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

Milind Agarwal, Daisy Rosenblum, Antonios Anastasopoulos







# Improving OCR for Kwakwala Texts

- OCR for legacy data in Kwakwala
- Why?
  - Rich Written Tradition

  - Digital files available only as scanned images

Crabapples.-Wä, la<sup>ε</sup>mē ăx<sup>ε</sup>ēdēda ts!Edā <sup>e</sup>ya, yîxs hë<sup>e</sup>maē ālēs tselx<sup>u</sup>ts!âlaxa tsely xwē giītslâq lāxa lāxidē gugedzoyosēx la<sup>e</sup>mē <sup>e</sup>nāxwaEm hë gwēx <sup>e</sup>īdxa hëlomāg ○ High value for community researc naāgemē lexa<sup>c</sup>ya. Wä, g'îl<sup>c</sup>mēsē lā loi lāxa tselxwē lā k'ladzālītaxa lēswasyē. k!wāg alīl lāx hēłk !õdenwalīlasa nānaag lā<sup>c</sup>wŭnemas k!wāg alīt lāx hëlk !õdenwalīt Wä, laem gēgemxagawalīła laelxa<sup>s</sup>yē lā dāgē Le<sup>e</sup>wis lā<sup>e</sup>wŭnemē. Wä, lāx da<sup>e</sup>xwē lāxa tselxwē qa's ēp!extē māg înodāla yîsēs hēlk lots lāna<sup>e</sup>yē. 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



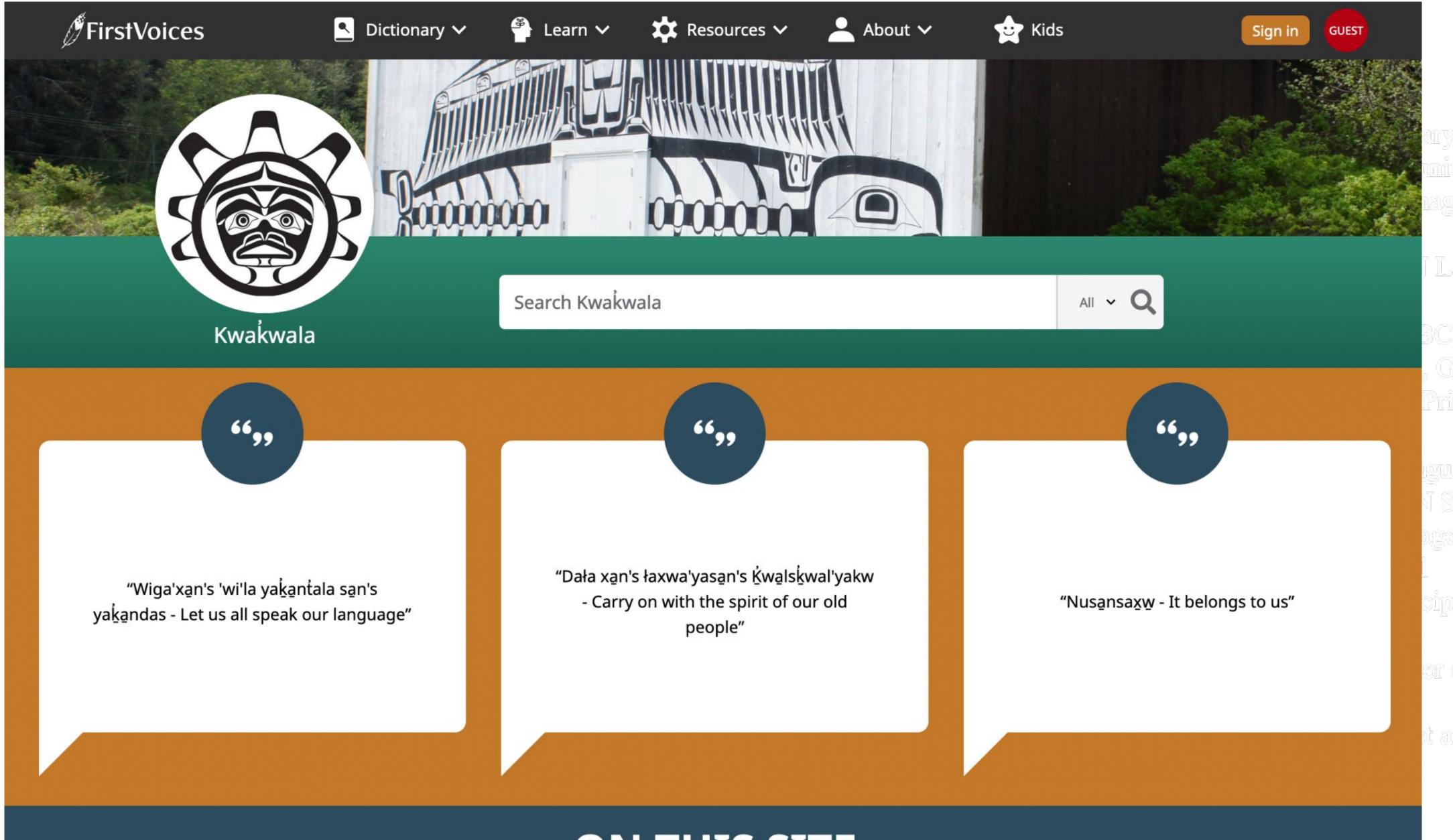
3

### Kwakwaka'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



### Kwakwaka'wakw communities, Kwak'wala language reclamation



### ON THIS SITE



# 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...)



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

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



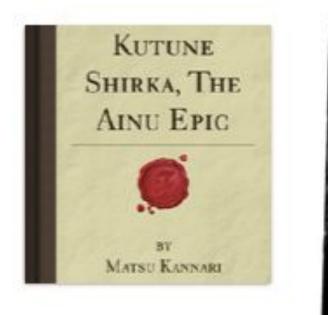
Elder Speakers' Gathering, Gwa'sala-'Nakwaxda'xw Elders' Centre, November 12, 2009





### Machine-readable text has many benefits

Mat	tiaxh	Xhunik	jos.om	maring
Mat	hias	John	work wood (tv).agent	marimb
Mat	iaxh	Xhunik	wood-worker	(of) m
3.	max.	W. al.on	cham	max
	comp.	3.say (tv	).PIV respected man (	NC1) co
	he sa	id		we





Digitizing these texts to machinereadable formats

ranuai wa rantukwa ní kittóoraspilikaan rhadro navier wa nantrikwa



Build NLP systems to support language revitalization efforts



### Input data

240

### BOAS: XNAKIUTL GRAMMAR

TRANS. AMER. PHIL. SOC.

### pEnqw+(a pe'nca) pe'nua) peki(a1) . 1 lāxa

vîsēs

- -ouixo in mouth (see -ou into): talege ulexed to throw into mouth III 359.13 -indzim throat (c -ins-im): to bindzim speck in throat. -ap! neck: «axa piala to have on neck III 19.6; g'splaislod to tusk into neckpiece III
  - Also: following, behind: (Swapiclis to atand behind on beach.
- -[x']at sysp! shoulder and arm above elbow: vix'stys plala to carry on shoulder III

embrace around neck X 121.38.

- -[x']telens hend: tl8'semx'talans atone-handed III 131.32; farat!& nend to put on hand. -Ibo. -Ipila chest: gremo tiabler left side of chest C II 48.17; t8 kivepils to have hanging on chest R 208.10.
- -ok'els, -og'e@ back: @adb'g'e@ back sinev Y 487.4; menë'g'end to strike back (mex'-) Also: lô'g's to follow; <u>ilô'beg's</u> to roast afterwards; na gek'sla to drink afterwards III 41.25.
- -k'islo[sls] front of body, lsp: g'i'k'!slgand to put in lap V 478.25.
- -Abedzes leg below knee: 48 dabedzes sinews at
- -<u>Obgo</u> penie:  $\underline{m5}^{\prime}\underline{x}^{B}\underline{s}\underline{s}\underline{g}\underline{s}\underline{v}\underline{s}\underline{k}^{B}$  with tied---III 138.11 ( $\underline{m5k}^{B-}$  to tie).
- -plog's shin: <u>gax pld'g'and</u> to put ring around shin III 89.37.
- -k'&x'e\* knee: Lenk'&'x'ex with scabby knees III 154.11.
- -[x']ais foot: 58 paidzend to pinch foot III 96.3; <u>ti8 paisels</u> to wear shoes CX 281.32. -k'lies, wes in body, in front of body; g'1'k'liedged what is in body C II 42.4;
- ge'enek':Me'sgla or gë'enese'sgla salmon with spawn in body (see egs). -[g']it body: no stat to sprinkle body III
- 105.38; 8 gvitet body III 202.24; te k itels well grown (tree) V 496.6; tid same it stone-bodied III 200.9.
- -[x']dzawmo in front of body; k's'x'dzawmod to place dish in front of, V 429.23. -[k'] isn body (of man, log, etc.) [relating
- rather to surface of body): "me'lk' isn white-bodied; <u>le'nlessévenz'eid</u> body gets dry all over ¥ 483,6; <u>g'I'k'iendals</u> to put on a log III 272.33; <u>sa'lak'ien</u> able-bodied III 208.39; -g'eren along a line; ma'g'edwarmet close to a line R 67.95. <u>d8'g'edward</u> to hold (a rope) C 26: 202.97.
- leg in mind: <u>5'k' legels</u> to feel good III 123.12; g'1'g'ie gels to think; <u>Sne'Enk'</u>!-
- mx\*id to begin to say in mind (i.e. to think) III 184.3.

4. Limitations of Form

Generally used with numerols sok" human beings: ms \$15'k" two persons III

- 48.21; g'and'k" how many persons? -xea flat: fne mxea one (day) III 18.2; gl8'xea
- many (leaves) N 298.51. -talag long: m5 talag four long ones III 10.12; 64 xataingas number of long ones C III 162.15.

<u>tsiaxsta</u>, <u>g'edtsiaxsta</u> movement in a long path; <u>hā baftsiaxsta</u> it goes right through C 26:20.113; <u>ent mpieratsiaxsta</u> (to call) only once along street of village C III 218.21.

- -agem, -asm round, on surface: ha magend to hold round thing in mouth C 26:13.6; darse and to put down round thing R 485.38; pie'ixelangem woolen blanket (fog om surface) H 691.8.
- aggemak" tribe: ock la aggemakives five tribes in one village.
- -den finger-width: <u>ene mdenx'ss</u> one finger-width through V 491.6.
- plank' fathom, span: "Liebo plank' seven spans 8 110.34.
- -plen(a) times: fnf mplena once; one mplanic to stay in house one day.
- -<u>zvats</u> days: <u>hb dop:snavats</u> the right number of days III 355.26. -[x']asyuk<sup>u</sup> bundle: <u>fnr'mx'asyuk<sup>u</sup></u> one bundle
- [x'] sok" five pairs of blankets: matiteo'k" two five pairs (i.e. ten pairs of blankets); also the tenths in each hundred:
- Eng'mg'acgwig'iu one hundred and ten. Also: EF'g'ack handsome III 48.29 -okiwema paira: ark'i&'kiwema five pairs
- [< ak<sup>2</sup>-!eas). -Xia, "ORLA dish: ma@Zixia' two dishes N 516.14; Entm8 Xia one dish Y 434.3.
- -x tyu, strings of fish: ent ax tyu one---, -dztg hole, ent mdztg one---; eva dztg size of ---V 332.24.
- jr] dzzkważk layers; <u>Ens'nz'dzskwałk</u> one---. -<u>[x] dzskw(sla)</u> one way: <u>ens'nz'dzsz<sup>6</sup>w1'd</u> tide
- begins to run one way. -lenk's at a time: "n&' 2"ran% mk's one at a
- -ibex" price (also verb); giacyo'ex" high (many) priced.

### 5. Temporal Suffixes

-of, -vel remote past: gry5 2 long ago III 12.4; 05 mp6 vel the late father; g'd x<sup>6</sup>vel n I came long ago III 142.19.

- -x aid recent past: Caxes xeid place where he had been III 42.4; <u>c6 cax 61d</u> he went (about a week ago); <u>c6 s61dex 61d</u> he took a walk.
- "L future: 10 gade, he will have a name III 19.1; swa kiwana, a future cance III 83-33.
- -x de transition from present to past: wa' idenx'de what he had said III 25.4;
- 120 al name slow match 098.20, pEngayu 1 I 49.15, blow off steam y steam

### Bilingual (English & Kwakwala) 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

la<sup>e</sup>mē <sup>e</sup>nāxwaEm hë gwēx<sup>.e</sup>īc naāgemē lexa<sup>s</sup>ya. Wä, g'îl<sup>s</sup> lāxa tselxwē lā k'!adzâlīłaxa k!wag alīl lāx hëłk !õd Enwalīl  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!

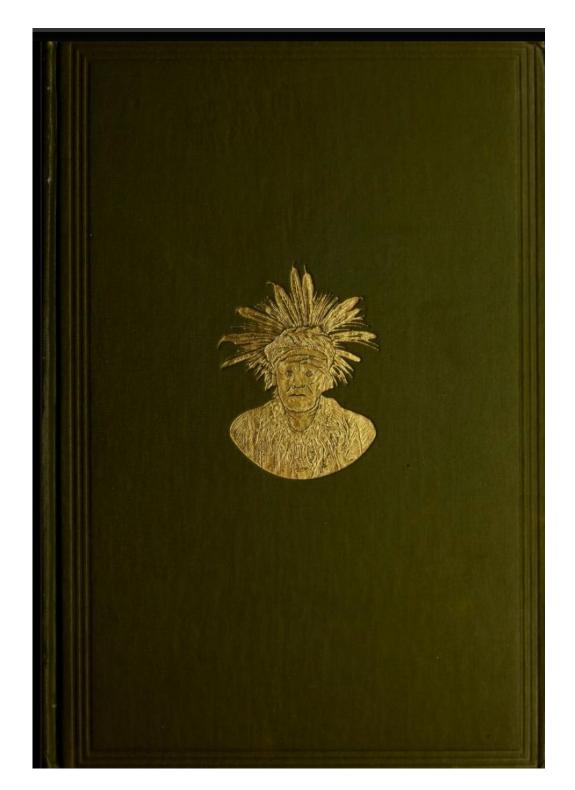


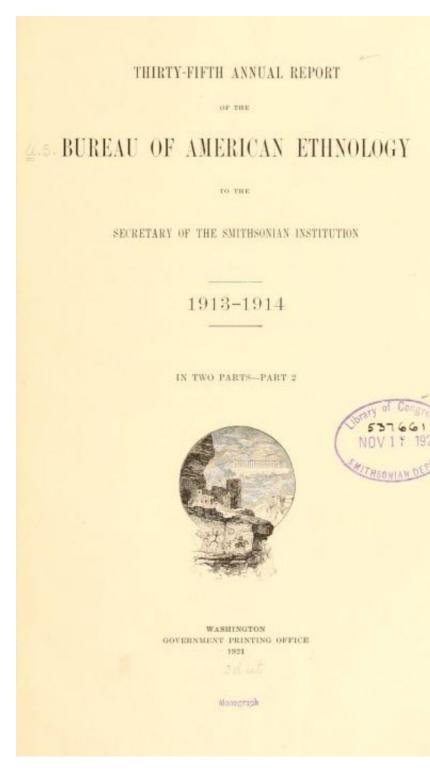




### Text Selection: Community-determined

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.





CONTENTS	
CONTENTS	
I. INDUSTRIES	
The making of disks	Page
The making of dishes. Dish for pounding salal-berries.	
The making of boxes	
The making of oil-boxes.	
Sewing with cedar-withes.	
Care of canoe	
Wooden sail	97
Mat sail and mast	
The making of horn spoons (1)	
The making of horn spoons (2)	
Cedar-bark breaker	
Bag of sea-lion hide Spruce-roots and cedar-withes	
Cedar-withes	
Spruce-roots (1).	
Spruce-roots (2)	
Cedar-withes	
Cedar-bark (1)	120
Cedar-mats	
Shredding cedar-bark	
Yellow cedar-bark	
Cedar-bark (2)	
Shredding cedar-bark Open-work basket	
Cedar-bark basket (1)	
Basket for viburnum-berries	
Basket for wild carrots.	
Cedar-bark basket (2)	
Huckleberry-basket	
Box for picking almon-berries	140
Tump-line	
Back-protector	
Belt.	
Implement for peeling cedar-bark	
Digging-stick for clover	
Digging-stick for roots.	
Digging-stick for cryptochiton	
Hook for devil-fish (1)	
Hook for devil-fish (2)	152
Spear for sea-oggs	
Hook for picking elderberries	
Pole for gathering eel-grass	
Flounder-spear	
Fishing-tackle for flounders	
75052—21—35 вти—рт 1—4	49

### I. INDUSTRIES

The Making of Dishes. - The dish-maker takes | along his ax 1 when he goes into the woods. When he | reaches a patch of aldertrees, 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 5 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 10 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 15 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 20 the thumb. | The bottom is one short span | wide and three spans

The Making of Dishes (Lõqwēläxa lõq!wē). - Wä, hëEm daaxusa 1 lõqwēlaēnoxwaxa lõq!wēs sõbayowaxs laē lāxa ār.!ē. Wä, gil-\*mēsē lāg an lāxa L'āsmadzexekulaxs laē dog lūx idxa ek ētelaxa kleåsē Liznāka. Wā, hētmēsēxs klēsaē klipela gats hētmaē döqwasösēda nāq!zqē lax kūxsentsz'wē. Wā, gilimēsē q!āxa 5 ēk axs laē sop!exodxa q!et !ep !enx sē sta lāxens q!wāq !wax ts !ānatyēx, yix wag it extaasas. Wā, giltmēsē tlax tidexs laē tem k'ödxa 'nemp!enk'ē lāxens bāLax qa lawāyēs t!emgülts!exLatyas Wä, lä bäl'idxa mõp!enk'as wäsgemas läxens q!wäq!wax'ts!äna-'yēxs laē temx"sendeq. Wā, gil'mēsē lāx sexs laē kūxsendeq 10 nāq!eqax domaqas. Wā, g'îl'mēse kūxsaakūxs laē sopālax domaqas qa "nemdenës läwoyâs häyäqaxa domaqë. Wä, lä aëk la sopâlaq qa neqelës. Wä, hë"mis qa k lësës selgwasnokwa qaxs hë'maë awabewësa loq!wës domaqë. Wa, g'll'mësë gwalexs laē sösebenödzendeq qa lēxoyowēs yîxs 'nemdenaē wādzexg iwa- 15 sasa öba'yasa lõq!wē. Wä, la mõdenbalēda 'nemp!enk'ē lāxens g wag wax ts lana vēx vix wadzegovūwasa laxes k ak ilx alaena vē. Wā, lā öxsg'iwa'yas 'nemp!enk'östå lāxens ts!ex"ts!āna'yaxsens q!wāq!wax ts!āna'yēx. Wā, laemxia laxs 'wī'laen q!wāq!wax ts!ānatyēx Ļetwens qomax yix twalagāk ilasas. Wā, la tnempteng apa 20 ăwabâyasêxEns ts!Exºts!äna'yasEns g!wäg!wax'ts!äna'yêx vîx "wädzegabasas. Wä, lä möden läxens q!wäq!wax'ts!äna'yes yix

### 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, hëem daax<sup>u</sup>sa lõqwēlaēnoēwaxa lõq!wēs sõbayowaxs laē lāxa ālē. Wä, g-îl- <sup>e</sup>mēsē lāg-aa lāxa Lāsmadzexekūlaxs laē dõqŭx-<sup>e</sup>īdxa ēk-ētelaxa k-!eâsē L!enāk-a. Wä, hē<sup>e</sup>mēsēxs k-!ēsaē k-!îpela qa<sup>e</sup>s hë<sup>e</sup>maē dõqwasõsēda nāq!eqē lāx kŭxsentse<sup>e</sup>wē. Wä, g-îl<sup>e</sup>mēsē q!āxa ēk-axs laē sõp!exõdxa q!eL!ep!enx-se<sup>e</sup>sta lāxens q!wāq!wax-ts!ā- na<sup>e</sup>yēx, yîx wāg-it!ex,aasas. Wä, g-îl<sup>e</sup>mēsē t!ax-<sup>e</sup>īdexs laē tem- k-!õdxa <sup>e</sup>nemp!enk-ē lāxens bāLax qa lawäyēs t!emgūlts!ex,a<sup>e</sup>yas. Wä, lä bāl<sup>e</sup>īdxa mõp!enk-as <sup>e</sup>wāsgemas lāxens q!wāq!wax-ts!āna- <sup>e</sup>yēxs laē temx<sup>u</sup>sendeq. Wä, g-îl<sup>e</sup>mēsē lax-sexs laē kŭxsendeq nāq!eqax-dõmaqas. Wä, g-îl<sup>e</sup>mēsē k!ūxsaakūxs laē sõpâlax dõmaqas qa <sup>e</sup>nemdenēs lāwoyās hayaqaxa dõmaqē. Wä, lä aëk-!a sõpâlaq qa neqelēs. Wä, hē<sup>e</sup>mis qa k-!ēsēs selgwasnõkwa qaxs hë<sup>e</sup>maē ăwâbe<sup>e</sup>wēsa lõq!wēs dõmaqē. Wä, g-îl<sup>e</sup>mēsē gwālexs laē sõsebenõdzendeq qa lēxoyowēs yîxs <sup>e</sup>nemdenaē <sup>e</sup>wādzexg-iwa- sasa õba<sup>e</sup>yasa lõq!wē. Wä, la mödenbalēda <sup>e</sup>nemp!enk-ē lāxens q!wāq!wax-ts!āna<sup>e</sup>yēx yîx <sup>e</sup>wādzegoyuwasa lāxēs k-āk-îlxalaēna<sup>e</sup>yē. Wä, lä õxsg-iwa<sup>e</sup>yas <sup>e</sup>nemp!enk-õstâ lāxens ts!ex<sup>u</sup>ts!āna<sup>e</sup>yaxsens q!wāq!wax-ts!āna<sup>e</sup>yēx. Wä, laemx,a lāxs <sup>e</sup>wā laem q!wāq!wax-ts!ā- na<sup>e</sup>yēx µš<sup>e</sup><sup>e</sup>wins qõmax yîx <sup>e</sup>wālagak-îlasas. Wä, la <sup>e</sup>nemp!eng-apa ăwābâ<sup>e</sup>yasēxens ts!ex<sup>u</sup>ts!āna<sup>e</sup>yasens q!wāq!wax-ts!āna<sup>e</sup>yēx yîx <sup>e</sup>wālagak-îlasas. Wä, la <sup>e</sup>nemp!eng-apa

The Making o'o Dihg.). We, he'am da'axwsa lukwila'ino'iwaxa luk'wis subayuwaxs la'i laxa ali. We, gil-'misi laga'a laxa tlesmadzaxakulaxs la'i dukux'idxa ekitalaxa ke'osi t'lanaka. We, he'misixs kisa'i kipala ka's he'ma'i dukwasusida nakaki lax kuxsantsa'wi. We, gil'misi kaxa ekaxs la'i subayudxa kat'lapanxsa'sta laxans k'wak'waxt'sa- na'yix, yix wagitaxdla'asas. We, gil'misi tax'idaxs la'i tamkudxa 'nampanki laxans batlax ka laweyis tamgult'saxdla'asas. We, le bal'idxa mupankas 'wasgamas laxans k'wak'waxt'sana- 'yixs la'i tamxwsandak. We, gil'misi laxsaxs la'i kuxsandak nakakaxdumakas. We, gil'misi k!uxsa'akuxs la'i supolax dumakas ka 'namdanis lawuyos hayakaxa dumaki. We, le a'eka supolak ka nakalis. We, he'mis ka kisis salgwasnukwa kaxs he'ma'i awoba'wisa luk'wis dumaki. We, gil'misi gwalaxs la'i susabanudzandak ka lixuyuwis yixs 'namdana'i 'wadzaxgiwa- sasa uba'yasa luk'wi. We, la müdanbalida 'nampanki laxans k'wak'waxt'sana'yix yix 'wadzaguyuwasa laxis kakilxala'ina'yi. We, le uxsgiwa'yas 'nampankusto laxans t'saxwt'sana'yaxans k'wak'waxt'sana'yix. We, la 'amxda laxs 'wa'a'am k'wak'waxt'sa- na'yix dla'wins kumax yix 'walagakilasas. We, la 'nampangapa awabo'yasixans t'saxwt'sana'yasans k'wak'waxt'sana'yix yix 'wa-dzagabasas. We, le mudan laxans k'wak'waxt'sana'yis yix

### **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 Kwala 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

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

Mark only one oval

C	2	Yes
C	þ	No

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:

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



## Secondary Challenges

- Revising 1921
- Restore line numbers
- Restore missing titles
- How to make this model transferable to ther texts?
- Essential:
  - How to handle diverse formatting, layouts & arrangement of English an Kwak'wala?
- Nice to have:
  - Images & Figures

### Geographical Names of the Kwakiutl Indians

20.99;  $\bar{o}'gwiw\check{u}L\bar{e}^{\varepsilon}$  ( $\bar{o}-g\cdot iu-L\bar{e}^{\varepsilon}$ ) forehead on water; bay in front of  $X\check{u}'mdasb\bar{e}^{\varepsilon}$ , Hope Island — M 370, 391;  $\bar{o}'gwitEm\bar{e}^{\varepsilon}$  ( $\bar{o}-g\cdot it-$ 

14

TL. 473 5'ts lā'xa g'î'ldasē yîx la g'īts!E-

### $[g]_{Em-\bar{e}}$ head of body of round thin 22.135; ō'gwitemē<sup>e</sup>s nō'mas (ō-g·it-g. body of Old Man 7.112; $\bar{o}'gwitem \bar{e}^{\epsilon}s$ gag·ē'was 6.30; ō'gwitema<sup>e</sup>lis (ō-g·it-g of body 8.15; 8a.70; ō'gwitemēLē<sup>e</sup> (ō-g' on water 10.59; $\bar{o}k!\check{u}'n\bar{e}^{\varepsilon}$ $(\bar{o}-k!\hat{n}-\bar{e}^{\varepsilon})$ , Deserter's Island 6.47 – CXXVI 220; beach body, sand bar in river 13.50 CXXVI 104; ō'gŭma<sup>ɛ</sup>lis (ō-gem-a-l=is) $(\bar{o}-g_{Em}-l-!a)$ front rock 3.47 — R 1221. is) beach at river mouth 7.59; $\bar{o}x^{u}s\bar{i}'$ river 8.109; 22.136; $\bar{o}' x^{u} s \bar{i} dz \bar{e}^{\varepsilon} (\bar{o} - x \cdot s \bar{i} s =$ $22.67; \bar{o}' x^{u} s \bar{i} dz a^{\varepsilon} laa' (\bar{o} - x \cdot s \bar{i} s = a - l - a)$ row $(\bar{o}-x\cdot ts/\bar{e}(?)-l-a)$ front of long, steep i $\bar{o}' x^{u} s \hat{a} (\bar{o} - x s \hat{a})$ passage through 11.32; through 10.74; 14.4; $\bar{o}' x^u s \hat{a} l \bar{e} s E l a$ ( $\bar{o}$ -xs) through 14.115; $\bar{o}x\bar{a}'w\bar{e}^{\varepsilon}$ ( $\bar{o}$ -!xo- $\bar{e}^{\varepsilon}$ ) nec — III 149.22; $\bar{o}xsd\bar{e}'^{\varepsilon}$ ( $\bar{o}$ -xsd- $\bar{e}^{\varepsilon}$ ) hind 14.9; 18.103; 20.27; ōxsde<sup>e</sup>las (ō-xsdinside CXXVI 95; ōxsde<sup>ɛ</sup>li's (ō-xsd-8.3, 18; 21.4; 22.134; ō'xsdēlē<sup>e</sup> (ō-xs 10.119; $\bar{o}xsd\bar{e}L\bar{e}^{\varepsilon}sn\bar{o}'mas$ ( $\bar{o}$ -xsd-e- $L\bar{e}^{\varepsilon}$ -s of Old Man 7.108; $\bar{o}' x^{u} L \ddot{a}$ top of head 25 head beach 3.107; 18.7; 20.43 - $(\bar{o}-!x_{La}-ato-a-l=is)$ beach at hind end 20.89; $\bar{o}x_{La}' \in laa (\bar{o} - |x_{La} - l - |a)$ rock : 16.50; 17, between 15 and 16; ox1a<sup>6</sup> hind end 3.6; 6.93; 8.10, 19; 10.28,

ō'yaa (ōya-a) (ōya Nak!wax da<sup>ɛ</sup>x<sup>u</sup>; gwa rock 7.51; 10.69; ō'yaēlbē<sup>ɛ</sup> (ōya=ēl-b 10.215; ō'yabaa' (ōya-b-!a) outside r (ōya-b-ē<sup>ɛ</sup>) outward point 18.9; ō'yagE outward rock 4.7; 15.61; 17, between ō'ya-xsta<sup>ɛ</sup>lis (ōya-ăxst-a-l=is) beach a

 $\bar{o}'bEk^u$  ( $\bar{o}p$ -ku?) whispered (?) 15.143

- $(\bar{o}'manis)$ , Nootka name; 4.63 M 329
- ō'maxsdelis (ōm-exsd-el=is) 13.36, 43
- ōmaxsela 13.25
- ō's<sup>e</sup>Eq<sup>u</sup> (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-e 14.128; ō'sagŭmlis 16.3

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 deathbringer 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 Tewi'x i<sup>§</sup>lak<sup>u</sup>, and said, "O friend! take care. When they sing again, you must enter. They will name you at once  $D\bar{a}'bEnd$ .<sup>1</sup> 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  $D\bar{a}'bEnd$ .' 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.

â'Emxat! q!wē'ł <sup>®</sup> ēda. Wä, lā' <sup>®</sup> laē ē't!ēd yā'q!eg alēda yā'yaq!Ente <sup>®</sup> mē'lasēda g ō'kwē. only again they stopped. Well, then it is again spoke the speaker of the house of the house.
Wä, lā' <sup>8</sup> laē <sup>8</sup> nē'k·a: "Gē'lag·a, Hä' <sup>8</sup> la <sup>8</sup> mâlag, qa <sup>8</sup> s lā'ōs ē't!ēd dō'x <sup>8</sup> wida. Wë'g·a Well, then it is he said: "Come here, Mouse (woman), that you go again to see. Go on said
â'lax <sup>.e</sup> īd dō'qwax hä'łä g'ā'xEns; lā'g'iłasîk <sup>.</sup> k'!ēs hë'łē'lag a'da halā'yuk <sup>u</sup> ." really look for what us; the reason for not being right this death-bringer affects this this."
Wä, hë'x <sup>.</sup> <sup>g</sup> IdaEm <sup>g</sup> Iā'wisē Hä' <sup>g</sup> Ia <sup>g</sup> mâ'laga la qā's <sup>g</sup> ēd qa <sup>g</sup> s lä lā'wEls lā'xa Well, at once it is said the Mouse (woman) then walked, that went to go out at the she
ts!ä'gats!ä'sēda <sup>8</sup> mē' <sup>8</sup> mE'lxLOWē. Wä, lā' <sup>8</sup> laē hë <sup>8</sup> nā'kulaEm lāx <sup>8</sup> Wu'ndzasas ts!ē'ts!ēqa house of mountain-goats. Well, then it is she went at once to the hiding-place the said of
TEWI'x'i <sup>8</sup> lakwē. Wä, lā <sup>7</sup> 8laē <sup>8</sup> nē'k'a: " <sup>8</sup> ya, qāst, wë'g'a yā'L!âx, laE'ms lāL TEWI'x'i <sup>8</sup> lak <sup>u</sup> . Well, then it is she said: "O friend! go on take care, then you will said go
laë'LOL qa <sup>g</sup> xO ë't!EL dE'nx <sup>g</sup> ELO. Hë' <sup>g</sup> maā'qO laL Ļē'x <sup>g</sup> ELXES ĻE'gEmOSE Dā'bEnde, you will when will will sing. At once they will name your your name Dā'bEnd, enter, again will sing.
wä, lā'LES dEWI'L lā'xwa t!êx·î'läxsa g·ō'kwēx. Wä, lā'LES dā'x· <sup>8</sup> iLxwa well, then you jump in at the door of the house. Well, you will will take the will
ts!E'lts!Elk·ēx; lā'LES <sup>8</sup> nē'x·LõL: 'Nō'gwaEm Dā'bEnd,' qa <sup>8</sup> sō lāL dā'laLEq. feather here; then you you will 'I am Dā'bEnd,' when then hold it. will say: you will
Wä, hë'x <sup>8</sup> idaEmłwisōxda <sup>8</sup> wā'lasēx bEgwā'nEm <sup>8</sup> mEx <sup>8</sup> ē'LEq <sup>u</sup> . Wä, lāLaLES 1 Well, at once will this large this man will let go of it. Well, then however you will
dā'lax sä Em Lałxa ts!E'lts!Elk ē. Hë' <sup>8</sup> mēq." Wä, lā' <sup>8</sup> laē qā's <sup>8</sup> ēdē Hä' <sup>8</sup> la <sup>8</sup> mâlaga will but only hold the feather. That is it." Well, then it is walked the Mouse (woman) said
qa <sup>8</sup> s lē laē'L lā'xa g'ō'kwē. Wä, lā' <sup>8</sup> laē <sup>8</sup> nē'k'a: " <sup>8</sup> ya <sup>8</sup> nē <sup>8</sup> nEmōkwä'i, laE'mx'dEn that went to in the house. Well, then it is she said: "O friends! I have
she enter said
lē' <sup>g</sup> stalī'sEla lā'xwa awī' <sup>g</sup> stäxsEns <sup>g</sup> nā'lax," <sup>g</sup> nē'x <sup>g</sup> laē Hä' <sup>g</sup> la <sup>g</sup> mâ'laga. gone around the at this around of our world," said it is said the Mouse (woman).
world

<sup>1</sup> That means "to take hold of end."

46

3

L

5

10

## **First-Pass OCR**

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**

	179 INTERNATIONAL JOURN	AL OF AMERICAN LINGUISTICS NOL VI
	mäs- to get cockles at half-tide	doxwa nettle
- 00	mäsı/ zk-ala slanting (instead of ması/ zk-a	
	met- to spear	temx- to bud, to sprout
	mêx-îdê (Kwakiutl), mêx-ît (Koskimo) pore	u- lesa'la to press against (instead of tsesala)
	pine	tr'g ek ila a small bird (not teg eg ila)
	mêx"xwê'walis Myxocephalus (instead	of Igk-ala to fit tight (not Ignk-ala); stem Igk-
	mē <sup>a</sup> x <sup>u</sup> xwēwalis)	legulexs to pull out of canoe (instead of
222	"mg"wa salmon weir (instead of "mg"wd)	tegülexs)
0 - 0(	māx*- to nibble (fish)	tek"- to hang (instead of tek"-)
	must stand to push into water (instead	
	max <sup>*</sup> stend)	t zkova)
	mo'mx*sEm dried whole meat; stem mo	
	to til	l Equia)
	mõquala to stand aside	$l_{Eq}$ - mud (= $l_{Eq}$ -)
	möxüla not to mind dirt	trge's mud on beach
0 -	mõ'lex*la to stir	tg'x mas bile, gall (not tgx mas)
~	dem exclamation of surprise at success	
	some difficult act, (I hit it! he beat him	
	demā'k-āla to be out of sight	in excitement as, tglkūla to run (Koskimo) (instead of tglkwala)
	demgulas armpit (instead of demga <sup>4</sup> )	låk-ap/a to strike together; stem Igk-
	dEmx white eelgrass	da'gusa to eat before starting
	dena' to pull a rope	tenok" to pole canoe
0	dents/ek dancing-board (not dents/eq)	<i>lik</i> - to snap stones with fingers
00 -	denz*- to stand in a row	<i>ldx</i> - animal passes water
	dgnöxtäla to walk single file	14- to sprinkle
	dex- owl	tlox wid to spin (instead of tox wid)
	dgx- to spread wood, legs	to'lasa to chop in blocks
	d gl <sup>a</sup> yāla to lisp	tlemak to lash up (not tlem[a])
	d glà'las shovelnose canoe	tlemset wart (instead of tlemsadt)
22	dēk-ints!ēs gla fish jumps down beach (inste	
00 -	of dak-ints/esgla)	Rydb.)
	tlewāna liver (not dewāna)	tlesa' to pound, crack shells
	dēstötlk" dolphin (Newettee) (instead	
	děstô*lax)	t'entego to marry in one's own family
	daplenk squid bones	t'ensila to give to one's own relatives
	do" mag heart of wood (instead of doma	
	dögwig ? Ulna barbata; död gwig ? plu	
00 -	stem dok"- long extended line	tlegu'n a kind of canoe
	L'ost End to cut into a fluid (instead	
	döst end)	t/gq- to blame
	dösdgk/wa Rubus macropetalus Dougl.,	trlq"- soft (instead of t'rlk")
	-ursinus (= strung out on rock)	t/glq/a*a spawn in fish
	dök*- to be long extended, as a line, a v	
	dõgwil spinal column of porpoise dõx*dxgwis Menyanthes trifoliata	(= V/as- to break stone
0	strung out on beach, meadow)	(= <u>Uas- to break stone</u> <u>Uaq- round things lie on ground (not Uaq-)</u>
00 -	strung out on beach, meadow)	I had tonne emills ue on ground (not sad-
		15.160.110 on Wed, 23 Nov 2016 19:36.09 UTC http://about.istor.org/terms

- Advantages

Google Cloud Vision OCR - Text Detection Module

- 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'x<sup>3</sup>waLElōdayu lāx o'g'iwa<sup>3</sup>yas la hë'bendāla lax ō'xĻa<sup>3</sup>yas, lāĻaqēxs 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'x<sup>3</sup>waLElōdayu lāx o'g'iwa<sup>3</sup>yas la hë'bendāla lax ō'xĻa<sup>3</sup>yas, lāĻaqēxs hë'nałaēda xwā'xwagumLē ga gwälega).

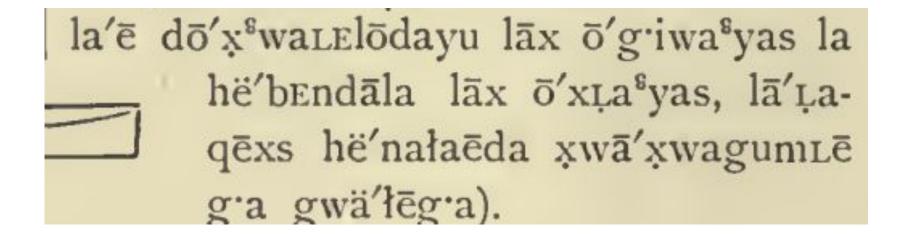
English text and other masked tokens are reintroduced at the appropriate indices. Orthographies can be converted (U'mista or SD-72)

Stretched rom therow right through To the stern but when the little Canoe is upright it is this way la'e do'x<sup>3</sup>waLElōdayu lāx o'g'iwa<sup>3</sup>yas la hë'bendāla lax ō'xĻa<sup>3</sup>yas, lāĻaqēxs hë'nałaēda xwā'xwagumLē ga gwälega).

### Evaluation

la'e do'x<sup>3</sup>waLElōdayu lāx o'g'iwa<sup>3</sup>yas la hë'bendāla lax ō'xĻa<sup>3</sup>yas, lāĻaqēxs hë'nałaēda xwā'xwagumLē ga gwälega).

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



## **Mixed-Methods Pipeline**

off, and the block falls on its back. k!waxia, nelele'le te'mkwa<sup>s</sup>yas.

Now the whole branch-side of the Wä, la<sup>s</sup>mē <sup>s</sup>wī'làwē L!E'nk'!ōt!Ena

wants to measure the middle of the bow of the canoe he is building and the middle of the 

to the stern; but when the little canoe is upright it is this way.)

under the knot. And at once it comes | nā'k·ē. Wä, lē hë'x·<sup>s</sup>idaem â'em nā'<sup>s</sup>na- 10

cedar is off. Then a cedar-bark rope yîsa wî'lkwē. Wä, hë'smis lā axsē'tsosa is taken by the canoe-builder, for he Lē'q!ēnoxwēs <sup>8</sup>mE'nyayuwē de'nsen dene'ma, qaxs hë'smaē sme'nse- 15

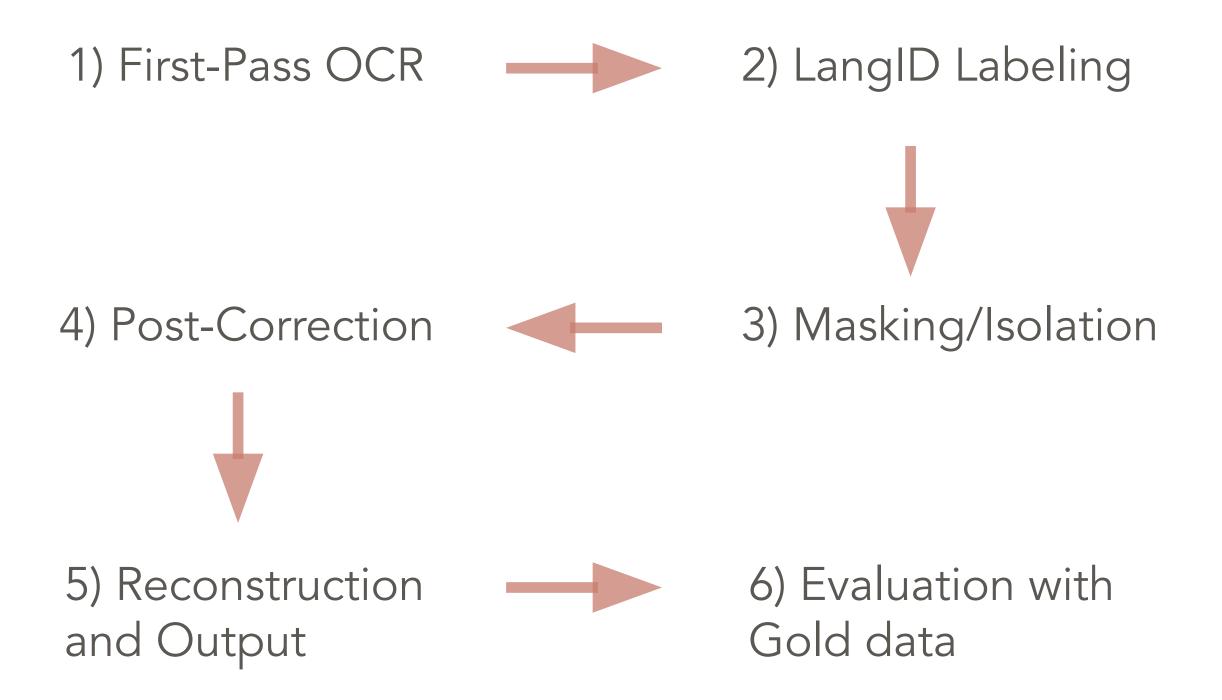
lasēxa <sup>s</sup>ne'gexslasa ö'g iwa<sup>s</sup>yasēs Lē'qase<sup>8</sup>wē Ļō<sup>8</sup> <sup>8</sup>negexsĻasa ō'xua gwä'leda wi'l-

I have unce it over in the drawing, | kwega, (yixs qo'qwate the latek qass so that you can see how the rope is do'qwalaosax gwa'laasasa dene'maxs 20 stretched from the bow right through | la'ē do'x<sup>s</sup>walelodayu lax o'g iwa<sup>s</sup>yas la

hë'bendāla lāx ō'xĻa<sup>s</sup>yas, lā'Ļaqēxs hë'nalaēda xwā'xwagumtē g·a gwä'łēg·a).

Then he marks the place where the | Wä, le xu'ltledex xi'ma'laasasa de- 25 rope is stretched from the bow to the ne'mē lā'xa ō'g'iwa<sup>\$</sup>ē Ļe<sup>\$</sup>wa ō'xĻa<sup>\$</sup>ē. stern. As soon as he has finished | Wä, g'í'lsmēsē gwā'lē mā'malt!ēk'!amarking it, he takes off the measuring- syasexs la'e la'wiyodxes me'nyayowe rope. Then he measures, beginning at dene'ma. Wä, le ba'l'idxa g'ä'g'iLela





### 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)

			5.1, 1902 SER		utl, 1909 SER
С	First Pass Corrected		25 2	0.33 0.15	18 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 <u>langID</u> modeling approaches for Kwak'wala (other than fastText)
- Create more <u>gold reference texts</u> for better evaluation
- More extensive <u>benchmarking</u> across more books/collections
- With consent of the language community, <u>share digitized texts</u> 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)
- (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

Transliteration into two community-preferred orthographies

