

*ARMY RESEARCH LABORATORY*



## **Introduction of Automation for the Production of Bilingual, Parallel-aligned Text**

**by Will Tanenbaum, Steve LaRocca, John Morgan,  
and Ghulam Hazrat Jahed**

**ARL-TR-5798**

**October 2011**

## **NOTICES**

### **Disclaimers**

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

# **Army Research Laboratory**

Adelphi, MD 20783-1197

---

---

**ARL-TR-5798**

**October 2011**

---

## **Introduction of Automation for the Production of Bilingual, Parallel-aligned Text**

**Will Tanenbaum, Steve LaRocca, John Morgan,  
and Ghulam Hazrat Jahed  
Computational and Information Sciences Directorate, ARL**

# REPORT DOCUMENTATION PAGE

*Form Approved*  
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

**PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE (DD-MM-YYYY) October 2011		2. REPORT TYPE		3. DATES COVERED (From - To)	
4. TITLE AND SUBTITLE Introduction of Automation for the Production of Bilingual, Parallel-aligned Text				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Will Tanenbaum, Steve LaRocca, John Morgan, and Ghulam Hazrat Jahed				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) U.S. Army Research Laboratory ATTN: RDRL-CII-T 2800 Powder Mill Road Adelphi, MD 20783-1197				8. PERFORMING ORGANIZATION REPORT NUMBER  ARL-TR-5798	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited.					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT As the study and application of statistical machine translation (SMT) grows, progress is often circumscribed by a lack of data. The statistical models that govern SMT engines rely on many large bilingual text corpora, each comprised of vast numbers of bilingual text segments. For certain languages, corpora already exist and help to power translation engines. Regrettably, this is not the case for every language the Army is interested in, making the creation or acquisition of such data a priority. To this end, a language expert in Dari and Pashto was hired, who collected, prepared, and ensured the quality of bilingual text. To explore ways to aid the expert, a variety of the steps performed by the expert and necessary to the process were automated. The hypothesis was that automation of selected processes would improve efficiency, measured in terms of both speed of production, and quantity of data produced, even when time to correct automation-caused errors was accounted for. As predicted, the net result of introducing automation was an increase in both the rate of producing correct bilingual segments and the number produced. The implications of these results for improving larger bilingual data creation and acquisition efforts are discussed.					
15. SUBJECT TERMS Statistical machine translation, data mining, automation					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT  UU	18. NUMBER OF PAGES  18	19a. NAME OF RESPONSIBLE PERSON Will Tanenbaum
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified			19b. TELEPHONE NUMBER (Include area code) (301) 394-5627

---

## Contents

---

<b>List of Figures</b>	<b>iv</b>
<b>Acknowledgments</b>	<b>v</b>
<b>1. Introduction and Background</b>	<b>1</b>
1.1 The Need for Bilingual Data .....	1
1.2 The Production of Bilingual Data for SMT.....	1
<b>2. Pipelines and Experiments</b>	<b>3</b>
2.1 The Original Pipeline .....	3
2.1.1 Initial Harvest.....	3
2.1.2 Working Storage.....	3
2.1.3 Extraction: Tag Stripping .....	3
2.1.4 Return to Storage.....	4
2.1.5 Text Clean-up.....	4
2.1.6 Segmentation and Alignment.....	4
2.2 The Enhanced Pipeline.....	4
2.2.1 Punkt Pickle Segmentation.....	4
2.2.2 Bilingual Sentence Aligner Alignment .....	4
2.3 The Experimental Procedure .....	5
<b>3. Results and Discussion</b>	<b>5</b>
<b>4. Summary and Conclusions</b>	<b>6</b>
<b>5. Software</b>	<b>7</b>
<b>List of Symbols, Abbreviations, and Acronyms</b>	<b>8</b>
<b>Distribution List</b>	<b>9</b>

---

## List of Figures

---

Figure 1. The pipeline process. ....	2
--------------------------------------	---

---

## **Acknowledgments**

---

The Pipeline project was conceived and directed by Dr. Stephen A. LaRocca of the Multilingual Computing Branch (MLCB). The Pipeline, itself, was originally developed by Dr. Mark Arehart of the MITRE Corporation for the U.S. Army Research Laboratory (ARL), with support from Dr. Sherri Condon of MITRE. Subsequent additions to the Pipeline were made by John J. Morgan and the author. The author would like to thank Dr. LaRocca and Mr. Morgan for their direction and encouragement, as well as Dr. Melissa Holland and Dr. Michelle Vanni for editing this publication. Thanks also to Ghulam Hazrat Jahed for serving as both language expert and translator. My sincere thanks to all others in MLCB, without whom this experiment could not have been conducted.

INTENTIONALLY LEFT BLANK.

---

# 1. Introduction and Background

---

Computational linguistics, as a discipline, centers on enhancing the capability of computers in translating one natural language into another, which is called machine translation (MT). Originally, it was assumed that MT would be as simple as compiling a multilingual lexicon; however, such methods met with only limited success. Today, MT relies on the ever-increasing capacity of computers to ingest and learn from large amounts of bilingual data from human translators, or ground truth data. This method, statistical machine translation (SMT), models patterns of translation by assigning weighted probabilities to bilingual correspondences derived from ground truth data.

## 1.1 The Need for Bilingual Data

To build an SMT engine, huge amounts of bilingual data are required, both to serve as ground truth and to properly calibrate the governing heuristics; more data generally serves to make an SMT engine more accurate. For some languages, bilingual data that pairs English with another language are numerous and easy to come by. These languages include French, German, and Spanish, all of which are Indo-European. Some non-Indo-European languages, like Japanese and Mandarin Chinese, show increasing amounts of bilingual data with English. However, there is little available, bilingual data between English and languages of remote, less developed regions of the world—Afghanistan (Dari and Pashto), for instance. Yet it is regions like this where the Army goes and, hence, where it needs MT. To compensate for the lack of data for building SMT, the Army has invested resources in the production of bilingual text-data for Dari-English and Pashto-English.

## 1.2 The Production of Bilingual Data for SMT

The production of high-quality, bilingual text that is usable for building SMT is long and involved. The primary processes are shown in figure 1.

**Finding Parallel Text** It begins with locating a source of text-data written with fairly equivalent versions in both English and the target language, (as opposed to, say, a Dari text and an English précis); this is called parallel text.

**Correction and Normalization** After location, the data must be copied into a word-processor so that a language expert may correct any errors. The expert must employ a standard so as to ensure uniform stylistic and character formats.

**Segmentation** Then, the expert must divide the text into segments small enough to be useful to an SMT engine. Again, a general standard must be adopted and applied to every sentence. This can present a particular challenge in the alignment phase if the sentences of the source language and target language texts do not enjoy a one-to-one correspondence.

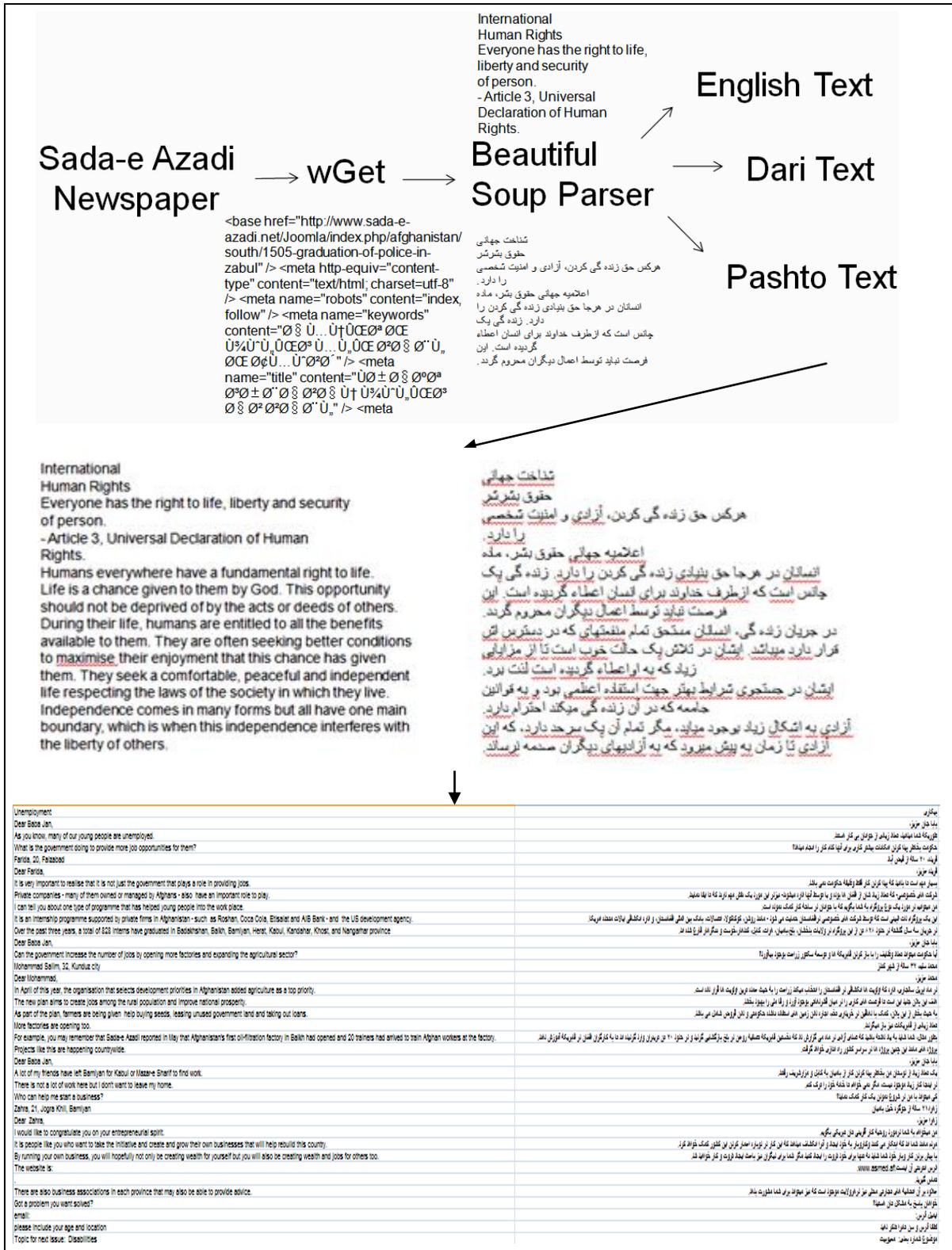


Figure 1. The pipeline process.

**Alignment Formatting** After that, the segments of one language must be aligned with those of the other. The aligned text is must then be converted into Translation Memory eXchange (TMX) format to be compiled into a cohesive corpus. TMX data constitutes a pair of aligned segments and a probability.

This entire procedure is painstakingly carried out by a single language expert. Understandably, this method is not conducive to maximum output from the expert. To optimize the use of the Army's time and resources, it was decided that automation should be introduced to aid the expert whenever appropriate.

---

## **2. Pipelines and Experiments**

---

We refer to the series of automated processes, designed to increase the amount of bilingual parallel \*.tmx text produced and authenticated by the language expert, as the *Pipeline*. Originally, the Pipeline featured a mix of open-source programs, as well as program code written by Mark Arehart (MITRE), and it was tailored to extract data only from a single source—the online newspaper, *Sada-e Azadi*. An enhanced version of the Pipeline incorporated additional open-source content, in addition to program code written by John Morgan and Will Tanenbaum.

### **2.1 The Original Pipeline**

The web site, Sada-e Azadi, displays the online version of an International Security Assistance Force (ISAF) publication. Its features include topics like current events, politics, economy, media, entertainment, and health. It also offers a question-and-answer column, called “Baba Jan.” Each article is offered in three languages: Dari, English, and Pashto.

#### **2.1.1 Initial Harvest**

To acquire this data, we used the open-source program, *Wget (1)*, which downloads html-annotated text from web sites.

#### **2.1.2 Working Storage**

The Pipeline then accessed a routine to write the html-annotated text of each Sada-e Azadi article accessed to one of three directories, based on the natural language of the article.

#### **2.1.3 Extraction: Tag Stripping**

From there, the Pipeline passed control to the open-source parser, *Beautiful Soup (2)*, which filtered out the html annotation, as well as any non-text data, such as pictures, sound files, videos, and links.

#### **2.1.4 Return to Storage**

The Pipeline then saved the text from each article, in block paragraph form, as a text data file in the appropriate language directory.

#### **2.1.5 Text Clean-up**

Prior to segmentation, the Pipeline accesses a routine for adjusting selected properties, especially punctuation, of the resulting text. This routine standardizes certain language-specific characters, such as the Arabic period and colon, and replaces stylized smart-quotation marks with the generic version.

#### **2.1.6 Segmentation and Alignment**

Finally, the Pipeline accessed a sentence-level segmenter, which output the text in a format acceptable as input to SMT engines. When applied to both sides of the parallel corpus data, the segmentation step contributed to the alignment process, which matched segments by order of occurrence, before the paired segments were presented to the language expert for data quality control processes.

### **2.2 The Enhanced Pipeline**

The original Pipeline featured a fairly simple segmenting algorithm based on end-of-sentence punctuation. Essentially, the segmenter would create a segment break at periods or similar end-of-sentence symbols, such as question marks. Despite efficient handling of notable exceptions to this rule, particularly title abbreviations (Dr., Mr., Ms., etc.), this was still a fairly inelegant solution to the problem. Sentences neither always contained the exact same information from language to language, nor were they of consistent length.

#### **2.2.1 Punkt Pickle Segmentation**

To better segment the text, the *Natural Language ToolKit (3)* (NLTK), particularly the *Punkt* tool, was used. When calibrated using a critical mass of data in a given language, Punkt generates a language-specific segmenter called a *Pickle*. The designers of Punkt included 12 Pickles with the NLTK, including one for English. A Dari Pickle was generated using a corpus to which the Army already had access. Thus far, a Pashto Pickle has yet to be created due to an insufficient volume of ground-truth data. The segments created by the Pickles tended to be far more sensible and coherent than those created by the original segmenter.

#### **2.2.2 Bilingual Sentence Aligner Alignment**

An open-source Perl script called the *Bilingual Sentence Aligner (4)* (BSA) was employed to align the data in parallel segments. The BSA's accuracy improves with ever greater volumes of data. Thus, given the large number of bilingual data segments produced from the first six steps in the Pipeline, we expected improved alignment accuracy and a speeding up of the language expert's data quality control process.

### **2.3 The Experimental Procedure**

Admittedly, the changes in processes constituting both the Original Pipeline (OP) and the Enhanced Pipeline (EP) introduce new opportunities for error. Nonetheless, for the first experiment, it was hypothesized that the time saved with automated Harvesting, alone, would more than compensate for any additional errors. Moreover, to confirm that decreases in time and increases in efficiency were caused by automation in Traditional (T) versus OP data preparation, and by Segmentation and Alignment improvements in OP versus EP data preparation, an experimental framework was designed to observe the time and efficiency of work performed under the three conditions—T, OP, and EP—and a second experiment (OP versus EP) was also conducted.

To compare efficiencies of work performance under the three conditions, 10 articles were selected from the Sada-e Azadi web site. The articles were of similar length, about 25 lines, plus or minus one line. Four articles were about politics, four involved health, and two discussed media and entertainment. Five articles were randomly chosen for each version of the Pipeline: two from politics, two from health, and one about media and entertainment. Both versions of the Pipeline featured the same Harvest and Extraction processes; they first diverged at the Segmentation stage. As noted in previous sections, the OP used a basic segmenter and aligned each language's segments solely according to the order in which they appeared. The EP used Punkt and the BSA for Segmentation and Alignment. To determine which Pipeline's use effected greater efficiencies, the language expert was timed from start to finish in the performance of his data quality control work of aligning and correcting any mistakes in the bilingual parallel versions of the 10 articles, five of which were processed by OP and five by EP. The versions were alternated to mitigate possible learning effects based on order of presentation.

---

## **3. Results and Discussion**

---

The value of automation was confirmed by the results of both experiments. As suspected, without any automation, the expert harvested and aligned only a tiny proportion of what he aligned in the same amount of time with automated assistance. The difference in volume of data aligned was immense, such that it invalidated any need for statistical comparison. In comparing the two versions of the Pipeline, the differences were also quite pronounced. We calculated the mean time required by the language expert to correct the paragraphs processed by the OP and the EP. The times, in minutes, were 24.036 and 1.978, respectively. A t-test found a statistically significant difference between groups,  $t = 7.257$  with 4 degrees of freedom ( $P = .002$ ).

While not constituting empirical evidence, the subjective opinion of the expert about the two versions of the Pipeline was, nevertheless, solicited and validated our experimental results. He found a marked increase in difficulty when attempting to reconcile the lines produced with the

OP, when compared with that found using the EP, due to spurious Segmentation and Alignment. It was also his opinion that the OP performed drastically worse on articles with a relatively greater number of sentences.

---

#### **4. Summary and Conclusions**

---

The results of the experiments indicate an undeniable advantage using automation for harvesting and processing bilingual parallel text data. Whereas full automation is not yet feasible, the addition of automated tools has proven an invaluable aid for language experts. A marked increase in efficiency, similar to that gained through use of the OP and the EP, can lead to a comparable growth in SMT capability.

---

## 5. Software

---

1. Scrivano, Giuseppe; Nikšić, Hrvoje. The GNU Project (Version 1.12) [Software]. Available from <http://ftp.gnu.org/gnu/wget/>, 2009.
2. Richardson, Leonard. Crummy (Version 3.1.0.1) [Software]. Available from <http://www.crummy.com/software/BeautifulSoup/#Download>, 2009.
3. Willy; Bird, Steven; Loper, Edward; Nothman, Joel. Natural Language Toolkit (Version 2.0) [Software]. Available from <http://www.nltk.org/download> Algorithm: Kiss, Tibor and Strunk, Jan (2006): Unsupervised Multilingual Sentence Boundary Detection. Computational Linguistics 32: 485–525, 2006.
4. Moore, Robert. Microsoft (Version 1.0) [Software]. Available from <http://research.microsoft.com/en-us/downloads/aafd5dcf-4dcc-49b2-8a22-f7055113e656/>, 2003.

---

## List of Symbols, Abbreviations, and Acronyms

---

ARL	U.S. Army Research Laboratory
BSA	Bilingual Sentence Aligner
EP	Enhanced Pipeline
ISAF	International Security Assistance Force
MLCB	Multilingual Computing Branch
MT	machine translation
NLTK	Natural Language ToolKit
OP	Original Pipeline
SMT	statistical machine translation
T	Traditional
TMX	Translation Memory eXchange

NO. OF COPIES	OGRANIZATION
1 ELEC	ADMNSTR DEFNS TECHL INFO CTR ATTN DTIC OCP 8725 JOHN J KINGMAN RD STE 0944 FT BELVOIR VA 22060-6218
1 CD	OFC OF THE SECY OF DEFNS ATTN ODDRE (R&AT) THE PENTAGON WASHINGTON DC 20301-3080
1	US ARMY RSRCH DEV AND ENGRG CMND ARMAMENT RSRCH DEV & ENGRG CTR ARMAMENT ENGRG & TECHNLOGY CTR ATTN AMSRD AAR AEF T J MATTS BLDG 305 ABERDEEN PROVING GROUND MD 21005-5001
1	US ARMY INFO SYS ENGRG CMND ATTN AMSEL IE TD A RIVERA FT HUACHUCA AZ 85613-5300
1	COMMANDER US ARMY RDECOM ATTN AMSRD AMR W C MCCORKLE 5400 FOWLER RD REDSTONE ARSENAL AL 35898-5000
1	US GOVERNMENT PRINT OFF DEPOSITORY RECEIVING SECTION ATTN MAIL STOP IDAD J TATE 732 NORTH CAPITOL ST NW WASHINGTON DC 20402
6	US ARMY RSRCH LAB ATTN IMNE ALC HRR MAIL & RECORDS MGMT ATTN RDRL CII T W TANENBAUM ATTN RDRL CII T J J MORGAN ATTN RDRL CII T S LAROCCA ATTN RDRL CIO LL TECHL LIB ATTN RDRL CIO MT TECHL PUB ADELPHI MD 20783-1197

INTENTIONALLY LEFT BLANK.