.1 PARTICIPANTS AND FUNDING SOURCES. Setting up a collaborative project involves entering into a partnership, typically with some particular sector of the community, and identifying relevant stakeholders and participants. Within some communities there are people viewed as holding authority in relation to language. These may include, for example, senior speakers or school teachers from the language community who have been delegated as representatives by older, more fluent speakers and act as intermediaries in developing partnerships. In other communities where there is no one who can be readily identified as holding the relevant authority, appropriate partnerships may be more difficult to establish. In the case of the Tsimshian community, academics wishing to conduct research on Sm'algyax are required to consult with the Ts'msyeen Sm'algyax Authority, which is authorized to act in all matters related to its language. The Authority works closely with the Sm'algyax Language Program and is comprised of elders who speak the language fluently and represent all dialect groups along with members of the Language Program and other interested members of the community. That is, the project is envisaged here as a cooperative effort between Stebbins and the Authority acting on behalf of the Tsimshian community to identify individual community members to perform various project tasks.

What both the linguist and the community can invest in language retrieval projects is finite and spread across a range of activities and responsibilities. The amount of time and energy participants can give is typically balanced against other demands and responsibilities. The community are the group who are likely to make the most use of the corpus over the long term, but they must often juggle commitments relating to the care of other family members, health, housing, education, and so on with time spent on language work. On the other hand, linguists based within universities may have limited periods of time, during breaks from teaching for example, when it is possible to focus on their research. It is important that expectations in this area are realistic and are based on clear communication from both sides. For example, in relation to Stebbins' work with the Sm'algyax Language Program on the *Sm'algyax Learner's Dictionary*, there were periods in which the language teachers had to prioritize curriculum planning, and the dictionary project was temporarily set aside.

In the proposed project, Stebbins hopes to have ongoing involvement at some level. As a linguist and an outsider to the community, her participation raises both ethical and practical issues. In terms of communication, a significant issue is the fact that Stebbins lives and works in Melbourne, Australia, while the community is located on the other side of the Pacific. Another is that her other work responsibilities will limit her capacity to respond to the needs of the community (in comparison with a full time linguist dedicated to the project).[10]

[10] We have written this section on the assumption that Stebbins could continue to be involved in the project but also want to acknowledge that the Tsimshian community may well decide to embark on this project with a linguist who can be more readily available.

Visits to a community are often expensive, and time away from other means of earning income is likely to affect other areas of life for the linguist concerned. The converse, of course, is also true; it is usually even more difficult for community members to come to where the linguist lives. In our experience, it may be possible to continue to work effectively at a distance, provided a partnership is built on a firm foundation established through extended time in the community at the beginning of the relationship.

A further set of questions addresses how the work is to be funded. There are now a number of international and national (Canadian) initiatives that support documentation and revitalization efforts. In addition, as an academic, Stebbins is able to participate in the project as long as certain conditions are met. Her university will support her involvement in the project provided she publishes academic papers associated with her activities. These publications are potentially good for the project, since they provide an opportunity for her to reflect and invite feedback from colleagues and peers on the work.

This need for linguists working in university-based research positions to publish scholarly papers potentially conflicts with a key concern for many communities: the reclamation of intellectual property associated with language. For some this means that no one else should be allowed to publish information about the language. Different perspectives on ownership and authorship can create difficulties for linguists and tensions in their relationships with language communities. Although linguists often speak of language as a unique representation of human thought, this view is not readily relatable to the idea that any one person or group can "own" or "hold copyright" over a language (see also Newman 2007 on potential tensions of copyright law and community views).

One possible compromise is to allow the linguist to publish scholarly material associated with the process of preparing the corpus and to make the corpus itself, as well as material derived from it the intellectual property of the community. Whether this compromise is acceptable and where the boundaries around this delineation should be set would have to be negotiated by the linguist and community concerned.

For the Sm'algyax Learners' Dictionary project, the division was made in roughly this way: The dictionary itself (see Stebbins 2001) is the property of the community (and has since been converted to an online format with the assistance of Margaret Anderson: http://smalgyax.unbc.ca/). Stebbins was involved in this project in conjunction with working on her dissertation (later heavily revised and published as Stebbins 2003a). A small number of other papers stemming from her experiences with the Tsimshian community and with her analysis of the language have also appeared based on this division.

It is important to note that such arrangements are not open-ended. Although Stebbins continues to reflect on these formative experiences and would like to continue research on the language, it is now many years since she has had direct contact with the community. Her relationships with the community require renewal before anything further can legitimately be done.

4.2 WORK PLAN. One strategy for dealing with data entry and analysis would be for Stebbins to take responsibility for the bulk of the work and to consult periodically with members of the Tsimshian community on issues such as the spelling of words not yet incorporated into the Sm'algyax Learner's Dictionary and on the details of any revisions to Beynon's free translations of texts.

Given the potentially limited resources available, the sheer volume of texts that could be incorporated into the corpus, and the poor quality of current microfilm format, it may make sense for community members to be more selective about where they direct their energies (e.g., by continuing their focus on teaching the language and developing materials). In an academic context this approach can seem relatively simple. However, there are a number of problems with an approach that relies on internet communications technology when working with communities. Some issues we know of from our own experiences or from colleagues include the following:

- Access to technology is likely to be limited to certain individuals within the community, such as senior teachers. These people are already likely to be carrying considerable leadership roles and may not be able to act as effective conduits for the project. Access is limited in terms of (1) ownership of or access to appropriate equipment, and (2) familiarity with computers. (This is particularly a barrier for the older members of many communities—the people whose knowledge and experience are most important to the project.)
- Difficulties in maintaining relationships are more likely to develop when face-to-face contact is not regular. Many communities speaking endangered languages place a very high value on the processes of joint work and the relationships that underlie these processes. In this view, progress through the project also involves ongoing learning in relation to the project. Absence from the community is likely to exacerbate misunderstandings and repeatedly stall work in this type of setting.
- Authority over the project is compromised. This approach takes away some of the agency from the community. Although drafts of the database entries can be sent back and forth, it is sometimes difficult for people within communities to persist in making corrections to the work of an outside expert. Not only is correcting a database intrinsically a difficult, time consuming, and unrewarding task, but it is also very frustrating to repeatedly correct mistakes that are the result of a more general misunderstanding that could have easily been dealt with in a face-to-face interaction.

For these reasons, although it may seem efficient for the linguist to digitize the manuscript materials as a relatively independent project, a more integrated approach to data entry and analysis could ultimately be more worthwhile.

In fact, funding bodies increasingly advocate the training aspect of language documentation programs (Dimmendaal 2004; Florey 2004; Foley 2004; Lastra 2004; Mosel 2006; Woodbury and England 2004). That is, the linguist's role could be conceived of as that of a trainer or adviser, and much of the practical work could be done by a member of the Tsimshian community, perhaps a younger person with a strong commitment to the language and an interest in computers. This approach would also have to involve senior speakers of the language in decision-making about the spelling and translation of words not yet included in the Sm'algyax Learner's Dictionary and about revisions to Beynon's translations.

Aside from the training opportunities involved, an advantage of having members of the Tsimshian community doing this work would be the benefit of having someone in the community with expertise about all aspects of the project. The ongoing involvement of

senior speakers of the language will significantly increase the credibility of the project within the community as well as the quality of the materials it produces. And, ultimately, it ensures that the community retains its agency and control over the project.

4.3 ENSURING USEABLE, TIMELY RESULTS. The project has a long time frame since, depending on how funding develops, the 252 texts may take many years to convert to electronic format. For the community to take advantage of this work in a timely fashion, we therefore expect that it will be necessary to complete the work in stages.

In addressing this issue it would be useful to identify any texts in which the community had a particular interest and to ensure that the corpus has a user-friendly interface for exporting the texts. The electronic corpus proposed here would facilitate the process of identifying texts appropriate for specific purposes by including metadata about each text and by making them searchable in various ways (keywords, text themes, authors, dialect groups and so on).

Such texts can already be identified since a preliminary step when setting up this project was to construct a smaller database containing the metadata from each text included in the microfilm. This information was drawn primarily from the microfilm collection guide (appearing in the first reel). It includes the location of the text in the notebooks and the microfilm, the title in both Sm'algyax and English, the name and basic biographic information about the author, background to the story, key themes in the story (as identified in the guide), and a set of key words that identify the area of Tsimshian life the story relates to.

5 CONCLUSION. There are three key steps to retrieving language materials for the community: (1) finding relevant manuscript materials, (2) developing these materials into a corpus, and (3) adapting the materials for use within the community. Linguists have specific skills to offer in tackling the second of these steps. How we undertake this work will have implications for how it is received by the community and whether or not it is subsequently used.

If a corpus is to be used by a community it must be trustworthy and adaptable, and in this paper we have examined some of the types of choices that contribute to these features. We have specifically explored these issues in relation to the design of an electronic corpus based on the Beynon manuscripts.

Most of the factors that determine the trustworthiness and adaptability of a corpus reflect the nature of the relationship between the linguist and the community more broadly. They are handled in the process of designing the corpus and of negotiating the principles applied in its construction. As for the contents of the corpus intended for use in language retrieval, it is particularly important that the decisions concerning the organization of the material be consciously made and clearly explained, so alternate viewpoints can be explored by members of the community as they assume ownership of the material and begin to turn it to their own ends.

References

Amery, Rob. 1995. It's ours to keep and call our own: Reclamation of the Nunga languages in the Adelaide region, South Australia. *International Journal of the Sociology of Language* 113. 63–82.

Austin, Peter K. 2006. Data and language documentation. In Jost Gippert, Nikolaus P. Himmelmann and Ulrike Mosel (eds.), *Essentials of language documentation*, 87–112. Berlin and New York: Mouton de Gruyter.

Barbeau, Mauris. 1917. Review of Franz Boas, *Tsimshian Mythology. American Anthropologist* 19. 548–563.

Beynon, William. 1932–1939. Tsimshian texts collected for Franz Boas. MS. Special Collections. Columbia University Archives.

Beynon, William. 2000. *Potlatch at Gitsegukla: William Beynon's 1945 Field Notebooks*. Ed. by Margaret Anderson and Marjorie Halpin (eds.), Vancouver: UBC Press.

Bird, Steven & Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79. 557–582.

Boas, Franz. 1916. Tsimshian mythology, based on texts recorded by Henry W. Tate. 31st *Report of the Bureau of America Ethnology 1909–1910*, 29–1037. Washington D.C.: Smithsonian Institution.

Cameron, Deborah, Elizabeth Frazer, Penelope Harvey, Ben Rampton, & Kay Richardson. 1992. *Researching language: Issues of power and method*. London: Routledge.

Cameron, Deborah, Elizabeth Frazer, Penelope Harvey, Ben Rampton, & Kay Richardson.1993. Ethics, advocacy and empowerment: Issues of method in researching language. *Language and Communication* 13(2). 81–94.

Coulmas, Florian. 2003. *Writing systems: An introduction to their linguistic analysis*. Cambridge: Cambridge University Press.

Craig, Colette. 1993. Fieldwork on endangered languages: A forward look at ethical issues. In André Cochetiere, Jean-Claude Boulanger, & Conrad Ouellon (eds.), *Proceedings of the XVth International Congress of Linguists*. Vol. 1, 33–42. Quebec: Presses de l'Université de Laval.

Dimmendaal, Gerrit J. 2004. Capacity building in an African context. In Peter K. Austin (ed.), *Language Documentation and Description* 2, 71–89. London: School of Oriental and African Studies.

Dobrin, Lise M. 2005. When our values conflict with theirs: Linguists and community empowerment in Melanesia. In Peter K. Austin (ed.), *Language Documentation and Description* 3, 42–52. London: School of Oriental and African Studies.

Dunn, John A. 1976. Tsimshian internal relations reconsidered (part 1). Paper presented at Northwest Studies Conference, Victoria, B.C.

Dwyer, Arienne M. 2006. Ethics and practicalities of cooperative fieldwork and analysis. In Jost Gippert, Nikolaus P. Himmelmann and Ulrike Mosel (eds.), *Essentials of language documentation*, 31–66. Berlin and New York: Mouton de Gruyter.

Easton, Catherine. 2007. *Orthography developments in Papua New Guinea: The interaction of linguistic structures and language attitudes*. La Trobe University dissertation.

Eira, Christina. 1998. Authority and discourse: Towards a model for orthography selection. *Written Language & Literacy* 1(2). 171–224.

Eira, Christina & Tonya Stebbins. 2008. Authenticities and lineages: Revisiting concepts of continuity and change in language. *International Journal of the Sociology of Language* 189. 1–30.

Evans, Nicholas & Hans-Jürgen Sasse. 2007. Searching for meaning in the library of Babel: Field semantics and problems of digital archiving. In Peter K. Austin (ed.), *Language Documentation and Description* 4, 58–99. London: School of Oriental and African Studies.

Florey, Margaret. 2004. Countering purism: Confronting the emergence of new varieties in a training program for community language workers. In Peter K. Austin (ed.), *Language Documentation and Description* 2, 9–27. London: School of Oriental and African Studies.

Foley, William A. 2004. Language endangerment, language documentation and capacity building. In Peter K. Austin (ed.), *Language Documentation and Description* 2, 28–38. London: School of Oriental and African Studies.

Foley, William A. 2007. Reason, understanding and the limits of translation. In Peter K. Austin (ed.), *Language Documentation and Description* 4, 100–119. London: School of Oriental and African Studies.

Garfield, Violet E. 1939. *Tsimshian clan and society*. University of Washington dissertation.

Grenoble, Lenore A. 2007. The importance and challenges of documenting pragmatics. In Peter K. Austin (ed.), *Language Documentation and Description* 4, 120–135. London: School of Oriental and African Studies.

Grenoble, Lenore A. & Lindsay J. Whaley. 2006. *Saving languages: An introduction to language revitalization*. Cambridge: Cambridge University Press.

Grinevald, Colette. 2003. Speakers and documentation of endangered languages. In Peter K. Austin (ed.), *Language Documentation and Description* 1, 52–72. London: School of Oriental and African Studies.

Hale, Ken. 2001. Ulwa (Southern Sumu): The beginnings of a language research project. In Paul Newman & Martha Ratliff (eds.), 76–101. *Linguistic Fieldwork*. Cambridge: Cambridge University Press.

Hale, Ken, Michael Krauss, Lucille J. Watahomigie, Akira Y. Yamamoto, Colette Craig, LaVerne Masayesva Jeanne, & Nora C. England. 1992. Endangered languages. *Language* 68(1). 1–42.

Hallett, Darcy, Michael J. Chandler, & Christopher E. Lalonde. 2007. Aboriginal language knowledge and youth suicide. *Cognitive Development* 22. 392–399.

Halpin, Marjorie. 1978. William Beynon, ethnographer: Tsimshian, 1888–1958. In Margot Liberty (ed.), Indian Intellectuals. *Proceedings of the American Ethnological Society* 1976, 242–256. St Paul: West Publishing Co.

Henderson, John. 2008. Capturing chaos: Rendering handwritten language documents. *Language Documentation & Conservation* 2(2). 212–243.

Hinton, Leanne & Kenneth Hale (eds.). 2001. *The green book of language revitalization in practice*. San Diego: Academic Press.

Hutchingson, S & Sm'algyax Language Teachers. (eds.) 1992. *Suwilaay'msga na ga'niiyatgm [Teachings of our Grandfathers]*. 7 vols. Prince Rupert, B.C.: Tsimshian Nation and School District 52.

Lastra, Yolanda. 2004. The need for capacity building in Mexico: Misión de chichimecas, a case study. In Peter K. Austin (ed.), *Language Documentation and Description* 2, 108–121. London: School of Oriental and African Studies.

Leech, Geoffrey N. 1997. Introducing corpus annotation. In Roger Garside, Geoffrey Leech & Tony McEnry (eds.), 1–18. *Corpus Annotation*. London: Longman.

McEnery, Tony. 2003. Corpus linguistics. In Ruslan, Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*, 448-463. Oxford: Oxford University Press.

Mosel, Ulrike. 2006. Fieldwork and community language work. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 67–85. Berlin and New York: Mouton de Gruyter.

Nettle, Daniel & Suzanne Romaine. 2000. *Vanishing voices: The extinction of the world's languages*. Oxford: Oxford University Press.

Newman, Paul. 2007. Copyright essentials for linguists. *Language Documentation & Conservation* 1(1). 28–43.

Otsuka, Yuko & Andrew Wong. 2007. Fostering the growth of budding community initiatives: The role of linguists in Tokelauan maintenance in Hawai'i. *Language Documentation & Conservation* 1(2). 240–256.

Rogers, Henry. 2005. *Writing systems: A linguistic approach*. Oxford: Blackwell.

Roth, Christopher F. 2008. *Becoming Tsimshian: The social life of names*. Seattle: University of Washington Press.

Seifart, Frank. 2006. Orthography development. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 275–299. Berlin and New York: Mouton de Gruyter.

Stebbins, Tonya N. (compiler). 2001. *Sm'algyax Learner's Dictionary*. Prince Rupert: Tsimshian Language Authority.

Stebbins, Tonya N. 2003a. *Fighting language endangerment: Community directed research on Sm'algyax (Coast Tsimshian). With an introduction by Fumiko Sasama*. Suita, Osaka: The Endangered Languages of the Pacific Rim Project.[11]

Stebbins, Tonya N. 2003b. On the status of intermediate form classes: Words, clitics, and affixes in Sm'algyax (Coast Tsimshian). *Linguistic Typology* 7(3). 383-416.

Warner, Natasha, Quirina Luna, & Lynnika Butler. 2007. Ethics and revitalization of dormant languages: The Mutsun language. *Language Documentation & Conservation* 1(1). 58–76.

Wilkins, David. 2000. Even with the best of intentions: Some pitfalls in the fight for linguistic and cultural survival. In Francisco Queixales & Odile Renault-Lescure (eds.), *As linguas amazonicas Hoje*. Sao Paolo and Paris: Instituto Ambiental and IRD.

Winter, Barbara J. 1984. William Beynon and the anthropologists. *The Canadian Journal of Native Studies* 2. 279–292.

Woodbury, Anthony C. 2007. On thick translation in linguistic documentation. In Peter K. Austin (ed.), *Language Documentation and Description* 4, 120–135. London: School of Oriental and African Studies.

[11] Not available for commercial sale, but copies may be obtained by contacting the author. Please email t.stebbins@latrobe.edu.au.

Woodbury, Anthony C. & Nora C. England. 2004. Training speakers of indigenous languages of Latin America at a US university. In Peter K. Austin (ed.), *Language Documentation and Description* 2, 122–139.
Yamada, Racquel-María. 2007. Collaborative linguistic fieldwork: Practical application of the empowerment model. *Language Documentation & Conservation* 1(2). 257–282.

Tonya N. Stebbins
La Trobe University
t.stebbins@latrobe.edu.au

Birgit Hellwig
University of Erfurt and La Trobe University
birgit.hellwig@uni-erfurt.de