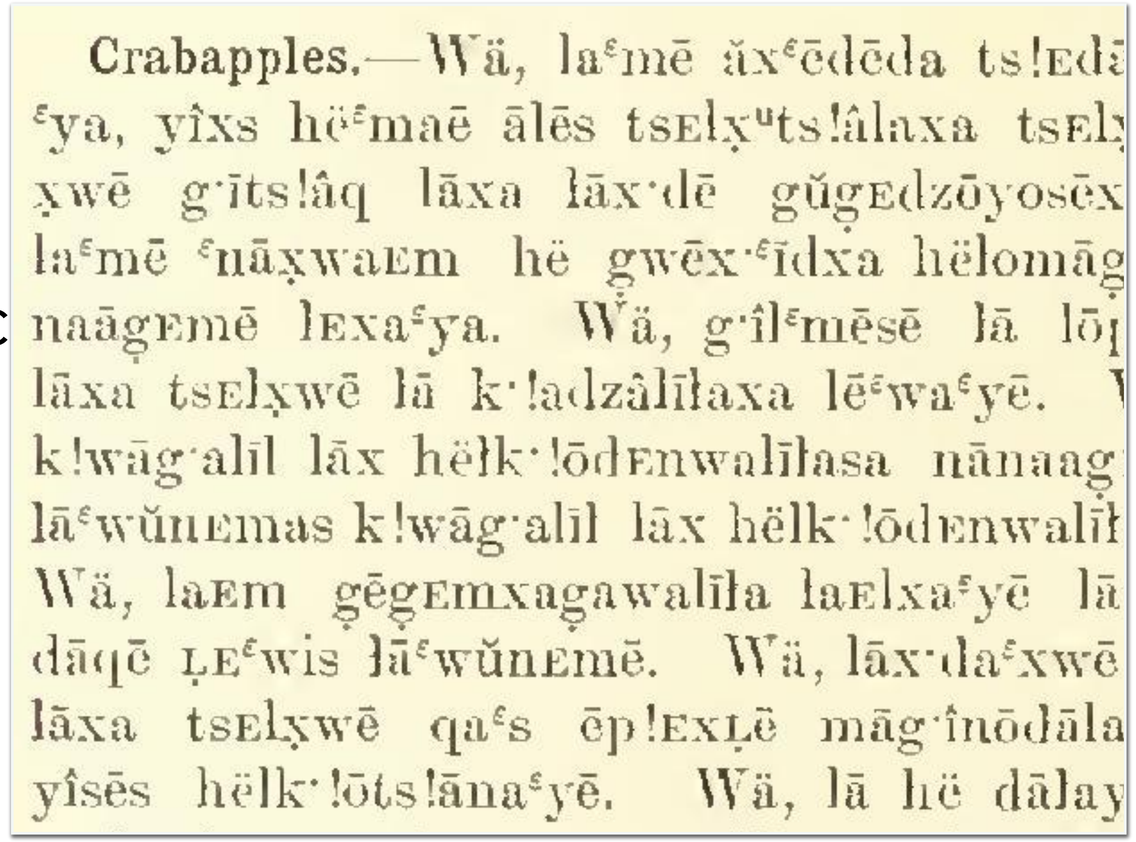
Developing a Mixed-Methods Pipeline for Community-Oriented Digitization of Kwakʼwala Legacy Texts

Milind Agarwal, Daisy Rosenblum, Antonios Anastasopoulos

Improving OCR for Kwak̓wala Texts

- OCR for legacy data in Kwak̓wala
- Why?
  - Rich Written Tradition
  - High value for community researc
  - Digital files available only as scanned images



Crabapples.—Wä, laᵋmē ăxᵋēdēda ts!Edā
ᵋya, yîxs hëᵋmaē ālēs tsElx̣ᵘts!âlaxa tsEl
x̣wē g·îts!âq lāxa łāx·dē gŭgEdzōyosēx
laᵋmē ᵋnāx̣waEm hë gwēx·ᵋīdxa hëlomāg
naāgEmē lExaᵋya. Wä, g·îlᵋmēsē lā lōp
lāxa tsElx̣wē lā k·!adzâlīłaxa lēᵋwaᵋyē.
k!wāg·alīl lāx hëłk·!ōdEnwalīłasa nānaag
lāᵋwŭnEmas k!wāg·alīl lāx hëlk·!ōdEnwalīł
Wä, laEm gēgEmxagawalīła laElxaᵋyē lā
dāqē ḷEᵋwis łāᵋwŭnEmē. Wä, lāx·daᵋxwē
lāxa tsElx̣wē qaᵋs ēp!ExḷĒ māg·înōdāla
yîsēs hëlk·!ōts!ānaᵋyē. Wä, lā hë dālay

Project Team

Milind Agarwal
PhD Candidate
George Mason University

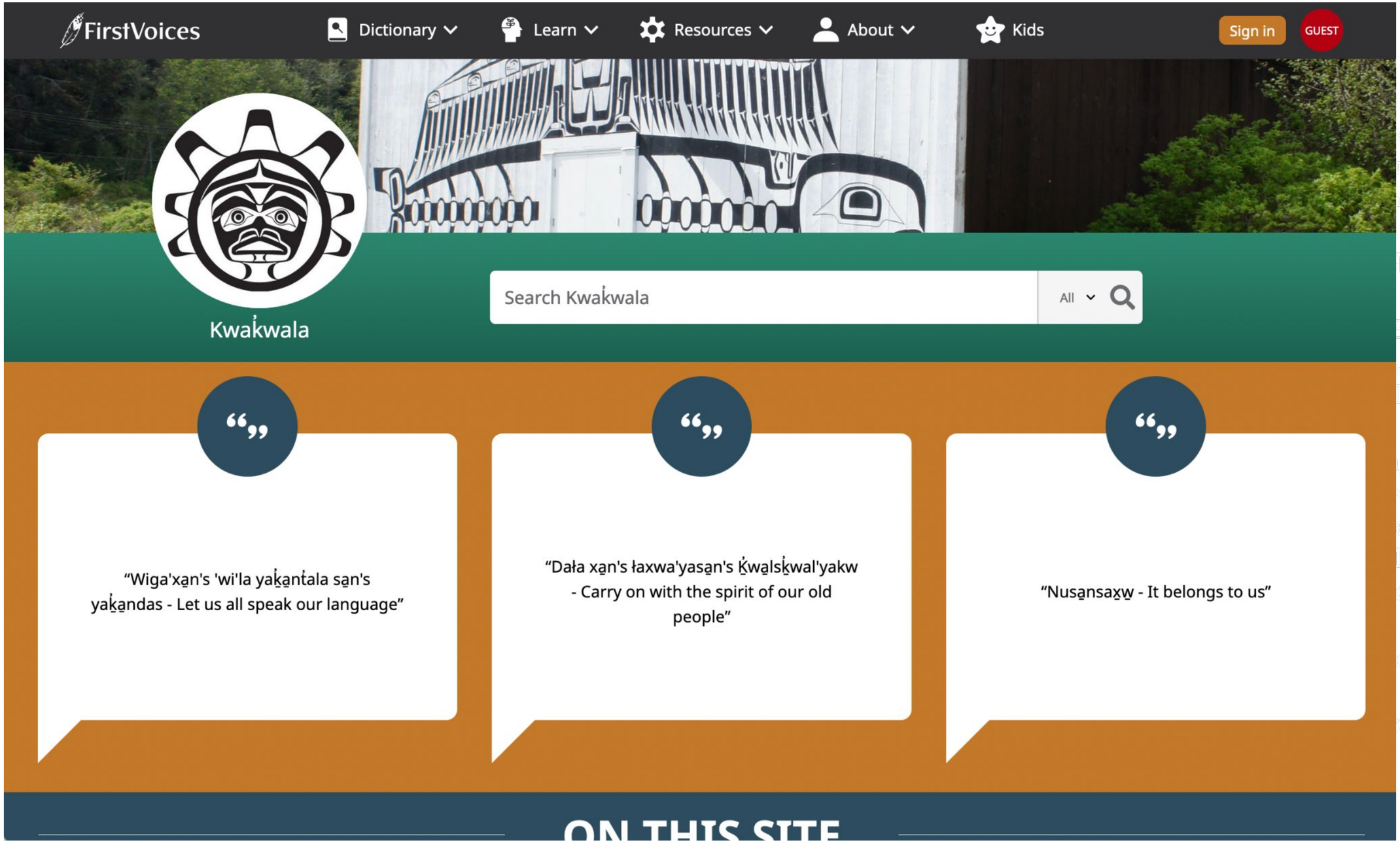
Antonios Anastasopoulos
Assistant Professor (CS, NLP)
George Mason University

Daisy Rosenblum
Assistant Professor (Indigenous Studies, Anthropology)
University of British Columbia

Kwakwa̱ka̱’wakw communities, Kwak’wala language reclamation

- Wakashan language indigenous to North Vancouver Island and opposing mainland in British Columbia (Canada)
- 18 distinct Nations, 5 dialects
- Robust learner community; multiple revitalization efforts.
- >3 orthographies, 2 most common
- Strong tradition of documentation, both internal and external
- Impacted by colonial occupation, forced relocation, removal from land, residential school system

Kwakwa̱ka̱’wakw communities, Kwak̓wala language reclamation



Value to Language Revitalization

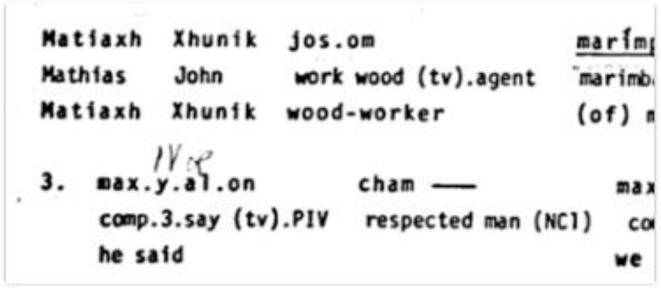
- Manual re-transcription is the norm
- If OCR works well, it can:
  - Reduce manual effort
  - Allow search, increase community access
  - Machine-translation enable NLP applications (language modeling, ASR, transliteration…)

Elder Speakers' Gathering, Gwa'sala-'Nakwaxda'xw Elders' Centre, November 12, 2009

Jaymyn LaVallée, Michayla King
Gwa'sala-Nak'waxda'xw Language Revitalization Program

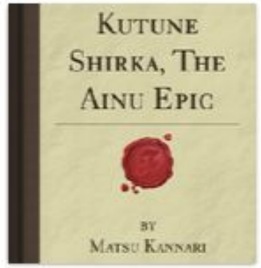
- Allow learners and speakers to spend more time together doing their most important work

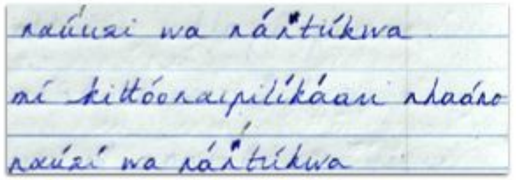
Machine-readable text has many benefits



Matiaxh Xhunik jos.om marfm
Mathias John work wood (tv).agent marimb
Matiaxh Xhunik wood-worker (of) n

3. max.y.al.on cham —— max
comp.3.say (tv).PIV respected man (NCl) co
he said we





Digitizing these texts to machine-readable formats

Support communities that speak these languages

Make native texts digitally accessible and searchable

Automatic orthography conversion

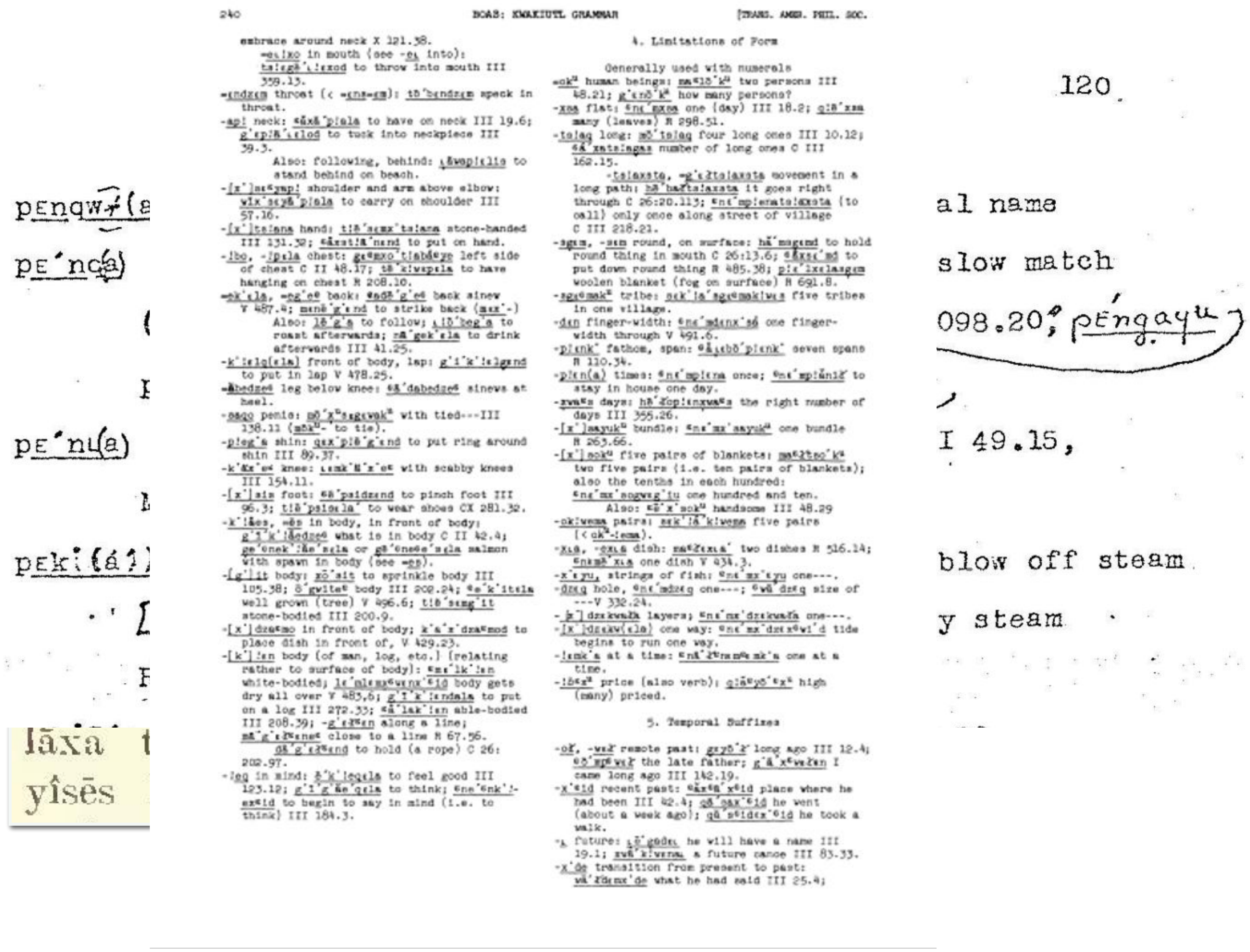
Aid language researchers, educatators, libraries…

Enable NLP for endangered languages

Annotate datasets for downstream NLP tasks

Build NLP systems to support language revitalization efforts

Input data



Bilingual (English & Kwak̓wala)

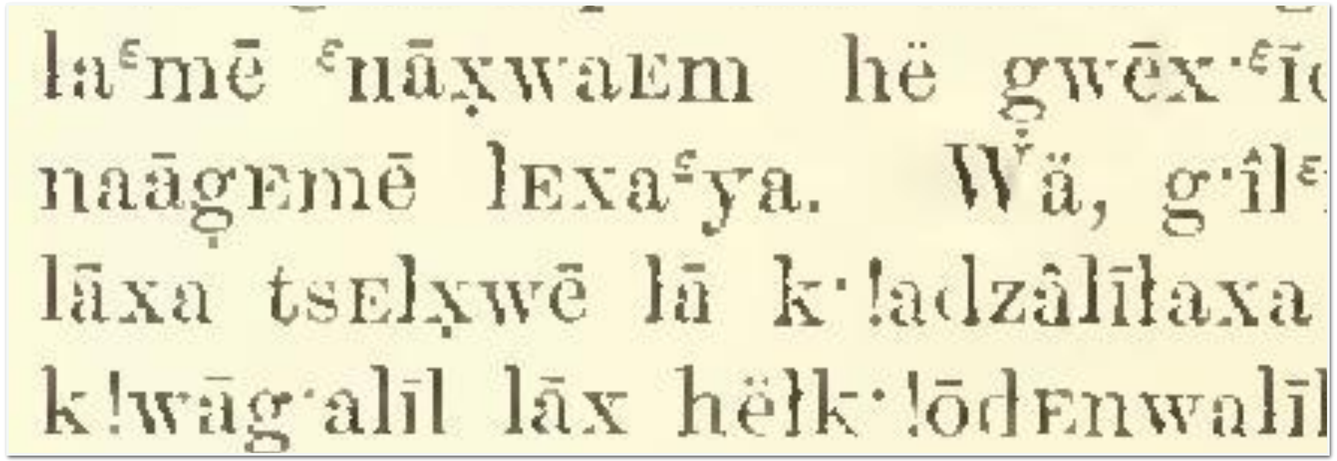
Boasian Trilogy: George Hunt & Franz Boas collaboration

1897-1948

over 11 published volumes (~10000 pages)

Additional trove of unpublished archival text (type- & hand-written)

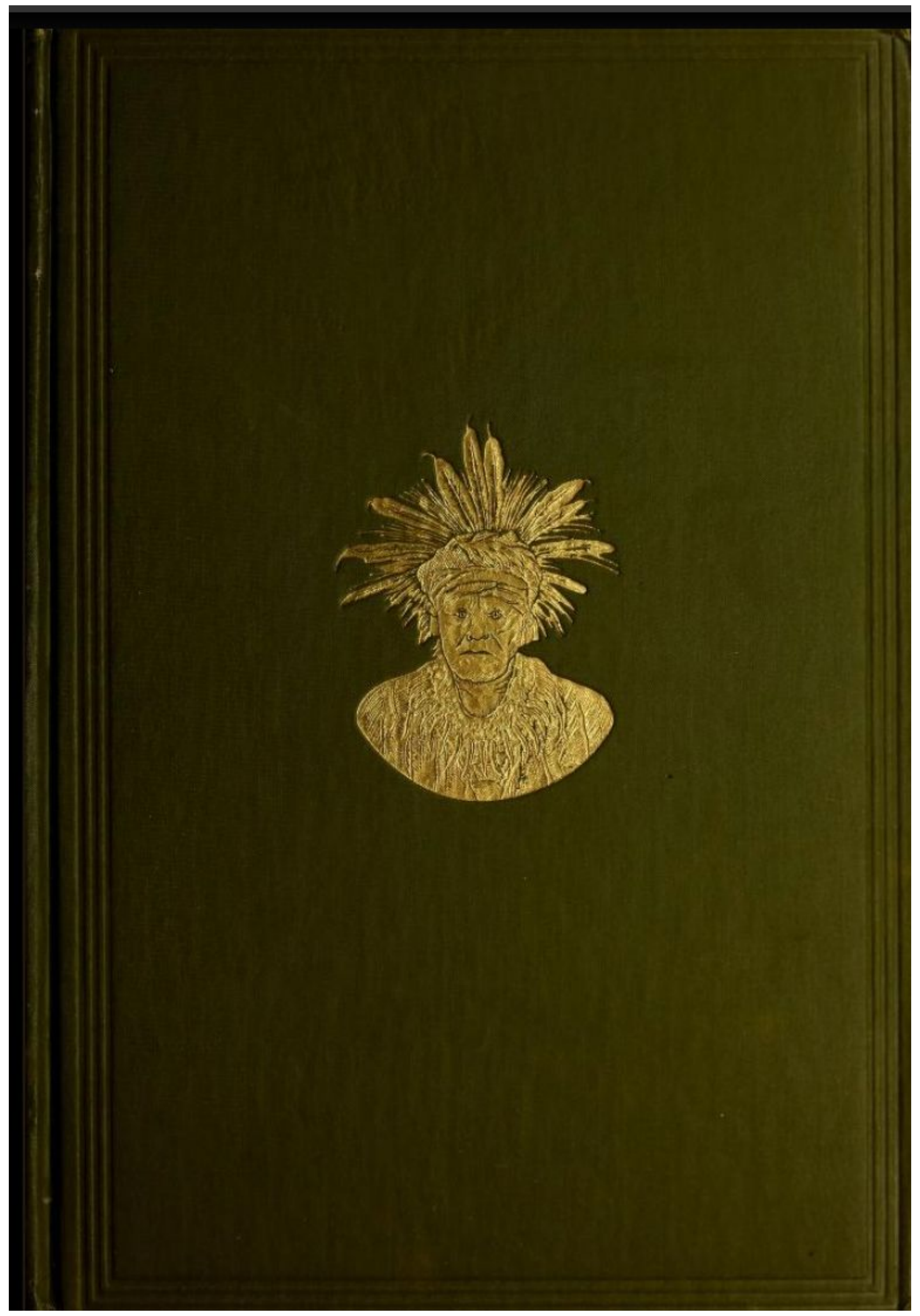
Initial Challenges

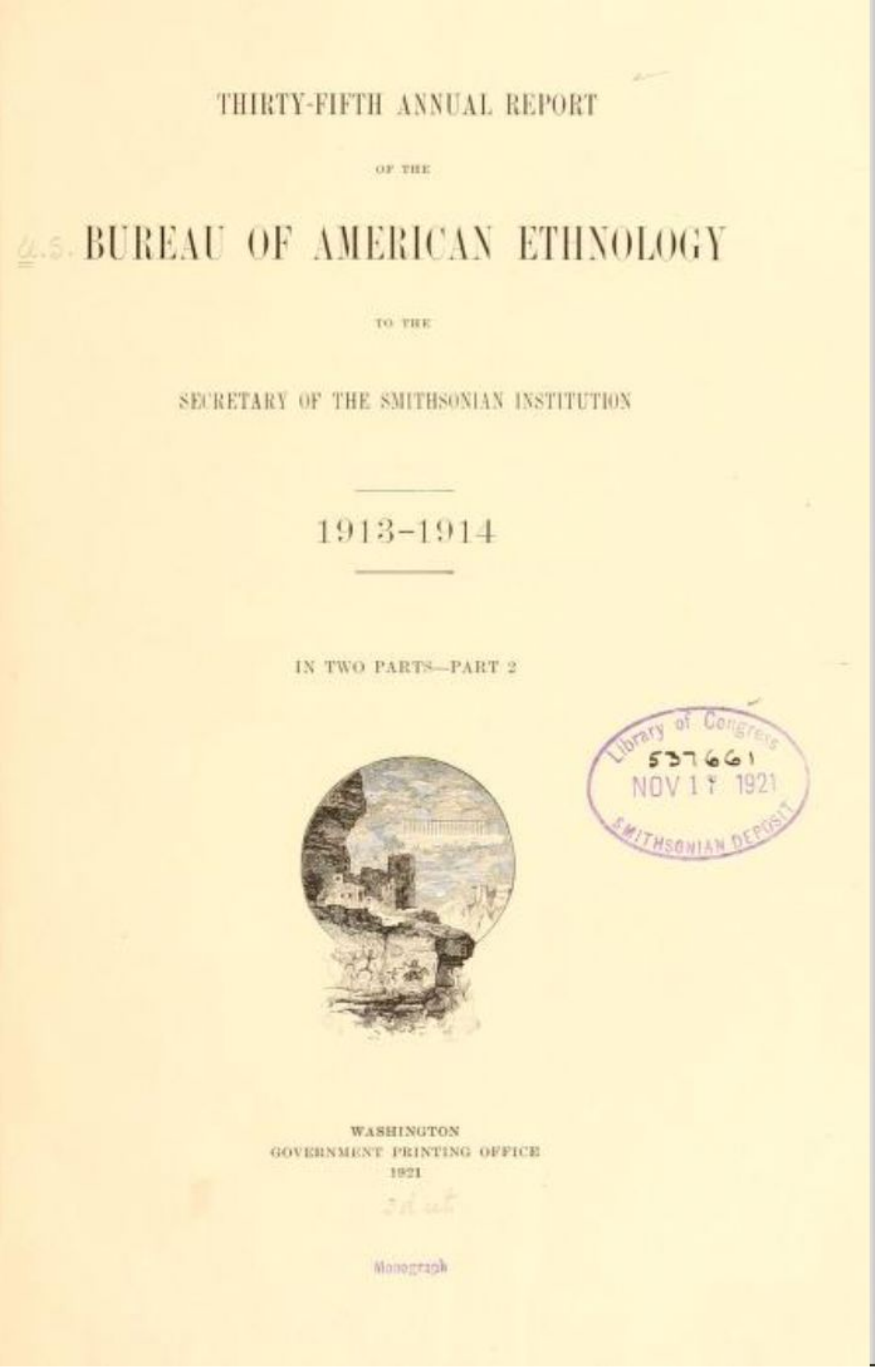


- Existing OCR systems don't work very well on many under-resourced languages
- Even pretrained OCR systems that support 30+ scripts do not support the early 20th c ethnographic writing systems
- More difficult on old books and low-quality scans
- Difficult to train a good model without data!

Text Selection: Community-determined

Boas, Franz, and George Hunt. 1921. “Ethnology of the Kwakiutl. Based on Data Collected by George Hunt.” In *Thirty-Fifth Annual Report of the Bureau of American Ethnology Presented to the Secretary of the Smithsonian Institution*, parts 1 and 2, 42–794, i–xi [index]. Washington, DC : Government Printing Office.





THIRTY-FIFTH ANNUAL REPORT

OF THE

BUREAU OF AMERICAN ETHNOLOGY

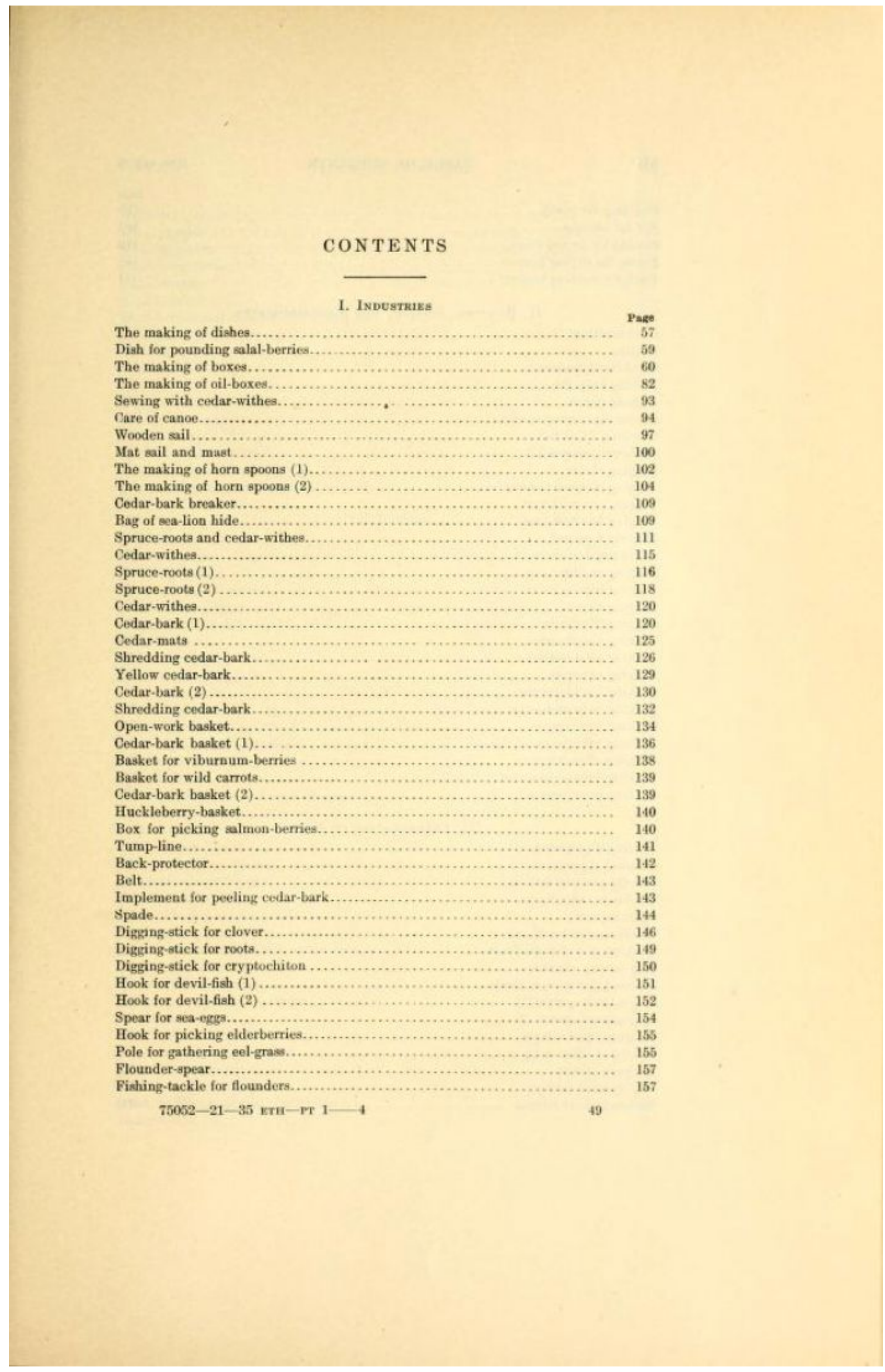
TO THE

SECRETARY OF THE SMITHSONIAN INSTITUTION

1913–1914

IN TWO PARTS—PART 2

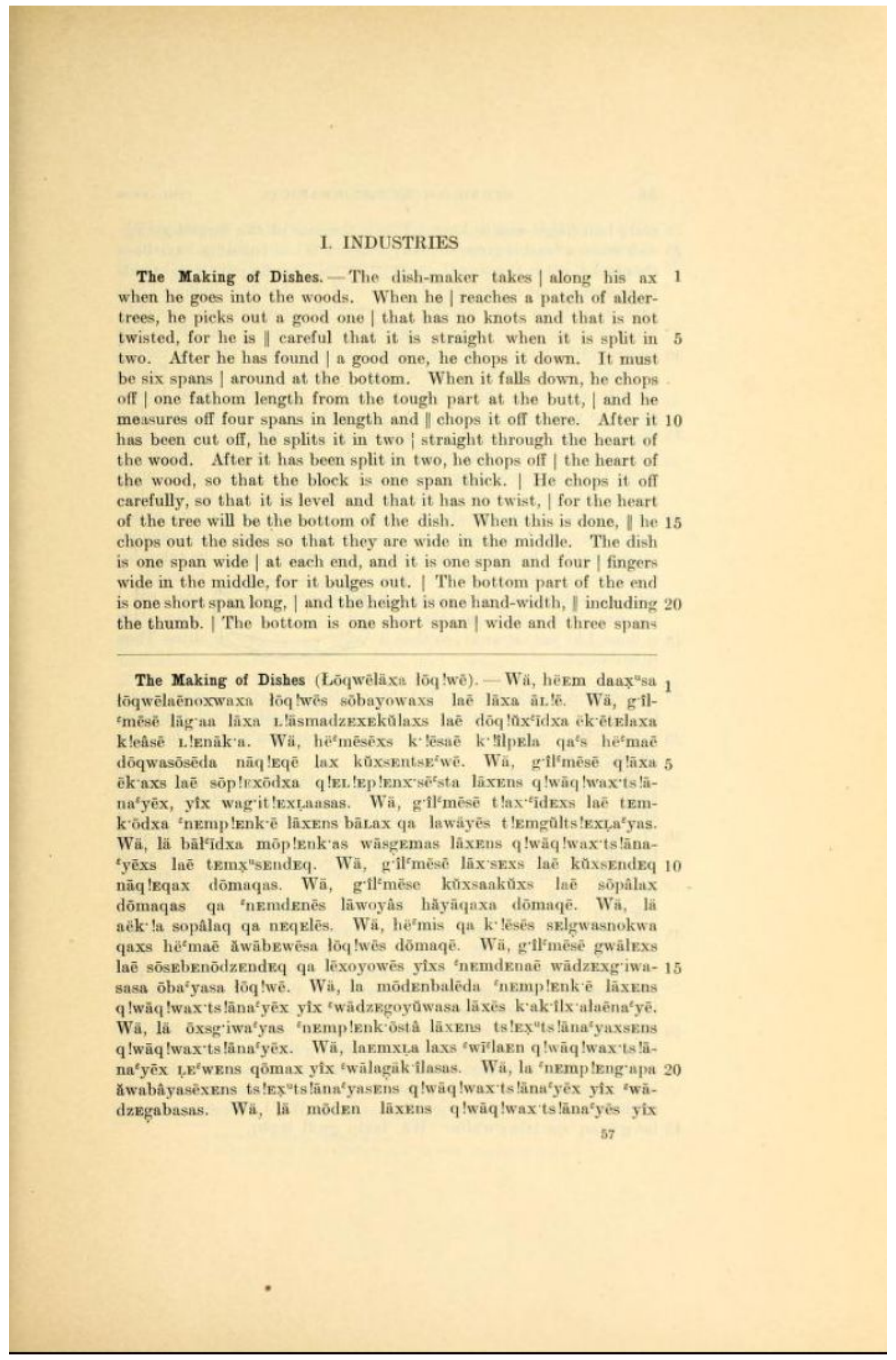
WASHINGTON
GOVERNMENT PRINTING OFFICE
1921



CONTENTS

I. Industries

75052—21—35 ETH—PT 1——4



I. INDUSTRIES

The Making of Dishes.—The dish-maker takes | along his ax when he goes into the woods. When he | reaches a patch of alder-trees, he picks out a good one | that has no knots and that is not twisted, for he is ‖ careful that it is straight when it is split in two. After he has found | a good one, he chops it down. It must be six spans | around at the bottom. When it falls down, he chops off | one fathom length from the tough part at the butt, | and he measures off four spans in length and ‖ chops it off there. After it has been cut off, he splits it in two | straight through the heart of the wood. After it has been split in two, he chops off | the heart of the wood, so that the block is one span thick. | He chops it off carefully, so that it is level and that it has no twist, | for the heart of the tree will be the bottom of the dish. When this is done, ‖ he chops out the sides so that they are wide in the middle. The dish is one span wide | at each end, and it is one span and four | fingers wide in the middle, for it bulges out. | The bottom part of the end is one short span long, | and the height is one hand-width, ‖ including the thumb. | The bottom is one short span | wide and three spans

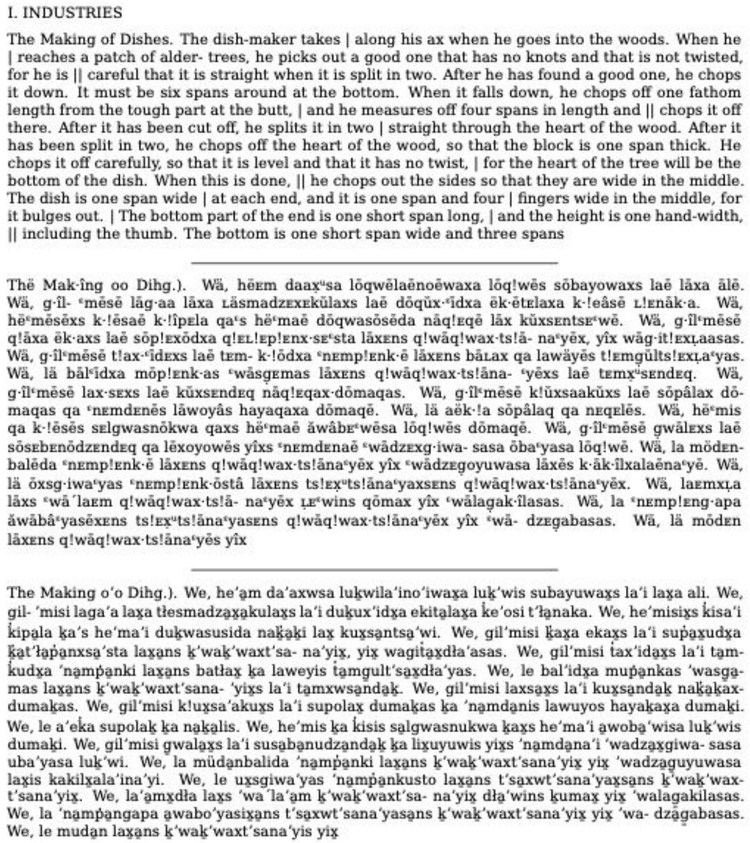
---

The Making of Dishes (Łōqwēlăxa łōq!wē).—Wä, hēᴇm daax̣ᵘsa łōqwēlaēnoxwaxa łōq!wēs sōbayowaxs laē lāxa ăʟ!ē. Wä, g·îlᶜmēsē lăg·aa lāxa ʟ!ăsmadzᴇxᴇkŭlaxs laē dōq!ŭxᶜīdxa ĕk·ētᴇlaxa k!eâsē ʟ!ᴇnāk·a. Wä, hēᶜmēsēxs k·!ēsaē k·!ilpᴇla qaᶜs hēᶜmaē dōqwasōsēda nāq!ᴇqē lax kŭxsᴇntsᴇᶜwē. Wä, g·îlᶜmēsē q!āxa ēk·axs laē sōp!ᴇxōdxa q!ᴇʟ!ᴇp!ᴇnx·sēᶜsta lāxᴇns q!wāq!wax·ts!ānaᶜyēx, yîx wag·it!ᴇxʟaasas. Wä, g·îlᶜmēsē t!ax·ᶜīdᴇxs laē tᴇmk·ōdxa ᶜnᴇmp!ᴇnk·ē lāxᴇns bāʟax qa lawäyēs t!ᴇmgŭlts!ᴇxʟaᶜyas. Wä, lä bālᶜīdxa mōp!ᴇnk·as wāsgᴇmas lāxᴇns q!wāq!wax·ts!ānaᶜyēxs laē tᴇmx̣ᵘsᴇndᴇq. Wä, g·îlᶜmēsē lāx·sᴇxs laē kŭxsᴇndᴇq nāq!ᴇqax dōmaqas. Wä, g·îlᶜmēse kŭxsaakŭxs laē sōpâlax dōmaqas qa ᶜnᴇmdᴇnēs lāwoyâs hăyāqaxa dōmaqē. Wä, lä aēk·!a sopâlaq qa nᴇqᴇlēs. Wä, hēᶜmis qa k·!ēsēs sᴇlgwasnokwa qaxs hēᶜmaē ăwābᴇwēsa łōq!wēs dōmaqē. Wä, g·îlᶜmēsē gwālᴇxs laē sōsᴇbᴇnōdzᴇndᴇq qa lēxoyowēs yîxs ᶜnᴇmdᴇnaē wādzᴇxg·iwasasa ōbaᶜyasa łōq!wē. Wä, la mōdᴇnbalēda ᶜnᴇmp!ᴇnk·ē lāxᴇns q!wāq!wax·ts!ānaᶜyēx yîx ᶜwādzᴇgoyŭwasa lāxēs k·ak·îlx·alaēnaᶜyē. Wä, lä ōxsg·iwaᶜyas ᶜnᴇmp!ᴇnk·ōstâ lāxᴇns ts!ᴇx̣ᵘts!ānaᶜyaxsᴇns q!wāq!wax·ts!ānaᶜyēx. Wä, laᴇmxʟa laxs ᶜwīᶜlaᴇn q!wāq!wax·ts!ānaᶜyēx ʟᴇᶜwᴇns qōmax yîx ᶜwālagāk·îlasas. Wä, la ᶜnᴇmp!ᴇng·apa ăwabâyasēxᴇns ts!ᴇx̣ᵘts!ānaᶜyasᴇns q!wāq!wax·ts!ānaᶜyēx yîx ᶜwādzᴇgabasas. Wä, lä mōdᴇn lāxᴇns q!wāq!wax·ts!ānaᶜyēs yîx

First Phase

- Post-correction OCR model
- CER improved from 30% > 4%!
- Automated conversion from Boas > U’mista
- 1401 separate txt files (66.txt, 67.txt…)

- Compiled first draft PDF for distribution



I. INDUSTRIES

The Making of Dishes. The dish-maker takes | along his ax when he goes into the woods. When he | reaches a patch of alder- trees, he picks out a good one that has no knots and that is not twisted, for he is || careful that it is straight when it is split in two. After he has found a good one, he chops it down. It must be six spans around at the bottom. When it falls down, he chops off one fathom length from the tough part at the butt, | and he measures off four spans in length and || chops it off there. After it has been cut off, he splits it in two | straight through the heart of the wood. After it has been split in two, he chops off the heart of the wood, so that the block is one span thick. He chops it off carefully, so that it is level and that it has no twist, | for the heart of the tree will be the bottom of the dish. When this is done, || he chops out the sides so that they are wide in the middle. The dish is one span wide | at each end, and it is one span and four | fingers wide in the middle, for it bulges out. | The bottom part of the end is one short span long, | and the height is one hand-width, || including the thumb. The bottom is one short span wide and three spans

---

Thë Mak·îng oo Dihg.). Wä, hëɛm daaχᵘsa lōqwēlaënoēwaxa lōq!wēs sōbayowaxs laē lāxa ālē. Wä, g·îl- ᶜmēsē lāg·aa lāxa ʟāsmadzɛxɛkūlaxs laē dōqūx·ᶜîdxa ēk·ētɬlaxa k·!eâsē ʟ!ɛnāk·a. Wä, hēᶜmēsēxs k·!ēsaē k·!îpɛla qaᶜs hēᶜmaē dōqwasōsēda nāq!ɛqē lāx kūxsɛntsɛᶜwē. Wä, g·îlᶜmēsē q!āxa ēk·axs laē sōp!ɛxōdxa q!ɛʟ!ɛp!ɛnx·sɛᶜsta lāxɛns q!wāq!wax·ts!ā- naᶜyēx, yîx wāg·it!ɛχʟaasas. Wä, g·îlᶜmēsē t!ax·ᶜîdɛxs laē tɛm- k·!ōdxa ᶜnɛmp!ɛnk·ē lāxɛns bāʟax qa lawāyēs t!ɛmgūlts!ɛχʟaᶜyas. Wä, lä bālᶜîdxa mōp!ɛnk·as ᶜwāsgɛmas lāxɛns q!wāq!wax·ts!āna- ᶜyēxs laē tɛmχᵘsɛndɛq. Wä, g·îlᶜmēsē lax·sɛxs laē kūxsɛndɛq nāq!ɛqax·dōmaqas. Wä, g·îlᶜmēsē k!ūxsaakūxs laē sōpâlax dōmaqas qa ᶜnɛmdɛnēs lāwoyâs hayaqaxa dōmaqē. Wä, lä aēk·!a sōpâlaq qa nɛqɛlēs. Wä, hēᶜmis qa k·!ēsēs sɛlgwasnōkwa qaxs hēᶜmaē āwâbɛᶜwēsa lōq!wēs dōmaqē. Wä, g·îlᶜmēsē gwālɛxs laē sōsɛbɛnōdzɛndɛq qa lēxoyowēs yîxs ᶜnɛmdɛnaē ᶜwādzɛxg·îwa- sasa ōbaᶜyasa lōq!wē. Wä, la mödɛn-balēda ᶜnɛmp!ɛnk·ē lāxɛns q!wāq!wax·ts!ānaᶜyēx yîx ᶜwādzɛgoyuwasa lāxēs k·āk·îlxalaēnaᶜyē. Wä, lä ōxsg·iwaᶜyas ᶜnɛmp!ɛnk·ōstâ lāxɛns ts!ɛχᵘts!ānaᶜyaxsɛns q!wāq!wax·ts!ānaᶜyēx. Wä, laɛmxɩa lāxs ᶜwā´laɛm q!wāq!wax·ts!ā- naᶜyēx ʟɛᶜwins qōmax yîx ᶜwālagak·îlasas. Wä, la ᶜnɛmp!ɛng·apa āwābâᶜyasēxɛns ts!ɛχᵘts!ānaᶜyasɛns q!wāq!wax·ts!ānaᶜyēx yîx ᶜwā- dzɛgabasas. Wä, lä mödɛn lāxɛns q!wāq!wax·ts!ānaᶜyēs yîx

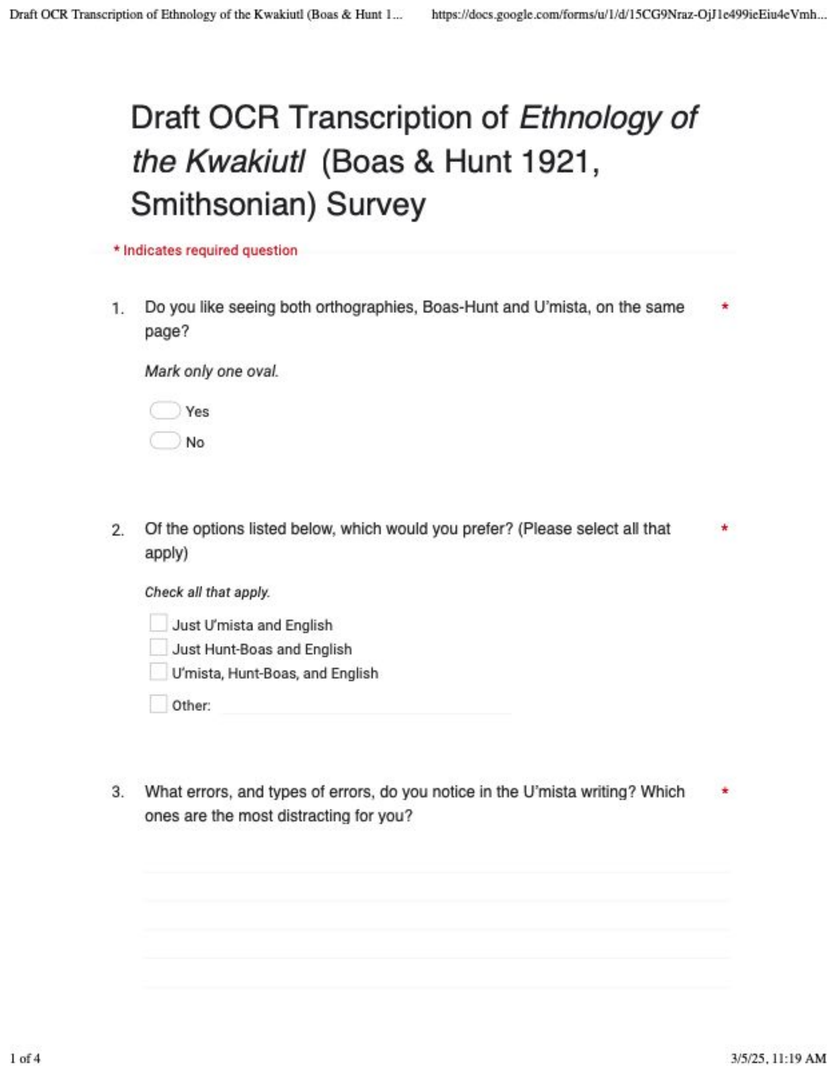
---

The Making o’o Dihg.). We, he’ą̈m da’axwsa luḵwila’ino’iwaxa luḵ’wis subayuwaxs la’i laxa ali. We, gil- ‘misi lagа’a laxa tɬesmadząxąkulaxs la’i duḵux’idxa ekitąlaxa ḵe’osi t’ą̈naka. We, he’misixs ḵisa’i ḵipąla ḵa’s he’ma’i duḵwasusida naḵąḵi lax kuχsąntsą’wi. We, gil’misi ḵąxa ekaχs la’i supąxudxa ḵąt’ą̈pą̈nxsą’sta laχąns ḵ’waḵ’waxt’sa- na’yiχ, yiχ wagitąxdła’asas. We, gil’misi ťax’idąxs la’i tąmkudxa ‘nąmp̓ą̈nki laχąns batłax ḵa laweyis tąmgult’sąxdła’yas. We, le bal’idxa mup̓ą̈nkas ‘wasgąmas laχąns ḵ’waḵ’waxt’sana- ‘yiχs la’i tąmxwsą̈ndąḵ. We, gil’misi laxsąxs la’i kuχsą̈ndąḵ naḵąḵaxdumaḵas. We, gil’misi k!uχsa’akuχs la’i supolax dumaḵas ḵa ‘nąmdą̈nis lawuyos hayaḵaxa dumaḵi. We, le a’eka supolaḵ ḵa naḵąlis. We, he’mis ḵa ḵisis sąlgwasnukwa ḵaχs he’ma’i ąwobą’wisa luḵ’wis dumaḵi. We, gil’misi ḡwaląxs la’i susąbą̈nudzą̈ndąḵ ḵa lixuyuwis yiχs ‘nąmdą̈na’i ‘wadzą̈xgiwa- sasa uba’yasa luḵ’wi. We, la müdą̈nbalida ‘nąmp̓ą̈nki laχąns ḵ’waḵ’waxt’sana’yiχ yiχ ‘wadzą̈guyuwasa laχis ḵakilχala’ina’yi. We, le uχsgiwa’yas ‘nąmp̓ą̈nkusto laχąns t’sąxwt’sana’yaχsąns ḵ’waḵ’waxt’sana’yiχ. We, la’ą̈mχdła laχs ‘wa´la’ą̈m ḵ’waḵ’waxt’sa- na’yiχ dłą’wins ḵumaχ yiχ ‘walagakilasas. We, la ‘nąmp̓ą̈ngapa ą̈wabo’yasiχąns t’sąxwt’sana’yasąns ḵ’waḵ’waxt’sana’yiχ yiχ ‘wa- dzą̈gabasas. We, le mudą̈n laχąns ḵ’waḵ’waxt’sana’yis yiχ

Communities and Consultations

First draft of text distributed widely to ~50 community experts and areal knowledge specialists/researchers for feedback

- GN Language Program
- Staff of the K̲wala adult immersion language program
- Language teachers in the GNN Elementary School
- Heritage language learners and teachers at UBC
- Elders working with the language
- Academic researchers, linguists



Draft OCR Transcription of Ethnology of the Kwakiutl (Boas & Hunt 1... https://docs.google.com/forms/u/1/d/15CG9Nraz-OjJ1e499ieEiu4eVmh...

Draft OCR Transcription of *Ethnology of the Kwakiutl* (Boas & Hunt 1921, Smithsonian) Survey

* Indicates required question

1. Do you like seeing both orthographies, Boas-Hunt and U'mista, on the same page? *

Mark only one oval.

- Yes
- No

2. Of the options listed below, which would you prefer? (Please select all that apply) *

Check all that apply.

- Just U'mista and English
- Just Hunt-Boas and English
- U'mista, Hunt-Boas, and English
- Other:

3. What errors, and types of errors, do you notice in the U'mista writing? Which ones are the most distracting for you? *

1 of 4 3/5/25, 11:19 AM

Secondary Challenges

- Revising 1921
  - Restore line numbers
  - Restore missing titles

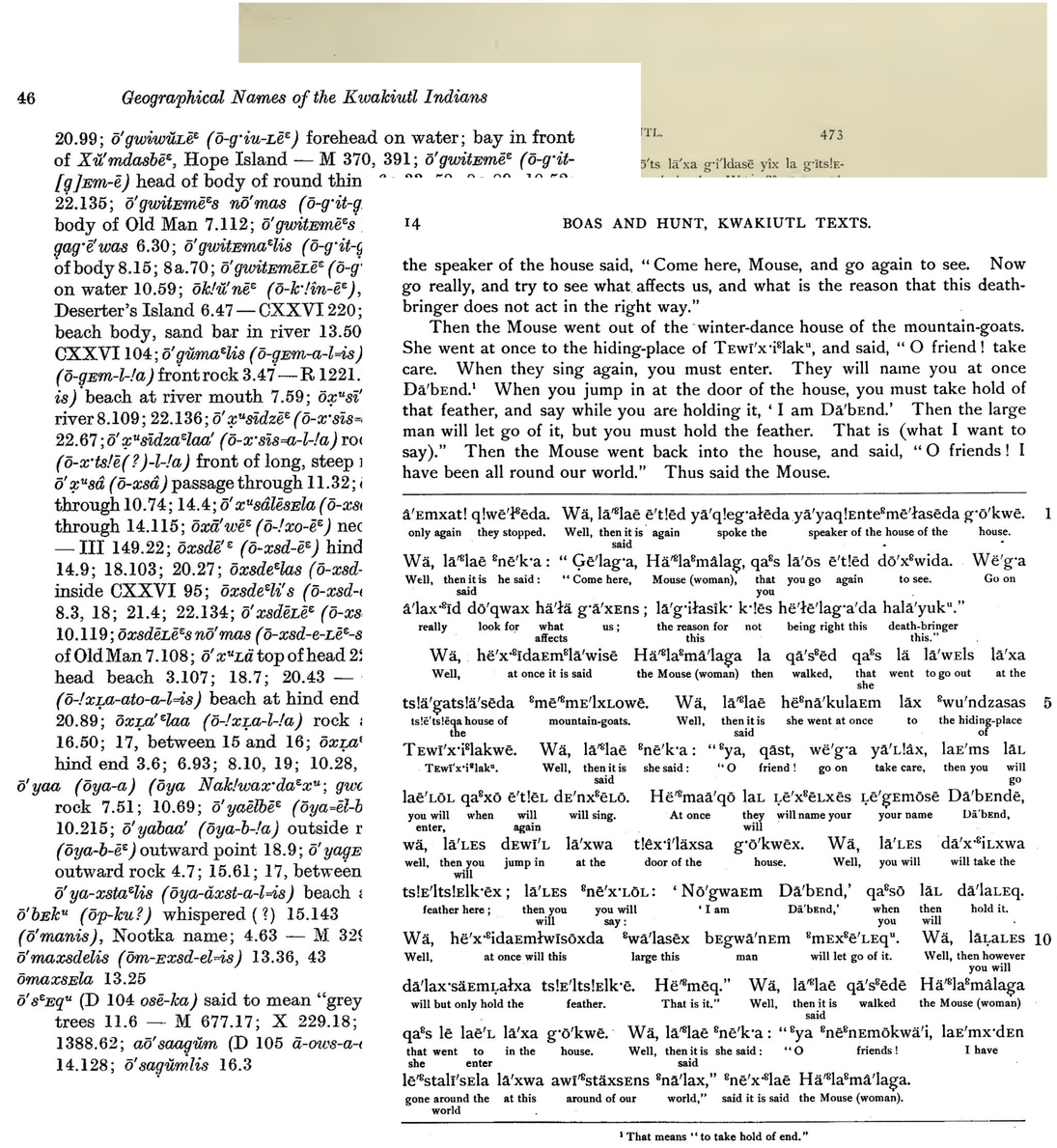
How to make this model transferable t other texts?

Essential:

  - How to handle diverse formatting, layouts & arrangement of English an Kwak'wala?

Nice to have:

  - Images & Figures



46 *Geographical Names of the Kwakiutl Indians*

20.99; *ō'gwiwŭlēᵉ (ō-g·iu-lēᵉ)* forehead on water; bay in front of *Xŭ'mdasbēᵉ*, Hope Island — M 370, 391; *ō'gwitemēᵉ (ō-g·it-[g]em-ē)* head of body of round thin … 22.135; *ō'gwitemēᵉs nō'mas (ō-g·it-g* body of Old Man 7.112; *ō'gwitemēᵉs gag·ē'was* 6.30; *ō'gwitemaˢlis (ō-g·it-g* of body 8.15; 8a.70; *ō'gwitemēlēᵉ (ō-g·* on water 10.59; *ōk!ŭ'nēᵉ (ō-k·!in-ēᵉ)*, Deserter's Island 6.47 — CXXVI 220; beach body, sand bar in river 13.50 CXXVI 104; *ō'gŭmaˢlis (ō-gem-a-l-is) (ō-gem-l-!a)* front rock 3.47 — R 1221. *is)* beach at river mouth 7.59; *ōx̣ᵘsĭ'* river 8.109; 22.136; *ō'x̣ᵘsīdzēᵉ (ō-x·sīs-* 22.67; *ō'x̣ᵘsīdzaˢlaa' (ō-x·sīs-a-l-!a)* ro *(ō-x·ts!ē(?)-l-!a)* front of long, steep *ō'x̣ᵘsâ (ō-xsâ)* passage through 11.32; through 10.74; 14.4; *ō'x̣ᵘsâlēsela (ō-xs* through 14.115; *ōxā'wēᵉ (ō-!xo-ēᵉ)* nec — III 149.22; *ōxsdē'ᵉ (ō-xsd-ēᵉ)* hind 14.9; 18.103; 20.27; *ōxsdeˢlas (ō-xsd-* inside CXXVI 95; *ōxsdeˢli's (ō-xsd-* 8.3, 18; 21.4; 22.134; *ō'xsdēlēᵉ (ō-xs* 10.119; *ōxsdēlēᵉs nō'mas (ō-xsd-e-lēᵉ-s* of Old Man 7.108; *ō'x̣ᵘlā* top of head 2 head beach 3.107; 18.7; 20.43 — *(ō-!xḷa-ato-a-l-is)* beach at hind end 20.89; *ōxḷa'ˢlaa (ō-!xḷa-l-!a)* rock 16.50; 17, between 15 and 16; *ōxḷa'* hind end 3.6; 6.93; 8.10, 19; 10.28, *ō'yaa (ōya-a) (ōya Nak!wax·daˢx̣ᵘ; gw* rock 7.51; 10.69; *ō'yaēlbēᵉ (ōya-ēl-b* 10.215; *ō'yabaa' (ōya-b-!a)* outside r *(ōya-b-ēᵉ)* outward point 18.9; *ō'yagᴇ* outward rock 4.7; 15.61; 17, between *ō'ya-xstaˢlis (ōya-ăxst-a-l-is)* beach *ō'bᴇkᵘ (ōp-ku?)* whispered (?) 15.143 *(ō'manis)*, Nootka name; 4.63 — M 32 *ō'maxsdelis (ōm-exsd-el-is)* 13.36, 43 *ōmaxsᴇla* 13.25 *ō'sᴇqᵘ* (D 104 *osē-ka)* said to mean "grey trees 11.6 — M 677.17; X 229.18; 1388.62; *aō'saagŭm* (D 105 *ā-ows-a-* 14.128; *ō'sagŭmlis* 16.3

TL. 473

ō'ts lā'xa g·i'ldasē yix la g·its!ᴇ-

14 BOAS AND HUNT, KWAKIUTL TEXTS.

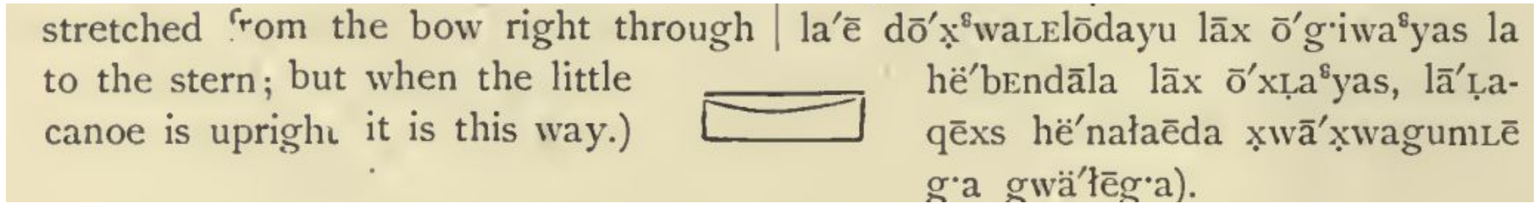
the speaker of the house said, "Come here, Mouse, and go again to see. Now go really, and try to see what affects us, and what is the reason that this death-bringer does not act in the right way."

Then the Mouse went out of the winter-dance house of the mountain-goats. She went at once to the hiding-place of Tᴇwī'x·i⁸lakᵘ, and said, "O friend! take care. When they sing again, you must enter. They will name you at once Dā'bᴇnd.[1] When you jump in at the door of the house, you must take hold of that feather, and say while you are holding it, 'I am Dā'bᴇnd.' Then the large man will let go of it, but you must hold the feather. That is (what I want to say)." Then the Mouse went back into the house, and said, "O friends! I have been all round our world." Thus said the Mouse.

á'ᴇmxat!	q!wē'ł⁸ēda.	Wä,	lā⁸laē	ē'tlēd	yā'q!eg·ałēda	yā'yaq!ᴇnte⁸mē'łasēda	g·ō'kwē.	1
only again	they stopped.	Well,	then it is said	again	spoke the	speaker of the house of the	house.	
Wä,	lā⁸laē	⁸nē'k·a:	"Ge'lag·a,	Hä⁸la⁸mâlag,	qa⁸s	lā'ōs	ē'tlēd	dō'x⁸wida.
Well,	then it is said	he said:	"Come here,	Mouse (woman),	that you	go again	to see.	
Wē'g·a								
Go on								
á'lax·⁸īd	dō'qwax	hä'łä	g·ā'xᴇns;	la'g·iłasīk·	k·!ēs	hē'łē'lag·a'da	halā'yuk".	
really	look for	what affects	us;	the reason for this	not	being right this	death-bringer this."	
Wä,	hē'x·⁶īdaᴇm⁸la'wisē	Hä⁸la⁸mâ'laga	la	qā's⁸ēd	qa⁸s	lā	lā'wᴇls	lā'xa
Well,	at once it is said	the Mouse (woman)	then	walked,	that she	went	to go out	at the
ts!ā'gats!ā'sēda	⁸mē⁸mᴇ'lxLowē.	Wä,	lā⁸laē	hē⁸nā'kulaᴇm	lāx	⁸wu'ndzasas		5
ts!ē'ts!ēqa house of the	mountain-goats.	Well,	then it is said	she went at once	to	the hiding-place of		
Tᴇwī'x·i⁸lakwē.	Wä,	lā⁸laē	⁸nē'k·a:	"⁸ya,	qāst,	wē'g·a	yā'l!âx,	laᴇ'ms	lāl
Tᴇwī'x·i⁸lakᵘ.	Well,	then it is said	she said:	"O friend!	go on	take care,	then you	will go	
laē'lōl	qa⁸xō	ē'tlēl	dᴇ'nx⁸ēlō.	Hē⁸maā'qō	lal	ḷē'x⁸ēlxēs	ḷē'gᴇmōsē	Dā'bᴇndē,	
you will enter,	when	will again	will sing.	At once	they will	will name your	your name	Dā'bᴇnd,	
wä,	lā'lᴇs	dᴇwī'l	lā'xwa	tlēx·i'läxsa	g·ō'kwēx.	Wä,	lā'lᴇs	dā'x⁸iLxwa	
well,	then you will	jump in	at the	door of the	house.	Well,	you will	will take the	
ts!ᴇ'lts!ᴇlk·ēx;	lā'lᴇs	⁸nē'x·lōl:	'Nō'gwaᴇm	Dā'bᴇnd,'	qa⁸sō	lāl	dā'laLEq.		
feather here;	then you will	you will say:	'I am	Dā'bᴇnd,'	when you	then will	hold it.		
Wä,	hē'x⁶idaᴇmłwīsōxda	⁸wā'lasēx	bᴇgwā'nᴇm	⁸mᴇx⁸ē'leqᵘ.	Wä,	lāḷaLᴇs		10
Well,	at once will this	large this	man	will let go of it.	Well,	Well, then however you will		
dā'lax·sämᴇmi̱ałxa	ts!ᴇ'lts!ᴇlk·ē.	Hē⁸mēq."	Wä,	lā⁸laē	qā's⁸ēdē	Hä⁸la⁸mâlaga		
will but only hold the	feather.	That is it."	Well,	then it is said	walked	the Mouse (woman)		
qa⁸s	lē	laē'l	lā'xa	g·ō'kwē.	Wä,	lā⁸laē	⁸nē'k·a:	"⁸ya	⁸nē⁸nᴇmōkwä'i,	laᴇ'mx·dᴇn
that she went	to enter	in the	house.	Well,	then it is said	she said:	"O	friends!		I have
lē⁸stalī'sᴇla	lā'xwa	awī⁸stäxsᴇns	⁸nā'lax,"	⁸nē'x·⁸laē	Hä⁸la⁸mâ'laga.					
gone around the world	at this	around of our	world,"	said it is said	the Mouse (woman).					

[1] That means "to take hold of end."

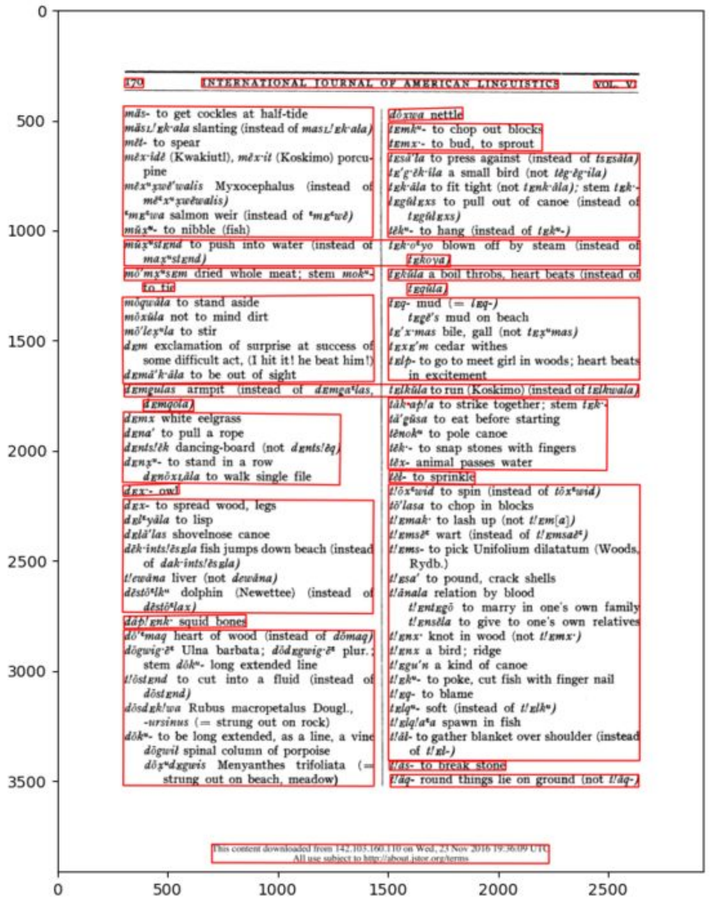
First-Pass OCR


stretched from the bow right through | la′ē dō′ẋ^{s}waʟᴇlōdayu lāx ō′g·iwa^{s}yas la
to the stern; but when the little hë′bᴇndāla lāx ō′xɬa^{s}yas, lā′ɬa-
canoe is upright it is this way.) qēxs hë′nałaēda ẋwā′ẋwagumʟē
g·a gwä′łēg·a).

Google Cloud Vision OCR - Text Detection Module

Stretched rom therow right through laid oxwalelodayu lax ogiwa la
To the stern but when the little he benda la lax o xlayas lala -
Canoe is upright it is this way)) qexs henaleda xwaxwagle
ga gwa lega).

First-Pass OCR



Google Cloud Vision OCR - Text Detection Module

- Advantages
  - Free to use, up to 1000 pages per month
  - Access through Google Cloud API services
  - Quick and easy setup, and rich metadata where available

- Disadvantage:
  - Does not allow finetuning
  - Language ID labels are often missing
  - Bounding boxes may not be structurally coherent

LangID

Stretched rom therow right through laid oxwalelodayu lax ogiwa la
To the stern but when the little he benda la lax o xlayas lala -
Canoe is upright it is this way)) qexs henaleda xwaxwagle
ga gwa lega).

English-language detector (fastText)

Stretched rom therow right through laid oxwalelodayu lax ogiwa la
To the stern but when the little he benda la lax o xlayas lala -
Canoe is upright it is this way)) qexs henaleda xwaxwagle
ga gwa lega).

LangID

fastText custom language identification training

- bilingual first-pass texts from Boas collection used for training
- 1000 sentences for each language (Kwak’wala and English)
- Trained using default fastText parameterizations
- Achieves 99+% accuracy on held-out test set
- We apply this to the first-pass text to potentially use language as a proxy for structure

Masking

Stretched rom therow right through laid oxwalelodayu lax ogiwa la
To the stern but when the little he benda la lax o xlayas lala -
Canoe is upright it is this way)) qexs henaleda xwaxwagle
ga gwa lega).

High-likelihood English spans are temporarily masked

laid oxwalelodayu lax ogiwa la
benda la lax o xlayas lala -
)) qexs henaleda xwaxwagle
ga gwa lega).

Masking

- Texts are diversely formatted with illustrations, figures, page numbers, line numbers etc.
- The post-correction model requires only Kwak’wala text and doesn’t handle non-Kwak’wala text very well
- We apply a masking layer that temporarily hides:
  - line numbers (that provide cross-reference information to other collections (very important for readers)
  - page numbers
  - punctuation
- Done with a simple deterministic Python script

Post-Correction

laid oxwalelodayu lax ogiwa la
benda la lax o xlayas lala -
)) qexs henaleda xwaxwagle
ga gwa lega).

Remaining Kwak’wala text is sent for post-correction, as per Rijhwani et al. (2020)

la'e do'x³waLElōdayu lāx o'g'iwa³yas la
hë'bendāla lax ō'xḶa³yas, lāḶa-
qēxs hë'nałaēda xwā'xwagumLē
ga gwälega).

Post-Correction

- Can automatically correct errors in very low-resource OCR settings
- Train a very small correction model on a sample of first-pass and gold reference sentences
- Multi-source neural architecture (based on Rijhwani et al. 2020) which was shown to reduce OCR character-level errors by 30-60%
- We use the model as is, with the lexically-aware decoding setting turned off as it was shown not to benefit Kwak'wala
- We train the model from scratch, replicate the CER results from the original paper, and then apply the model to our dataset

Reconstruction

la'e do'x³waLElōdayu lāx o'g'iwa³yas la
hë'bendāla lax ō'xḶa³yas, lāḶa-
qēxs hë'nałaēda xwā'xwagumLē
ga gwälega).

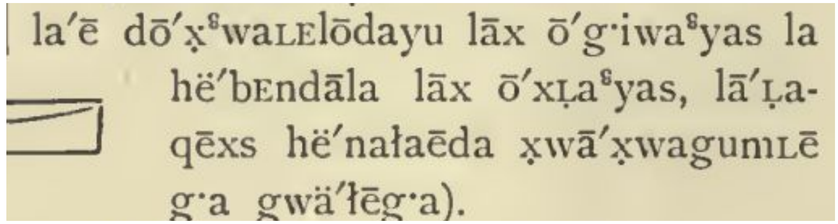
English text and other masked tokens are reintroduced at the appropriate indices. Orthographies can be converted (U’mista or SD-72)

Stretched rom therow right through
To the stern but when the little
Canoe is upright it is this way
la'e do'x³waLElōdayu lāx o'g'iwa³yas la
hë'bendāla lax ō'xḶa³yas, lāḶa-
qēxs hë'nałaēda xwā'xwagumLē
ga gwälega).

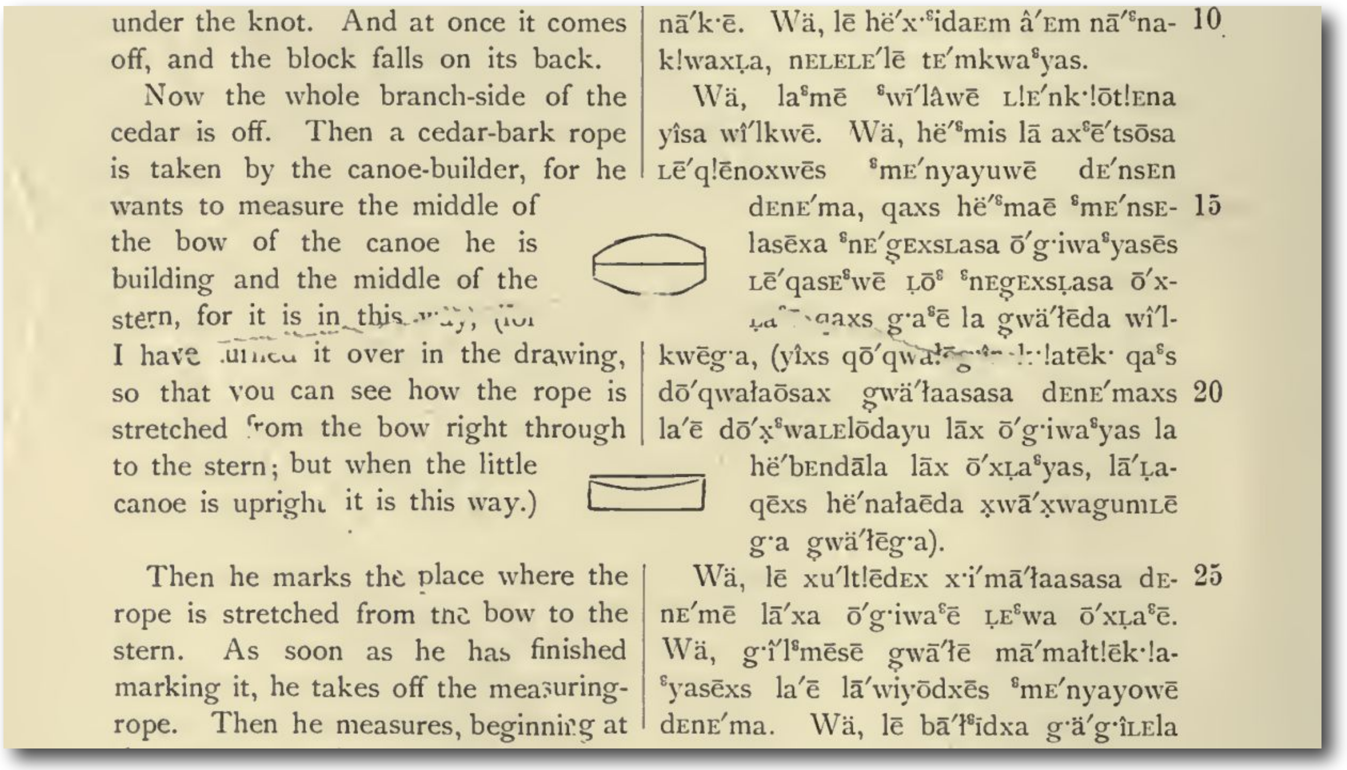
Evaluation

la'e do'x³waLElōdayu lāx o'g'iwa³yas la
hë'bendāla lax ō'xḶa³yas, lāḶa-
qēxs hë'nałaēda xwā'xwagumLē
ga gwälega).

Compare final corrected outputs to original gold labels, at the character level to obtain CER


la′ē dō′x̣ˢwaLElōdayu lāx ō′g·iwaˢyas la
hë′bEndāla lāx ō′xḶaˢyas, lā′Ḷa-
qēxs hë′nałaēda x̣wā′x̣wagumLē
g·a gwä′łēg·a).

Mixed-Methods Pipeline


under the knot. And at once it comes off, and the block falls on its back.

Now the whole branch-side of the cedar is off. Then a cedar-bark rope is taken by the canoe-builder, for he wants to measure the middle of the bow of the canoe he is building and the middle of the stern, for it is in this way, (for I have turned it over in the drawing, so that you can see how the rope is stretched from the bow right through to the stern; but when the little canoe is upright it is this way.)

Then he marks the place where the rope is stretched from the bow to the stern. As soon as he has finished marking it, he takes off the measuring-rope. Then he measures, beginning at

nā′k·ē. Wä, lē hē′x·ᵋidaEm â′Em nā′ᵋnak!waxḷa, nELELE′lē tE′mkwaᵋyas.

Wä, laᵋmē ᵋwī′lâwē L!E′nk·!ōt!Ena yîsa wî′lkwē. Wä, hē′ᵋmis lā axᵋē′tsōsa Lē′q!ēnoxwēs ᵋmE′nyayuwē dE′nsEn dEnE′ma, qaxs hē′ᵋmaē ᵋmE′nsElasēxa ᵋnE′gExsLasa ō′g·iwaᵋyasēs Lē′qasEᵋwē ḶōᵋᵋnEgExsḶasa ō′xḷa… qaxs g·aᵋē la gwä′łēda wî′lkwēg·a, (yîxs qō′qwa… …latēk· qaᵋs dō′qwałaōsax gwä′łaasasa dEnE′maxs la′ē dō′x̣ᵋwaLElōdayu lāx ō′g·iwaᵋyas la hē′bEndāla lāx ō′xḷaᵋyas, lā′ḷaqēxs hē′nałaēda x̣wā′x̣wagumLē g·a gwä′łēg·a).

Wä, lē xu′lt!ēdEx x·i′mā′łaasasa dEnE′mē lā′xa ō′g·iwaᵋē ḶEᵋwa ō′xḷaᵋē. Wä, g·î′lᵋmēsē gwā′łē mā′małt!ēk·!aᵋyasēxs la′ē lā′wiyōdxēs ᵋmE′nyayowē dEnE′ma. Wä, lē bā′lᵋīdxa g·ä′g·îLEla

1) First-Pass OCR

2) LangID Labeling

3) Masking/Isolation

4) Post-Correction

5) Reconstruction and Output

6) Evaluation with Gold data

Results

- Community-oriented prioritization of texts for digitization
- With our mixed-methods pipeline, including language identification, masking, and automatic post-correction, we achieve for our Kwak'wala texts:
  - ~50% decrease in character error rate
  - ~87.5% reduction in structural error
    - insertion, deletion, and maximal move operations required across the output page to make it identical with the reference text (Kanai et al, 1995)

	Jesup 5.1, 1902		Kwakiutl, 1909	
	CER	SER	CER	SER
First Pass	0.43	25	0.33	18
Corrected	0.18	2	0.15	3

Table 1: For both books, we find that using our pipeline greatly reduces not only textual errors (CER) but also greatly improves the layout and structure (SER)

Next Steps

- Stable OCR Packages for Kwak'wala:
    1. Developer friendly (repeatable & transferable)
    2. End-user friendly
- Investigation of multiple langID modeling approaches for Kwak'wala (other than fastText)
- Create more gold reference texts for better evaluation
- More extensive benchmarking across more books/collections
- With consent of the language community, share digitized texts with the data hosting institutions, such as the American Philosophy Society and Columbia University Rare Books and Special Collections

Gilakas’la! Thank you!

Developing a Mixed-Methods Pipeline for Community-Oriented Digitization of Kwak'wala Legacy Texts

Milind Agarwal, Daisy Rosenblum, Antonios Anastasopoulos

GitHub Repository

Contains code for all stages of the pipeline, as applied on Kwak’wala

Next Steps

- Text extraction from two volumes (Boas and Hunt, 1921)
- Transliteration into two community-preferred orthographies (U'mista, Liq'wala)
- Apply model to other languages written in same typeface
- Develop/improve OCR model for unpublished typescript
- Create dataset for developing other NLP tools