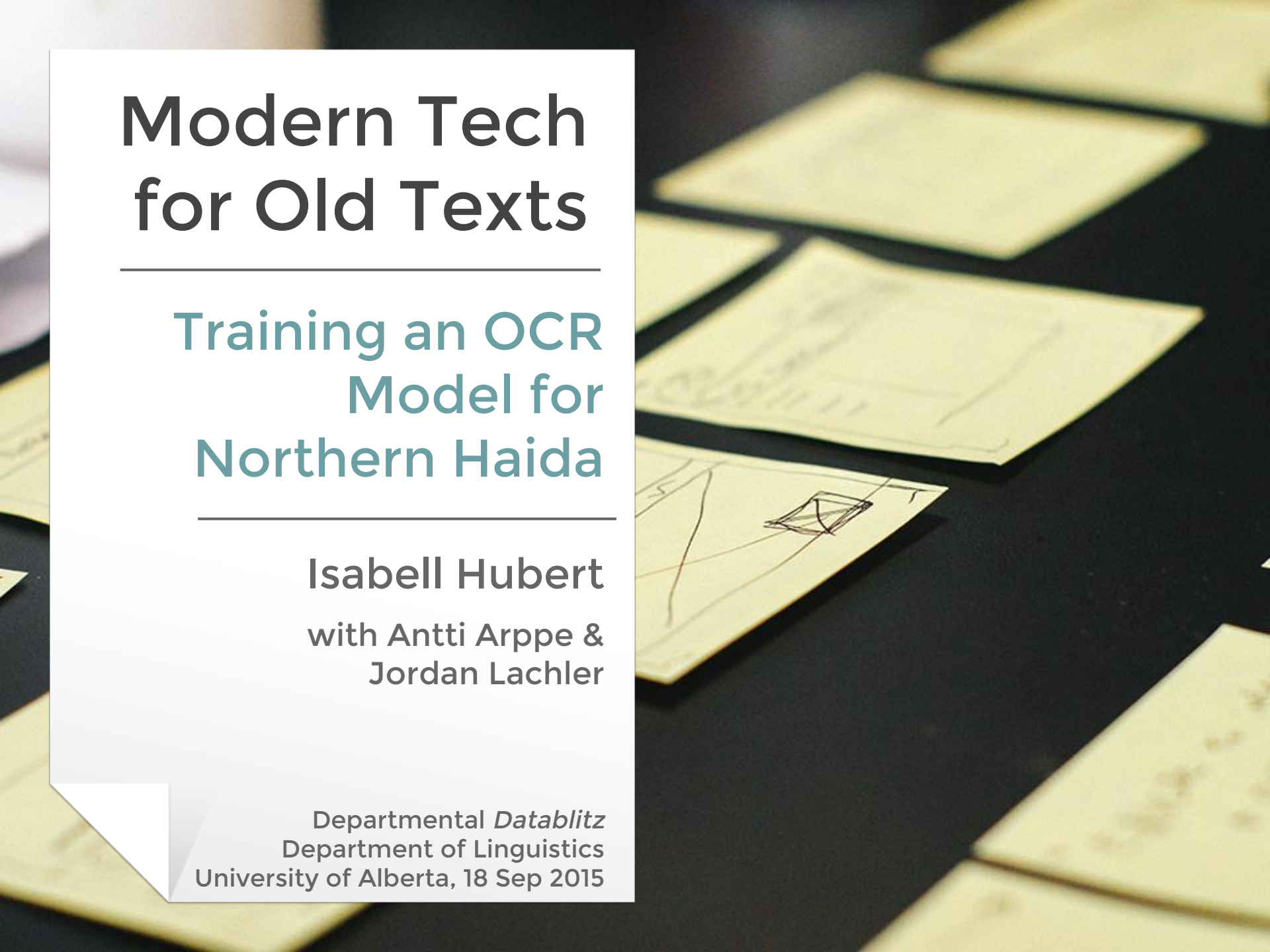


Modern Tech for Old Texts

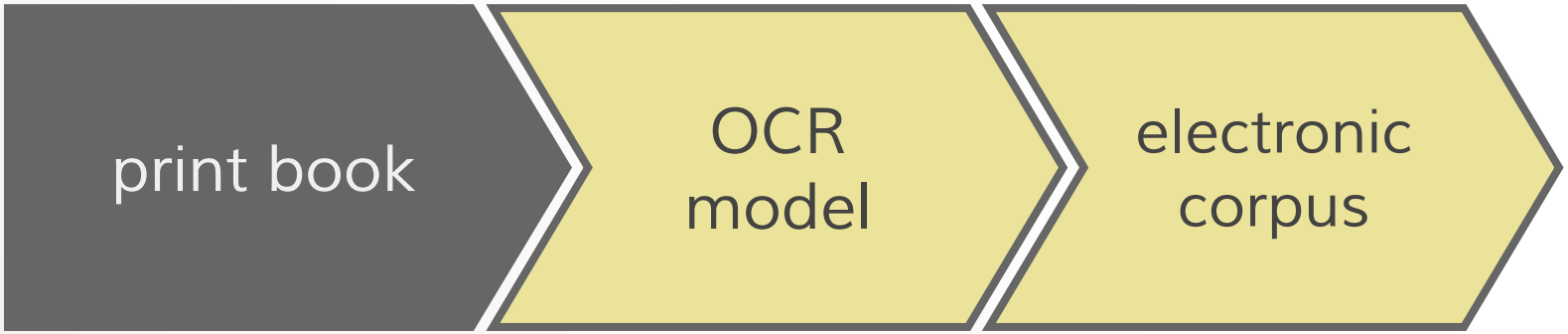
Training an OCR Model for Northern Haida

Isabell Hubert

with Antti Arppe & Jordan Lachler

Departmental *Datablitz*
Department of Linguistics
University of Alberta, 18 Sep 2015

The Overall Idea

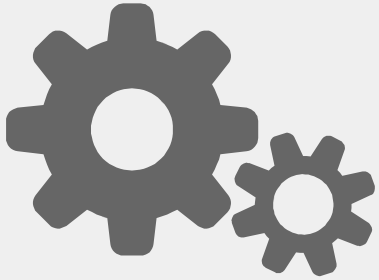


comparing & assessing different approaches

detailed documentation

OCR System

OCR Engine



- computational work
- language-independent
- *tesseract*

Language Definition/Model

ABC

- “knowledge”
- language-specific
- **this is what we do!**

OCR Model Training Issues

- no model yet
- lots of special characters
- no other OCR model in existence uses the same characters

⇨ pick the closest OCR model - in our case: Portuguese

- use this to OCR a couple pages

Fix Box Files

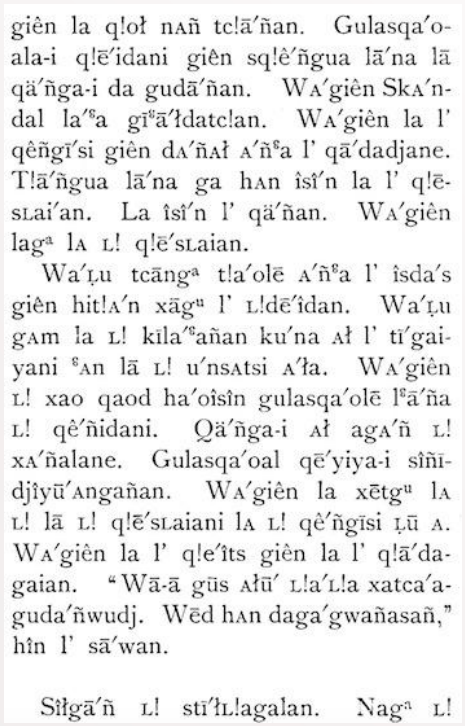
- box files define characters computationally
- you teach the model where it “went wrong”

⇨ **base model**

	Char	X	Y	Width	Height
1	1	702	657	12	36
2	'	720	657	6	12
3	s	766	670	17	23
4	ā	786	664	22	29
5	'	809	656	10	15
6	w	821	671	33	22
7	a	856	670	22	23
8	n	881	670	23	22
9	.	909	686	7	7
10	G	975	657	36	36
11	í	1013	659	11	34
12	ñ	1026	661	23	32
13	a	1051	670	22	23
14	'	1075	657	11	15
15	n	1088	670	23	23
16	L	1150	670	22	23
17	!	1176	658	6	36
18	w	1215	671	33	23
19	a	1250	670	23	23
20	'	1274	657	10	15
21	g	1286	671	27	34
22	a	1315	670	22	23
23	n	1340	670	23	23
24	.	1367	687	7	6
25	"	1436	660	14	15
26	H	1461	657	37	35
27	a	1501	670	23	23
28	-	1525	679	13	4
29	i	1539	657	12	35
30	d	704	718	24	36
31	i	730	725	11	28
32	g	770	731	27	35

OCR Larger Batches of Pages

- OCR 20-30 pages, validate & fix
- rinse & repeat until you have a whole corpus!

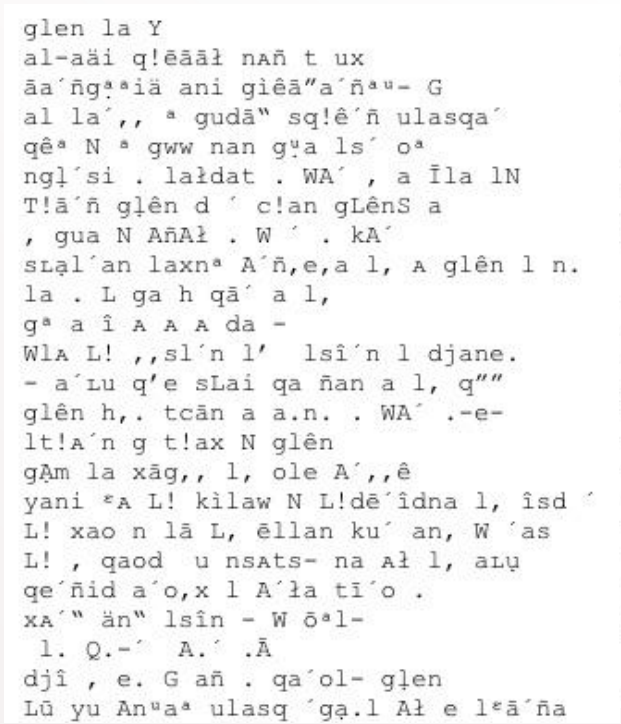


giên la q!oł nAñ tc!ā´ñan. Gulasqa´o-
ala-i q!ē´idani giên sq!ê´ñgua lā´na lā
qä´ñga-i da gudā´ñan. WA´giên SkA´n-
dal la´ᵉa gīᵉā´łdatc!an. WA´giên la l´
qêñgī´si giên dA´ñAł A´ñᵉa l´ qā´dadjane.
T!ā´ñgua lā´na ga hAn îsî´n la l´ q!ē-
sLai´an. La îsî´n l´ qä´ñan. WA´giên
lagᵃ lA L! q!ē´sLaian.

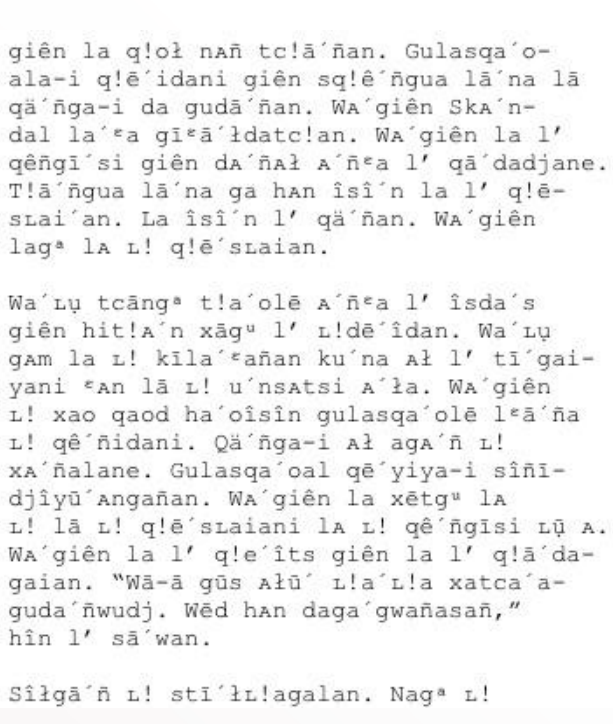
Wa´Lu tcāngᵃ t!a´olē A´ñᵉa l´ îsda´s
giên hit!A´n xāgᵘ l´ L!dē´idan. Wa´Lu
gAm la L! kīla´ᵉañan ku´na Ał l´ tī´gai-
yani ᵉAn lā L! u´nsAtsi A´ła. WA´giên
L! xao qaod ha´oîsîn gulasqa´olē lᵉā´ña
L! qê´ñidani. Qä´ñga-i Ał agA´ñ L!
xA´ñalane. Gulasqa´oal qē´yiya-i sîñī-
djîyū´Angañan. WA´giên la xētgᵘ lA
L! lā L! q!ē´sLaiani lA L! qê´ñgīsi Lū A.
WA´giên la l´ q!e´îts giên la l´ q!ā´da-
gaian. "Wā-ā gūs Ałū´ L!a´L!a xatca´a-
guda´ñwudj. Wēd hAn daga´gwañasañ,"
hîn l´ sā´wan.

Sîłgā´ñ L! stī´łL!agalan. Nagᵃ L!

Swanton (1908), p.184



glen la Y
al-aäi q!ēāāł nAñ t ux
āa´ñgᵃᵃiä ani gìêā"a´ñᵃᵘ- G
al la´,, ᵃ gudā" sq!ê´ñ ulasqa´
qêᵃ N ᵃ gww nan gᵘa ls´ oᵃ
ngL´si . lałdat . WA´ , a Īla lN
T!ā´ñ gLên d ´ c!an gLênS a
, gua N AñAł . W ´ . kA´
sLąl´an laxnᵃ A´ñ,e,a l, A glên l n.
la . L ga h qā´ a l,
gᵃ a î A A A da -
WlA L! ,,sl´n l´ lsî´n l djane.
- a´Lu q´e sLai qa ñan a l, q""
glên h,. tcān a a.n. . WA´ .-e-
lt!A´n g t!ax N glên
gĄm la xāg,, l, ole A´,,ê
yani ᵉA L! kìlaw N L!dē´îdna l, îsd ´
L! xao n lā L, ēllan ku´ an, W ´as
L! , qaod u nsAts- na Ał l, aLų
qe´ñid a´o,x l A´łа tī´o .
xA´" än" lsîn - W ōᵃl-
l. Q.-´ A.´ .Ā
djî , e. G añ . qa´ol- gLen
Lū yu Anᵘaᵃ ulasq ´gą.l Ał e lᵉā´ña



giên la q!oł nAñ tc!ā´ñan. Gulasqa´o-
ala-i q!ē´idani giên sq!ê´ñgua lā´na lā
qä´ñga-i da gudā´ñan. WA´giên SkA´n-
dal la´ᵉa gīᵉā´łdatc!an. WA´giên la l´
qêñgī´si giên dA´ñAł A´ñᵉa l´ qā´dadjane.
T!ā´ñgua lā´na ga hAn îsî´n la l´ q!ē-
sLai´an. La îsî´n l´ qä´ñan. WA´giên
lagᵃ lA L! q!ē´sLaian.

Wa´Lu tcāngᵃ t!a´olē A´ñᵉa l´ îsda´s
giên hit!A´n xāgᵘ l´ L!dē´îdan. Wa´Lu
gAm la L! kīla´ᵉañan ku´na Ał l´ tī´gai-
yani ᵉAn lā L! u´nsAtsi A´ła. WA´giên
L! xao qaod ha´oîsîn gulasqa´olē lᵉā´ña
L! qê´ñidani. Qä´ñga-i Ał agA´ñ L!
xA´ñalane. Gulasqa´oal qē´yiya-i sîñī-
djîyū´Angañan. WA´giên la xētgᵘ lA
L! lā L! q!ē´sLaiani lA L! qê´ñgīsi Lū A.
WA´giên la l´ q!e´îts giên la l´ q!ā´da-
gaian. "Wā-ā gūs Ałū´ L!a´L!a xatca´a-
guda´ñwudj. Wēd hAn daga´gwañasañ,"
hîn l´ sā´wan.

Sîłgā´ñ L! stī´łL!agalan. Nagᵃ L!

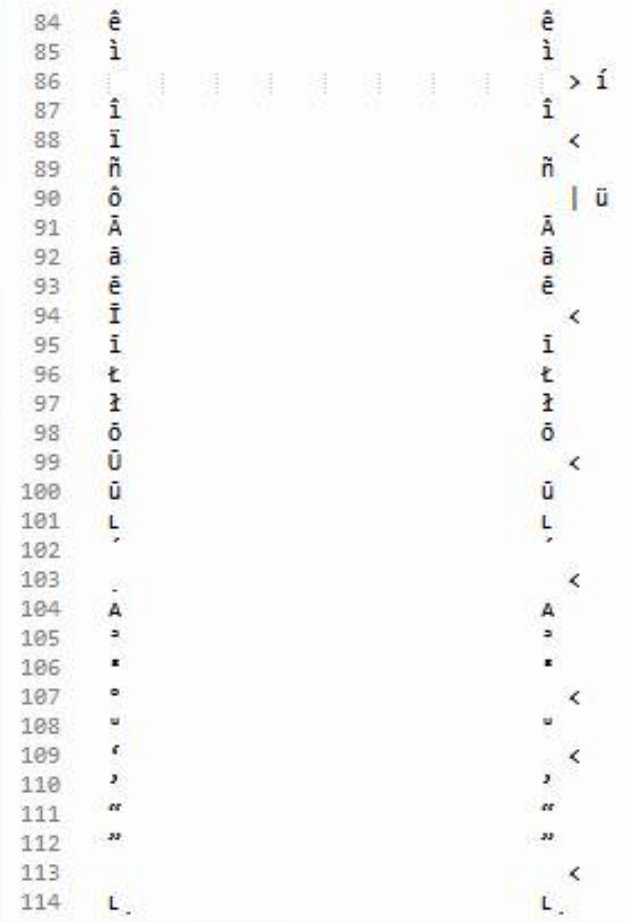
scan intermediate final

Assessing Initial Set Size

current stage

RQ: How many pages for the initial set are ideal?

- manual labour/ manpower/ accuracy tradeoff
- comparing models built from different set sizes to find sweet spot



Thanks!

References

Swanton, J. R. (1908). Haida Texts - Masset Dialect. In F. Boas (Ed.), The Jesup North Pacific Expedition, Memoir of the American Museum of Natural History (Vol. X). Leiden/New York: Brill & Stechert.

tesseract. Open source OCR Engine. Version 3.02.02. github.com/tesseract-ocr

Also, thanks heaps to:

Antti Arppe, Jordan Lachler, Megan Bontogon, Darren Flavelle, Catherine Ford, Evan Lloyd, Corey Telfer - and John R. Swanton!

template by SlidesCarnival

photos by unsplash & Death to the Stock Photo (license)