

DRAFT

ACCESSING AND MANAGING STONECORP

Prepared by: Jacob Straub

Rev. #: 0.3

Date: 04.22.19

Updated: 12.08.19

Approved: _____

Date: _____

Contents

Summary	3
Terms and Acronyms	3
Document Acronyms - Sources	3
Coding Acronyms - Perspectives	3
Coding Acronyms - Aspects of Policy	3
Accessing	3
Software	3
Folder Structure	4
ARCHIVE	4
ATLAS <i>ti</i> _ProjectFiles	4
Central Document	4
Corpus	4
TRASH	4
File and Corpus Structure	4
Status Tag	4
Terms	4
Codes	4
Codes - Statistics	5
The ‘Central’ Document	5
Central Document - Structure	6
Managing	7
Data Collection	7
Method of Annotation	7
Initial	7
Current	8
Coding	8
Additional Information	8
Archiving	8
Update Fall Semester 2019	9
Summary	9
Contact	9
Folders & Files	9
ITINERARY	9
TEXT LIBRARY	9
EDUCATORS	10
MAP	10
Work Done	10
Incomplete or Flawed Work	10
Folder Structure	11

Summary



This document serves to inform on accessing and managing the collection of texts comprising the Stonecorp corpus. Terminology section defines acronyms and cursory terms while accessing details the structure and location of working documents, and management details the methods of annotation, coding schema, and software currently applied to manipulate Stonecorp.

Terms and Acronyms

Corpus	Accumulation of texts pertaining to a central subject.
Coding	Word or short phrase that symbolizes an attribute for a selection of the corpus.
Concordance	List of words used in a text – every instance of each word with its immediate neighbors.

Document Acronyms - Sources

A/RBIC	Annual Report of the Board of Indian Commissioners
A/RCE	Annual Report of the Commissioner of Education
NS-DMS	A Day in the Mission at Sitka The North Star (Magazine)
A/RBENA	Report on the Work of the Bureau of Education for the Natives of Alaska
Book	...It's a book

Coding Acronyms - Perspectives

EA	External to Alaska—Federal government
SL	State-level administrator
T	Teacher
S	Student
IO	International observer

Coding Acronyms - Aspects of Policy

ER	Explicit rules (federal, state, classroom)
RP	Rationale for policy (civilization; Americanization; Industrialization; influences)
CP	Classroom practices & curricular materials
EE	Evaluation of effectiveness

Accessing

Software

[Adobe Acrobat](#) Used to export text file copies of the original scanned documents to be processed by AntConc's Concordancer or another text parser.

[AntConc's](#) Concordance feature and frequency word list were used to create a central document of phrasal structures.

[ATLAS.ti](#) Used to aid in the organization of annotations, coding, and facilitate connecting contextual excerpts along with a larger theme within larger collections of documents.

[7-Zip](#) Used for compression/archiving. Windows's native compression feature also suffices.

Folder Structure

ARCHIVE

Directory to store compressed ‘snapshots’ of Stonecorp’s comprising folders/files being worked on. See ‘archiving’ under ‘Additional Items’ for archiving process.

ATLAsTi_ProjectFiles

Stores the **exported** project file(s) from ATLAS.ti. It is not recommended to store the current working project file in a syncing as remote syncing from another team member editing the file may subsequently corrupt the project file. See ‘Archiving under ‘Additional Items’ for archiving process.

Central Document

Stores the ‘Central’ Document. See Below.

Corpus

Stores the folders and files containing the texts comprising Stonecorp.

TRASH

Files no longer pertinent to research progress – rather than being deleted they are cautiously moved to the current directory level’s *TRASH* directory.

File and Corpus Structure

The corpus follows a rudimentary folder/file organization:

STONECORP / Corpus / [Acronym - Publication body/Author / [Publication Year] - – [STATUS] / [[Publication Acronym] – [Year]].pdf
for example, ‘~\STONECORP\Corpus\RCE\1882\RCE-1882.pdf.

Status Tag

The status tag suffixes the publication year within the publication body/author directory within the corpus denoting the current working status of the sub-corpus contents. There are three values:

- Process** – The sub-corpora’s PDF file has not been annotated/highlighted through cross-referencing the ‘Central’ document returned excerpts; additionally, the contents have not been added to the corpus by querying the terms list through AntConc.
- Missing** – The documents within are incomplete – pertinent sections/pages are missing from the larger publication body of that year.
- Reprocess** – The sub-corpora’s PDF file produced an inaccurate text file export. An attempt should be made to re-acquire the source file(s).
- [Null]** – The files within are complete to the best of current knowledge and are annotated/highlighted.

Terms

A working text file containing the terms derived from AntConc’s word frequency list. Used to create the central document of phrases including the selected term(s). Refer to “Central’ Document” section for a table. Terms can is located in ~\ STONECORP\Schema\TERMS.txt

Codes

A working excel file - an export from ATLAS.ti containing all names of proposed codes. Located in ~\ STONECORP\Schema\STONECORP_CODES.xlsx

Codes - Statistics

A working excel file - an export from ATLAS.ti containing statistical information for each file within the sub-corpora: Contains the 'Document Name' and the amounts of each code encoded within. Located in ~\STONECORP\Schema\STONECORP_Stat_Export.xlsx

The 'Central' Document

The 'central' document is an excel spreadsheet – an initial export of the 'concordance' generated through AntConc; however, it by no means serves the role of a concordance – AntConc's simple wildcard characters allowed for the quick creation of a list of relevant contexts which could be exported in an organized manner. It goes beyond the purpose of a concordance which is to determine salient patterns of language use within the direct neighbors to a queried word.

This list of terms was created by referencing a word frequency list within AntConc; however, some terms generated a great number of false positives (e.g. 'school') and were omitted; additionally, After the initial filtering and analysis of the structures, it was discovered that the terms 'composition' and variations of 'speak' and 'Understand' were not included – this added an additional 500 results. Previous versions of these search terms may be referenced within the 'TERMS.txt' file within the root folder of Stonecorp. Note the highlighting of 'Tongus' in table 1 – this was a common misrecognition of 'tongue+' by Adobe Acrobat's character recognition.

Search Terms	
Language+	Conjunction+
Composition	Preposition+
Spell+++	Interjection+
Tongue+	Preposition+
Tongus	Pronoun+
Dialect+	Vernacular+
Speak+++	English++
Spoke+	Read+
Talk+++	Write
Understand+++	Wrote
Understood	Written
Utter+	Grammar+
Verb+	Grammatical+++
Adjective+	Instruct+++
Noun+	Russian
'+' denotes zero or more characters	

Table 1 - Search Terms Used in AntConc.

Individuals referencing this central document can sort by select sources, or by particular search terms, and follow the list of excerpts to determine their pertinence to the research subject matters; however, with the inclusion of qualitative data analysis software such as ATLAS.ti, the purpose of this document has generally run its course; however, it may be used as a quick reference tool of the pertinent excerpts found within the documents without referencing the PDFs. Context phrases are usually a couple of structures to a paragraph in length. The file can be accessed in: '~\STONECORP\Central Document\antconc_concordance_results*.xlsx'

ID	Condition	Excerpt	Source
1473 T		the Thinklet tongue. The pupils are making quite rapid progress in English through the teacher's ability to make his explanations in the dialect which they have learned at home. Those who for years have been making this highly commendable and necessary ARCE - 1886-7.txt	ARCE - 1886-7.txt
177 T		to their primers and to teach them something of reading, writing, figuring, to train them to say some prayers in their own dialect, and to sing some simple hymns. But these good intentions were not crowned with success. The Koloshes about Kiliisroo are pa ARCE - 1899-02.txt	ARCE - 1899-02.txt
243 T		1)1487 from 80 to 40 different groups of natives in this district, separated not only by long distances but by various dialects. Attendance-The school attendance, owing to the limited number of natives in the various communities, also to their migratory hnt RWBENA-1914-15.txt	RWBENA-1914-15.txt
453 T		paid for all their services. Language-The children speedily acquire an English-speaking vocabulary when strictly prohibited from using their native dialects. For five years English has been the, exclusive language of the school. Experience has removed all do ARCE - 1890-1.txt	ARCE - 1890-1.txt
1532 T		school, as they had often visited the school in company with older children. The first step was to give them a small English, 1487 vocabulary. This was done by teaching them the names of objects, of actions, and the use of short sentences. After ARBENA-1912-13.txt	ARBENA-1912-13.txt
17 T		the first lessons with perfect, accuracy, and can now understand the meaning of the greater part of what they read, and are talking English a little. They have been taught to analyze words phonetically, and when the teacher wants native words he can get the ARCE - 1888-9.txt	ARCE - 1888-9.txt
50 T		15. The teacher reports little progress with the natives, owing to their irregular school attendance and the difficulty they experienced in learning English after having learned Russian. Kodlak, 1407-141, and Mrs. C. J. Kerr, teachers, pupils, white and native, em ARCE - 1904-05.txt	ARCE - 1904-05.txt
175 T		school at Deering, within the Arctic Circle. There is no doubt in my mind that the Eskimo child can be taught to speak English and to acquire habits of cleanliness, industry, and morality. Though the advancement is necessarily slow, the Eskimo is a susceptible RCE-1905.txt	RCE-1905.txt
176 T		direction not leaves much to be desired. It is 14145 possible in the few hours to the schoolroom to teach enough English and to get it used enough to make it the medium of speech for the children and young people; the language used RWBENA-1914-15.txt	RWBENA-1914-15.txt
234 T		enrolled in the public school. The school population is Russian and Creole. The Russian language is the language of the people. English becomes a school language for the people. Our efforts to have them make English something more than the language of ARCE - 1901-02.txt	ARCE - 1901-02.txt
368 T		distinguishing between the consonants b, p, and d. Some of the more advanced pupils kept up their study and practice of English by writing a diary. A marked difference was noticed in the mental activity of the scholars during the sunless period, when ARCE - 1904-05.txt	ARCE - 1904-05.txt
318 T		the teaching of English. Writing (copy books) was enjoyed by nearly all the members of the two upper departments. Next in importance to English comes arithmetic. For the advanced section this has meant the gaining of an insight us to what numbers really ARCE - 1899-00.txt	ARCE - 1899-00.txt
323 T		, subtract, multiply, and divide. The boy was in higher fractions, had read some history and could read English with ease. He could write English compositions, draw, and construct maps and charts. Unalaska, 1497 Teachers, William A. Davis and Mary S. Davis, pi ARCE - 1904-05.txt	ARCE - 1904-05.txt
335 T		much interest. The teacher reports that he never knew children prouder or better satisfied than when pupils who had previously known nothing of English could sing a few verses or read a lesson nicely from their first reader; that just as soon as they were ab ARCE - 1886-7.txt	ARCE - 1886-7.txt
347 T		energy and zeal to accomplish in the way of teaching the pupils the English language, education, together with and the rudiments of an English diligent effort at improving their moral condition. The entire management of the institution deserves great credit ARCE - 1886-7.txt	ARCE - 1886-7.txt
356 T		, Belkofski, 1497. A. Golder, teacher, enrollment, 40, population, Aleut and creole. This is one of the very few summer schools intended to teach English during the summer as a supplement to the work of the Russian parochial school in the winter. The popular ARCE - 1901-02.txt	ARCE - 1901-02.txt
367 T		streams, the opening of mines, etc., this more liberal and broader education along both intellectual and industrial lines was necessary. A common English education is given. The girls are trained in the domestic industries so that they are fitted to become i ARCE - 1898-9.txt	ARCE - 1898-9.txt
377 T		trust that in the near future the Russian children will become regular attendants of the Government school and enjoy the privileges of the English education which it affords and which they so much need. Unalaska, 1497. R. McKinney, teacher, enrollment, 40, pop ARCE-1897-8.txt	ARCE-1897-8.txt
378 T		influence in shaping character, in connection it, this report, uO do more than speak situation, and then only has as is, with the English education which the contract requires to be imparted to the pupils, have had numerous opportunities of observing the d ARCE - 1886-7.txt	ARCE - 1886-7.txt

Figure 1- Sample of 'Central' Document

Central Document - Structure

Worksheets

The Central document includes multiple worksheets. This is due to the inclusion of additional terms after determination of excerpt relevance has begun as it was unknown whether

‘Excerpts_03.15.19’ – Contents include AntConc’s results from AntConc – See Table 1 for terms used.

Excerpts_OLD_02.17.19

The previous results of the initial compilation of terms searched through the corpus. This worksheet should remain as it contains the initial relevancy values (see Excerpt subsection of Columns).

Excerpts_OLD_02.17.19-Sandbox

A Working copy of the preceding. This worksheet was created to prevent any detrimental/irreversible/unintentional changes made so there would be no loss of data; additionally, it was created to find a method to merge or transition the ‘relevant’ values of columns matching ‘Excerpt_b’ values from ‘Excerpts_OLD_’ to ‘Excerpts_03.15.19’. It was created to find a method within excel to run a string comparison, find matching strings and then copy the row’s relevancy value; unfortunately, limited knowledge of Excel’s formula/ macro language prevents this.

Formula Sandbox

A Blank sheet which would include any cross-worksheet formulas.

Column Data

ID

The unique numerical ID assigned by AntConc as results were compiled. Its values hold little relevancy to the deeper understanding of the corpus save for the it reflects what order the excerpt was identified containing a relevant keyword.

Relevant

Values include T/rue, F/false, N/eutral, and the hyphen (‘-’) which denote the excerpt’s relevancy to the scope of research.

- T/rue reflected a excerpt which structure as is or in combination with its larger context was deemed relevant to the topic inquired by the research.
- F/false, of course, reflected the opposite nature of T/rue.
- N/eutral reflected a context which required the confirmation of Dr.Stone;
- Hyphen (‘-’) identified contexts which were redundant, either within or next to an excerpt which was already identified to be relevant or not. The intern assigned this task later got lazy and would skip the duplicate rows....

Excerpt/eab

Excerpt is the combination of the String values in ‘Excerpt_a’ and ‘Excerpt_b’ to aid in initial reading. It is an optional field if Excerpt_a is right justified and Exerpt_b is left justified

Source

Identifies the source file/sub-corpus the excerpt was found in – used to cross reference the larger context when the 150 characters did not provide enough.

Hidden Columns

Excerpt_a/ea

The half of the 150-character excerpt preceding the returned keyword.

Excerpt_B/eb

The following half of the 150-character excerpt including the returned keyword.

Source_File_Code

Redundant column which became irrelevant once sub corpora labeling followed current naming schemes - see 'File and Corpus Structure' section. Values are those included under 'Terms – Acronyms – Sources': (A/RBIC, A/RCE, NS-DMS, Book).

Perspective_Code

Redundant column to be used for coding of the excerpt - became irrelevant once coding began in ATLAS.ti as the feature is built into the coding aspects of the software.

Values are those included under 'Coding Acronyms – Perspective' - (EA,SL,T,S,IO).

Policy_Aspect_Code

Redundant column to be used for coding of the excerpt - became irrelevant once coding began in ATLAS.ti as the feature is built into the coding aspects of the software.

Values are those included under 'Coding Acronyms - Aspects of Policy' – (ER,RP,CP,EE)

Reference String

An unused column which would compile the preceding values of 'Source_File_Code, Perspective_Code, and Policy_Aspect_Code' to create an easy reference string. This became irrelevant once coding began in ATLAS.ti as the feature is built into the software.

Managing

Data Collection

Government documents were collected through the [Hathi Trust Digital Library](#). As the University of Alaska is not a partnered institution, collection was limited to opening individual tabs – for each page for the selected year of publication. These files were then downloaded to a temporary folder to be bound within Adobe Acrobat, labeled, and appropriately 'filed' according to the corpus structure; additionally, the newly bound document would need to be cropped of watermarks as they affect the accuracy of Adobe Acrobat's character recognition;

These PDF files would then be exported to a text file copy which could be used by AntConc to generate word frequency list to determine words which could be used to mass query the corpus to create a list of pertinent structures; however, for half of this process the original, unbound, files were not retained. This was amended should a future method require the original, un-cropped, files.

Method of Annotation

Initial

The relevant excerpts of the central document were then cross-referenced alongside their corresponding PDF file in order to highlight and annotate the larger context to be later coded

either within the central document or a qualitative data analysis software suite. Future annotation methods may either occur directly within ATLAS.ti's document view or by cross-referencing

Current

Current annotation takes place within ATLAS.ti. Like highlighting in an e-reader or PDF, this creates a 'quotation' (It's a highlight) which can then be assigned codes. These codes can have hierarchical relations – where codes can be a subset of larger themes or groups.

Coding

A 'code' in qualitative research is often a word or phrase that represents a characteristic or context of language. This can consist of transcripts, ethnography field notes, journals, documents, literature, photographs, video, online media such as email and webpages, and the like. Coding is relatedly straight forward – the highlights from the initial annotation within Adobe Acrobat transfer over to ATLAS.ti - as such, pertinent excerpts appear yellow and merely need to be coded by selecting the highlighted region and drag-n-dropping the relevant codes upon the highlight; however, in some older low-resolution or quality scans, the process of highlighting may inadvertently include larger structures which may not be pertinent to the current coding.

Originally, the ATLAS.ti codes included Enforcement, Explicit_Rules, Pedagogy, Policy_Justification, and Resistance; however, as a test, this coding schema was extended slightly to differentiate industrialization and westernization from rationals for policy; additionally, codes to include past Russian influence (Past_Russian) and the code to include descriptions of the native populace (Population_Description) were added; lastly, code groups were created to represent three tiers of causation.

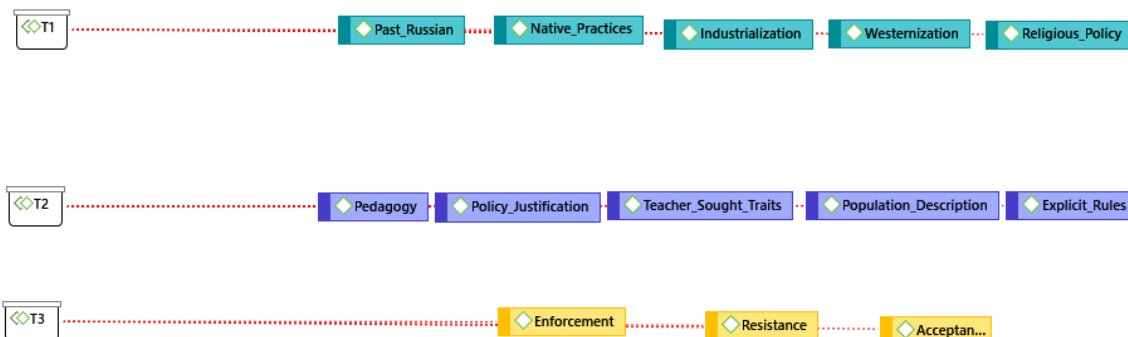


Figure 2 - Extended Coding Schema

Additional Information

Archiving

Ideally, in areas where large changes in structure or method have occurred, an additional folder labeled ARCHIVE or TRASH has been added. 'ARCHIVE' serves to store 'snapshots' of the current level's contents given a period of time which can alter be reviewed to identify. This does not serve as a backup solution – merely to maintain a 'timeline' of a folder/file's development

'Trash', in contrast, serves to store folders and files that have been deemed irrelevant – but understandably are not deleted as to maintain a sort of 'timeline' of the projects development.

- Fall Semester 2019 Update -

Summary

This section update serves to inform on the structure, utilization, and status of work furthering the information within StoneCorp. Updated work includes completion of a list Mission Educators stationed at schools across Alaska; additionally, work was done towards creating a robust archive repository tabulating pertinent archive, their locations, and additional sheets tabulating pertinent collections or entries - organized by denomination.

Contact

Should any questions arise regarding the structure or means of accessing the following information – contact Jacob Straub (907) 841-3263 / danestraub@gmail.com.

Folders & Files

Work folders have been moved to be managed under one folder titled ‘Sandbox’ - It serves as the working directory for all other project efforts.

At the end of Fall Semester 2019 all worksheets were flagged as ‘protected’ to avoid possible mis-entry of data. There is no associated password, one only need to go to ‘Review > ‘Unprotect Workbook’ to create changes to the documents.

IT IS HEAVILY RECOMMENDED THAT BACKUPS ARE CREATED ON A REGULAR SCHEDULE.

ITINERARY

‘ITINERARY’ has been the main working directory for Fall 2019. Its purpose is to manage a growing list of archive repositories, their locations, pertinent archive entries or collections, and their means of access for later StoneCorp research. The directory includes subdirectories labeled after denominations of interest, apart from ‘State’, ‘National’, and ‘Third Party’ which include repositories that are not directly affiliated with a specific denomination. These subdirectories contain relevant organizations, their contact information, and pertinent collection guides.

Further information on individual entries be accessed from the ‘ITINERARY.xlsx’ Excel Workbook – which is organized in a similar manner. The first worksheet is a table of Educators referenced in the Reports to the Commissioner of Education. The second worksheet is a table of pertinent archive repositories including major collections.

TEXT LIBRARY

This folder contains the digital copies of Reports to the Commissioner of education and other like entities. Near the end of the semester another online resource was found – the [Orbis Cascade Alliance’s Archive West](#). This archive has not been searched thoroughly but preliminary queries have been very promising. This archive consortium, *focusing on correspondence, diaries, and manuscripts*, had digital copies of Reports to the Governor of Alaska for download – these have been added to the Text Library under the ‘RGA’ directory. They are searchable; however, the low-quality scans left many artifacts; as such, character recognition has not been very accurate.

EDUCATORS

‘EDUCATORS’ is an Excel workbook containing a snapshot of Rose Kruger’s previous work in compiling a list of educators referenced within Reports to the Commissioner of Education. It also includes multiple sheets that were used by Jacob to compile additional information while keeping previous work separate as to minimize chance of mis-entering data within an existing field. While the Educator list can be accessed separately, it has also been included as a sheet within the Itinerary book. once added as a **sheet within ITINERARY.xlsx**, data was added and changed at a later date and **should be considered the more ‘up-to-date’ version.**

It is strongly recommended to retain EDUCATORS.xlsx for archival reference.

MAP

‘MAP’ includes a custom Google map presenting the information found within ITINERARY.xlsx’s ‘Repository Directory’ sheet as an interactive map - should one grow tired of looking at Excel rows. As this map is currently tied to Jacob’s University Google account, the information has been exported to a (Keyhole Markup Language) KML file. Should project members lose access to this map and wish to create another, the KML file may be imported to any Google Maps or Google Earth application – A guide can be found [here](https://support.google.com/mymaps/answer/3024836) - <https://support.google.com/mymaps/answer/3024836>

Work Done

Educators

Compiling a list of educators referenced within the previously accumulated Text Library and creating a list of archive repository locations which contain entries or references to thematic or aforementioned educators with the goal being to create an itinerary for future archive visits by Dr. Stone.

Itinerary

Continuing Rose Kruger efforts, a list was completed containing a list of educators as they appear with Reports to the Commissioner of Education – These were to be tabulated along with their denomination, place of origination, their station, years present, their position, and miscellaneous notes. Once rows, which included the same name but different years, were combined there were over 300+ unique entries.

Incomplete or Flawed Work

While searches for names in the education table were performed, they were not, in sum, exhaustive or complete. Work needs to continue querying names to archive repositories. This is due mostly to an error in Jacob’s search approach – educator name searches were far too restrictive in the date range – limiting possible results; as such, all work towards this goal should be considered incomplete, flawed, and redone.

DRAFT

Folder Structure

For an appreciation on the growth of StoneCorp from start – a folder tree.

