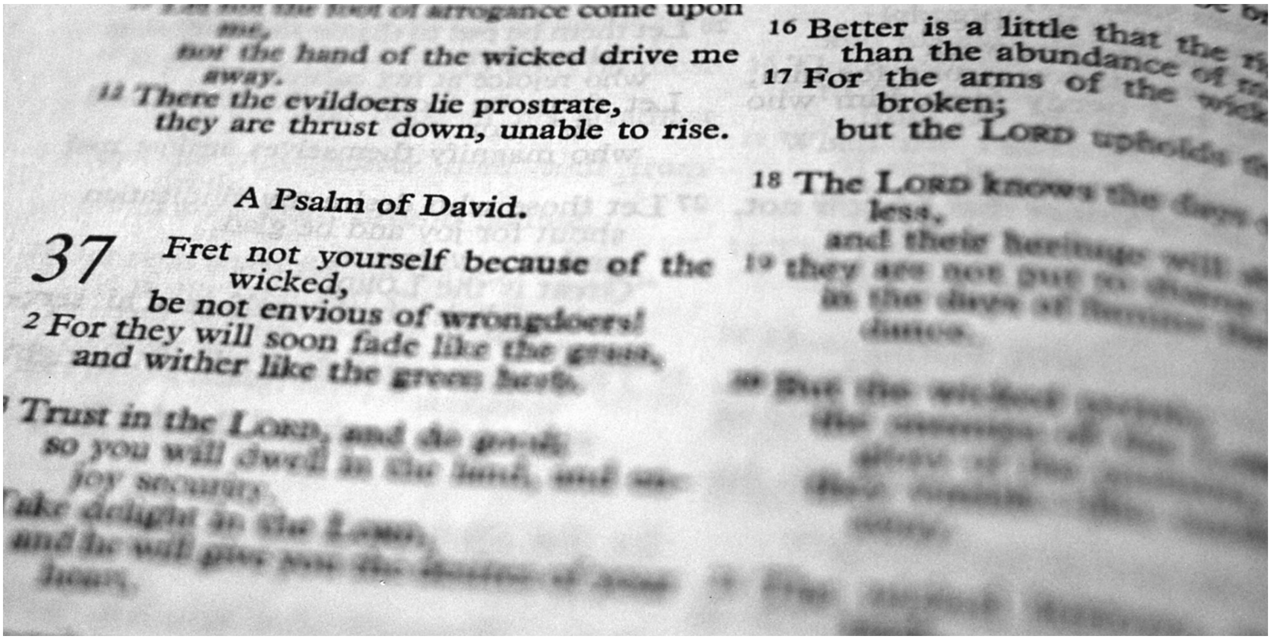
THE BIBLE AND OPTICAL CHAR RECOGNITION

The Bible—an unlikely resource for language technology research—proves ideal for evaluating OCR techniques.

As global online access to information becomes more common, the technology of multilingual optical character recognition (OCR) increases in importance as a way to convert paper documents into electronic, searchable, text. In OCR, as in any evolving technology, careful evaluation is an integral part of research and development. OCR evaluation is done by comparing a system's output for a dataset of document test images with the corresponding correct symbolic text, known as *ground truth*. Unfortunately, the usual way of obtaining ground truth is by manual data entry by humans, which is labor-intensive, time-consuming, expensive, and prone to errors. Worse, because no single set of ground truth evaluation data can be used in more than one language, there has until now been no way to conduct carefully controlled OCR experiments in a multilingual setting.

MULTILINGUAL ACTER

By Tapas Kanungo, Philip Resnik, Song Mao, Doe-Wan Kim, and Qigong Zheng


Let not the foot of arrogance come upon me,
nor the hand of the wicked drive me away.
12 There the evildoers lie prostrate,
they are thrust down, unable to rise.
A Psalm of David.
37 Fret not yourself because of the wicked,
be not envious of wrongdoers!
2 For they will soon fade like the grass,
and wither like the green herb.
Trust in the LORD, and do good;
16 Better is a little that the
than the abundance of
17 For the arms of the
broken;
but the LORD upholds
18 The LORD knows the days
less,

CAREN ROSENBLATT

To address this problem, we introduce the Bible as a dataset for evaluating multilingual OCR accuracy. Bible translations are closely parallel in structure, careful to preserve meaning, surprisingly relevant with respect to modern-day language, and widely available. These properties make the Bible attractive as a way to control document content while varying language, and we control document layout by using synthetically generated page image data.

When physical pages are processed through a scanner, it is challenging to unambiguously identify what characters and words appeared on the original page. Noise from the scanner, variation in font types and sizes, and inherent ambiguity (for example, the letter l being used as the digit 1 in older typed materials) lead to uncertainty in the recognized output text, and so developers and users of OCR must scientifically characterize algorithm performance in terms of a continuous measure of recognition accuracy. The competitive marketplace, however, has forced numerous commercial OCR system vendors to claim near-perfect text recognition (close to 99.9%). These accuracy rates are rarely achieved in practice; most systems break down when the input document images are highly degraded, such as scanned images of carbon-copy documents, low-resolution faxed documents, and *n*-th generation photocopies.

As a result, independent scientific experimentation with OCR systems and algorithms is needed in order to monitor progress in the field, identify areas that need improvement, and explain why a system performs at a particular level of accuracy. Furthermore, when OCR is a component of a larger system, such as machine translation or information retrieval, it is important to understand how the overall performance is related to the performance of individual subsystems.

The traditional experimental methodology for evaluating of OCR text recognition has several stages. First a corpus of paper documents is selected and scanned. Next, the text zones are delineated in each image. Then, for each text zone, the correct text string (ground truth) is keyed in manually. The process of delineating the zones and keying in the text is laborious, prohibitively expensive, and prone to human errors. Finally, the OCR algorithm is run on each text zone and the text strings produced are compared with the corresponding keyed-in ground truth text using a string matching routine.

TO MEANINGFULLY COMPARE SYSTEMS *for different languages, the contents of the documents must somehow be held fairly constant, yet by definition each system must receive its input in a different language.*

In theory, the corpus should be a representative sample of the population of images for which the algorithm was designed. In practice, however, factors such as time and cost force us to limit the size of the dataset. This process was adopted by the OCR evaluation program at the University of Nevada at Las Vegas [11] and the Arabic OCR evaluation process at the University of Maryland [4]. Since each evaluation had its own (often heterogeneous) sets of documents, ranging from business letters to technical journal articles, comparing accuracies across datasets is not very meaningful.

More recently, Kanungo et al. [2, 3] advocated the use of synthetically generated degraded images of entire documents for OCR evaluation. Documents are first typeset using a standard typesetting system with open file representation such as $\TeX$ [6]. Then a noise-free bitmap image of the document and the corresponding ground truth is automatically generated. The noise-free bitmap is degraded using a parameterized degradation model [2, 3], varying model parameters to control the degradation level. This method completely avoids the laborious processes of manual data entry and manual scanning, and is entirely independent of language, as far as the limits of the typesetting software. Moreover, since the experimenter controls the typesetting, the effects on OCR accuracy of page layout, font size, and type can be experimentally controlled.

THE BIBLE AS A CORPUS

While synthetically generated data and degradation models make it possible to control visual properties that affect OCR algorithm performance, we are faced with a conundrum as soon as we attempt to compare OCR systems that work in different languages. In order to meaningfully compare systems for different languages, the *contents* of the documents must somehow be held fairly constant, yet by definition each system must receive its input in a different language.

To address this problem, we have taken the unusual step of using the Bible as an OCR evaluation dataset. The Bible 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 discuss [9], the Bible is surprisingly relevant for research involving present-day language; for example, in domains such as cross-language information retrieval and machine translation for low-density languages. Resnik et al. evaluate the vocabulary of the New International Version (NIV) Bible against two benchmarks: the approximately 2,200-word control vocabulary for Longman's Dictionary of Contemporary English (LDOCE [7]), and the most frequent 2,000 words in the Brown corpus of present-day American English [1] (an oft-cited source of word frequency data for English).

Their analysis demonstrates that 78–85% of the items in the LDOCE control vocabulary are found in the NIV, including ample vocabulary representative of typical, everyday usage as well as a wide range of English orthography. A similar comparison focusing on frequently used words shows that, of the most fre-

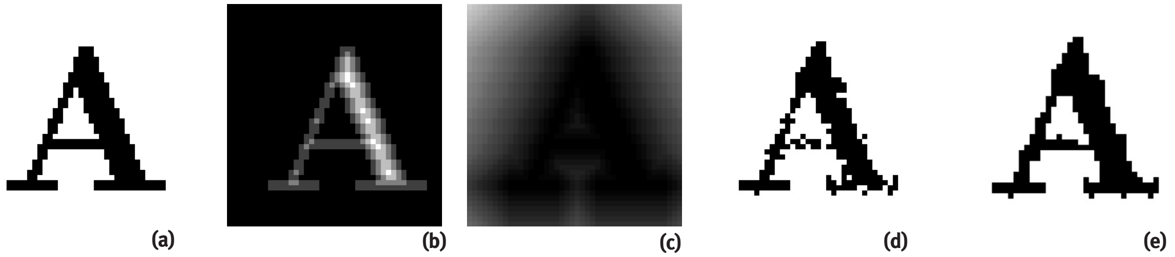


Figure 1. Local document degradation model: (a) Ideal noise-free character; Distance of the black pixels (b) and white pixels (c)—pixels farther away from the character boundary are brighter. (d) Result of the random pixel-flipping process. The probability of a black pixel flipping is $P(1|0, d, \beta_0, \beta, \eta) = \beta_0 e^{-\beta d^2}$ **and that of white is** $P(0|1, d, \alpha_0, \alpha, \eta) = \alpha_0 e^{-\alpha d^2}$**; (e) Blurring of the result in (d) by a disk of size** k**. The model parameter used is** $\Theta = (\eta, \alpha_0, \alpha, \beta_0, \beta, k)$ **= (0.0, 1, 2, 1, 2, 2).**

quent 2,000 words in the Brown corpus, fully 75% occur in the NIV.[1] 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 2,000 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.

This parallelism of content at a global level is matched by parallelism at a much finer grain. Indeed, Bible translations are done with great care to preserve nuances of meaning and they generally maintain verse-level parallelism. This permits fine-grained analysis for OCR evaluation—for example, 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 [8] and cross-language analysis of semantic patterning [10].

Complete Bible translations exist in over 383 languages, New Testament translations in 987, and at least one book of the Bible exists in 2,261 languages. Moreover, these numbers are increasing rapidly.[2] Bibles in hardcopy exist in print in various formats, fonts, and paper types, and Bibles in many languages are available online or in electronic form—often free or for a reasonable licensing cost—providing an alternative to manual entry of ground truth data. As a text corpus, the Bible is large 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 1,000 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 of more modern origin or simply outside its scope (for example, *atom, Buddhist, January, cat*). Formats for addresses, dates, and the like are also absent, as are complex layouts such as tables and graphics. (However, Biblical poetry and sometimes pictures do appear in some editions.) Thus, there is a trade-off: the Bible lacks some elements that might help distinguish OCR systems' lexicon coverage, page segmentation, or zone classification performance, but it provides an unmatched degree of consistency, availability, and parallelism.

Evaluation Using a Bible Image Dataset

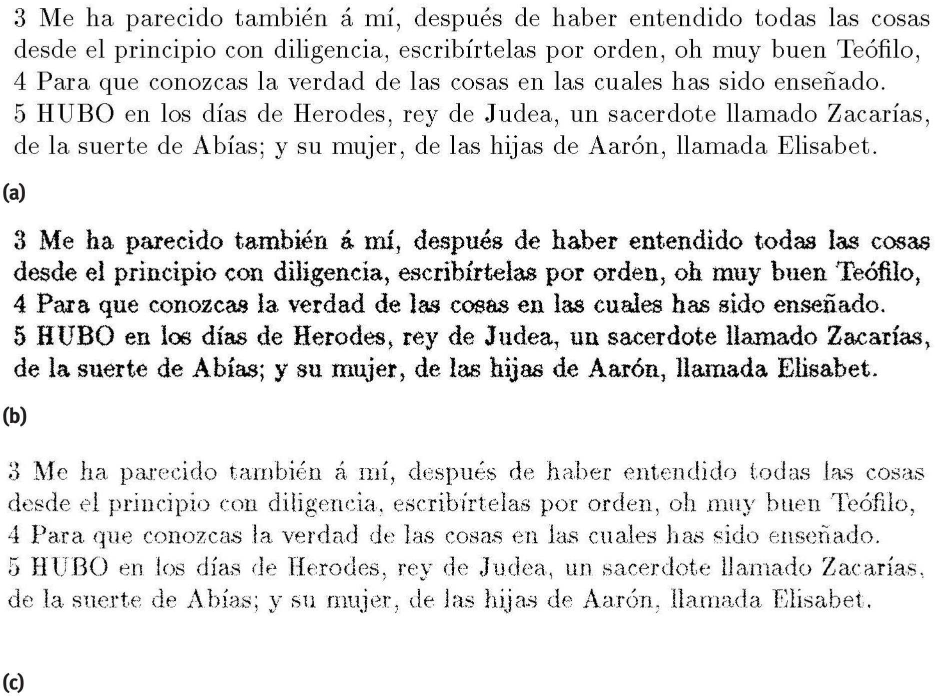
To create an evaluation dataset, we used a degradation model to generate synthetically degraded documents [2, 3]. The degradations produced by this model are local—the sort that appear while scanning a flat page.[3] Noise-free and degraded document images were generated for complete Bibles in seven languages, and 15 OCR systems were evaluated.

Local image degradation occurs for many reasons. For example, variation in light intensity, sensor sensitivity, and thresholding level can result in random pixel inversions (from black to white and vice versa). This is typically more pronounced near the boundary of the character. Transformations due to the point-

[1] Conversely, 62% of the NIV's vocabulary items appear in the Brown corpus. Based on a random-sample analysis of the 38% of NIV vocabulary not appearing in the Brown corpus, we find, not surprisingly, that the vast majority of Biblical terms missing from the Brown corpus are proper names (76%; for example, *Balaam, Manahathites*). Of the remainder, 11% are hyphenated words or numbers (*sun-scorched*, 721) and 13% are common words that don't happen to occur in the Brown sample (*drowsiness, slur, ravage, sightless*). Excluding Biblical proper names, therefore, Brown corpus coverage for the vocabulary of a modern language Bible is very high indeed.

[2] *Scripture Language Report 2000*, United Bible Society; www.biblesociety.org/wr 358/slr 2000.htm.

[3] A more general model that accounts for perspective distortions near the spine of a thick book is described in [3].



3 Me ha parecido también á mí, después de haber entendido todas las cosas
desde el principio con diligencia, escribírtelas por orden, oh muy buen Teófilo,
4 Para que conozcas la verdad de las cosas en las cuales has sido enseñado.
5 HUBO en los días de Herodes, rey de Judea, un sacerdote llamado Zacarías,
de la suerte de Abías; y su mujer, de las hijas de Aarón, llamada Elisabet.

(a)

3 Me ha parecido también á mí, después de haber entendido todas las cosas
desde el principio con diligencia, escribírtelas por orden, oh muy buen Teófilo,
4 Para que conozcas la verdad de las cosas en las cuales has sido enseñado.
5 HUBO en los días de Herodes, rey de Judea, un sacerdote llamado Zacarías,
de la suerte de Abías; y su mujer, de las hijas de Aarón, llamada Elisabet.

(b)

3 Me ha parecido también á mí, después de haber entendido todas las cosas
desde el principio con diligencia, escribírtelas por orden, oh muy buen Teófilo,
4 Para que conozcas la verdad de las cosas en las cuales has sido enseñado.
5 HUBO en los días de Herodes, rey de Judea, un sacerdote llamado Zacarías,
de la suerte de Abías; y su mujer, de las hijas de Aarón, llamada Elisabet.

(c)

Figure 2. Application of the degradation model on a Spanish Bible image. The layout was formatted using TEX and degraded using the model described here. (a) A small fragment of the entire noise-free text. (b) Artificially degraded version of (a) generated with the model parameters set to create a "blurry" image. (c) A degraded version of (a) with model parameters set to create "broken" image.

spread function of the scanner's optical system can, on the other hand, result in thick and joined characters, or thin and broken characters.

Various components of the degradation model [3] account are designed to account for these effects.[4] The flipping probability of a pixel is modeled as an exponential function of its distance d from the nearest boundary pixel. Parameters α_0 and α control the probability of a black pixel switching to white, and β_0 and β control the probability of a white pixel switching to black. The parameter η is the constant probability of flipping for all pixels. Finally, the parameter k, accounts for the correlation introduced by the point-spread function. Thus the degradation model has six parameters $\Theta = (\eta, \alpha_0, \alpha, \beta_0, \beta, k)^t$.

The model is used to degrade a noise-free binary image as follows. First the distance d of each pixel from the nearest character boundary is computed. Then each black pixel is randomly inverted with probability $p(0/1, d, \alpha_0, \alpha, \eta) = \alpha_0 e^{-\alpha d^2} + \eta$, and each white pixel with probability $p(1/0, d, \beta_0, \beta, \eta) = \beta_0 e^{-\beta d^2} + \eta$. Finally the resulting image is blurred with a disk of diameter k. Figure 1 illustrates the steps of the degradation model at the character level.

The noise-free documents are typeset using the TEX formatting system [6]. The files containing the text and the TEX typesetting information are then converted into a device-independent format (DVI) using TEX. A conversion program, dvi2tiff, is run to produce one-bit-per-pixel binary images in TIFF format from the DVI files. In addition to producing binary images of the documents, dvi2tiff produces character-by-character ground truth information for the document image. The implementation of the document degradation model takes as input an ideal binary document image in TIFF format and the degradation model parameter Θ, and produces the binary degraded images in TIFF format.[5] Figure 2 illustrates the application of the degradation model at the page level.

OCR Evaluation Results

We obtained modern-language electronic versions of the Bible in Arabic, Chinese, English, Japanese, Korean, Russian, and Spanish, and used TEX to typeset them in a standard page format. In order to compare OCR system performance under noise-free and noisy conditions, we used our degradation model to create 100 synthetically degraded images in each language. These page images were parallel in the sense that for each page image in one language, there was a corresponding page (with parallel text) in each of the other languages. The chosen font size for Latin scripts was 12-point. Chosen font sizes for other scripts were comparable to the Latin pick. The fonts used were the default fonts provided in the language packages publicly available from the TEX CTAN repository. Figure 2 shows a synthetically degraded image of a page from a Spanish Bible at 300dpi resolution. The same degradation model

[4] Issues regarding model validation and model parameter estimation are discussed elsewhere [2, 5].

[5] Both programs—dvi2tiff and the degradation model software, ddm—are implemented in the C language and are available in the University of Washington database.

parameters were used for each language.

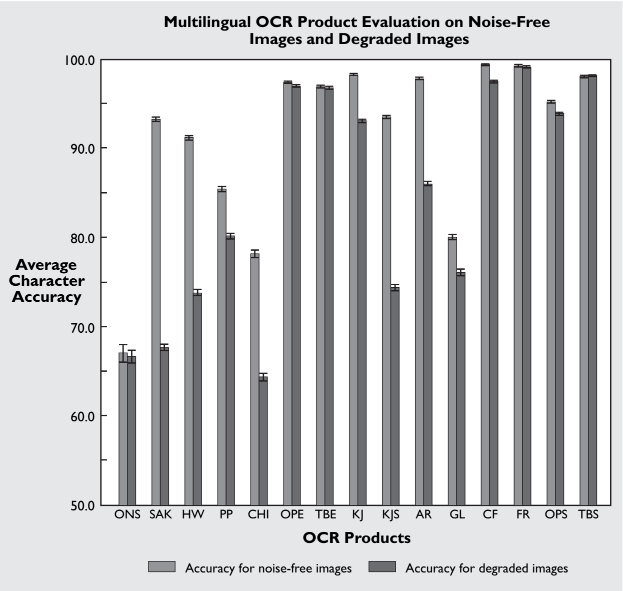
Figure 3 shows performance results for 15 commercial OCR systems. Unlike previous work on OCR evaluation, the simultaneous presentation of cross-language results here affords a meaningful comparison of the state of the art for OCR in different languages. In particular, by using the Bible dataset, we can be reasonably confident of having controlled for differences in page layout, proportion of names versus common words, proportion of function versus content words, and other variables related to document content.

Figure 3. Plot of performance results of OCR products in various languages. The table in (a) gives the product names, their abbreviated names, and the text language they recognize. The plot in (b) shows two average character recognition accuracy for each product. The light bar represents accuracy for noise-free images and the dark bar represents accuracy for degraded images. The values are in percentage.

(a)

Language	OCR Product Name	Abbr. Product Name
Arabic	Onset2	ONS
	Sakhr3	SAK
Chinese	Hw99	HW
	PenPower2.0	PP
	CHIOCR	CHI
English	OmniPage8.0	OPE
	TextBridge98	TBE
Japanese	KanjiOCR2.0	KJ
	KanjiScan1.0	KJS
Korean	Armi4.0	AR
	Glnun97	GL
Russian	Cuneiform2000	CF
	FineReader4	FR
Spanish	OmniPage8.0	OPS
	TextBridge98	TBS

(b)



We can see, for example, that Arabic OCR systems in general perform more poorly than the English and Spanish. While the number of Arabic characters is comparable to that of English, Arabic text has connected script, and the shape of the symbols change depending on the preceding and following symbols.

Recognizers that first segment the text and then classify it have poor performance due to segmentation errors. Algorithms for Chinese also perform poorly compared to English. While Chinese text is composed of isolated characters like in English, the number of symbols in Chinese is much larger than English. In fact, it is estimated that one needs to know about 3,000 symbols just to read a Chinese newspaper, and the official number of Chinese symbols is much larger. Thus, for comparable recognition accuracy, one needs a much larger training corpus for Chinese than for English. Furthermore, the accuracy drop between the degraded and noise-free images is larger for some systems than others, suggesting that some algorithms may be more robust to noise. Since the same noise model parameters were used across languages, and the contents used were translations, we can be confident the performance distinctions between OCR systems for different languages, and the directions they suggest, are based on meaningful comparisons rather than being artifacts of a heterogeneous document collection.

Summary

We have described a method for creating OCR evaluation datasets that permits a greater degree of experimental control than has previously been available. By starting with electronic documents and generating synthetically degraded documents, we exercise control over the visual properties of documents. By using the Bible as the text, we ensure control over conceptual content and a range of linguistic properties that otherwise might represent experimental confounds, and we avoid the impractical alternative of creating a dataset of this kind from scratch. These issues grow

BY USING THE BIBLE AS THE TEXT, WE ENSURE *control over conceptual content and a range of linguistic properties that otherwise might represent experimental confounds, and we avoid the impractical alternative of creating a dataset of this kind from scratch.*

in importance as OCR research proceeds in an increasingly multilingual setting. In future work we hope to generate similar datasets for more languages, and we are exploring alternative text corpora with similar properties.

References

1. Francis, W.N., and Kucera. H. *Frequency Analysis of English Usage: Lexicon and Grammar.* Houghton Mifflin, 1982.
2. Kanungo, T., Haralick, R.M., Baird, H.S., Stuetzle, W., and Madigan, D. Statistical, nonparametric methodology for document degradation model validation. *IEEE Trans. on Pattern Analysis and Machine Intelligence 22,* 11 (2000), 1209–1223.
3. Kanungo, T., Haralick, R.M., and Phillips, I. Nonlinear local and global document degradation models. *Int. Journal of Imaging Systems and Technology 5,* 4 (1994) 220–230.
4. Kanungo, T., Marton, G., and Bulbul, O. OmniPage vs. Sakhr: Paired model evaluation of two Arabic OCR products. In *Proceedings of SPIE Conference on Document Recognition and Retrieval.* D. Lopresti and Y. Zhou, Eds. (San Jose, CA, 1999), 109–120.
5. Kanungo, T., and Zheng, Q. Estimating degradation model parameters using neighborhood pattern distributions: An optimization approach. *IEEE Trans. Pattern Analysis and Machine Intelligence 26,* 4 (2004), 520–524.
6. Knuth, D.E.. *TEX: The Program.* Addison-Wesley, Reading, MA, 1988.
7. Proctor, P. *Longman Dictionary of Contemporary English* (LDOCE). Longman Group, 1978.
8. Resnik, P., Oard, D., and Levow, G. Improved cross-language retrieval using backoff translation. In *Proceedings of the Human Language Technology Conference* (San Diego, CA, Mar. 2001).
9. Resnik, P., Broman Olsen, M., and Diab, M. The Bible as a parallel corpus: Annotating the “Book of 2000 Tongues.” *Computers and the Humanities 33,* 1–2 (1999) 363–379.
10. Resnik, P., and Yarowsky, D. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering 5,* 2 (1999), 113–133.
11. Rice, S.V., Jenkins, F.R., and Nartker, T.A.. The fifth annual test of OCR accuracy. Technical Report TR-96-01. Information Science Research Institute, University of Nevada, Las Vegas, NV, 1996.

Tapas Kanungo (kanungo@us.ibm.com) is a research staff member at IBM Almaden Research Center, San Jose, CA.

Philip Resnik (resnick@umiacs.umd.edu) is an associate professor at UMIACS and the Department of Linguistics, University of Maryland, College Park, MD.

Song Mao (smao@mail.nih.gov) is a Postdoctoral Fellow at the National Library of Medicine, Bethesda, MD.

Doe-wan Kim (doewan@lge.com) is a senior research engineer at LG Electronics, Seoul, Korea.

Qigong Zheng (qzheng@lucent.com) is a member of the technical staff at Lucent Technologies, Westford, MA.