Training & Quality Assessment of an Optical Character Recognition Model for Northern Haida

Isabell Hubert, Antti Arppe, Jordan Lachler, Eddie A. Santos
University of Alberta
Edmonton, Alberta, Canada T6G 2E7
isabell.hubert@ualberta.ca, antti.arppe@ualberta.ca, lachler@ualberta.ca, easantos@ualberta.ca

Abstract

In this paper, we are presenting our work on the creation of the first optical character recognition (OCR) model for Northern Haida, also known as *Masset* or *Xaad Kil*, a nearly extinct First Nations language spoken in the Haida Gwaii archipelago in British Columbia, Canada. We are addressing the challenges of training an OCR model for a language with an extensive, non-standard Latin character set as follows: (1) We have compared various training approaches and present the results of practical analyses to maximize recognition accuracy and minimize manual labor. An approach using just one or two pages of Source Images directly performed better than the Image Generation approach, and better than models based on three or more pages. Analyses also suggest that a character's frequency is directly correlated with its recognition accuracy. (2) We present an overview of current OCR accuracy analysis tools available. (3) We have ported the once de-facto standardized OCR accuracy tools to be able to cope with Unicode input. We hope that our work can encourage further OCR endeavors for other endangered and/or underresearched languages. Our work adds to a growing body of research on OCR for particularly challenging character sets, and contributes to creating the largest electronic corpus for this severely endangered language.

Keywords: Optical Character Recognition, Endangered Languages, Indigenous Languages

1. Textual Source

This project uses John R. Swanton's pre-phonemic transcriptions of Northern Haida stories from the early 20th century (Swanton, 1908) as a basis, which – at more than 100,000 words and more than half a million characters – form the largest written corpus of Northern Haida in existence and are of excellent print quality: The established English OCR model for the Tesseract engine operated at a character recognition accuracy of 99.44% and a word recognition accuracy of 98.56%, close to (or above) accuracy levels reported for English in the literature (see Table 1 in Section 4.2.) Good image quality is crucial for any OCR effort as the "[a]ccuracy of single engines largely depends on the training sets created for each collection" (Boschetti et al., 2009, p.164), and as OCR engines "are known to be sensitive to the quality of the document images, with significant errors observed for even moderately degraded documents" (Dutta et al., 2012, p.1).

With native speakers only being found among the elderly (n.a., 2015a), and the language being far from fully described or analyzed, Swanton's books provide an excellent opportunity in language preservation, revitalization, and technological development efforts (cf. Section 5.) At the same time, our textual source is also of great interest to OCR research, where languages using special character sets still present challenges and where error rates are generally higher for old documents.

2. Comparison of Training Approaches

We have chosen to develop the Northern Haida language model for the *Tesseract* engine (Google, 2012), the most accurate open-source OCR engine available (Boschetti et al., 2009), sporting a transparent training system and being licensed as a GNU open-source project. While many sets of Tesseract language data are already available, indigenous languages are strongly underrepresented (n.a., 2015c); this can be remedied, however, due to the open source nature of the engine, making it accessible to researchers and language communities alike, especially considering that the use of Tesseract incurs no licensing fees and is not subject to copyright restrictions. In addition, the ability to modify the source code enables the development of further customizations and extensions, such as finite-state transducers (FSTs).

The training approach outlined in the Tesseract training pages,[1] an approach tested in the literature (Beusekom et al., 2008, for example), recommends generating a few fully correct pages of natural text in the language and, importantly, in the *font* that is used in the textual source. We will refer to this as the *Image Generation* approach. It requires the hand-typing of at least one page of text to generate the ground truth, and that the font used in the textual source be readily available. Our research has pointed us in the direction of *Mediæval* being the original print font (De Drukkerij E. J. Brill, 1932), with a digital representation being found in *Dutch Mediæval*; the latter, however, appears to be difficult to obtain, and also differs from the original font in the shape of its serifs and that of various capital letters. However, even if we had been able to retrieve the original font, text printed on a 100-year-old typesetting machine in comparison to that generated by modern typesetting software might differ too strongly in appearance to train a language model for the textual source (Nagy et al., 2000).

It is an ongoing debate whether generated images are more appropriate than real data in the training of an OCR model or not (Beusekom et al., 2008). Shapes in one font tend to resemble a different shape in another font more than a different shape in their own typeface (Nagy et al., 2000),

[1] See the Wiki on Google (2012).

which can result in recognition errors when the font used in the original text source is different from that in the generated images, as is the case in this project. This might be partially offset by the adaptive font classifier in Tesseract, but we have decided to compare the recommended Image Generation approach to an approach that makes use of scanned images of the original source text directly. Both approaches have been used in the creation of the *Project Gutenberg* (n.a., 2014) ebook database (Feng and Manmatha, 2006), but to this day there is no agreement as to which of those two approaches produces the better OCR models (Beusekom et al., 2008).

Instead of generating training images, the first step in this approach is to OCR a page from the original source text using an existing Tesseract language model. To minimize manual labor, it is advisable to choose a model with a character set similar to that used in the print source. In our case, the Portuguese language model came closest, also using diacritics present in Northern Haida (e.g. the circumflex and the tilde). After OCR'ing the respective number of pages using the Portuguese model, the resulting text files and box files were hand-validated using *jTessBoxEditor* (Nguyen, 2015) to generate the ground truth. The resulting language models were then being saved as the initial Haida models. We call this latter approach the *Source Image* approach. Both the Image Generation and Source Image models were trained on the same set of pages.

As almost all uses of OCR'ed text require post-processing (Nagy et al., 2000), an important aspect in the training of OCR language models is maximizing initial model accuracy while minimizing time-consuming manual labor. To our knowledge, there are currently no sources available that present the "sweet spot", in terms of pages required, between little manual labor and high accuracy. To amend, we have created twelve models in total, six in each of the two approaches outlined above. Within each approach, each of the six models is based on one to five or ten randomly selected pages (with about 1,200 characters per page) of generated/hand-validated text. Those twelve models were then tested for accuracy, as outlined in Section 4.

3. Accuracy Assessment Tools

Over the course of this project, we have worked with four quality assessment tools, all of which produce a measurement of OCR accuracy, either as a percentage value (Rice, 1996; Nagy et al., 2000), or its inverse, the error rate, on both the character (*character recognition accuracy - CRA*) and the word level *(word recognition accuracy - WRA*). While the CRA is the most detailed measure of model performance, the WRA can be considered a more important measure in corpus creation as only correctly recognized words are of use in search queries and information retrieval (Rice, 1996). In addition, the WRA can be very useful in determining whether errors are "clustered" and hence easier and quicker to fix for a human proofreader, or whether they are spread out more evenly across words, and hence harder to detect. It took considerable time to find assessment tools that would accept Unicode and provide the statistics we needed, so we initially explored implementing a Smith-Waterman string alignment algorithm based on the `swalign` python package (Breese, 2012), but abandoned this once the ISRI Unicode port (see below) was operational. We will present an overview over the tools below.

***ocrevalUAtion* (Carrasco, 2014)**
: This open-source tool, developed at the University of Alicante, deals favourably with Unicode text as input and computes both character and word accuracies in one go. Aggregate reports over a set of pages are available as well. The tool comes with a GUI, although we have found it buggy and resorted to the command line interface, which operates as intended. Out of all four tools we tested, this one produces the most "fancy" – interactive, even – output, where hovering over a character in the HTML output file highlights the corresponding character in the ground truth or the OCR output. The tool also provides per-character accuracy rates, but no character confusion matrices. Results obtained using this tool would also not have been immediately comparable to values in the literature, obtained using the ISRI tools.

***The ISRI Analytic Tools for OCR Evaluation* (Rice and Nartker, 1996)**
: This toolkit was used to compare the performance of various OCR engines in the *Annual Tests of OCR Accuracy* between 1992 and 1996 (Rice et al., 1994; Rice et al., 1995; Rice et al., 1996; Rice and Nartker, 1996). While it does not come with a GUI, it can be run easily from the Unix command line. Three aspects that make this toolkit suite stand out, and that – to our knowledge – no other tool or toolkit can provide, are that (1) it presents confusion matrices and accuracy values for single characters and words, (2) it comes with an extensive set of separate tools that each assess and highlight different performance metrics, and (3) it is the only toolkit suite in existence to have been used as a *de facto* standardized assessment tool. However, non-ASCII text was causing troubles to this toolkit from the mid-nineties, when OCR operations were still focused on purely ASCII texts (Beusekom et al., 2008). This could not be amended through character replacement operations, which would in addition skew character and word counts, so that we eventually decided to port the toolkit to a Unicode version (described below.)

***The ISRI Analytic Tools –Unicode Version* (Rice and Nartker, 2016)**
: We decided to adapt the once *de facto* standardized ISRI toolkit so that it would cope with Unicode input, and could hence cover most – if not all – languages of the world (see also Section 5.). At this point, the port is fully functional with regard to the `accuracy` and `wordacc` tools. In addition to the overall word and character accuracy rates, the tool also provides per-character accuracy rates and confusion matrices. Like its predecessor, it does not come with a GUI, but its command line usage is very straightforward. As it provides all the functionality we needed to asses model quality in this project, and as it makes com-

paring our values to values reported in the literature possible, all results reported in this paper will have been obtained using this toolkit.

4. Model Quality Assessments

In this section we present OCR accuracy analyses as instrumentalized through a number of variables, such as character coverage by types and tokens, and character and word recognition accuracies.

4.1. Character Coverage

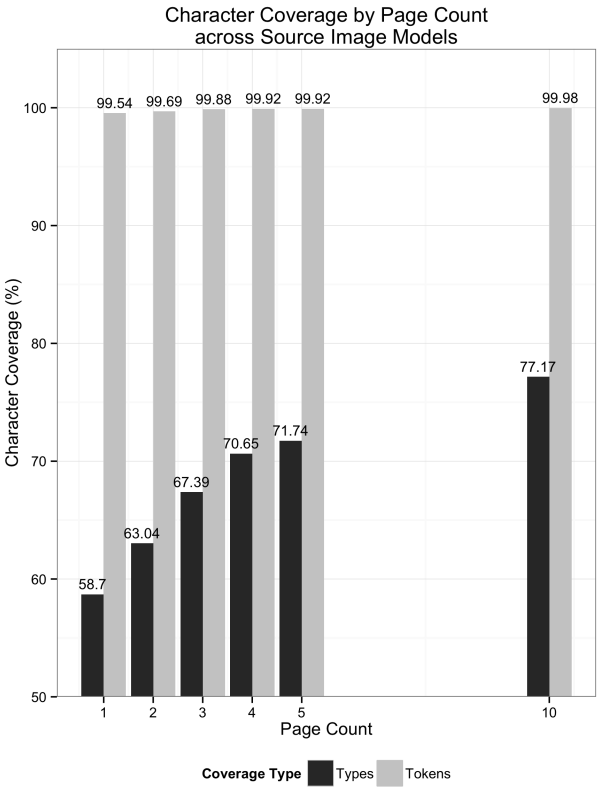

Figure 1: A visualization of *type* vs. *token* character coverage across the five different Source Image models.

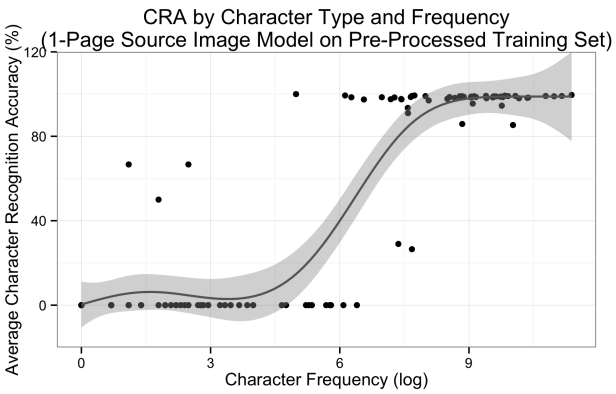

Figure 2: Visualization of generalized additive modelling of mean weighted CRA by log frequency.

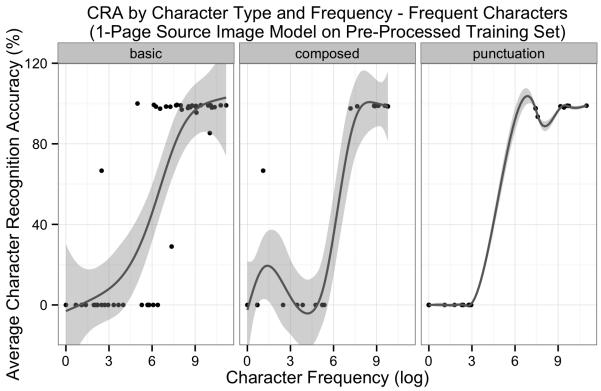

Figure 3: Visualization of generalized additive modelling of mean weighted CRA by character class and log frequency.

As a first measure, to gauge whether the initial models are accurate enough to identify most of the Haida character set, and also to see whether there might be persistent errors in the models (resulting from e.g. a wrong Unicode description in a box file), we first turned to creating character lists and counts for both the ground truth and the OCR output files.

It is important to note that the Swanton source texts are interspersed with some English comments. These comments introduce non-Haida characters into the OCR system, and present characters that can never be attained by a pure Haida model. For this reason, all pages containing English comments, 65 in total, were excluded from the below character coverage calculations, so that the character lists and counts created from the model outputs were not being held to unattainable standards.

The character lists were then compared to each other in a standard Unix `diff` operation. Character type coverage rose from 58.7% in the 1-page Source Image model to 77.17% in the ten-page model (cf. Figure 1), suggesting that adding more pages to a model makes the model recognize more target characters overall, hence making it perform better. As will be shown below, this is a false conclusion for several reasons (cf. Section 4.2.).

We have to keep in mind that this relatively low ceiling score of 77.17% describes *type* coverage. Looking at *token* coverage (cf. also Figure 1), we find that exactly half of the 92 character types in the text occur less than 1,000 times. Those 46 character types together make up 0.93% of text, with the other half of character types, namely those that occur more than 1,000 times each, making up 99.07% of text. With this in mind, we can illustrate the stark difference between *type* coverage – our earlier measure, where we discovered a ceiling of about 77% – and *token* coverage, which arguably is the more relevant measure: While type coverage seems low, actual token coverage is higher than 99% for all six Source Image models. This shows that (a) type coverage is a misleading measure in this context, and that (b) adding more pages to a model is not worthwhile from the perspective of character coverage.

On the level of individual characters, we worked with the

OCR output for all 499 pages generated by the one-page Source Image model for its superior performance, as will be described below. We then grouped characters into eight naturally occurring types (*basic, numbers, punctuation, dot below, smallcaps, superscripts, composed*, and *others*) to investigate whether CRA differed significantly depending on character class. A one-way ANOVA revealed that this was not the case ($p = 0.08$).

Unlike membership in a character class, the frequency of a character's occurrence was found to significantly influence its recognition accuracy, replicating the findings of the *Fifth Annual Test* (Rice et al., 1996, Tables 4a-4g). We observed a natural boundary at 403 occurrences, where 93% of all characters that occurred fewer than 403 times were never recognized correctly, and where characters that occurred more than 403 times had a 83% chance of being recognized correctly more than 95% of the time. Generalized additive modelling using the `mgcv` package (Wood, 2011) for `R` consequently indicated a significant effect of frequency on CRA across all character types ($F(5.37, 106.63) = 79.01, R^2 = 0.82, p < 0.001$, cf. Figure 2), but also within the basic ($F(3.51, 45.49) = 24.84, R^2 = 0.69, p < 0.001$), composed ($F(5.22, 15.78) = 31.47, R^2 = 0.91, p < 0.001$), and punctuation ($F(8.96, 11.04) = 21384, R^2 = 1, p < 0.001$) classes (cf. Figure 3) – all those subtypes with enough members to warrant the fitting of a non-linear model.

4.2. Recognition Accuracy

We expected:

1 The Source Image approach to result in more accurate models than the Image Generation approach overall, as we assumed that the font and typsettting used for Image Generation was too different from what was used in the print source; and

2 Those models based on more pages to be more accurate than those based on fewer pages, as we assumed prototypes to become more refined with more instances the engine had seen for a character.

We assessed the quality of the twelve initial models by measuring their character and word accuracies across the *assessment set* consisting of 499 pages, the full set of 509 pages minus the ten pages used for training, using the ISRI Unicode port introduced previously. Visualizations of the results can be found in Figures 4 and 5. Pages that exhibited abysmal recognition rates had to be pre-processed, including cropping (to remove black edges that would confuse the engine) and rotating (to counteract skewing and bending effects that remained from scanning the bound print book.) A set of t-tests showed that image pre-processing resulted in significant improvements for all Source Image and Image Generation models: Pre-processing improved the CRA by up to 1.8%, and the WRA by up to 1.42% (for detailed results see Table 2), so that all analyses below were obtained using the pre-processed image set.

The models that performed most accurately overall, as measured for accuracy on the pre-processed page set, were the two-page model for character recognition accuracy (*CRA* = 96.47%) and the one-page model for word recognition accuracy (*WRA* = 89.03%), both trained using the Source Image approach. In a set of t-tests, the difference in CRA between 96.47% for the two-page model and 96.30% for the one-page model was not found to be significant ($p = 0.77$); the same was found for the difference in WRA between 89.03% for the one-page model and 88.22% for the two-page model ($p = 0.11$). As adding a second page to the model improves the CRA less than it decreases the WRA, we recommend basing the initial model off a one-page model trained in the Source Image approach, as committing to the additional manual labour required to hand-validate another page does not seem warranted under those circumstances. While these results are not in line with our expectations, they certainly are good news for research projects where only a small amount of ground truth and/or manpower is available.

Language	Recognition Accuracy in %	
	Character	Word
English	99.53	93.60
Italian	99.46	94.59
Russian	99.33	94.43
Hebrew	96.80	89.42
Northern Haida	**96.47**	**89.03**
Japanese	95.74	81.28
Vietnamese	94.94	80.61
Thai	78.69	19.47

Table 1: Recognition accuracies of various established Tesseract language models (cf. Table 1 in Smith (2013, p.11)) in comparison to the most accurate Source Image approach model for Northern Haida.

To gauge whether our models perform appropriately, we have compared their performance to general accuracy rates reported for Tesseract and other engines, but also to error rates for old documents and documents written in a non-Latin script. Error rates generally seem to rise with increasing age of the document (cf. Table 1 in Mihov et al. (2005, p.3)), and performance on non-Latin scripts trails behind that of Latin scripts (cf. Table 1). Of the languages in this table, only Vietnamese uses diacritics as extensively as Northern Haida, rendering our model promising.

As is evident from these results, and again running counter to our expectations, the Image Generation approach did *not* consistently perform worse than the Source Image approach; in fact, in models based on four or more pages, it outperformed those models generated using the Source Image approach (cf. Figs. 4 and 5): In a simple linear regression model, adding one page to the Image Generation model was found to result in a 0.2% increase in WRA ($F(1, 2992) = 18.64, R^2 = 0.006, p < 0.001$), whereas adding pages had no significant effect on CRA. However, as determined by a one-way ANOVA, even the best model trained in the Image Generation approach performed significantly worse than the best Source Image model (CRA $F(1, 996) = 13.32, p < 0.001$; WRA $F(1, 996) =$

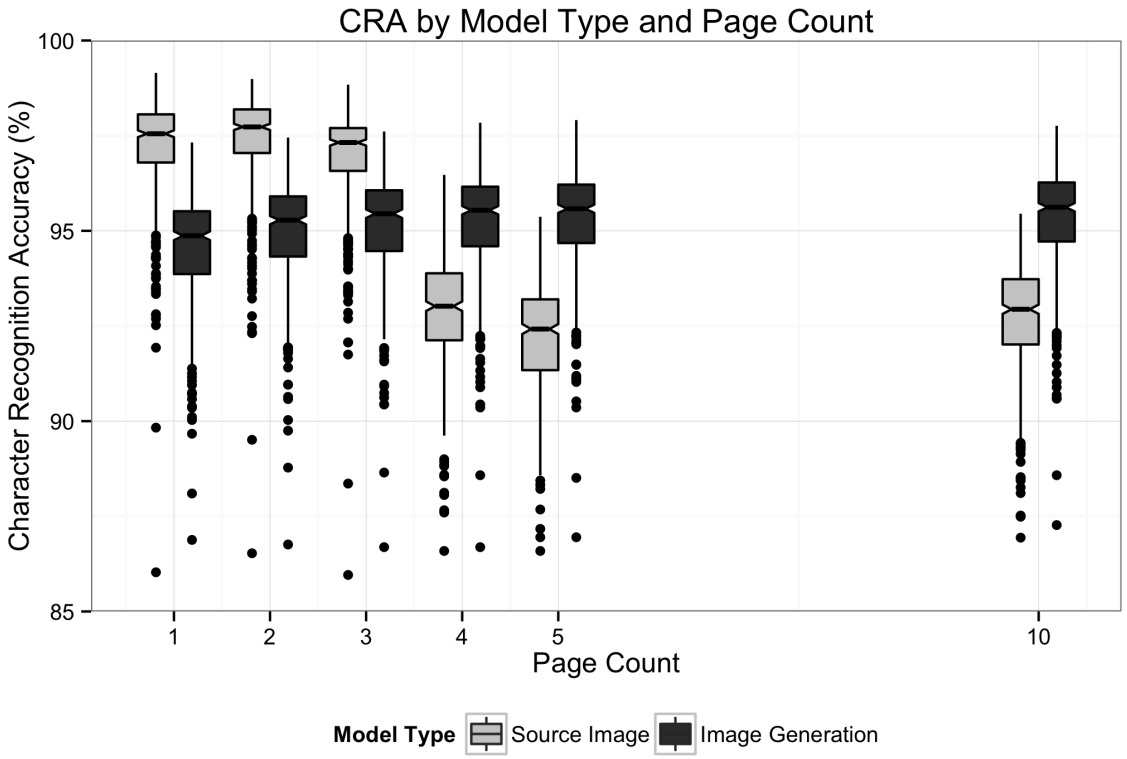

Figure 4: A visualization of character recognition accuracy by model type and page count. Note that some outlying data points are outside of the viewing area.

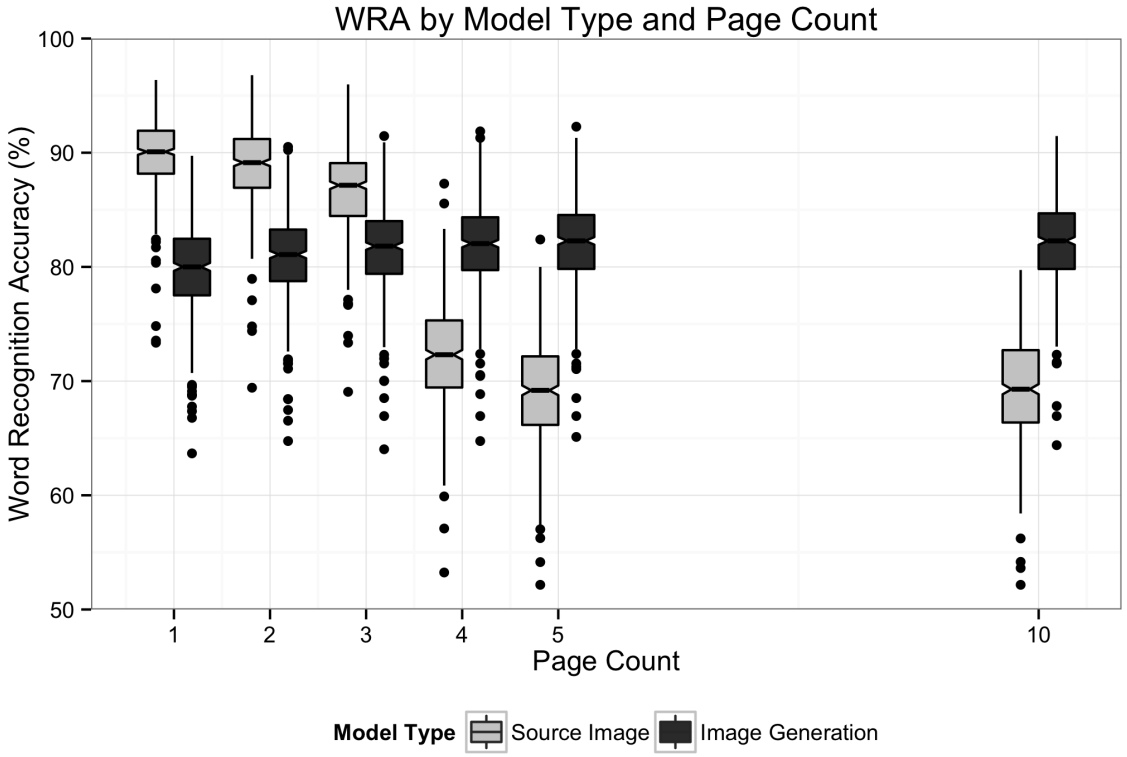

Figure 5: A visualization of word recognition accuracy by model type and page count. Note that some outlying data points are outside of the viewing area.

		Source Image Models			*Image Generation* Models		
		Original	*PP*	*t-test*	Original	*PP*	*t-test*
CRA	M	92.17%	93.97%	$t(4881.6) = -5.5$	92.38%	94.18%	$t(4809.4) = -5.65$
	SD	15.42	9.19		15.06	8.75	
WRA	M	77.44%	78.79%	$t(5679.4) = -3.86$	79.35%	80.77%	$t(5034.1) = -5.34$
	SD	14.95	11.80		12.34	7.74	

Table 2: The effects of image pre-processing (*PP*) through cropping and de-skewing on model accuracy rates as compared to the original, not pre-processed images. All values are significant at $p < 0.001$.

$226.2, p < 0.001$).

As the most accurate models overall were the one- and two-page Source Image models, basing a Source Image model on more pages crucially did *not* result in more accurate model performance. Quite the opposite was true, as shown by linear modelling: Adding one page to a Source Image model resulted in the CRA decreasing by 0.59% ($F(1, 2992) = 108.1, R^2 = 0.03, p < 0.001$) and in the WRA decreasing by 2.45% ($F(1, 2992) = 1809, R^2 = 0.38, p < 0.001$), effectively being detrimental to model performance. We currently have no adequate explanation for this behavior, but we suspect that a "fuzzying out" of the respective prototypes when too many exemplars are added might play a role. The resulting U-shaped graph is reminiscent of phonological acquisition in children (see e.g. Fig. 7 in Becker and Tessier (2011)), and other general learning behaviors (Carlucci and Case, 2013).

To investigate whether the high accuracy for the one-page Source Image model was an artifact of random page selection, we trained ten one-page models in the Source Image approach, each model based on a different page of the training set. A one-way ANOVA revealed that the choice of page indeed had a significant influence on CRA ($F(9, 4980) = 34.14, p < 0.001$) and WRA ($F(9, 4980) = 471.8, p < 0.001$). It is hence advisable to ensure that the page selected for training is of good print quality; In addition, if possible, it is also advisable to train two to three one page models, and then select the one that performs best. This approach is likely to result in a very accurate model from the start, reducing the possibility of accidentally basing one's model on a page not apt for training, while at the same time not requiring excessive amounts of manual labor.

To test our claim that our Northern Haida model would be a good starting point for further OCR endeavors for other indigenous language, we have begun comparing its performance on Southern Haida text with that of the out-of-the box English and Portuguese models. Southern Haida, like Northern Haida a member of the Haida macrolanguage, only shows borderline intelligibility with Northern Haida (n.a., 2015b). Initial results seem promising: Using the approach outlined above, we were able to train an OCR model for Southern Haida within just two days (including page selection, image pre-processing, and producing ground truth for the testing set) that shows recognition accuracies around those of the established Thai model (cf. Table 1). These results certainly seem promising, suggesting that the Northern Haida model might be an appropriate starting point for other OCR endeavors if the character set is similar.

5. Research Contributions & a Look into the Future

In addition to the above findings illustrating that those Source Image models based on just one or two pages seem to perform best overall, our efforts will contribute to the existing body of research and the linguistic community in a number of ways:

1. We have ported the once *de facto* standardized ISRI toolkit to a Unicode version. This is crucial as "[l]arge-scale, automated tests are needed in which expressive and precise measures of performance are computed" (Rice, 1996, p.71), and as "[p]rogress in page-reading technology depends on thoughtful, multi-faceted evaluation" (Rice, 1996, p.73).

2. The OCR model resulting from our efforts will be among the first for a North American indigenous language and can be made freely available.

3. Our work can encourage other researchers and community members to develop OCR models for other underresearched and/or indigenous languages, similar to the approach taken in e.g. Mihov et al. (2005).

4. Initial results obtained by applying our model to Southern Haida suggest that it can potentially serve as a good base model for other OCR endeavors if the character sets are similar.

5. The electronic corpus of Northern Haida that will result from our OCR efforts will be the largest to date (at around 107,000 words and 620,000 characters), and will enable the extraction of e.g. word and letter frequencies and co-occurrences of elements in Northern Haida.

6. This corpus would then allow for quantitative linguistic research, informing the development of electronic dictionaries, translation applications, and even interactive learning materials to assist ongoing documentation and revitalization processes.

7. As a further immediate real-world application, this project also contributes to recording Haida culture and national heritage by transforming traditional stories and myths into a searchable, shareable, and widely accessible electronic format.

Further plans for the project include combining the OCR model with an FST to potentially improve accuracy, thus addressing the problem of spell-checkers and simple word lists being inadequate in increasing recognition accuracy, especially in morphologically rich languages (Boschetti et al., 2009; Smith et al., 2009; Smith, 2013). This FST could then enable researchers, but also students and language learners, to e.g. highlight a word in the corpus, and immediately receive information on the word's morphological makeup. We think that this will be an invaluable resource in language revitalization and teaching, especially in a morphologically rich language such as Northern Haida. A further extension of this FST could be its use as a transcriptor, to automatically transform Swanton's pre-phonemic orthography into the modern Haida orthography.

6. Acknowledgements

This research has been funded by the *Social Sciences and Humanities Research Council (SSHRC) Partnership Development Grant (890-2013-0047), 21st Century Tools for Indigenous Languages,* and the *Kule Institute for Advanced Study (KIAS) Research Cluster Grant: 21st Century Tools for Indigenous Languages*, the financial support of which is gratefully acknowledged. The authors would like to thank Megan Bontogon, Darren Flavelle, Catherine Ford, and Evan Lloyd for their efforts in hand-validating and hand-typing Swanton text; Darren Flavelle for his efforts in tracking down the original source text font; Corey Telfer for his Haida orthography charts and his remarks with regards to U-shaped learning; Dustin Bowers for helpful tips on string alignment algorithms and their Python code skeletons; Anne-Michelle Tessier, Kaidi Lõo, and Miikka Silfverberg for various helpful suggestions along the way; and the members of ALTLab for their suggestions for the future of this project.

7. Bibliographical References

Becker, M. and Tessier, A.-M. (2011). Trajectories of faithfulness in child-specific phonology. *Phonology*, 28:163–196.

Beusekom, J. V., Shafait, F., and Breuel, T. M. (2008). Automated OCR Ground Truth Generation. In *The Eighth IAPR International Workshop on Document Analysis Systems*, pages 111–117.

Boschetti, F., Romanello, M., Babeu, A., Bamman, D., and Crane, G. (2009). Improving ocr accuracy for classical critical editions. In *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL'09, pages 156–167, Berlin, Heidelberg. Springer-Verlag.

Carlucci, L. and Case, J. (2013). On the Necessity of U-Shaped Learning. *Topics in Cognitive Science*, 5(1):56–88.

De Drukkerij E. J. Brill. (1932). *Letterproef*. Brill Archive, Leiden, Netherlands.

Dutta, S., Sankaran, N., Sankar, K. P., and Jawahar, C. V. (2012). Robust Recognition of Degraded Documents Using Character N-Grams. In *IAPR International Workshop on Document Analysis Systems*, pages 130–134, mar.

Feng, S. and Manmatha, R. (2006). A hierarchical, HMM-based automatic evaluation of OCR accuracy for a digital library of books. pages 109–118.

Mihov, S., Schulz, K., Ringlstetter, C., Dojchinova, V., Nakova, V., Kalpakchieva, K., Gerasimov, O., Gotscharek, A., and Gercke, C. (2005). A corpus for comparative evaluation of OCR software and post-correction techniques. *Proceedings of the 8th International Conference on Document Analysis and Recognition*, pages 162–166.

Nagy, G., Nartker, T. A., and Rice, S. V. (2000). Optical Character Recognition: An Illustrated Guide to the Frontier. In *Proceedings of the SPIE Conference on Document Recognition and Retrieval*, pages 58–69.

Rice, S. V., Kanai, J., and Nartker, T. A. (1994). The third annual test of OCR accuracy. http://stephenvrice.com/images/AT-1994.pdf, accessed 3 Mar 2016.

Rice, S. V., Jenkins, F., and Nartker, T. (1995). The fourth annual test of OCR accuracy. http://stephenvrice.com/images/AT-1995.pdf, accessed 3 Mar 2016.

Rice, S. V., Jenkins, F. R., and Nartker, T. A. (1996). The fifth annual test of OCR accuracy. http://stephenvrice.com/images/AT-1996.pdf, accessed 3 Mar 2016.

Rice, S. V. (1996). *Measuring the Accuracy of Page-reading Systems*. Ph.D. thesis, Las Vegas, NV, USA.

Smith, R., Antonova, D., and Lee, D.-S. (2009). Adapting the Tesseract Open Source OCR Engine for Multilingual OCR. In *Proceedings of the International Workshop on Multilingual OCR*, pages 1–8.

Smith, R. (2013). History of the Tesseract OCR engine: what worked and what didn't. In Richard Zanibbi et al., editors, *Proceedings of the Conference on Document Recognition and Retrieval*.

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

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36.

8. Language Resource References

Breese, M. (2012). swalign v0.2. https://pypi.python.org/pypi/swalign/0.2.

Carrasco, R. C. (2014). ocrevalUAtion 1.3. https://sites.google.com/site/ textdigitisation/ocrevaluation.

Google. (2012). tesseract v3.2.2. https://github.com/tesseract-ocr, accessed 21 Nov 2015.

n.a. (2014). Project Gutenberg. www.gutenberg.org, accessed 22 Nov 2015.

n.a. (2015a). Northern Haida. In Lewis M. Paul, et al., editors, *Ethnologue: Languages of the World*. SIL International, Dallas, Texas. Online version: http://www.ethnologue.com.

n.a. (2015b). Southern Haida. In Lewis M. Paul, et al., editors, *Ethnologue: Languages of the World*. SIL International, Dallas, Texas. Online version: http://www.ethnologue.com.

n.a. (2015c). tesseract Language Data. https://github.com/tesseract-ocr/langdata, accessed 21 Nov 2015.

Nguyen, Q. (2015). jTessBoxEditor 1.4. http://sourceforge.net/projects/vietocr/files/jTessBoxEditor/, accessed 21 Nov 2015.

Rice, S. V. and Nartker, T. A. (1996). The ISRI Analytic Tools for OCR Evaluation.

Rice, S. V. and Nartker, T. A. (2016). The ISRI Analytic Tools for OCR Evaluation - Unicode Version. Ported and extended by Eddie Antonio Santos. https://github.com/eddieantonio/isri-ocr-evaluation-tools, accessed 05 Jan 2016.