Producing Data for Under-Resourced Languages: A Dari-English Parallel Corpus of Multi-Genre Text

Sherri Condon
The MITRE Corporation
7525 Colshire Dr.
McLean, VA
scondon@mitre.org

Luis Hernandez
Army Research Laboratory
2800 Powder Mill Rd.
Adelphi, MD 20783
luis.hernandez2@us.army.mil

Dan Parvaz, Mohammad S. Khan
The MITRE Corporation
7525 Colshire Dr.
McLean, VA
dparvaz,mskhan@mitre.org

Hazrat Jahed
Army Research Laboratory
2800 Powder Mill Rd.
Adelphi, MD 20783
ghulam.h.jahed@us.army.mil

Abstract

Text that is available for data to train machine translation engines tends to be limited to a narrow range of genres and linguistic features, such as news text. This paper describes an effort to produce a Dari-English parallel corpus containing text in a variety of styles and genres that more closely resemble the kinds of documents needed by government users than do traditional news genres. The process began with a survey of Dari documents catalogued in a repository of material obtained from Afghanistan so that documents with similar linguistic features could be selected from two university library collections. The paper also reports improvements to Dari-English translation of multi-genre text when translation systems are trained and tested using the new 253,363-word corpus.

1 Introduction

Developers producing language technology for under-resourced languages often find relatively little machine readable text for data required to train machine translation (MT) systems. Typically, the kinds of text that are most accessible for production of parallel data are news and news-related genres, yet the documents that require translation for analysts and decision-makers reflect a broad range of forms and contents. Consequently, a team from the Army Research Laboratory and The MITRE Corporation proposed to produce a parallel corpus of documents from multiple genres.

The language selected for the data development effort was Dari, the primary written language of Afghanistan. Dari is a high priority language for the United States government and especially the Department of Defense due to the ongoing military mission in that country. Dari is related to Farsi, the Persian language of Iran, but the two languages are too different for Farsi resources to be used with Dari materials without significant adaptation. Machine-readable Dari text is increasingly available from Dari language websites, but the form and content of web text is limited compared to the many different document types that occur in the field (see section 2).

For the collection described in this report, a broad range of text contents and linguistic features were obtained from library collections of Afghan documents. Section 2 describes the survey of sample Afghan documents that was conducted to identify contents and linguistic features that occur in texts that have been translated for government purposes. Section 3 describes the library collections and the documents that were acquired for the corpus. Section 4 explains the processes and guidelines developed to produce the 250,000-word parallel corpus and present some lessons learned from the endeavor.

The corpus development effort included the goal of demonstrating the usefulness of the data by showing that the performance of MT engines improves when they are trained on the additional data. A demonstration of this sort is not a simple experiment: many factors contribute to MT performance, including the training data, the language model, the tuning data, the test data, and other parameters in the process. In some respects, this is a question of domain adaptation, which is receiving increasing attention from MT researchers (Bertoldi and Federico, 2009; Civera and Juan, 2007; Koehn and Schroeder, 2007). Section 5 presents the results of some experiments conducted to examine the effects of training and tuning with the new multi-genre parallel corpus.

2 Survey of Dari Document Types

A government database of documents provided examples of the types of Dari documents that require translation. Search parameters were used to identify documents that were classified as Persian-Dari language and "electronic documents." The total number of documents returned with those features was 7572, most of which were document images in portable document format (pdf).

The database also provides classifications of document types to access specific documents. For each document type, the total number of documents associated with that type was recorded, and a sample of the documents was opened. The contents of the sample documents were recorded along with their linguistic features.

The content of Dari documents in the database is highly diverse. Figure 1 lists each document type and the number of Dari electronic documents returned for that type. The frequencies should be viewed as very rough estimates of the frequencies of those contents in the database because the classifications include errors and inconsistencies. The documents are submitted to the database from several different sources, and contributors may have different interpretations of the categories. In addition, each document may be associated with any number of classes. An example of the unreliability of the classifications in the database is the category labeled "equipment passport." Documents in this classification include a vehicle registration form, an investment license, and a school diploma.

The largest group of documents is in the "handwritten" classification, which includes any kind of document that is handwritten. The second and third largest groups of documents are "administrative publications" and "government form" respectively, which contain the same types of documents: requisitions, meeting minutes, money transfer receipts, job announcements, organization charts, certificates of recognition, and election candidate lists. The fourth most frequent document type is "address/phone book" and the fifth most frequent document type is "financial records." The later include receipts, accounting ledgers, requests for government aid, deeds, expense forms, and price lists.

The most salient linguistic feature encountered in the documents was the presence of frozen politeness expressions. These occurred in letters, including official interagency communications, witness statements, and requests/requisitions. The latter had a typical form in which the body consisted of two or three columns. The document would begin with a frozen address form such as "To the respected passport authority" followed by a salutation on a separate line consisting of محترما "respectfully." Below the salutation, the right hand column contains the request or complaint, while the left hand column is labeled "Disposition" and provides space for the addressee to indicate any decisions or actions taken with respect to the request. Often a narrow middle column is labeled "Date" so that the date of the disposition can be recorded. We rarely encountered documents with more than the right column completed.

Other frozen politeness expressions include politeness phrases that elevate the recipient and place the sender in a humbler light, such as این بنده "this servant", جنابعالی "your excellency", حضرت

500 – 900 documents	201-499 documents	101-200 documents	21 – 100 documents
Address/phone book Administrative publication Financial record Government Form Handwritten note Letter Photograph	Article Contract documentation Electronic communication (email/chat) Personal ID Study	Business card Commercial form Logbook Military inventory Military orders Miscellaneous loose paper Passport certificate Personnel roster	Book Documents list Evidence Lecture/briefing Maps Medical record Meeting documentation Operating instructions Organizational chart Parts list Personal file Poetry Religious text Technical description bulletin

0-20 documents

Biography	Handbook	Patent
Brochure	Handwriting samples	Periodical
Cargo documents	ICRC (Int'l Red Cross/ Crescent)	Police report
Catalogue	Immigration	Press release
Collection	Incoming letter	Radio frequency list
Combat Instructions	Incoming postcard	Radiogram
Conference proceedings	Interview	Range firing table
Credit bank card	Issue	Repair/maintenance manual
Diary/daily planner	Message	School records
Dictionary	Military ID	Tactical handbook
Dissertation/thesis	Miscellaneous serial	Tags/labels
Equipment guide	Outgoing letter	Technical specifications
Equipment passport	Outgoing postcard	Travel document
Facsimile		

Figure 1: Types and Frequency Ranges of Documents in the Survey

تعالی "respected sir", عرضم به حضور شما "I submit to your presence", این عسکری "this soldier", and the general strategy of having the sender refer to himself in the plural (which contrary to the English "royal We" is calculated to reduce individuality – indeed, to make the individual disappear) .

Another frequently encountered feature is text that does not occur as part of a full sentence. Many documents have tabular formats so that terms are not even part of a phrase or sentence fragment. These include candidate lists, financial and other record-keeping ledgers, price lists, rosters, expense reports, lists of martyred officers, and names from SIM cards. Many of the terms in these documents are named entities such as person names, locations, organization names, dates, monetary amounts, and radio frequencies.

Forms containing entity descriptors and entity names such as "Father: Daoud" are also common in registration and identification documents, job applications, deeds, bank receipts, and medical records. A related feature is very high proportions of named entities in all of the documents, which may be set apart from running text in letterheads and logos, as addressees in letters and requisition forms, on business cards, in notes, or on invitations.

Another context with high frequencies of named entities occurring in full sentences is letters of introduction. These typically contain biographic information about the bearer, including place of birth, address, tribe, parentage, education, and work experience. Letters of introduction are also used for commercial entities, specifying previous work and principal personnel. A similar type of letter referred to as an "investment license" certifies the bearer to trade in investment commodities and contains some biographic information as well as names of commercial entities. All of these types of letters would provide good training material for translation of named entities without the challenging formatting found in tables and forms.

Highly stylized fonts are characteristic of logos, letterheads, and business cards. There is an immense variety of font styles and sizes in the Dari documents. It is not uncommon for names and addresses on business cards and commercial or government letterheads to be presented in both Dari and English, and some documents contain

both Dari and Pashto. Most documents appear to be machine print, though many forms contain machine print descriptors (e.g., "Name:") with handwritten information. Signatures may be the only handwritten elements of a document, though thumbprints are also used in place of signatures.

Aside from frozen politeness expressions, first and second person pronouns are relatively rare in the formal documents and records, but they do occur in more personal letters, descriptions of personal experiences, and email. In formal documents, first person singular tends to be avoided or replaced with a circumlocution such as "this person." This can be compared with a letter titled "One Mujahid," which is replete with first and second person pronouns:

> *Because the government of Mr. Karzai has extended peace and reconciliation offers to the Mujahidin. As a Mujahid, I [1st singular] welcome this proposal and want to live alongside my [1st singular] brothers. But the following conditions must be met: First, release all prisoners. Second, our [1st plural] lives and the lives of our [1st plural] fellows must be guaranteed [...] If the government does not accept our [1st plural] proposal, we [1st plural], for ourselves [1st plural] are ready to follow our [1st plural] way and you [2nd plural] for yourselves [2nd plural] the government may follow yours [2nd plural]).*

A full range of tenses, aspects, and modalities occur in the documents. Promissory notes, organization planning, and scheduled events provide contexts for future time relations. Reports of meetings, status reports, witness statements, and biographic text provide contexts for past time relations and perfective aspect. Person descriptions, letters of introduction, investment licenses, instructions, and training manuals provide contexts for present time relation and timeless present references. Modal contexts occur in requests, instructions, promissory notes, letters like the one above, and invitations.

3 Selecting and Acquiring Documents

The plan for producing the corpus was to leverage two library collections of documents from Afghanistan. Documents were chosen based on the likelihood that they would contain linguistic features that are similar to features identified in the documents from the database survey. Hardcopy documents were captured using a digital camera.

3.1 Afghanistan Document Collections

One of the library collections that was used to produce the Dari corpus is an online resource created by the University of Arizona and the University of Kabul. The *Preserving and Creating Access to Unique Afghan Records* project[1] aims to catalog, digitize and create metadata for Afghan documents, focusing on the Jihad period from 1989 to 2006. The 2727 documents are primarily in Dari, Pashto, and English, though the language classification is not always reliable.

All of the document images in the University of Arizona collection are pdf and most are carefully handwritten. The variety of content is broad, but not as broad as the documents in the survey. Examples of the documents include information about Mujahidin and Communist forces fighting in Jalalabad (23 pages), accounts of the Soviet occupation of Afghanistan (134 pages, 33 pages, 139 pages), a text about the Loya Jirga in Afghanistan (66 pages), a speech by Mohammad Asmayil during the regime of the Mujahidin (36 pages), and a speech by Mujahidin fighter Azizullah Mubashir (36 pages).

The Arthur Paul Afghanistan Collection at the University of Nebraska at Omaha includes books and microfiche copies of documents. Most of the texts are Dari, Pashto, and English, but there is scholarship in other languages, too. It is not possible to search the collection according to the language of the text, but the collection's librarian kindly provided a copy of her (out of print) bibliography of the 803 Dari and Pashto holdings (Wahab, 1995).

The largest groups of documents in the Arthur Paul collection are periodicals/newspapers and texts in education, history/political science, religion and philosophy. Other content types include reference texts (bibliographies, dictionaries, encyclopedias, almanacs, and atlases), biographies, and texts in economics, language/literature, arts/music/folklore, and law.

3.2 Document Selection Criteria

Library collections are not likely to contain many of the everyday kinds of documents encountered in the survey described in section 2 (e.g., business

[1] University of Arizona. Preserving and Creating Access to Unique Afghan Records. http://184.73.243.18/about.html

cards, receipts, purchase orders, work schedules). Consequently, the criteria for selecting documents to include in the corpus focused on linguistic features and on contents which are likely to include vocabulary similar to the documents in the survey.

The linguistic generalizations that guided selection of documents are as follows:

1. References to everyday contents like those in Figure 1 may be found in fiction or biography.
2. History/political science and biography texts are likely to be rich in named entities.
3. History/political texts may contain some language of official/government documents.
4. Directives and requests may be embedded in biographies or stories.
5. Letters may appear in biographies or histories.
6. Instructions/training materials may be available for any content. Military and medical contents were selected.
7. Documents describing cultural practices may contain everyday vocabulary.
8. Descriptions of military history/practice are likely to contain military vocabulary and named entities.
9. First and second person pronouns and frozen politeness expressions may occur in speeches and fiction.
10. Documents with biographic information are likely to be rich in named entities and to resemble the biographic letters of introduction and investment licenses found in the survey.
11. Recent publications should be preferred in order to have examples of modern usage.

With these guidelines, 50 documents were selected for the corpus, and 1200 pages were selected from those documents. Figure 2 shows the number of pages of each document content that were transcribed to produce the corpus.

3.3 Image Capture

The method for acquiring document images depended on the source of the image. Most of the documents (39) were from the digital Arizona collection and were easily downloaded from the project website. However, it was necessary to decompose the multi-page pdf documents into single image files. Microfiche images were printed to pdf files by library software at a rate of about 1 minute per frame. The time required to produce each image was increased by the need to manually frame the page image by adjusting the position of the film in the microfiche reader.

Page images of books were captured using a 21.1 Megapixel Digital Single Lens Reflex AF/AE full frame camera (Canon EOS 5D Mark II) with a standard zoom lens that ranges from 24 – 105 millimeters. Camera capture has the advantage that

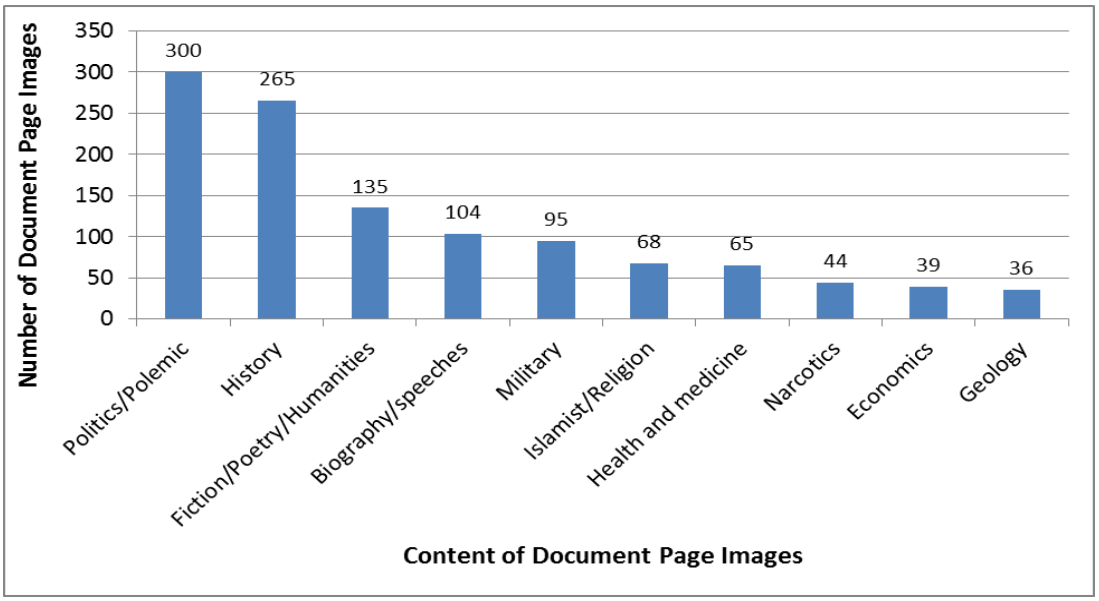

Figure 2: Content Distribution of Dari Document Page Images

delicate books need not be pressed flat, as required for scanning devices. Luis Hernandez designed an apparatus consisting of a book cradle that holds the book open at 90 degrees and a camera mounted on a tripod which is positioned parallel to the page. In addition, software on a laptop running Windows XP SP3 allowed remote camera operation and automated saving the images to files. Figure 3 illustrates the book capture apparatus.

4 Processing from Image to TMX

The document images were transformed into a parallel corpus in two steps. First, the images were transcribed into UTF-8 text documents that matched the images line by line. Transcribers were asked to transcribe the document in this way rather than directly into translation units for three reasons: (1) this method maximized the likelihood that the pages were transcribed accurately by not demanding that transcribers perform more than one task at a time, (2) matching transcription to the page image makes it easier to check accuracy for quality control, and (3) the transcriptions can be used to compute error rates for evaluations of optical character recognition accuracy, which effectively produces a second corpus from the investment.

In the second step, the transcribed text was arranged into translation units and translated. Both steps were performed by a contractor with extensive experience in data production. However, these were new tasks for the staff, and they reported that they underestimated the effort required to convert the transcribed text into sentence units that were suitable for translation.

We developed a 10-page document of transcription and translation guidelines and reviewed it with the translation contractor over a 2-month period. The transcription guidelines included instructions for transcribing complex layouts such as columns and tables. Conventions were also needed for handwritten characters that cannot be reproduced using a standard keyboard (such as ligature errors), for kashida, and for character spacing. Errors in the original document were preserved.

For translation, the guidelines specified the translation unit and translation style. Conventions were provided for English spelling, proper names, numbers, and *abjad* ordering. Errors in the transcribed text were corrected if the changes required only changes of a few characters in a word, such as spelling errors, or minimal changes in words such as insertion or deletion. Translations were produced as if the transcribed text were free of errors.

The texts were challenging for translators, which underscored the difference between the multiple genres in the library corpus and the general news genres that translators have typically contended with. Specialized vocabulary was difficult if translators were not familiar with the content, especially military and medical topics, which required the contractor to find translators who could cope with the subject areas. The breadth of content made it necessary to employ more translators than anticipated.

It was also challenging to manage the data production process. For most documents, only portions of the text were selected for transcription and translation in order to maintain the breadth of genres in the corpus. As a result, the accounting required to assign tasks to transcribers and translators and to track the data for quality control processes was time-consuming. Finally, consolidating all of the separate translation files into a single TMX file required additional effort.

5 Translation Experiments

In addition to creating the corpus, we performed some experiments to estimate the effects of adding the data to training sets that varied in size and genre.

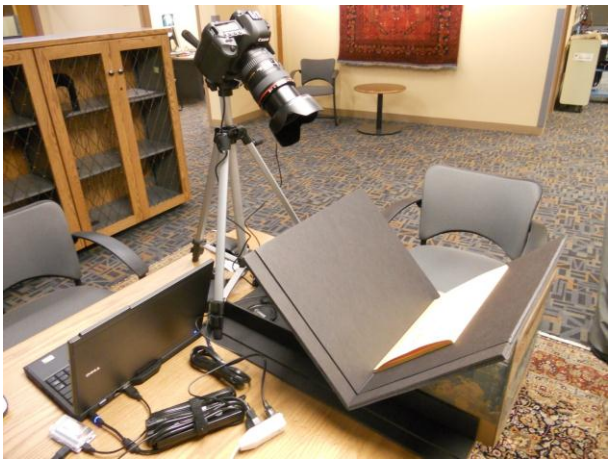

Figure 3: Book Cradle and Camera Imaging Apparatus

	Training Set		Tuning Set		Test Set	
Content	**Segments**	**Words**	**Segments**	**Words**	**Segments**	**Words**
Multi-genre	8725	193,347	1,351	35,275	930	24,741
News	40,514	691,675	1,284	20,928	1,684	47,796
Military	28,915	355,194	1,648	24056	2,786	43,289
Medical	15,978	103,588	1,237	15,147	4,482	68,894
Legal	16,623	264,357	1,736	41,966	1,056	21,724
Total (without multi-genre)	102,030	1,414,814				
Total	110,755	1,608,161				

Table 1: Quantities of Training, Tuning, and Test Data for Corpora Used in the Experiments

5.1 Data for Translation Experiments

The 253,363 word multi-genre corpus was divided into three datasets for training, tuning, and testing. Table 1 presents the quantities of data in each set of the multi-genre corpus as well as the quantities of data in several other corpora used for the experiments. The "military" corpus is from field manuals that were translated into Dari for the U.S. Army. The "medical" corpus is a training manual, *Fundamental Critical Care Support (FCCS)*, published by the Society for Critical Care Medicine and used as a textbook for training doctors and nurses for work in Intensive Care Units. The "legal" corpus is from *An Introduction to the Law of Afghanistan*, published by the Afghanistan Legal Education Project at Stanford Law School. Most of the text in the "news" corpus is from the archives of Sada-e-Azadi, which is a trilingual (English, Dari, and Pashto) news website sponsored by the International Security Assistance Force, which is the NATO command in Afghanistan[2].

Several pre-processing operations were applied to the Dari text in order to reduce orthographic variation. Unicode mappings were normalized so that each character corresponded to a single code point. Pashto and other non-Dari characters were replaced with the equivalent Dari characters, and diacritics were removed. In addition, a customized tokenizer was used to tokenize the Dari text.

5.2 Evaluation Procedure

For each combination of training, tuning, and test data, the experimental procedure began by training the MT engine using the training set. The open source Joshua 4.0 toolkit was used to build each hierarchical, phrase-based machine translation engine (Li et al., 2009; Ganitkevitch et al., 2012)[3]. Aside from the Dari tokenizer, default settings and resources in the Joshua toolkit were used: word alignment was performed by the Berkeley Word Aligner (Haghighi et al., 2009)[4] and the English language model was built using KenLM (Heafield, 2011)[5]. No additional monolingual data were used to train the language model. Tuning was performed using Z-MERT (Zaidan, 2009)[6], and for most of the results presented here, each MT engine was tuned 3 times in order to observe the variability of the results of Z-MERT's maximization operations.

Tuning and testing were performed with the Bilingual Evaluation Understudy (BLEU) metric (Papenini et al., 2002) using the BLEU scoring function provided in Joshua. Scores were computed with a single reference translation.

5.3 Results of the Experiments

For an initial set of comparisons, the effects of adding the multi-genre corpus to the set of all training corpora were observed for each domain. The first comparison in Table 2 is an example: all corpora described in Table 1 except for the new multi-genre corpus were used to train the baseline MT engine, and this was compared to training with all the corpora including the multi-genre set. Both were tuned with the tuning set from the news corpus and tested with the test set from the news

[2] http://www.sada-e-azadi.net/

[3] Download: http://joshua-decoder.org/ More documentation: http://www.cs.jhu.edu/~ccb/joshua/index.html

[4] http://code.google.com/p/berkeleyaligner/

[5] http://kheafield.com/code/kenlm/

[6] http://cs.jhu.edu/~ozaidan/zmert.)

corpus. Therefore, the comparison demonstrates little more than increasing the size of the training set by about 9%. Similar comparisons were performed using the military, medical, and legal tuning and test sets with similar results: the BLEU scores for engines built with the multi-genre corpus were higher than those without the new corpus but the difference was less than 0.01.

For the remaining comparisons, the multi-genre test set was used, reflecting the use case in which documents with a variety of contents are translated by the MT system. Table 2 demonstrates the results of these experiments. In one case, we supposed that the MT engine of a low resource language has been trained and tuned on a relatively small amount of news data, but is used to translate multi-genre documents like those represented in the multi-genre test set. When the MT system is trained and tuned on the news data only, testing with the multi-genre test set results in a BLEU score (average) of 0.096. Adding the multi-genre data to the training set, but still tuning the engine on news data increases the score by nearly 0.044 to almost 0.14 (see Comparison 2 in Table 2). This experiment demonstrates the significant impact that the new multi-genre corpus can have for a low-resource language.

Of course, the MT system scores higher on translations of the multi-genre test set when tuning is performed using the multi-.genre tuning set. If the news training set is tuned with the multi-genre tuning set, the BLEU scores average 0.109. In comparison, when the engine is trained on news data plus the multi-genre corpus, then tuned on the multi-genre tuning set, the average BLEU score is 0.143 (see Comparison 3 in Table 2). Although the scores are higher when the engine is tuned on the multi-genre tuning set, they do not increase as much as they increase when the multi-genre training data is tuned to the news data. The multi-genre training set is about one fourth the size of the news training set, so it is not surprising that adding the additional data significantly increases the scores. For the next set of comparisons, we used the training sets from all of the domains in order to examine the case in which the engine is trained with much more data.

For the experiments using all of the available training data, we investigated two tuning options. In one set of comparisons, the engine was tuned using the news tuning corpus, and in the other set of comparisons, we sought to maximize performance on the multi-genre test set by tuning with the multi-genre tuning set. The BLEU scores on the multi-genre test set for the engine that was trained with all of the corpora except the multi-genre training data and tuned with the news tuning set average 0.112, which is 0.016 higher than the

Comparison	Training Data	Tuning Data	Test Data	BLEU Scores*			
				Trial 1	Trial 2	Trial 3	Mean
1	All previous corpora	News	News	0.1210	0.1243	0.1231	0.1228
	All + multi-genre	**News**	**News**	**0.1318**	**0.1319**	**0.1281**	**0.1306**
2	News corpus	News	Multi-genre	0.0954	0.0983	0.0934	0.0957
	News + multi-genre	**News**	**Multi-genre**	**0.1431**	**0.1391**	**0.1374**	**0.1399**
3	News corpus	Multi-genre	Multi-genre	0.1082	0.1074	0.1103	0.1086
	News + multi-genre	**Multi-genre**	**Multi-genre**	**0.1416**	**0.1423**	**0.1453**	**0.1431**
4	All previous corpora	News	Multi-genre	0.1147	0.1066	0.1134	0.1116
	All + multi-genre	**News**	**Multi-genre**	**0.1458**	**0.1415**	**0.1437**	**0.1437**
5	All previous corpora	Multi-genre	Multi-genre	0.1252	0.1282	0.1255	0.1263
	All + multi-genre	**Multi-genre**	**Multi-genre**	**0.1472**	**0.1473**	**0.1475**	**0.1473**

*For each trial, the MT engine was re-tuned in order to demonstrate the variability of Z-MERT tuning

Table 2: BLEU Scores Achieved with Different Training and Tuning Sets Showing Effects of the Multi-Genre Data

scores for the MT system that was trained and tuned with only the news data. This demonstrates the effect of greatly increasing the size of the training data set. However, when the multi-genre training set is added to the training data and tuning is still performed using only the news tuning set, the BLEU scores for the multi-genre test set increase even more to 0.144, which is more than 0.03 higher (see Comparison 4 in Table 2)). Although the multi-genre training set is only about one tenth the size of the other corpora combined, adding the data resulted in a significant increase in the BLEU score on the multi-genre test set.

Finally, the maximum performances on the multi-genre test were obtained by training with the combined corpora and tuning with the multi-genre tuning set. When the training corpora did not include the multi-genre training corpus, BLEU scores averaged 0.126, and when the training corpora included the multi-genre training corpus, the scores averaged 0.147, which is an increase of more than 0.02 (see Comparison 5 in Table 2). Although this score is the highest average obtained on the multi-genre test set in all of the experiments, it is only 0.003 – 0.004 higher than the average scores obtained in two other experiments reported in Table 2. When the training set included all of the corpora and the engine was tuned with the news tuning corpus, the average score was 0.144 (see Comparison 4 in Table 2), which suggests that for the multi-genre test set, the effects of the tuning corpus were relatively small if the system was trained with a large set of combined corpora from multiple domains.

It is tempting to suggest that this result might be due to the fact that the combined training set included multiple domains, making it more similar to the multi-genre test set. However, the other high score suggests that this is not the case: when the engine was trained only on news data plus the multi-genre training set, then tuned on the multi-genre tuning set, nearly the same average BLEU score of 0.143 was obtained (see Comparison 3 in Table 2). This result suggests that news corpora are actually the best foundation on which to add the multi-genre training set, even though doubling the size of the training set by using all of the corpora may provide a slight increase in performance.

6 Conclusions and Further Work

The research reported here describes the process of building a multi-genre data set for the low-resource Dari language. The effort required to obtain document page images from hard copy texts and to produce a parallel corpus from document page images is recorded. The enterprise began with a survey of document images that had been collected in Afghanistan, and the linguistic features of those documents were described along with the content types of the multi-genre corpus.

Given the effort required to produce the corpus, it was informative to examine the value of the corpus by investigating the effects of building MT engines using the additional multi-genre data. Significant improvements in the performance of the MT engines were obtained when the engines were tested using a test set that had been held out from the multi-genre corpus. However, we believe that the most important test of the value of the corpus remains to be performed. We are currently producing a test set using document images drawn from the aforementioned database of documents that were collected in Afghanistan. Because those are the documents that have genuine operational significance, testing with those documents is needed to truly judge the success of the corpus creation effort.

Acknowledgments

We would like to thank the personnel of University of Nebraska's Arthur Paul Afghanistan Collection for allowing us to access the collection and produce the camera images in their facility. We especially thank Professor Shaista Wahab, curator of the Arthur Paul collection for all of the time and assistance that she provided. We would also like to express our appreciation to Michelle Vanni, Melissa Holland, and Steve LaRocca in the Multilingual Computing Branch of the Army Research Laboratory for their support and assistance in the creation of this resource.

The research reported in this document was performed by the U.S. Army Research Laboratory in collaboration with MITRE Corporation. The views and conclusions contained in this document are those of the authors and should not be interpreted as presenting the official policies or

position, whether expressed or implied, of the U.S. Army Research Laboratory or the U.S. Government unless so designated by other documents. Citation of manufacturers or trade names does not constitute an official endorsement or approval of the use thereof.

References

Nicola Bertoldi, Marcello Federico. 2009. Domain Adaptation for Statistical Machine Translation with Monolingual Resources Proceedings of the Fourth Workshop on Statistical Machine Translation , pages 182–189.

Jorge Civera and Alfons Juan. 2007. Domain adaptation in statistical machine translation with mixture modeling. In Proceedings of the Second Workshop on Statistical Machine Translation, pages 177–180.

Juri Ganitkevitch, Yuan Cao, Jonathan Weese, Matt Post, and Chris Callison-Burch, 2012. Joshua 4.0: Packing, PRO, and Paraphrases. In Proceedings of the Seventh Workshop on Statistical Machine Translation.

Aria Haghighi, John Blitzer, John DeNero, and Dan Klein. 2009. Better word alignments with supervised ITG models. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2, ACL '09, pages 923–931.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In Proceedings of the Sixth Workshop on Statistical Machine Translation.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In Proceedings of the Second Workshop on Statistical Machine Translation, pages 224–227.

Zhifei Li, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Sanjeev Khudanpur, Lane Schwartz, Wren N. G. Thornton, Jonathan Weese and Omar F. Zaidan. 2009. Joshua: An Open Source Toolkit for Parsing-based Machine Translation. In Proceedings of the Workshop on Statistical Machine Translation (WMT09).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In Proceedings of ACL 2002, pp. 311-318.

Shaista Wahab. 1995. Arthur Paul Afghanistan Collection Bibliography. Volume I: Pashto and Dari Titles. Omaha: University of Nebraska.

Omar F. Zaidan. 2009. Z-MERT: A Fully Configurable Open Source Tool for Minimum Error Rate Training of Machine Translation Systems. The Prague Bulletin of Mathematical Linguistics, No. 91:79-88.