The Bible, Truth, and Multilingual OCR Evaluation

Tapas Kanungo[1,3] and Philip Resnik[2,3]

[1]Center for Automation Research
[2]Department of Linguistics
[3]Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742
Email: kanungo@cfar.umd.edu, resnik@umiacs.umd.edu
Web: www.cfar.umd.edu/~kanungo, www.umiacs.umd.edu/~resnik

ABSTRACT

Multilingual OCR has emerged as an important information technology, thanks to the increasing need for cross-language information access. While many research groups and companies have developed OCR algorithms for various languages, it is difficult to compare the performance of these OCR algorithms across languages. This difficulty arises because most evaluation methodologies rely on the use of a document image dataset in each of these languages and it is difficult to find document datasets in different languages that are similar in content, layout, and fonts.

In this paper we propose to use the Bible as a dataset for comparing OCR accuracy across languages. Besides being available in a wide range of languages, Bible translations are closely parallel in content, carefully translated, surprisingly relevant with respect to modern-day language, and quite inexpensive. A project at University of Maryland is currently implementing this idea. We have created a scanned image dataset with groundtruth from an Arabic Bible. We have also used image degradation models to create synthetically degraded images of a French Bible. We hope to generate similar Bible datasets for other languages, and we are exploring alternative corpora with similar properties such the Koran and the Bhagavad Gita. Quantitative OCR evaluation based on the Arabic Bible dataset is currently in progress.

Keywords: Bible, OCR, performance evaluation, datasets, groundtruth, parallel corpus, linguistics.

1. INTRODUCTION

To evaluate any OCR algorithm we need (i) datasets of scanned document images and (ii) the corresponding symbolic groundtruth. Obtaining manually generated groundtruth, as is generally done, is labor-intensive, time consuming, prohibitively expensive, and prone to errors. Furthermore, collecting datasets across languages in a way such that they all have similar content is a non-trivial task.

In this paper we propose to use the Bible as a corpus for evaluating OCR accuracy across languages. As an example, a French version of the Bible provides us with a test set and groundtruth for a French OCR system, with far less effort than one would typically expend in obtaining groundtruth for a 1000-page text. In addition, using the French dataset also allows us to make a far more controlled comparison to our Arabic OCR system than if we were using groundtruth from an unrelated text. At University of Maryland, we have collected the electronic groundtruth for Arabic, English, and French Bibles and have scanned the paper version of the Arabic Bibles. We have also obtained, but not processed as datasets, versions for complete Bibles or New Testaments in 17 other languages. Notwithstanding licensing restrictions, all our datasets will be made freely available to OCR researchers.

In Section 2 we give a survey of OCR evaluation methods and programs that have been pursued in the past and discuss their limitations. Next, in Section 3 we explain why the Bible is a good corpus for the purposes of OCR evaluation. We then report on a project[1] recently started at University of Maryland for creating Bible datasets with groundtruth in various languages. Finally, in Section 5 we describe how the Bible corpus can also be used for generating synthetically degraded multilingual datasets.

2. PERFORMANCE EVALUATION METHODOLOGIES

Numerous commercial OCR systems claim that their products have near-perfect recognition accuracy (close to 99.9%). In practice, however, these accuracy rates are rarely achieved. Most systems break down when the input document images are highly degraded, such as scanned images of carbon-copy documents, documents printed on low-quality paper, and documents that are n-th generation photocopies.

Characterizing the performance of OCR systems is important for many reasons:

- Performance prediction: Typically OCR is part of a bigger system, e.g., an information retrieval (IR) system or a machine translation (MT) system. Since the overall performance depends on the performances of the individual subsystems, the overall performance of the MT/IR system is a function of the OCR recognition rate. Knowledge of end-to-end performance as a function of OCR accuracy rate will allow us to predict the minimum recognition rate required for achieving a specified overall MT/IR system performance rate.
- Monitor progress: In order to monitor progress in research/development of OCR systems, we need quantitative measures. Periodic quantitative performance evaluation of the OCR systems will allow us to assess progress in the field.
- Provide scientific explanations: Understand the contributions to the accuracy improvement by various specific submodules. That is, explain *why* a OCR system achieves a particular accuracy.
- Identification of open problems: To identify areas that need improvement/research.

OCR evaluation can be broadly categorized into two types: i) blackbox evaluation and ii) whitebox evaluation. In blackbox evaluation an entire OCR system is treated as one indivisible unit and the end-to-end performance of the system is characterized. The performance of the system is evaluated as follows. First a corpus of scanned document images is selected. Next, the text zones are delineated. Then for each text zone, the correct text string is keyed in by humans. The process of delineating the zones and keying in the text is very laborious, expensive, and prone to errors. Finally the OCR algorithm is run on each text zone and the results are compared against the keyed in groundtruth text using a string matching routine. In theory the corpus should be a representative sample of the population of images for which the algorithm is designed. In practice, however, factors like time and cost forces us to limit the size of the dataset to something feasible. This process was adopted by the UNLV OCR evaluation program,[2] the UW evaluation process,[3] and the UMD Arabic OCR evaluation process.[4] The UNLV evaluation corpus consisted of English annual reports, documents from department of Energy, magazines, business letters, legal documents, Spanish newspapers, and German business letters. The UW dataset[5] consisted of English technical journals. The UMD evaluation used the DARPA/SAIC Arabic dataset,[6] which consists of book chapters, magazine articles, and computer generated documents. UNLV evaluation reported average character and word accuracies, while UW and UMD evaluations reported average character accuracy. Since the English, Spanish, German and Arabic datasets have very different content and layout, comparing accuracies across languages is not very meaningful.

Whitebox evaluation, on the other hand, characterizes the performance of individual submodules. Most OCR systems have submodules for skew detection and correction, page segmentation, zone classification, and text extraction. Zone segmentation evaluation has been attempted earlier by Vincent *et al.*[7,8] Whitebox evaluation is possible only if the evaluator has access to the input and output of the submodules of the OCR system. Thus for segmentation evaluation, access to coordinates of zones produced by OCR is crucial. While blackbox evaluation does not require access to intermediate results, it does not provide performance analysis at the submodule level. Furthermore, the blackbox evaluations described above do not take into account the errors due to segmentation.

More recently, researchers have advocated the use of synthetically generated data for OCR evaluation. In this methodology (see Kanungo *et al.*[9,10]) documents are first typeset using a standard typesetting system such as LaTeX or Word. Then a noise-free bitmap image of the document and the corresponding groundtruth is automatically generated. The noise-free bitmap is then degraded using a parametrized degradation model.[11,9,10] The degradation level is controlled by varying the parameters of the model. This methodology has the advantage that the laborious process of manually typing in the data is completely avoided. Furthermore, no manual scanning is required, and the process is entirely independent of language (up to the limits of the typesetting software). Since the typesetting software is available to us, the effects of page layout, font size and type, on OCR accuracy can be studied by conducting

controlled experiments. A variant of the above methodology proposed by Kanungo and Haralick[12,10] generates real degradations by printing the ideal document, scanning them, and then transforming the ideal groundtruth to match the real image This process allows a researcher to generate groundtruth at a geometric level (character bounding boxes, identity, font, etc.) in any language, which is essential for building classifiers.

Although the degradation model can be applied to documents in any language, the contents of the documents in the corpus again become crucial if we need to compare OCR accuracies across languages. Otherwise the comparisons are not very meaningful. In the next section we propose to use the Bible as a source of such documents and discuss why it is a good dataset, as well as the potential limitations.

3. THE BIBLE AS A CORPUS

Text corpora — bodies of naturally occurring text — have in the past 5–10 years become a focal point of research in computational linguistics.[13] The Bible, at first blush, seems like an unlikely resource for research in language technology, conjuring up images of archaic syntax, atypical vocabulary, and religion-specific subject matter. However, as Resnik, Olsen, and Diab[14] discuss, the Bible turns out to be surprisingly relevant for research involving present-day language, if one begins with a modern language translation such as the New International Version (NIV) for English. Resnik *et al.* evaluate the vocabulary of the NIV against two benchmarks: the approximately 2200-word control vocabulary for Longman's Dictionary of Contemporary English (LDOCE[15]), and the most frequent 2000 words in the Brown corpus of present-day American English.[16]

Since LDOCE is a learner's dictionary, its control vocabulary — that is, the set of words used in its definitions — can be viewed as a list of particularly basic or useful words in English, as determined through an extensive process of lexicography. Resnik *et al.* find that 78–85% of the items in the LDOCE control vocabulary are found in the NIV.* Examination of the LDOCE vocabulary also found in the NIV shows that Bible text contains ample vocabulary representative of typical, everyday usage, not to mention being representative of a wide range of English orthography, as illustrated in the following 50-word random sample:

> *anyone ashamed at baby behave bit bite black blame build calm circular clay cliff cloth contain control damage ditch educate finish fit heart insect its lid neither particular presence press price pronounce rent seem soap stand strength strike take thick throw tonight undo vote weave west wheel wine within worst*

A similar comparison with frequent words in the Brown corpus provides evidence that a modern-language Bible provides good coverage not only of "useful" words, but of words that occur frequently — the Brown corpus is an oft-cited source of word frequency data for English, e.g. when controlling for word frequency in psychological experiments, and it is also one of the most widely used corpora in natural language processing research. Resnik *et al.* show that of the most frequent 2000 words in the Brown corpus, fully 75% occur in the NIV. A 50-word random sample includes the following:

> *achievement address arrive brief building call climb conclude dream dry exchange extent family fast fat happen here hide impression increase lady narrow nine observation officer opportunity plan please pleasure public reflect relative requirement road satisfy search select silence simple single spread straight test tragedy watch wave west wind wine work*

Because the Brown corpus spans multiple genres, it is also possible to assess vocabulary coverage as a function of text type. Resnik *et al.* show that even for texts in genres far removed from Biblical material, such as science fiction, theater and music reviews, and science writing, the NIV text covers at least two thirds of the most frequent 2000 words in each genre.

Although we have not conducted a similar comparison for non-English versions of the Bible, it is reasonable to expect the results to carry over: because the underlying *content* is the same, one can expect similar patterns of vocabulary content in a modern-language version of the Bible, regardless of the language in which that content is expressed.

*The exact percentage depends on how certain cases are handled, e.g. whether or not the word *practice*, found in the American-published NIV, is counted as an instance of the word *practise* in the British-published LDOCE.

This parallelism of content at a global level is matched by parallelism at a much finer grain: unlike other parallel document collections, e.g. parallel bilingual corpora used in research on automatic machine translation, translations of the Bible contain verse-level parallelism. This will permit a fine grained level of analysis for OCR evaluation — e.g., some verses may be difficult in *any* language owing to the presence of Biblical names. It also provides valuable parallel data for multilingual language processing applications, such as the automatic discovery of term translations[17] and construction of cross-language information retrieval systems.[18] Indeed, OCR work on the Bible has the potential to be of great benefit to the language technology community as a whole, by providing data in electronic form for languages in which corpus data is otherwise difficult to obtain. Uses of the text go well beyond simple acquisition of vocabulary: because it contains text by a large set of authors, in a variety of text styles, and touching on range of content areas, the Bible is also a rich data source for cross-language data on syntactic structure and semantic patterning.

In summary, the advantages of using the Bible include the following:

1. It exists in a huge number of languages: as of this writing, complete Bible translations exist in over 360 languages, New Testament translations in over 900, and at least one book of the Bible in nearly 2200 languages.[†]
2. These figures are increasing rapidly: within the last year 13 new Bible translations were completed, 25 New Testaments were completed, and 180 new translations were initiated.
3. Translations are verse by verse, providing a reliably parallel corpus.
4. Bibles exist in print in various formats, fonts, and paper types.
5. Coverage of everyday modern language is surprisingly high: approximately 80% of the defining vocabulary of Longman's Dictionary of Contemporary English (which has controlled-vocabulary definitions) can be found in a modern-English translation of the Bible (the New International Version).
6. Bibles in most common languages are available on-line or in electronic form, often free or for a reasonable licensing cost. This easily available groundtruth data frees us of much of the manual work.
7. The Bible is a large corpus by the standards of work in OCR, and non-trivial by the standards of corpus-based work in natural language processing. For example, our French version has over 1000 pages, comprising on the order of 800,000 words.

Use of the Bible as a language resource is not without its limitations, of course. Many elements of modern day documents are missing from its pages, such as technical terminology, many modern proper names, and everyday words that are of more modern origin or simply outside its scope (e.g. *atom*, *Buddhist*, *January*, *cat*). Formats for addresses, dates, and the like are also clearly not present. Furthermore, complex layouts such as those found in newspapers and magazines, tables and graphics, are also absent from the Bible corpus. Thus there is a tradeoff: as a corpus the Bible provides an unmatched degree of consistency, availability, and parallelism, at the cost of some elements that might help distinguish between OCR systems, e.g. on the basis of the coverage of their lexicons, or page segmentation and zone classification performance.

4. REAL IMAGE DATASET: THE SCANNED BIBLE

At University of Maryland we have started a project, which we internally refer to as Project Gutenberg,[1] to create scanned Bible image and groundtruth datasets for OCR evaluation in various languages. At the time of writing this article we have been able to scan the entire new testament of the Arabic Bible.[19] The scanning was done at 600dpi resolution and both binary 1bit/pixel and grayscale 8bits/pixel images were scanned. Two pages of the Bible were scanned at a time. There are 198 scanned binary images, which are saved in TIF format using Group 4 compression, and 198 grayscale images stored in TIF format without compression. The zone groundtruth was generated manually using the PinkPanther software.[7,8] The attributes of the zones — id, bounding box, type (body, running head, section title, page number), etc. – are stored in a zone groundtruth file. We did not have to type in the electronic text groundtruth — these were available on the International Bible Society web page *http://www.gospelcom.net/ibs*.

[†]http://www.biblesociety.org/trans-gr.html

The text encoding format is CP1256. The text groundtruth corresponding to a particular zone, however, had to be extracted from the electronic text groundtruth files separately. We are currently working on a method to make this process automatic.

In Figure 1 we show a binary scanned image from the Arabic Bible. Manually delineated zones for this image are shown in Figure 2. The zones are non-overlapping rectangles and are ordered in Arabic reading order. A section of the zone groundtruth file storing the attributes of a zone is shown in Figure 3. The details of the database design, file formats, directory structures, etc. is available in a separate article.[1]

Figure 1. Scanned image of an Arabic Bible page.

5. SYNTHETIC IMAGE DATASET: MODEL-BASED DEGRADATION

We first describe a degradation model proposed by Kanungo *et al.*[9,20,10] that can be used to generate synthetically degraded documents. The degradations produced by this model are local – typical degradations that appear while scanning a flat paper. Next we use the model to generate degraded images of French Bible pages. The same process can be used for generating degraded Bible images in any language. A more general model that accounts for perspective distortions near the spine of a thick book is described in Kanungo *et al.*[9] and not discussed in this article.

5.1. A Degradation Model

In this section we present a model that accounts for (i) pixel inversion (from foreground to background and vice-versa) that occurs independently at each pixel due to light intensity fluctuations, sensitivity of the sensors, and the thresholding level, and (ii) blurring that occurs due to the point-spread function of the scanner optical system.

The degradation model has six parameters: $\Theta = (\eta, \alpha_0, \alpha, \beta_0, \beta, k)^t$. We model the pixel-flipping probability of a pixel as an exponential with the decay rate as a function distance d from the nearest boundary pixel. The

Figure 2. Zone groundtruth for the scanned image shown in Figure 1. Notice that the zones are rectangles and non-overlapping. The zones were created using the PinkPanther groundtruthing tool.[7,8]

foreground and background 4-neighbor distance of pixels is computed using a distance transform algorithm (see Borgefors[21]). The parameters α_0 and α control the probability of a foreground pixel switching to a background pixel and β_0 and β control the probability of a background pixel switching to a foreground pixel. The parameter η is the constant probability of flipping for all pixels. Finally, the last parameter k, which is the size of the disk used in the morphological closing operation, accounts for the correlation introduced by the point-spread function of the optical system. These parameters are used to degrade an ideal binary image as follows.

1. Compute the distance d of each pixel from the character boundary.
2. Flip each foreground pixel with a probability

$$p(0|1, d, \alpha_0, \alpha) = \alpha_0 e^{-\alpha d^2} + \eta.$$

3. Flip each background pixel with a probability

$$p(1|0, d, \beta_0, \beta) = \beta_0 e^{-\beta d^2} + \eta.$$

4. Perform a morphological closing operation with a disk structuring element of diameter k. (See Haralick *et al.*[22] for an introduction to morphological image processing.)

The noise-free documents are typeset using the LaTeX formatting system.[23,24] The ASCII files containing the text and the LaTeX typesetting information are then converted into a device independent format (DVI) using LaTeX.

```
BEGIN_IMAGE
FILENAME /tapas/Gutenberg/arabic/bw/tif600/ar061.tif
IMAGE_WIDTH 6600
IMAGE_HEIGHT 5096
IMAGE_XRES 600
IMAGE_YRES 600
END_IMAGE
     .
     .
     .

BEGIN_REGIONS

R_TYPE TEXT
R_SUBTYPE RUNNING_HEAD
R_NUMBER 1
R_ATTACHMENT 0
R_PARENT 0
R_DRAW_TYPE Box
R_NAME Zone1
R_ATTRIBUTES BODY
REGION_POLYGON 500 632 5712 6408

R_TYPE TEXT
R_SUBTYPE PAGE_NR
R_NUMBER 2
R_ATTACHMENT 0
R_PARENT 0
R_DRAW_TYPE Box
R_NAME Zone2
R_ATTRIBUTES BODY
REGION_POLYGON 516 612 4952 5124

R_TYPE TEXT
R_SUBTYPE BODY
R_NUMBER 3
R_ATTACHMENT 0
R_PARENT 0
R_DRAW_TYPE Box
R_NAME Zone3
R_ATTRIBUTES BODY
REGION_POLYGON 672 1160 5076 6424
```

Figure 3. Part of the zone groundtruth file for the zones shown in Figure 2. This file was generated by the PinkPanther groundtruthing tool.[7,8]

A software program called DVI2TIFF – which is a modified version of a DVI file previewer called XDVI[25] – is run to produce one bit/pixel binary images in TIFF format from the DVI files. Besides producing the binary images of the documents, DVI2TIFF also produces the groundtruth information regarding each character on the document image.

The local document degradation model itself is another software program called DDM. This program takes as input an ideal binary document image in TIFF format, and a file containing the degradation model parameter values, and produces the binary degraded images in TIFF format. Both programs – DVI2TIFF and DDM – are implemented in the C language and have been tested on SUN and IBM machines running the UNIX operating system. The software is available on the UW CD-ROM-1.[5] Software for simulating noisy documents using the above degradation model is available from University of Washington English Document Database I and the model itself has appeared in the

literature.[9,20,10] The application of the various steps of our model is shown in Figure 4.

Issues regarding model validation and model parameter estimation have been discussed by Kanungo *et al.* elsewhere.[9,26,27]

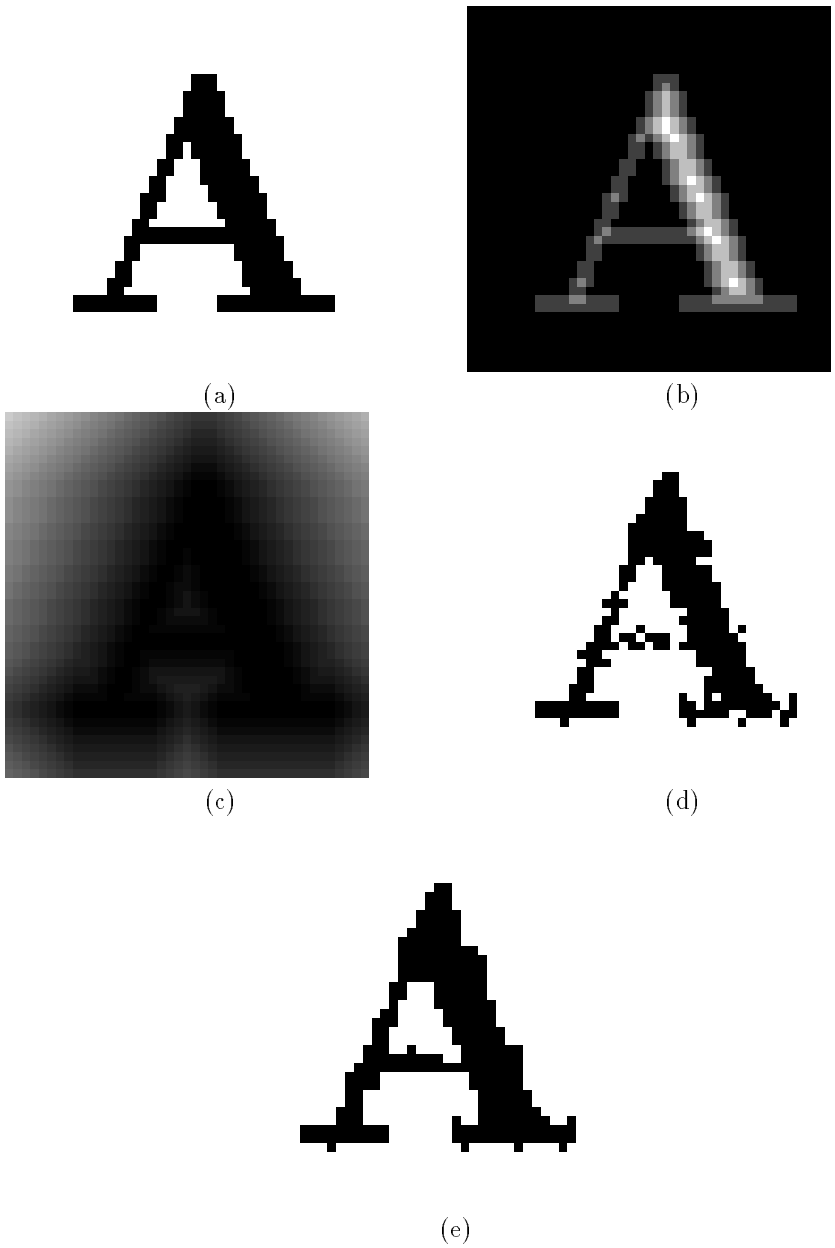
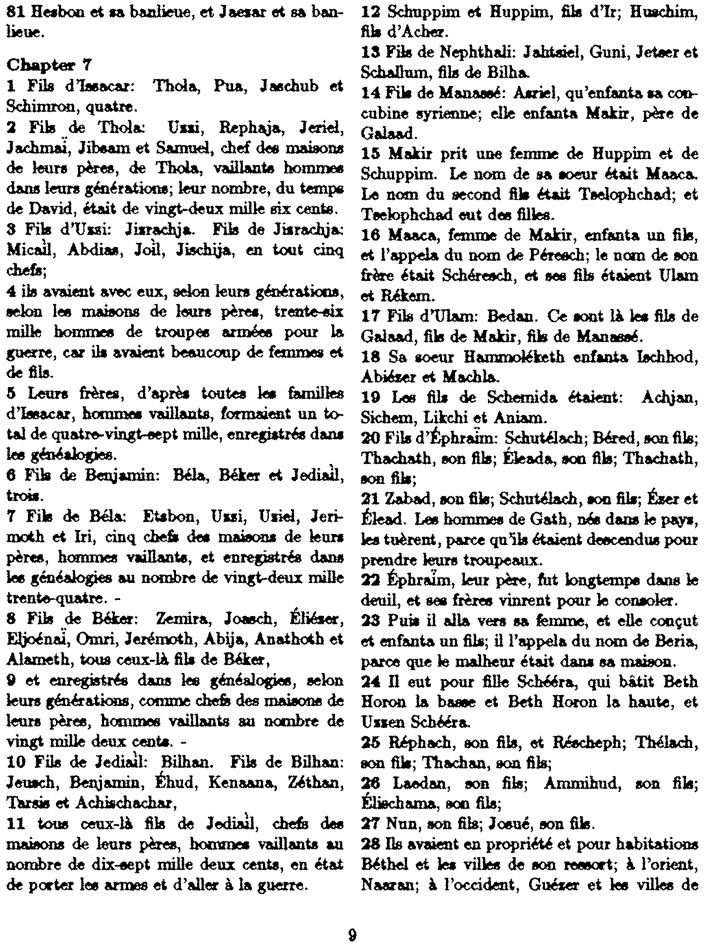


Figure 4. Local document degradation model: (a) Ideal noise-free character; (b) Distance transform of the foreground; (c) Distance transform of the background; (d) Result of the random pixel-flipping process. The probability of a pixel flipping is: $P(0|d,\beta,f) = P(1|d,\alpha,b) = \alpha_0 e^{-\alpha d^2}$ here $\alpha = \beta = 2$, $\alpha_0 = \beta_0 = 1$; (e) morphological closing of result in (d) by a 2×2 binary structuring element.

5.2. Application to Bibles

Since the electronic text files for some versions of the Bible are available free of cost, we decided to format the French Bible using LaTeX[23,24] and then generated a typeset Bible. We degraded this image synthetically: In Figure 5 we show a synthetically degraded image of a page from a French Bible. In Figure 6 the output of the OCR processing



81 Hesbon et sa banlieue, et Jaezar et sa banlieue.

Chapter 7

1 Fils d'Issacar: Thola, Pua, Jaschub et Schimron, quatre.
2 Fils de Thola: Uzzi, Rephaja, Jeriel, Jachmaï, Jibsam et Samuel, chef des maisons de leurs pères, de Thola, vaillants hommes dans leurs générations; leur nombre, du temps de David, était de vingt-deux mille six cents.
3 Fils d'Uzzi: Jizrachja. Fils de Jizrachja: Micaël, Abdias, Joël, Jischija, en tout cinq chefs;
4 ils avaient avec eux, selon leurs générations, selon les maisons de leurs pères, trente-six mille hommes de troupes armées pour la guerre, car ils avaient beaucoup de femmes et de fils.
5 Leurs frères, d'après toutes les familles d'Issacar, hommes vaillants, formaient un total de quatre-vingt-sept mille, enregistrés dans les généalogies.
6 Fils de Benjamin: Béla, Béker et Jediaël, trois.
7 Fils de Béla: Etsbon, Uzzi, Uziel, Jerimoth et Iri, cinq chefs des maisons de leurs pères, hommes vaillants, et enregistrés dans les généalogies au nombre de vingt-deux mille trente-quatre. -
8 Fils de Béker: Zemira, Joasch, Éliézer, Eljoénaï, Omri, Jerémoth, Abija, Anathoth et Alameth, tous ceux-là fils de Béker,
9 et enregistrés dans les généalogies, selon leurs générations, comme chefs des maisons de leurs pères, hommes vaillants au nombre de vingt mille deux cents. -
10 Fils de Jediaël: Bilhan. Fils de Bilhan: Jeusch, Benjamin, Éhud, Kenaana, Zéthan, Tarsis et Achischachar,
11 tous ceux-là fils de Jediaël, chefs des maisons de leurs pères, hommes vaillants au nombre de dix-sept mille deux cents, en état de porter les armes et d'aller à la guerre.
12 Schuppim et Huppim, fils d'Ir; Huschim, fils d'Acher.
13 Fils de Nephthali: Jahtsiel, Guni, Jetser et Schallum, fils de Bilha.
14 Fils de Manassé: Asriel, qu'enfanta sa concubine syrienne; elle enfanta Makir, père de Galaad.
15 Makir prit une femme de Huppim et de Schuppim. Le nom de sa soeur était Maaca. Le nom du second fils était Tselophchad; et Tselophchad eut des filles.
16 Maaca, femme de Makir, enfanta un fils, et l'appela du nom de Péresch; le nom de son frère était Schéresch, et ses fils étaient Ulam et Rékem.
17 Fils d'Ulam: Bedan. Ce sont là les fils de Galaad, fils de Makir, fils de Manassé.
18 Sa soeur Hammoléketh enfanta Ischhod, Abiézer et Machla.
19 Les fils de Schemida étaient: Achjan, Sichem, Likchi et Aniam.
20 Fils d'Éphraïm: Schutélach; Béred, son fils; Thachath, son fils; Éleada, son fils; Thachath, son fils;
21 Zabad, son fils; Schutélach, son fils; Ézer et Élead. Les hommes de Gath, nés dans le pays, les tuèrent, parce qu'ils étaient descendus pour prendre leurs troupeaux.
22 Éphraïm, leur père, fut longtemps dans le deuil, et ses frères vinrent pour le consoler.
23 Puis il alla vers sa femme, et elle conçut et enfanta un fils; il l'appela du nom de Beria, parce que le malheur était dans sa maison.
24 Il eut pour fille Schéera, qui bâtit Beth Horon la basse et Beth Horon la haute, et Uzzen Schéera.
25 Réphach, son fils, et Réscheph; Thélach, son fils; Thachan, son fils;
26 Laedan, son fils; Ammihud, son fils; Élischama, son fils;
27 Nun, son fils; Josué, son fils.
28 Ils avaient en propriété et pour habitations Béthel et les villes de son ressort; à l'orient, Naaran; à l'occident, Guézer et les villes de

9

Figure 5. An artificially degraded text document image. The text is a page from a Bible in the French language. The layout was formatted using LaTeX and degraded using the model described in the article.

is shown. OmniPage8.0 with the French lexicon was used for generating this text. The ASCII text generated by this process resembles OCR results on real images much more closely than those generated by simulating single character errors.

We have created degraded images of the entire New Testament of the French Bible and are in the process of creating degraded images of Bibles in other languages.

81 Hesbou et sa banlieue, et 3 aesar et sa banlieue.

chapter 7
1 Fils d1wacar: 17hola, Pua, Jwchub et Schimron, quatre.
2 Fils de Thola: Usai, Rephaja, Jeriel, Jachmaî, Jibeara et Samuel, chef de6 maisons de leurs pères, de 1101a, vaillants honunes dam leurs générations; leur nombre, du tempa de David, était de v deuX mille six cents.
3 Fils dTssi: Jisrad ja Fils de Jisrad ja: Micà, Abdias, Joà, JÎs , en tout cinq Chxâ;
4 ils avaient avec eux, selon leurs générations, 9" lets maisons de lem pères, trente-six mille bomm de troupes armes pour la guerre, car il avaient beaucoup de fesmues et de fils.
5 Ieurs ftères, d'après toutes let funilles dImacar, honunea vaillants, formaient un t¡¿. W de quatre,-vingt-eept mille, enregistrée dam lm généaktw
6 Fils de Benjamin: Béla, Béler et Jedià, trOW
7 Fils de Ba&- Etsbon, Uui, Usiel, Jeri-

Figure 6. OCR text for the synthetically degraded French Bible page image shown in Figure 5. We see that image degradation models can be used to simulate text output produced by OCR systems. This methodology allows us to easily create such simulations for any language text, at varying degradation levels, arbitrary layouts, and any font.

6. SUMMARY

We described a project to create datasets for multilingual OCR evaluation. The key idea is to use the Bible as a corpus in each language. The Bible is well suited for this purpose for a variety of reasons: the printed Bible is available in a huge range of languages, the text groundtruth is already available for many languages, the linguistic properties of the modern Bible are close to that of the current day language, Bibles exist in a variety of layouts and fonts, and the corpus is quite large. We also showed several scanned images from our Arabic Bible image dataset and the corresponding zone groundtruth. It was also demonstrated that synthetically degraded images of the Bible can be generated by using a model-based degradation approach. We hope to generate similar Bible datasets for other languages, and we are exploring alternative corpora with similar properties such the Koran and the Bhagavad Gita.

7. ACKNOWLEDGEMENT

We would like to thank Judy Day at International Bible Society for providing us with free copies of the Arabic Bible, Masaaki Kamiya for scanning and zoning the Arabic Bible dataset, Osama Bulbul for creating the text groundtruth, Greg Marton for writing automation scripts, and Mari Broman Olsen for discussions. This research is supported in part by Army Research Lab (ARL 01-5-29294), the Department of Defense (DOD 01-5-29177 and MDA90496C1250), DARPA/ITO Contract N66001-97-C-8540, and Sun Microsystems Laboratories.

REFERENCES

1. T. Kanungo, M. Kamiya, and O. Bulbul, "Project Gutenberg: Bible Image Corpus for Multilingual OCR Evaluation and Training," 1998. Forthcoming technical report, University of Maryland, College Park, MD.

2. S. V. Rice, F. R. Jenkins, and T. A. Nartker, "The fifth annual test of OCR accuracy," Tech. Rep. TR-96-01, Information Science Research Institute, University of Nevada, Las Vegas, NV, 1996.
3. S. Chen, S. Subramaniam, R. M. Haralick, and I. T. Phillips, "Performance evaluation of two OCR systems," in *Proc. of Annual Symp. on Document Analysis and Information Retrieval*, pp. 299–317, April 1994.
4. T. Kanungo, G. Marton, and O. Bulbul, "OmniPage vs. Sakhr: Paired model evaluation of two Arabic OCR products," in *Proc. of SPIE Conf. on Document Recognition and Retrieval VI*, D. Lopresti and Y. Zhou, eds., (San Jose, CA), 1999.
5. R. M. Haralick, I. Phillips, *et al.*, "UW-CDROM-I."
6. R. Davidson and R. Hopely, "Arabic and Persian OCR training and test data sets," in *Proc. of Symp. on Document Image Understanding Technology*, April 30 – May 2 1997.
7. B. A. Yanikoglu and L. Vincent, "PinkPanther: a complete environment for ground-truthing and benchmarking document page segmentation," *Pattern Recognition* **31**, pp. 1191–1204, September 1998.
8. S. Randriamasy and L. Vincent, "Benchmarking page segmentation algorithms," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, June 1994.
9. T. Kanungo, R. M. Haralick, and I. Phillips, "Non-linear local and global document degradation models," *Int. Journal of Imaging Systems and Technology* **5**(4), 1994.
10. T. Kanungo, *Document Degradation Models and a Methodology for Degradation Model Validation*. PhD thesis, University of Washington, Seattle, WA., 1996. http://www.cfar.umd.edu/ kanungo/pubs/phdthesis.ps.Z.
11. H. S. Baird, "Document image defect models," in *Structured Document Image Analysis*, Springer-Verlag, New York, 1992.
12. T. Kanungo and R. M. Haralick, "An automatic closed-loop methodology for generating character groundtruth for scanned images," *IEEE Transactions on Pattern Analysis and Machine Intelligence* , December 1998.
13. K. W. Church and R. Mercer, "Introduction to the special issue on computational linguistics using large corpora," *Computational Linguistics* **19**(1), pp. 1–24, 1993.
14. P. Resnik, M. B. Olsen, and M. Diab, "The Bible as a parallel corpus: Annotating the 'Book of 2000 Tongues'," *Computers and the Humanities* , in press.
15. P. Proctor, ed., *Longman Dictionary of Contemporary English (LDOCE)*, Longman Group, 1978.
16. W. N. Francis and H. Kučera, *Frequency Analysis of English Usage: Lexicon and Grammar*, Houghton Mifflin, 1982.
17. I. D. Melamed, "Automatic construction of clean broad-coverage translation lexicons," in *Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas*, (Montreal, Canada), 1996.
18. T. K. Landauer and M. L. Littman, "Fully automatic cross-language document retrieval using latent semantic indexing," in *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, pp. pages 31–38, (UW Centre for the New OED and Text Research, Waterloo, Ontario), October 1990.
19. IBS, *Arabic Bible: New Arabic Version*, International Bible Society, 1988.
20. T. Kanungo, R. M. Haralick, and I. Phillips, "Global and local document degradation models," in *Proc. of Second Int. Conf. on Document Analysis and Recognition*, pp. 730–734, (Tsukuba, Japan), October 1993.
21. G. Borgerfors, "Distance transforms in digital images," *Computer Vision, Graphics, and Image Processing* **34**, pp. 344–371, 1986.
22. R. Haralick, S. Sternberg, and X. Zhuang, "Image analysis using mathematical morphology," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **PAMI–9**, pp. 532–550, July 1987.
23. L. Lamport, *LATEX: a document preparation system*, Addison-Wesley, Reading, MA, 1986.
24. D. E. Knuth, *TEX: the program*, Addison-Wesley, Reading, MA, 1988.
25. P. Vojta *et al.*, "XDVI Software," 1990.
26. T. Kanungo, R. M. Haralick, H. S. Baird, W. Stuetzle, and D. Madigan, "Document degradation models: Parameter estimation and model validation," in *Proc. of Int. Workshop on Machine Vision Applications*, (Kawasaki, Japan), December 1994.
27. T. Kanungo, H. S. Baird, and R. M. Haralick, "Validation and estimation of document degradation models," in *Proc. of Fourth Annual Symp. on Document Analysis and Information Retrieval*, (Las Vegas, NV), April 24-26 1995.