Document Image Collection Using Amazon's Mechanical Turk

Stephanie Strassel
Linguistic Data Consortium
3600 Market Street, Suite 810
Philadelphia, PA 19104, USA
strassel@ldc.upenn.edu

Audrey Le, Jerome Ajot, Mark Przybocki
National Institute of Standards and Technology
100 Bureau Drive, Stop 8940
Gaithersburg, MD 20876, USA
{audrey.le|jerome.ajot|mark.przybocki} @nist.gov

Abstract

We present findings from a collaborative effort aimed at testing the feasibility of using Amazon's Mechanical Turk as a data collection platform to build a corpus of document images. Experimental design and implementation workflow are described. Preliminary findings and directions for future work are also discussed.

1 Introduction

The National Institute of Standards and Technology (NIST) and Linguistic Data Consortium (LDC) at the University of Pennsylvania have a strong collaborative history of providing evaluation and linguistic resources for the Human Language Technology (HLT) community[1]. The NAACL 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk[2] presents an interesting opportunity to extend this collaboration in a novel data collection task. This collaborative experiment will occur in the context of the NIST Open Handwriting Recognition and Translation Evaluation (OpenHaRT) (NIST, 2010), which requires a collection of Arabic handwritten document images. While some Arabic handwritten document collections do exist (Combating Terrorism Center, 2006, 2007; University of Colorado at Boulder, 1998) these resources are inadequate to support an open technology evaluation. Some existing corpora are not publicly accessible, while others are very small or limited in scope/content, or contain features (e.g. Personal Identifying Information) that prevent their use in a NIST evaluation. New data collection for OpenHaRT using traditional methods of recruiting human subjects is cost-prohibitive and time consuming.

Recent studies (Callison-Burch, 2009) have demonstrated the viability of Amazon's Mechanical Turk as a data collection platform for tasks including English translations for foreign text sources. We propose to build on the success of previous studies, and expand data collection to target highly variable samples of foreign handwritten texts along with their English translations. While data collected from this effort will be donated to the workshop and larger research community, our hope is that this collaboration will also provide a means to explore the feasibility of this approach more generally, and that it will result in protocols that can be used to collect substantial volumes of handwritten text to support the NIST OpenHaRT evaluation. The remainder of this paper documents this pilot study, describing both the experimental design and implementation workflow followed by our findings and hypotheses.

2 Data Collection Experimental Design

Our data collection targets images of handwritten foreign language text, prepared according to a pre-defined set of characteristics. Text from the images is transcribed verbatim. English translations of the text are provided. Collected data is verified for accuracy[3] and image quality.

[1] Since 1987 NIST has conducted public evaluations of human language technologies and has collaborated with LDC to collects much of the data used in support for these evaluations.

[2] http://sites.google.com/site/amtworkshop2010

[3] Due to time constraints, transcript and translation verification was conducted offline by LDC staff.

2.1 Collection Approach

For this pilot study we collected data in two primary languages, Arabic and Spanish[4]. These languages are of interest for a number of reasons. Linguistically, they show large typological and orthographic differences. Strategically, Arabic is of high interest to a number of ongoing HLT evaluations, while Spanish is important to U.S. commercial interests. Practically, we also hoped to take advantage of a likely pool of Spanish-speaking Turkers[5] whose facility with the written language may vary. We defined two categories or genres for collection – *shopping list* and *description of the current weather*. The rationale for selecting the general shopping list was to elicit text with a potentially large set of vocabularies while the description of the current weather was included to elicit text with a narrow set of vocabularies.

To simplify the collection process and to make the tasks as natural as possible, we placed no artificial constraints on the writers. That is, we do not regulate writing implements (e.g., pen, pencil, crayon, marker, etc.), paper types (e.g., lined, unlined, graphed, colored, etc.), orientation of the handwriting (e.g., straight, curved, etc.), handwriting speed, etc. To sample naturally-occurring variation in digital images, we placed no constraints on image quality (resolution, lighting, orientation, etc.) Many of these features could be labeled as subsequent Mechanical Turk HITs[6].

2.2 Collection Tasks

The collection has three types of tasks:

- *Image Collection* – This task requires the Turker to perform a writing assignment given a specified topic and source language. The writing assignment is electronically scanned or photographed and uploaded to our repository.
- *Image Transcription* – For each handwritten image, the corresponding text is transcribed verbatim.
- *Image Translation* – For each transcribed foreign language text, an accurate and fluent English translation is provided.

2.3 Task Implementation and Quality Control

Each task listed above corresponds to a single HIT. Initial payment rates for each HIT type were established after reviewing comparable HITs available between 2/19/10 – 2/23/10. Payment rates were finalized after additional review of comparable HITs in mid-March; HIT payments were also adjusted to encourage rapid completion for some tasks. Arabic HITs were priced higher than Spanish HITs because we wanted to investigate the price dimension when we compared Arabic to a language that is more widely spoken by the population at large[7].

Image Collection

We targeted collection of 18 images per language (Spanish, Arabic) per genre (weather report, shopping list) for a total of 36 per language. Three English shopping list images were also collected as a control set for comparison of Turker performance. HIT instructions were brief:

> 1. Take a piece of paper, and write down {a brief description of today's weather | a shopping list} in {Spanish | Arabic}. You can use any type of paper and writing implement (pen, pencil, etc.) you have handy, but only write on the front of the page. Use your normal handwriting.
> 2. Using a digital camera or scanner, take a picture or scan a copy of the {weather report | list} you just created. Make sure you don't cut off any of the handwriting.
> 3. Upload the image file[8].

[4] A very small English collection was undertaken to provide a baseline control set for comparison.

[5] "Turkers" is the term used to refer to people who perform tasks for money at the online marketplace Amazon's Mechanical Turk.

[6] Coined by Amazon, HIT stands for Human Intelligence Task and refers to a task that a person can work on and be compensated for completing the work.

[7] http://www.nationsonline.org/oneworld/most_spoken_languages.htm

[8] Turkers were not given instructions about how to name the uploaded file; such instructions could have facilitated task/workflow management and should be implemented in future efforts.

HIT instructions were written in English for the Spanish-language task; a note at the top of the HIT specified that the task should be performed by Spanish speakers. For the Arabic-language task, initial instructions were also written in English; this was later revised to use instructions written in Arabic to better target Arabic-speaking Turkers. We did not require Turkers to take a language qualification test. The HITs remained open for one week, and time allocated per HIT was 30 minutes. Payment for the image collection task was set at $0.11 per image for Spanish and $0.15 for Arabic.

Quality control for the image collection task involved annotators at LDC reviewing each submitted image and determining whether it was in fact in the targeted language and genre (weather report or shopping list).

Image Transcription

Each image was then transcribed. We targeted two unique transcripts per collected, approved image[9]. HIT instructions were as follows:

> •The image below contains {Spanish | Arabic} handwriting. Your job is to transcribe exactly what you see. Type out all the words and punctuation you see, exactly as they are written. Do not correct any spelling mistakes or other errors in the handwriting. If the image contains any punctuation, copy that exactly using the punctuation character on your keyboard that is closest to what was written.
> •If the handwritten image is a list on multiple lines, transcribe one line at a time, inserting a line break (by hitting the "Enter" key) after each new line.
> •For any words that you cannot read, or if you're just not sure what the writing says, just do the best you can and transcribe as much of the word as you can make out. Add ?? to the beginning of any word you are not sure of.
> •Before submitting your transcript, please double check to make sure you have transcribed every line in the image, without leaving anything out or adding anything that isn't in the image.

Instructions for the Spanish transcription task were provided in English, whereas the Arabic instructions were written in Arabic. For this task we required Turkers to have an approval rating of 95% or higher. The HITs remained open for four days for Spanish and one week for Arabic; time allocated per HIT was 30 minutes. Payment for the transcription task was set at $0.20 per image for Spanish and $0.25 for Arabic.

Quality control on the transcription task involved fluent Spanish or Arabic annotators at LDC reviewing the transcripts against the image and making a three-level accuracy judgment: perfect transcript (no errors); acceptable transcript (minor errors in transcription or punctuation); unacceptable transcript (major errors). Transcripts judged as "perfect" or "acceptable" were passed on to the final translation task.

Image Translation

We targeted one unique translation per collected, approved transcript. HIT instructions were as follows:

> Below is a brief shopping list or weather report in {Spanish | Arabic}. Your job is to provide an English translation of this document. Your translation should be accurate, and should use fluent English.
> •Translate every sentence, phrase or word, without leaving anything out or adding any information.
> •If there are spelling mistakes or other errors in the {Spanish | Arabic} text, just translate what you think the intended word is.
> •If there is any punctuation in the {Spanish | Arabic} text, copy that over exactly into the English translation.
> •Try to follow the same document layout and line breaks as in the original {Spanish | Arabic} text.
> •Some {Spanish | Arabic} words may have ?? at the beginning. You should copy the ?? over onto the beginning of the corresponding English translated word.
> •Put !! at the beginning of any English word whose translation you're not sure of.

[9] The total number of images assigned for transcription was lower than the number collected in some cases, due to timeline and task staging constraints.

•NOTE: Do not use automatic translations from the web for this task. Such submissions will be rejected.

Because this task targeted fluent English translations, instructions were written in English for both the Spanish and Arabic translation HITs. For this task we required Turkers to have an approval rating of 95% or higher. The HITs remained open for two days for Spanish and four days for Arabic; time allocated per HIT was 1 hour. Payment for the translation task was set at $1.25 per image for Spanish and $1.50 for Arabic[10].

Quality control on this task involved LDC bilingual annotators checking the translation against the transcript, and making a judgment of "acceptable" or "unacceptable". Perfect translation was not required but the translation had to be a generally adequate and fluent rendering of the foreign language text. Translation QC annotators were permitted to consult the image file for context, but were not permitted to penalize a Turker based on information only available in the image file, since Turkers working on translation HITs did not have access to the image file.

3 Collection Yield and Results

Table 1 summarizes the total number of image, transcription and translation HITs made available, submitted and approved for each language and genre. As originally planned, our study would have produced a total of 36 images per language, with two transcripts per image (for a total of 72 per language) and one translation per transcript (72 per language). Actual yields for the image collection task were considerably lower, and targets for the subsequent tasks were adjusted.

In the case of Arabic, all approved images were made available for transcription. For Spanish some images were submitted and approved after the transcription HITs had been assigned; time constraints did not permit creating additional transcription HITs for these later images. For both languages, all approved transcription HITs were made available for subsequent translation.

Note too that the number of submitted HITs actually exceeds the number of available HITs in some cases; this is because rejected HITs were made available for completion by new Turkers.

		Avail. HITs	Submtd	Aprvd
Spanish Shopping List	*Images*	18	13	10
	Transcripts	14	16	14
	Translations	14	21	12
Spanish Weather Report	*Images*	18	7	5
	Transcripts	6	6	6
	Translations	6	13	5
Arabic Shopping List	*Images*	18	11	3
	Transcripts	6	9	6
	Translations	6	7	0
Arabic Weather Report	*Images*	18	6	2
	Transcripts	4	5	4
	Translations	4	5	1
English Shopping List	*Images*	3	3	3
	Transcripts	6	6	6
	Translations	n/a		

Table 1: Collection Summary

Proof of Concept: English Control Set

Collection of the small English-language control set was entirely successful: Turkers quickly completed the image collection task and provided accurate transcripts of each English image. Though small in number, the submitted images show considerable variation in image quality (lighting, rotation, resolution, scan versus photo) and handwriting quality (paper type and writing implement).

All images were approved during the quality control pass. Transcript collection was extremely fast: all six transcripts (two copies per image) were collected within minutes of posting the HITs. Transcript quality was uniformly acceptable. As a baseline, the English control task demonstrates the feasibility of using MTurk for at least some kinds of image collection and transcription.

[10] These rates were set in part based on need for rapid completion of these HITs.

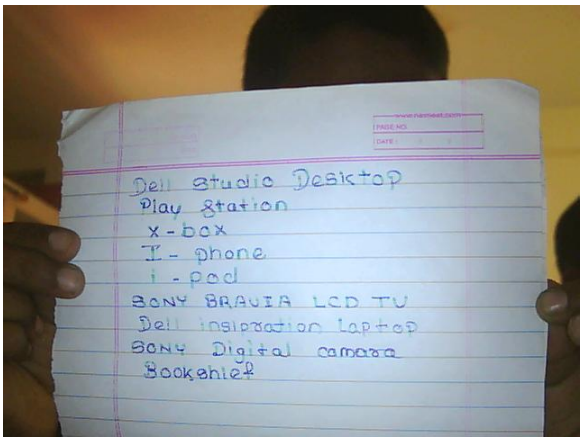

Figure 1: Handwritten English shopping list

Spanish Results

The Spanish language collection was largely successful. The first challenge was finding fluent Spanish speaking Turkers. No special effort was made to advertise the task to Spanish speakers beyond posting the hits on MTurk. Each HIT's title, description and associated keywords included the term "Spanish" but did not contain any Spanish language content.

While we targeted a total of 36 Spanish images, only 20 were submitted, of which 15 were approved. Images were rejected largely because of fraud, principally stemming from duplicate copies of the same handwritten image being submitted under different Worker IDs. Image and handwriting quality showed a great deal of variation. Turkers used plain unlined paper, lined paper and graph paper with a variety of writing implements. Some submitted printed handwriting while others used cursive. We observed some interesting document formatting issues; for instance some Turkers provided multi-column shopping lists. Image quality ranged from a clean, high resolution scan with the image perfectly centered, to low-quality bitmap files with edges of the paper bent or wrinkled and the page skewed off-center. Other image artifacts included lighting variation within a single image due to the use of a flash while photographing the image.

The transcription task was completed largely as planned, with two transcripts acquired for all images. Two transcripts HITs were rejected, in both cases because the Turker provided a translation instead of a transcript; these images were made available for re-assignment to other Turkers and accurate transcripts were eventually obtained. The transcription task presented several difficulties that, while anticipated, were not fully addressed in this limited pilot study. While Spanish handwriting contains numerous diacritics (e.g., the tilde in piñata) these were variably rendered in the transcription task. Some transcribers tried to incorporate the diacritics directly, whereas others used plain ascii for transcription resulting in either missing diacritics, or non-standard symbols standing in for diacritics. For instance, "piñata" might be alternately transcribed as "piñata", "pinata", "pin~ata" or something else. The issue of input and rendering for non-English characters is a well-known problem in corpus creation, but in this pilot study no special effort was made to control for it. Similarly, special formatting characters (e.g. for bullet-pointed shopping lists) were variably rendered by Turkers and did not always display as intended in the resulting output file. During transcription QC, LDC annotators made an effort to standardize rendering of such characters to facilitate the translation task. Future MTurk data collection efforts will need to devote more attention to character encoding, input and rendering issues.

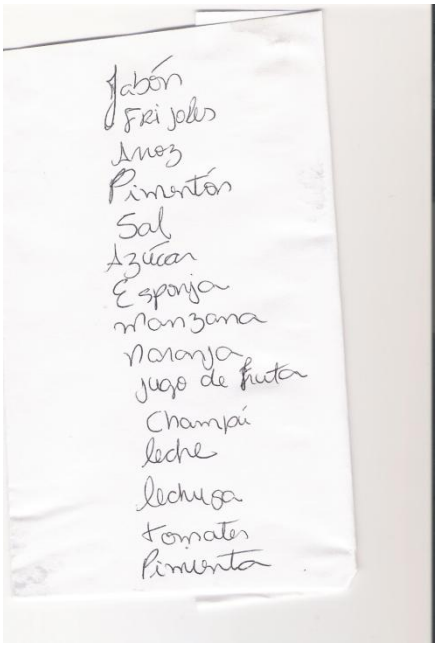

Figure 2: Handwritten Spanish shopping list

Translation proved to be the most difficult of the three tasks. We targeted collection of one

translation per approved transcript, for a planned total of two translations per image. We fell somewhat short of this collection goal, in part because timeline constraints meant the batch of translation HITs was only available for a few days. The rejection rate on translation HITs was also much higher than the rate for images or transcripts. Rejected translation HITs fell into two categories: an obvious machine translation from the web (typically Google Translate)[15]; or an apparent human translation that did not constitute fluent English.

Because the collected images were short and simple with little or no formatting or document structure, and because transcripts were QC'd prior to translation, it was believed that Turkers could create an accurate translation without making reference to the original image file. Therefore, the image was not displayed during translation; instead Turkers were given only a plain text version of the transcript. This approach did contribute to some translation difficulties especially for special characters (like the Celsius symbol, °C, frequently used in the weather reports).

Based on the rejection rate for individual HITs, some images proved harder to translate than others. This appears to be largely an issue of translation difficulty due to specialized terminology (e.g. brand names and abbreviations in a shopping list) rather than influence from errors in transcription.

Arabic Results

Not surprisingly, data collection for Arabic proved quite challenging. Locating fluent Arabic speakers among Turkers was extremely difficult. As noted elsewhere our pilot study was limited to using Amazon's default MTurk infrastructure and so we did not undertake any special efforts to direct Turkers to our HITs beyond posting them on MTurk. Instructions for the image collection and transcription HITs were written in Arabic, and keywords for all tasks contained the words "Arabic", written in both Arabic and English. The HIT titles also contained the word "Arabic" written in both languages.

[15] Each suspect translation was submitted to Google Translate during the QC/review process.

As with Spanish we targeted a total of 36 Arabic images (18 per genre). While nearly as many images were submitted as in the corresponding Spanish task (17 compared to 20 for Spanish), only 5 Arabic images were approved. Reasons for rejection included the image being in English rather than Arabic; the image being typed instead of handwritten; and several cases of identical images being submitted under multiple WorkerIDs. Among the approved images we again observed an exciting range of image and handwriting variation, including several cases of out-of-focus photos; an example is provided in Figure 3.

The Arabic transcription task proved to be fairly straightforward, and we successfully collected two independent transcripts for each approved image. A handful of transcripts were rejected because they were grossly inaccurate (the Turker simply copied the instructions or image URL into the transcript rather than providing an actual transcript).

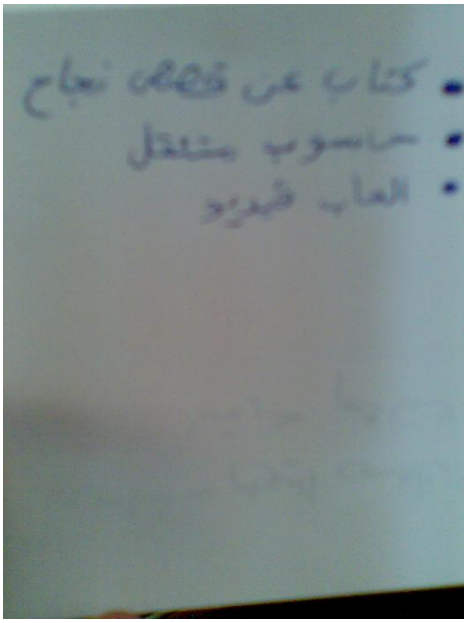

Image 3: Handwritten Arabic shopping list

There was an initial concern about whether Turkers would encounter difficulties inputting Arabic text into the transcription HIT interface but this did not seem to affect performance. One unanticipated difficulty was creation of the HITs themselves; the Amazon HIT management interface had some difficulties rendering bi-directional text. This is a common problem in

annotation tool design, and is especially problematic when left-to-right and right-to-left reading order is required in a single line, for instance when characters like parenthesis or ascii numbers are interspersed with Arabic text. A similar difficulty emerged when viewing batch-wide results of the transcription task. The default file output format (.csv) is intended for viewing in a tool like Excel. However, the output does not appear to be natively Unicode-compliant and therefore Arabic characters are not rendered correctly. No straightforward solution presented itself within the Amazon HIT management interface, and the scope of this pilot project did not permit exploration of solutions using third-party APIs. Instead, results were extracted individually for each HIT using the GUI, which proved to be very time consuming and resulted in a loss of some formatting information (like line wrapping).

Unsurprisingly, the Arabic translation task proved to be the most difficult. Of twelve submitted translation HITs, only one resulted in an acceptable translation. The low success rate is likely due to a number of factors. As discussed, there appear to be few Arabic speakers (and even fewer fluent Arabic-English bilinguals) among the Turker population at large. Second, the translation task was available for only a few days. To offset this, the payment per HIT for Arabic (and Spanish) was quite high, though this may have contributed to the final challenge: fraudulent submissions. Rejected HITs followed the normal pattern. Most were machine translations from the web (again, primarily from Google Translate) while others appeared to be highly disfluent human translations, of the type that might be expected from a first year Arabic student working without a dictionary.

4 Discussion and Future Plans

As a feasibility study, our experiment can be called a qualified success. With respect to image collection MTurk seems to be a viable option at least for some languages and genres. While there was some "fraud", most submitted images were usable and the image properties were highly variable, suggesting that this task is well-suited to MTurk and that the HIT instructions were adequate. A wide range of writing surfaces emerged including lined, unlined and graph paper, as well as colored paper. Less variation was observed in writing implements (for instance it appears that no one used pencil, crayon or fat marker). There were some surprising features of the handwriting itself beyond the expected quality variation; for instance someone writing perpendicular to the lines on ruled paper. Image quality ranged along the expected dimensions of resolution, skew and slant, and scanning artifacts like bent corners or wrinkled pages. There were also unexpected artifacts of image photography including uneven lighting due to a flash going off and out-of-focus images. The content submitted for each genre showed considerable variation as well. For instance, Turkers submitted shopping lists for not just groceries, but also electronics and a combined shopping/to-do list for an upcoming vacation or trip. The results are promising for future image collection efforts at least for English or other languages for which Turkers are readily accessible.

The transcription task was moderately successful. Apart from finding Turkers with appropriate language skills, the primary challenges are technical, in terms of character encoding for input and display/rendering. The basic Amazon MTurk interface does not provide adequate support to fully resolve this and future efforts will need to explore other options. Quality control is a bigger issue for the transcription task, and adequate resources must be devoted to addressing this in any future collection. Multiple transcripts were generated for each image to facilitate using MTurk to collect comparative judgments on transcription quality, although time constraints prevented this from being implemented. In future we envision incorporating three kinds of MTurk QC for transcription: simple judgment of transcript accuracy; comparison of multiple transcripts for a single image; and correction of transcripts judged inadequate. It is expected that project staff (as opposed to Turkers) will still need to be involved in some degree of QC. We anticipate that the transcription task would be substantially harder for images collected in other genres, particularly in cases where reading order is not obvious or explicit in the image. For instance in a collection of images of complex multi-column forms that have been completed by hand, one transcriber might work from top to

bottom in column 1 then proceed to column 2, whereas another transcriber might proceed left to right (or right-to-left for Arabic) without respect to columns. It is unclear whether MTurk could be productively used for these more complex transcription tasks that typically require a customized user interface and significant annotator training.

Unsurprisingly, translation was the most difficult and least successful task, largely because of the shortage of Arabic and Spanish Turkers and the compressed timeline for translation. Still, translation of general content is a feasible task for MTurk given appropriate quality control measures. Future efforts will need to explore other options for locating appropriate Turkers. As with the transcription task, we also anticipate adding more quality control steps to the MTurk pipeline including acquisition of multiple translations with staged quality judgments, comparision and correction. We will also revisit the question of whether translation HITs should include both the transcript and the image file. While this adds complexity to the translation task, it may also help to improve the overall translation quality, and for more complex types of handwritten images translation may be impossible without reference to the image.

In future efforts we also anticipate needing to have dedicated project staff to facilitate HIT construction and approval, data processing, and interactions with either the Amazon or third party APIs. We encountered some practical challenges in this pilot study with respect management of the results across tasks. As noted earlier, naming conventions were not specified in the HIT instructions for image collection, so images had to be manually renamed to make them unique and readily identifiable by genre and language. Extracting transcription output from the results table and presenting it for the translation HIT with document formatting and character encoding intact was another challenge that requires additional exploration.

Future efforts should also revisit the cost model, using information about actual time required to complete each type of HIT. In all cases, HITs were completed in just a few minutes. We also need to further explore cost/quality tradeoffs, since high-paying tasks (like translation) are also the most prone to fraud and therefore require additional QC measures.

In conclusion, we have used MTurk to produce a small pilot corpus of handwritten, transcribed and translated images in three languages and two genres. This study has provided evidence that MTurk is a viable option for image corpus creation at least for some languages, and has suggested avenues for task refinement and future work in this area. The data collected in this study will be distributed to workshop participants, and portions will be selected for use in the NIST Open HaRT evaluation.

Disclaimer

Certain commercial products and software are identified in this paper in order to explain our research. Such identification does not imply recommendation or endorsement by NIST, nor does it imply that the products and software identified are necessarily the best available for the purpose.

References

Chris Callison-Burch. 2009. *Fast, Cheap and Creative: Evaluating Translation Quality with Amazon's Mechanical Turk*, in Proceedings of Empirical Methods in Natural Language Processing 2009.

Combating Terrorism Center at West Point, United States Military Academy. 2007. *CTC's Harmony Reports*, http://ctc.usma.edu/harmony/harmony_menu.asp (accessed March 2, 2010).

Combating Terrorism Center at West Point, United States Military Academy. 2006. *The Islamic Imagery Project: Visual Motifs in Jihadi Internet Propaganda*, Combating Terrorism Center at West Point, West Point, NY.

NIST. 2010. *NIST 2010 Open Handwriting Recognition and Translation Evaluation Plan*, http://www.nist.gov/itl/iad/mig/upload/OpenHaRT2010_EvalPlan_v2-7.pdf (accessed February 25, 2010).

University of Colorado at Boulder Office of News Services. 1998. *CU-Boulder Archives Acquires Iraqi Secret Police Files*, http://www.colorado.edu/news/releases/1998/33.html (accessed March 2, 2010).