Handwriting Recognition Without Human Annotation

Patrick Schone
patrickjohn.schone@familysearch.org
FamilySearch, 50 E North Temple, Salt Lake City, UT

ABSTRACT

Handwriting recognition is needed in many languages, but the cost to create it is usually prohibitive. Therefore, we explore the ability to create handwriting recognition systems which are built without human annotation. Our process starts by leveraging existing data from other languages in the same script and coupling this with image synthesis and iteratively-discovered phrase redundancy. We illustrate a case example of moving performance from 0% word accuracy to 81% word accuracy in Hungarian by taking advantage of the whole process we describe here. We then demonstrate the same principles and effects on four other languages: Dutch, Polish, Icelandic, and Estonian.

1. BACKGROUND

FamilySearch is a non-profit genealogical organization which grants free access to billions of historical images and to a family tree of the human family. These historical images are particularly useful because they help provide the source information that patrons can use to extend and refine the family tree. However, for patrons seeking to use this data, easy access to the records is critical.

For many years, FamilySearch has had a massive crowd-sourcing effort to "index" these records – that is, to extract key facts and associations from the documents – which involves hundreds of thousands of volunteers each year. However, there is a sparsity of volunteers in the majority of languages, and even for those where there are volunteers, the work to do is enormous. Barring some sort of access to the content of the document, the only other option for find one's ancestors would be to search one image after another – an extremely tedious and time-consuming effort.

Consequently, to enable faster access for patrons and to reduce the workload for volunteers, FamilySearch, since 2015, has been automating extracting the semantic content of these images. Handwriting recognition, or in other words, the automatic transcription of handwritten documents, along with its sister technology – OCR -- for transcribing print are primary technologies used for performing this extraction process. FamilySearch, supported also by partners such as BYU, has built a system for jointly transcribing printed and handwritten images [1] and has gone on to apply it to hundreds of millions of images and provide patrons with the associated semantic access -- especially in English, Spanish, and Portuguese.

To date, Family Search has capabilities in varying degrees to transcribe dozens of different languages including some that are now extinct (like Manchu). However, a majority of the languages that FamilySearch currently supports operationally could benefit from further work *and* there are at least 70-100 more languages that are needed but have no current capability. Development of recognizers for new languages from the ground up can cost tens if not hundreds of thousands of dollars. Such a large expense is mostly intractable for a non-profit organization like ours.

Yet as one considers all the languages of the world for which the genealogical community has record access, it turns out that a majority of languages use common scripts such as Arabic, Cyrillic, Devanagari, and *especially*, Latin. Thus a question arises: might it be possible to leverage existing handwriting recognition capabilities, the creation of synthetic images, plus existing resources or the Web to to create new recognition capabilities without having any human annotation? If we focus just on Latin-scripted languages, for example, FamilySearch has substantial capability in Italian, French, Swedish, Norwegian, Danish, Vietnamese, and German in addition to the three others mentioned earlier. Can we use these to bolster the creation of recognizers for languages we do not yet support?

In this paper, we demonstrate that if sufficient numbers of untranscribed historical images exist, it is *indeed* possible to build recognizers without any specific annotation for other same-script languages. Moreover, the accuracies of recognizers in those languages can approach or rival those where human annotation was even extensively involved. In particular, we will here apply these techniques to the following languages for which we have had *zero* training data: Hungarian, Dutch, Polish, Estonian, and Icelandic. It should be mentioned that our training languages come almost exclusively from Romance or Germanic languages; but the demonstration languages includes other families such as Uralic, West Slavic, and Finnic. This shows the portability of this approach beyond language borrowings.

2. TRANSCRIPT-LEARNING PROCESS

2.1. System Components

The elements of our system are depicted in Figure 1. It has four main components: (a) image synthesis; (b) access to a trainable recognizer (visual and possibly language modeling) that uses the same script as the target language; (c) universal line segmentation; and (d) web-mining (or, if it exists, index mining) for acquiring language-specific lists of names, places, and numeric expressions which will be used for filtering potentially spurious results. These four components are used in a process that will start to hypothesize transcriptions in the target language and which will then be iteratively refined.

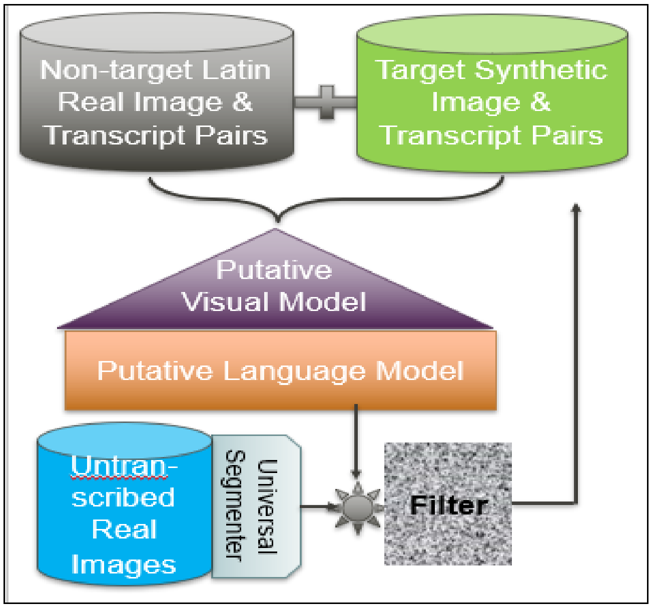
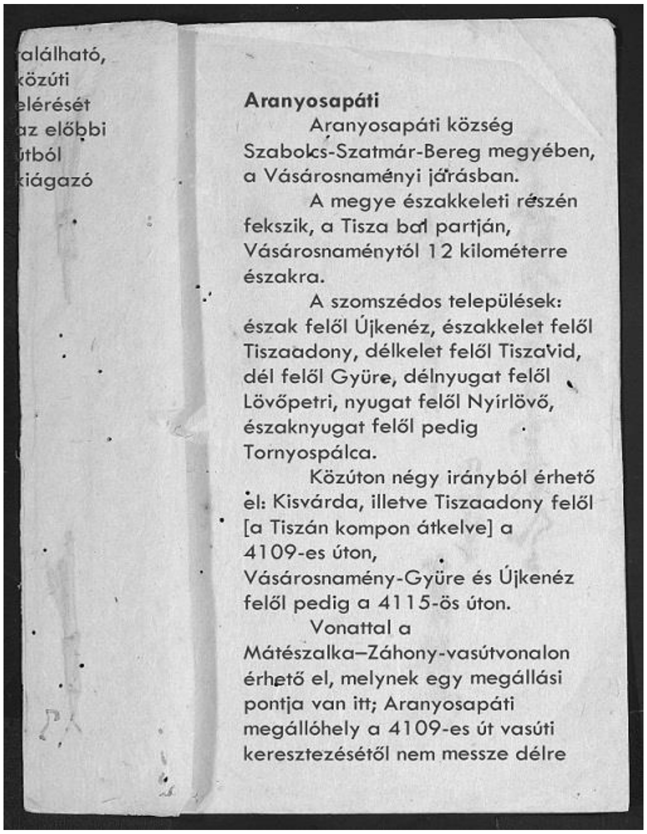

Non-target Latin Real Image & Transcript Pairs
Target Synthetic Image & Transcript Pairs
Putative Visual Model
Putative Language Model
Untran-scribed Real Images
Universal Segmenter
Filter

Figure 1. Overview of the full process

2.1.1. Image synthesis: We have created a simple capability to generate images in any language of choice. These synthetic images in many cases look authentic but, of course, are fictitious (though the text thereon is legitimate). Such an image is shown in Figure 2. We begin this process by acquiring thousands of fonts from around the world and we especially focus on fonts that appear to be handwritten in nature. We then use images from FamilySearch's collections that are authentic but are blank. We mark those images with regions where texts could ostensibly be added. Next, we download large samples of Wikipedia in the target languages and distill out the text. We then leverage that "wikitext" to populate the image's blank regions with rendered text. The rendered texts use the myriad fonts while also automatically incorporating character drift, background noise, bolding, character resizing, and other phenomena.

2.1.2. Same-script Recognizer: We gather together our existing training data from all the other languages that we have in the language script of interest, and we augment that data collection with synthetically-produced images and their associated wikitext-based transcripts in the target language (about 500-1000 such images/transcript pairs). We then build two initial putative models in the language of interest: (a) a visual model of what the system thinks it is seeing, and (b) a language model for the character sequences that makes most sense in the target language based on the transcript collection. The visual model trainer has been equipped with functionality which lets us weight the training samples from each language and modality so that there is significant breadth in each recognition batch and no one language or mode dominates the training process. The language model can be built using any subset of the training data, so we allow to see all languages, but we limit the quantity of non-target languages.



található,
közúti
elérését
az előbbi
útból
kiágazó

Aranyosapáti
Aranyosapáti község
Szabolcs-Szatmár-Bereg megyében,
a Vásárosnaményi járásban.
A megye északkeleti részén
fekszik, a Tisza bal partján,
Vásárosnaménytól 12 kilométerre
északra.
A szomszédos települések:
észak felől Újkenéz, északkelet felől
Tiszaadony, délkelet felől Tiszavid,
dél felől Gyüre, délnyugat felől
Lövőpetri, nyugat felől Nyírlövő,
északnyugat felől pedig
Tornyospálca.
Közúton négy irányból érhető
el: Kisvárda, illetve Tiszaadony felől
[a Tiszán kompon átkelve] a
4109-es úton,
Vásárosnamény-Gyüre és Újkenéz
felől pedig a 4115-ös úton.
Vonattal a
Mátészalka–Záhony-vasútvonalon
érhető el, melynek egy megállási
pontja van itt; Aranyosapáti
megállóhely a 4109-es út vasúti
kereszteződésétől nem messze délre

Figure 2. Synthetic Hungarian image

Universal Line Segmentation: We have an image segmentation system that was developed using samples from scores of languages – but it was not customized in any way for the target languages of this paper. Even so, the same model can be used for almost any language, whether it be vertically-oriented, horizontally-oriented, or both. The system identifies about two dozen different kinds of image phenomena, but in particular, it finds the lines of text in images. For each such line, the segmenter will emit a png file that represents a normalized form of that text line (which we will here call a "snippet") as well as any associated metadata. The snippets can be auto-transcribed using the handwriting recognition system. The associated metadata can be used after recognition for determining bounding boxes of the hypothesized words and phrases in each line.

Web or Index Scraping: As a final system component, we mine websites and/or any available indexes for word lists of several kinds. We are not using these lists to train the recognizer per se, but to assess the validity of recognizer predictions. We identify valid **people names** -- both given names and surnames -- and we record various combinations of names in the order they are likely to appear in the specific language. For example, a list of Hungarian baby names may say that "Istvan" and "Gavrila" are commonly-occurring male and female names, respectively [2]. We also distill from the various data sources **place names** that are likely to occur in the target language (countries, states, cities, mountains, etc). Next, we work to find **numeric expressions** (such as “twenty-fifth”) and **date elements** (like “March”) in the language since these are highly likely to occur in genealogical documents and are often fully spelled out (eg., “twenty-third day of December in the year of our Lord one thousand eight-hundred and five”). Lastly, we optionally mine **occupations** since these have genealogical value and may also occur in average documents.

2.2. Using the Elements Iteratively

To begin the process of getting the recognizer to adapt to the language, we select thousands of previously-untranscribed target images in the language of choice (about 5000 is a typical starting amount). We run our image segmentation process on these images to extract the snippets that are believed to be text lines. These snippets are then passed to the recognizer (visual modeling coupled with language model) to produce a first-pass transcription for the untranscribed files. Our expectation is *not* that we will initially have good results across the board. Rather, we hope there will be some emitted strings and patterns that are reoccurring or that match words from the lists derived from web scraping. Repetition, especially if it is long, often indicates correctness or near correctness. We hope to be able to exploit these repetitions to get the recognizer to produce higher-quality transcripts in the target language through a cyclical process of recognition, followed by selection and reinforcement, and then retraining.

So our process proceeds as follows: after the first wave of transcription, we identify these repetitive phrases (where we will call them ‘repetitive’ if they coincide with our person, place, date, occupation, or number lists) *or* if they occur multiple times in the documents. We then using the recognition metadata to discover where those repetitive phrases were amongst the original image snippets. Although a whole snippet could have been transcribed correctly, mostly we will find repetition as sub-snippets. Therefore, we will cut down the snippets to those areas that correspond to repetition, and we will treat those subsnippets and their corresponding hypothesize transcripts as if they were valid target-language training data. We add these to the original collection of non-target, same script data plus the synthetic data and “rinse and repeat” – train the whole process again, recognize new data, produce new snippets, etc. This process continues as many times as desired, but usually a few iterations is sufficient.

3. AN ASIDE: SYNTHETICS ONLY?

Now before we actually look at the whole process that was just described, it might be beneficial to ask: what if we *just* use the synthetic images? Do we actually need the whole process shown in Figure 1, or could we merely use synthetic images as has been done by others (eg., [3])?

We will take a brief look at a synthetic-only situation and we will use one of the languages for which we have many transcribed files as a test set (in particular, we will look at our Spanish test set). The question is: what if we produce *N* synthetic images and we use only them to train a visual model and we use *M* of those for training a language model: how well can we learn to transcribe our Spanish test set? Moreover, we will look at how well this process works from printed documents from the Spanish test versus handwritten ones.

#Synth Images / #in LM	**Word Acc on Synth Test**	**Word Acc Real Print Only**	**Word Acc of Real HW Only**
1K/1K	97.4%	70.7%	-6.5%
2K/1K	98.2%	73.5%	3.5%
2K/2K	98.4%	74.1%	3.8%
4K/2K	98.7%	74.9%	10.7%
4K/4K	98.8%	78.2%	11.6%
8K/4K	98.9%	80.3%	17.1%
8K/8K	99.0%	**80.8%**	**17.3%**

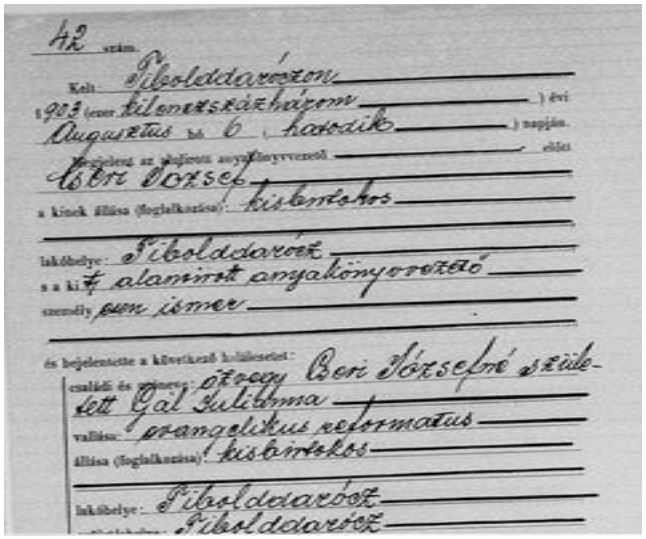
Table 1. Spanish: What if we just use synthetic data?

Table 1 reveals the kind of performance one might be able to derive from synthetic images *alone*. In its first column, we indicate the number of images used for the visual model and the number of those whose transcripts are used for building the language model. “K” in this case means a thousand. The second column of the table shows the word accuracy that could be derived if we use the visual and language models applied to a synthetically-built test set. The third column shows the accuracies of the models as applied to actual printed documents of our true test set, and the last column shows performance on our real handwritten documents. As can be quickly seen, training on synthetic images and testing on synthetics works well even with relatively few images in the training set. It also continues to improve well as additional synthetic images are added.

Likewise, and perhaps surprisingly, the system works quite well on the printed documents in Spanish. With 1000 synthetic images in both the visual and language model training, we can get 70.7% accuracy on our historical Spanish print data. It is usually the case that for every doubling of data, one often sees a 3-4% absolute performance gain. We see that this rule-of-thumb holds true here as well since after three successive doublings in the number of synthetic images, we have accuracies of 80.8%. One might assume that if we were again to double three more times, maybe we could even get to 90%.

However, handwritten documents seem mostly immune to the synthetic images. With 1000 such training documents, we have performance that is worse than having done nothing. Even after 8000 images, we only have 17.3% accuracy which may be a lot of punctuation and numbers. This suggests that using synthetic images alone is insufficient -- which makes our process all the more relevant.

4. THE WHOLE PROCESS IN PRACTICE



42 szám.

Kelt: Tiboldaróczon

1903 (ezer kilenczszázhárom) évi

Augusztus hó 6 (hatodik) napján.

Megjelent az aláirott anyakönyvvezető előtt

Béri József

a kinek állása (foglalkozása): kisbirtokos

lakóhelye: Tiboldaróc

s a ki az alamirott anyakönyvvezető

személyesen ismer

és bejelentette a következő haláleset: Béri Józsefné szüle-

családi és utóneve: özvegy

tett Gál Julianna

vallása: evangelikus református

állása (foglalkozása): kisbirtokos

lakóhelye: Tiboldaróc

Tiboldaróc

Figure 3. Region of a Hungarian birth record

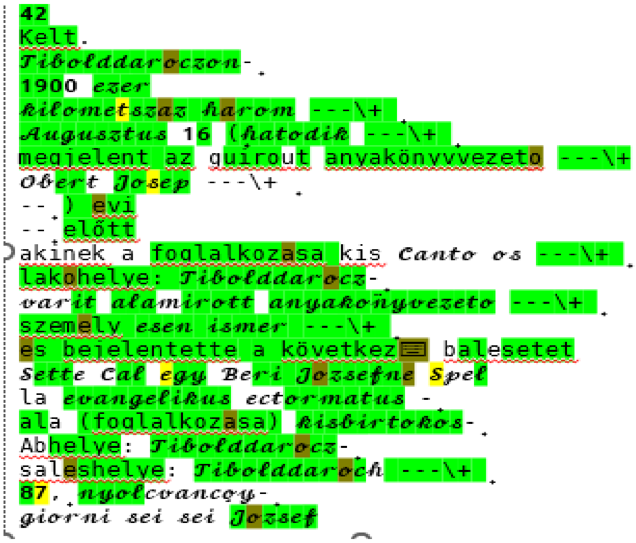
We will now demonstrate the elements from Figure 1 on actual data. We will focus on Latin-script languages and initially demonstrate our findings on Hungarian, but we show later the performance on other Latin-script languages. We believe this technique should also work for non-Latin-script. Figure 3 is an example (among the first, actually) of the ~5000 untranscribed Hungarian images we use here.

As was indicated previously, we will attempt learning in an iterative fashion by first augmenting our "All Known Latin" data with synthetic images and then building a recognizer (visual and language model). If the initial Latin-script + synthetics recognizer is of decent quality, we expect many phrases in the target language will get transcribed correctly and, because of volume, we will find valid repeats. We will use those repeats as additional training data, rebuild the models, retranscribe – and continue this process iteratively until we hit a maximum.

4.1. All Known Latin + Synthetics Phase

For Hungarian, we produced about 1000 synthetic images to add to our 'All Known Latin" training data and built our first recognizer. When we run this recognizer on the snippets from this document in Figure 3, we get the output text shown in Figure 4a (where green highlights indicate words that are actually correct in the initial phase, and where yellow and brown highlights are partially correct).

We measure system performance in terms of word accuracy, which is measured as 100% minus the word error rate. Word error rate is the sum of insertions, deletions, and substitutions all divided by the total number of words in the correct transcript. Partially-correct words are still counted as errors, but we do ignore casing. If we ignore accents, the full-word accuracy of the text in Figure 4b is about 46.4% but if we do not ignore accents, the number is lower: **31.0%**. This is not really usable yet for the kinds of tasks we want, namely, distilling out the genealogical information and/or image search type applications.



```
42
Kelt.
Tiboldda roczon-
1900 ezer
kilometszaz harom ---\+
Augusztus 16 (hatodik ---\+
megjelent az quirout anyakönyvvezeto ---\+
Obert Josep ---\+
-- ) evi
-- előtt
akinek a foglalkozasa kis Canto os ---\+
lakohelye: Tiboldda rocz-
varit alamirott anyakoñyvezeto ---\+
szemely esen ismer ---\+
es bejelentette a következ▤ balesetet
Sette Cal egy Beri Jozsefne Spel
la evangelikus ectormatus -
ala (foglalkozasa) kisbirtokos-
Abhelye: Tiboldda rocz-
saleshelye: Tiboldda roch ---\+
87, nyolcvancgy-
giorni sei sei Jozsef
```

Figure 4a. Latin+Synthetic-pass recognition of Figure 3.

4.2. Phases 1 & 2: Latin + Synthetic + repeats

We seek to make that word accuracy higher by leveraging duplicates that we observe in the recognition output and then by treating those observations as training data for the next wave. We also might treat a word/phrase as a duplicate if it appears only once but happens to be a list of valid words or phrases that we have scraped from the Web (or, possibly, from indexes). For Hungarian, Table 2 indicates some of the words and phrases we might think are valid based on web and/or indexed data. It should be mentioned that, since we would prefer longer repeats, the 'Valid Names" column will include full names in given + surname order, surname + given order, etc. Likewise, longer numeric representations are also helpful.

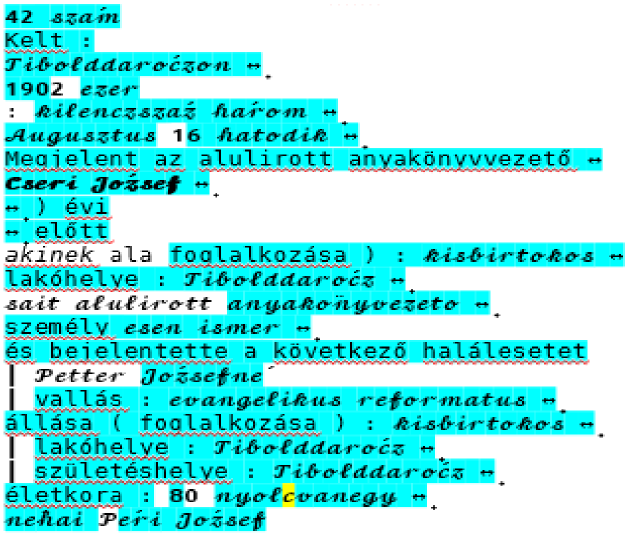
Valid Names	Valid Places	Valid Dates/Numbers
Mária	Budapest	megnevezés
József	Pest-Pilis-Solt-Kiskun	napok száma
Erzsébet	Magyarország	január
János	Hungary	Október
István	Szabolcs	szám
Ilona	Miskolcz	kettő
Anna	Mezőkövesd	Három

Table 2. Hungarian scraped word lists

After this first wave of output, the system discovered such repeated words and phrases as ***September***, ***az aldbbi***, *a kinek allasa* , lakóhelye -,) napján, ***alóliross***, vallos : ***római katholikus***, *Chazastarsanak*, ***ismer*** , and many thousands of others. Note that italics, cursive, underlining, etc. have also been captured by this process of looking for repeats, and we add these words/phrases into the training data in whatever style in which they emerge.

When we augment the original training data with these newly-discovered repeats and/or in-list phrases, the training data increases by 89K lines and 430K words. This is substantial growth in the target language's potential training data. We use this data and train again and then re-recognize. Amazingly, the accuracy of the system after this new round of training, when applied to the data in Figure 3, moved from 31.0%to **76.2%**! (or 77.4% if we ignore accents). From experience with other languages, we can say that this level of performance is comparable to using our current system to train a language that has seen roughly 750-1000 human-provided transcripts. Yet aside from mining the web lists and/or indexes, there has been no specific human annotation.

We can repeat this selection and retraining process as long as we have interest. If we perform another wave of training, since the recognizer is now better, we discover 34K more lines of data and roughly 200K more words which we add to the training data and start the training process again. This yet newer "Phase 2" recognizer yields **81.0%** accuracy (82.0% if accents are ignored) on the data in Figure 3. We do not get additional gains beyond that due to image segmentation issues rather than training issues. Even so, we have moved the accuracy needle from ZERO to 31.0%. then to 76.2% and on to 81.0%! The resultant transcription is depicted below in Figure 4b, where blue color now indicates *final* correct words.



42 szain
Kelt :
Tiboldaroczon ↔
1902 ezer
: kilencszaz harom ↔
Augusztus 16 hatodik ↔
Megjelent az alulirott anyakönyvvezető ↔
Cseri Jozsef ↔
↔) évi
↔ előtt
akinek ala foglalkozása) : kisbirtokos ↔
lakóhelye : Tiboldaroćz ↔
sait alulirott anyakoñyvezeto ↔
személy esen ismer ↔
és bejelentette a következő haláleseted
| Petter Jozsefne
| vallás : evangelikus reformatus ↔
állása (foglalkozása) : kisbirtokos ↔
| lakóhelye : Tiboldaroćz ↔
| születéshelye : Tiboldaročz ↔
életkora : 80 nyolcvanegy ↔
nehai Peri Jozsef

Figure 4b. Second-pass recognition of Figure 3.

Though this second-pass recognition still has errors, 81.0% is enough to do some reasonable levels of genealogical processing. For example, in Figure 4b, we can note such things such as the fact that the document is talking about an individual named József Czeri who is from Tibolddarócz and is a member of the Reformed Evangelical Church.

5.APPLICATION TO OTHER LANGUAGES

The fact that this process worked in one language, Hungarian, does not yet signify that it works generally. Therefore, we briefly explore its application to four other Latin-scripted languages: Dutch, Polish, Icelandic, and Estonian. These were chosen because they are languages for which FamilySearch has a reasonably high number of records. In the cases of these four languages, we used our synthetic and discovered Hungarian transcripts as part of the "All Known Latin" data. Also, we did Dutch before the latter three, so any "discovered" Dutch is included in the initial training sets of the last three languages. Like Hungarian, there were 1000 synthetic images used for Dutch; but the other three languages only have 250-350 synthetic images (although their training datasets each have access to the others' synthetic images).

Method	Dutch	Polish	Icelandic	Estonian
OutOfBox on HW	54.6%	6.6% (7.4%)	14.8% (17.1%)	34.6% (40.7%)
+Synth on HW	68.5%	10.7%	26.2% (29.3%)	38.3% (51.9%)
+Pass1 on HW	67.9%	18.2%	29.3% (33.5%)	54.3% (69.1%)
+Pass 2 on HW	72.8%	19.0%	29.3% (33.8%)	51.9% (66.7%)
OutOfBox on PR	90.6%	65.2% (67.8%)	75.2% (17.1%)	52.0% (56.8%)
+Synth on PR	94.3%	84.3% (86.1%)	94.5% (95.0%)	79.2%
+Pass1 on PR	93.8%	86.1% (89.6%)	96.6% (97.2%)	80.8%
+Pass 2 on PR	94.6%	86.1% (89.6%)	97.3% (97.6%)	81.6%
2H+P	66.6 => **80.1%**	26.1 (27.5) => **41.4%** (42.5%)	34.9 (36.9) => **52.0%** (55.1%)	40.4 (46.1) => **61.8%** (71.7%)

Table 3. Performance in other Latin-script languages

Table 3 shows the results of using this process in each of the four additional languages, which are indicated in the column headers. The row headers "Out of Box" indicates performance one could get by running the same system as used in Figure 4b (i.e., the auto-learned Hungarian recognizer) on the indicated images. "+Synth" is performance after synthetic images have been added. "+Pass1" and "+Pass2" illustrate performance once the repeats have been added the first time and second times, respectively.

Items in light yellow have to do with test on handwritten documents, and this is also indicated by the row header phrase "on HW". Those in light blue are tests made on printed documents ("on PR"). The numbers indicated are the word accuracies on the corresponding test sets. If there are parenthetical smaller numbers in the cell, those indicate the performance on the task if one could ignore accents.

Since we did not have available data, we had to create our own test sets. Consequently, the test sets used for Table 3's scores are very small – typically one side of one image each. Yet the numbers should hopefully be indicative of the process's ability to improve performance.

The last row of the table is an estimate of the system's general performance when one assumes that print is usually half as voluminous in practice as is handwriting (so (2*HW + 1*PR)/3). When we see "X=>Y," the "X" indicates the estimate of general performance one would get from the OutOfBox system, whereas the "Y" indicates the estimate after the whole process has been run.

If we analyze the table, we see that the accuracies all increase through this process – and that increase is usually substantial. In Dutch, what started as a 67%-level system by just using the All-Latin model became, in the end, a very usable 80% system – a 40% relative reduction in the error. That 40% error reduction actually came equally from the handwriting and the print, though it should be noted that much of the print gains came from adding the synthetic images.

If we now look at the other three languages, we see, too, that they have improved. Each produces a weighted result after the full process has run that is at least a 50% relative improvement over what it was at the start. The results on handwriting for these languages still leave something to be desired; but the Estonian handwritten system seems possibly useable and the Icelandic recognition has extremely high accuracy on print. It should be noted, too, that the Estonian appears to have gotten worse in Pass 2, but this may be due to inconsistent handling of the training data between Pass 1 and Pass 2. Even so, we used Pass 2's scores for the overall numbers displayed in Table 3.

The Polish handwritten performance appears to be much worse than the other languages at the start. This may be because the image that was chosen for transcription was not amongst the 5000 that were used for the overall process since there was an image in a related dataset that was already transcribed and could be used for evaluation. Possibly if that image could have been included in the 5000, its results might have been somewhat higher. Even so, its improvements seem consistent with those of the other languages.

6. COMMENTS AND DISCUSSION

Clearly, the ability to train recognizers with little or no transcription as was done here has great potential for accelerating automatic image processing in many languages. We have shown that this process has been able to make it so some languages (such as Hungarian and Dutch) which were previously non-existant can now produce results that on par with other languages with 500-1000 images in their training sets. We also showed that in other languages, even though their results may not be of the same caliber, they definitely improved significantly over merely using a All-Latin model and they may have applications where they can provide high-quality outcomes (such as Icelandic on print). We are confident that our process here provides a way forward for handling scores of more languages – at least to some reasonable degree.

REFERENCES

[1] Schone, P., Hargraves, C., Morrey, J., Day, R., Jacox, (2018) M. Neural text line segmentation of multilingual print and handwriting with recognition-based evaluation. *ICFHR*, Niagara Falls, NY.

[2] Author unknown, https://babynames.com/hungarian-baby-names

[3] Journet, N., Mansencal, B., Visani, M. Massive, free and reproducible grountruthed document image databases generation with DocCreator. 1st International Workshop on Open Services and Tools for Document Analysis, 14th IAPR International Conference on Document Analysis and Recognition, OST@ICDAR Nov 2017, Kyoto, Japan. pp.1139-1143