New language resources for the Pashto language

Djamel Mostefa [1] **, Khalid Choukri** [1] **, Sylvie Brunessaux** [2] **, Karim Boudahmane** [3]

[1] Evaluation and Language resources Distribution Agency, France
[2] CASSIDIAN, France
[3] Direction Générale de l'Armement, France

E-mail: mostefa@elda.org, choukri@elda.org, sylvie.brunessaux@cassidian.com, karim.boudahmane@dga.defense.gouv.fr

Abstract

This paper reports on the development of new language resources for the Pashto language, a very low-resource language spoken in Afghanistan and Pakistan. In the scope of a multilingual data collection project, three large corpora are collected for Pashto. Firstly a monolingual text corpus of 100 million words is produced. Secondly a 100 hours speech database is recorded and manually transcribed. Finally a bilingual Pashto-French parallel corpus of around 2 million is produced by translating Pashto texts into French. These resources will be used to develop Human Language Technology systems for Pashto with a special focus on Machine Translation.

Keywords: Pashto, low-resource language, speech corpus, monolingual and multilingual text corpora, web crawling.

1. Introduction

There are very few corpora and Human Language Technology (HLT) services available for Pashto. No language resources for Pashto can be found in the catalogues of LDC[1] and ELRA[2].

Pashto is a very low-resource language. Google doesn't support Pashto in its search engine or translation services. Microsoft doesn't provide a language pack for Pashto on Windows 7 or Vista operating systems. There is no style file and font for Pashto in LaTeX, which makes impossible to write this article with LaTeX. In the scope of a multilingual data collection project, three large corpora are being collected for Pashto. Firstly a monolingual text corpus of 100 million words is collected by identifying, negotiating, crawling and cleaning Pashto websites. Secondly a speech database of 100 hours is recorded and orthographically transcribed. Finally a bilingual Pashto-French parallel corpus of around 2 million words is produced by translating Pashto texts into French. The project started in August 2011 and we expect to complete the data collections by October 2012. These resources will be used to develop HLT systems for Pashto with a special focus on Machine Translation.

2. The Pashto language

Pashto is an indo-iranian language spoken by the Pashtun people mainly in Pakistan (Khyber Pakhtunkhwa and Balochistan regions) and Afghanistan (east, south and southwest of the country). It is also spoken by the Pahstun diaspora around the world with significant population in United Arab Emirates, Iran, United Kingdom, Canada, India, United States, Malaysia and Singapore.

It is one of the two official languages of Afghanistan (the other one being Dari) and one regional language in Pakistan.

The code assigned to the language by the ISO 639-3 standard is [pus].

According to the Ethnologue.com website, it is spoken by around 20 million people and three main dialects are to be considered:

- **Northern Pashto.** Spoken by 9 million people in Pakistan (Afghanistan border, Khyber Pakhtunkhwa province) and Afghanistan (Central Ghilzai area). ISO 639-3 code: [pbu].
- **Central Pashto.** Spoken by 8 million speakers in Southern Pakistan (Wazirstan, Bannu, Karak and adjacent regions). ISO 639-3 code: [pst].
- **Southern Pashto.** Spoken by 2,6 million speakers in Pakistan (Balochistan and Quetta area) and in Afghanistan (Kandahar area, Afghanistan border east of Qa'en). ISO 639-3 code: [pbt].

Pashto (**پښتو**) can be transliterated with many variants such as Pakhto, Pushto, Pukhto, Pashtu, Pushtu, Pukhtu, etc. The variation in spelling reflects the different pronunciations in different regions (first vowel pronounced as *a* or *u*, the fricative pronounced as *sh* or *kh* and the final vowel pronounced as *o* or *u*).

There are many difficulties when dealing with Pashto.

Like Arabic, it is written cursively right-to-left but includes Arabic numbers written left-to-right. Letters have different shapes depending on their position in a word (begin, middle, end or isolated) and there are significant differences in the use of shared Arabic letters between Pashto and Arabic. Some vowels are missing and therefore Pashto texts are phonologically underspecified.

The main issue of the Pashto writing system is the lack of standard orthography (Kathol 2005).

Two different writing systems coexist for Pashto. The first one, called Yousufzaï, is used in Pashto regions of Pakistan and is influenced by Urdu and English.

[1] Linguistic Data Consortium www.ldc.upenn.edu
[2] European Language Resources Association www.elra.info

The second one, called Afghani Pashto, is used in Afghanistan and is influenced by Dari. There is a great difference between the vocabularies of Yousufzaï and Afghani Pashto. A common Yousufzaï Pashto mother-tongue speaker would not know the exact meaning of an Afghani-specific Pashto word in isolation, Even if there are established dictionaries and conventions for orthography, one word can be spelled out in different ways as depicted in Table 1 and Table 2. Moreover different words can be spelled out in the same way (homographs).

Word	Romanization	Variant 1	Variant 2
Blackbird	/xaro/	خارو	ښارو
Breeze	/vagma/	وړمه	وګمه
Our	/zmung/	ځمونږ	زمونږ

Table 1Spelling variations in Pashto for the words Blackbird, Breeze and Our.

English	Yousufzaï	Afghani
World	دنيا	نړۍ
Country	ملک	هيواد
Decision	فيصله	پريکړه
Support	ملګرتيا	مرسته
Patrolling	ګشت	ګزمه
Airport	ائرپورټ	هوائي ډګر
Airplane	جهاز	الوتکه
Missile	ميزائل	توغندي
Refrigerator	فريج	يخچال
University	يونيورسټي	پوهنتون
Professor	پروفيسر	پوهاند
Culture	کلچر	کلتور

Table 2 Lexical variants between Yousufzai and Afghani Pashto

Word tokenization is an issue in Pashto. As for other languages like Urdu space omission and insertion errors might occur between words and inside words. Words are made of ligatures which are sequences of joined characters. A ligature is ended by either a non joining character or a space. Space is not mandatory to separate two consecutive words. Word boundaries are identified in a text by the reader. Space is used to get appropriate character shapes and thus it may even be used within a word to break the word into constituent ligatures (Akram 2010), (Durrani 2010). Other issues of producing language resources for Pashto are depicted in section 4.1.

3. Monolingual text corpus

The aim of this task is to collect a 100 million text corpus for the Pashto language. The corpus can then be used for language modelling in automatic speech recognition, machine translation or other natural language processing. Several steps have to be implemented to collect the data and make it usable for research and development. The first step is to identify the sources from which the text data can be collected. Then, once the sources have been identified, the negotiation with the copyright owners has to be carried out. Once the data have been crawled, they must be cleaned and formatted properly.

3.1 Identification

The first phase to produce the monolingual text corpus is to identify the sources from which the text material can be collected. Since the beginning of the project in August 2011, we identified several thousands of URLs with Pashto text material including websites, blogs, SMS and forums.

Important sources are international media who publish Pashto news on a daily basis. These sources are Voice of America (VOA), British Broadcasting Corporation (BBC), Deutsche Welle (DW), China Radio International (CRI), Voice of Russia (VOR) and Turkish Radio Television (TRT). More important text materials are available in a number of Afghan and Pakistan websites. We list here the most important sources we have crawled and processed so far. The size in terms of word tokens are given for each site between parenthesis and are calculated after the data have been cleaned and formatted properly in XML.

- www.tolafghan.com (11M words)
- www.baheer.com (10M words)
- www.rohi.af (8M words)
- www.benawa.com (5M words)
- www.sporghay.com (3 M words)
- www.spenghar.com (3 M words)
- www.taand.com (2.5 M words)

At the beginning of the project, we felt that it might be difficult to collect 100 million words of Pashto from the web. The Internet penetration is quite low in Afghanistan and Pakistan (3% of the total population in Afghanistan and 10% in Pakistan)[3] and we thought there might not be enough available data in Pashto on the web. But after a few weeks of work on the collection of the texts from the web, we were surprised that there are sufficient data nowadays to constitute technically a 100 million words corpus for Pashto.

Nevertheless, if it's not the case or if we fail to obtain the rights from copyright owners to use the data from sufficient number of web sources, we envisage, as a fallback solution, to use Pashto texts from other media such as archives, newspapers or books.

So in any case we will be able to reach the size of 100 millions words without any problem.

[3] http://www.internetworldstats.com/asia.htm

3.2 Negotiation of intellectual property rights

Obtaining the right to use the texts collected from web sites is not an easy task and requires a lot of time and negotiation. We have to face a multiplicity of ownership interlinked between different source providers. For instance, publications of Voice of America (VOA) are in the public domain and can be used or redistributed without any restriction[4]. But some materials published by Voice of America on its website are copyrighted news publications coming from press agencies such as Associated Press, Agence France Presse, Reuters, etc. We therefore have to make the distinction between the data owned by VOA and publications owned by others or alternately negotiate with all copyright owners. Moreover the more sources we use in our collection process the more licenses are to be included with the corpus. For instance Pashto texts from Wikipedia are available under the Creative Commons Creative Commons Attribution/Share-Alike License[5] which states that the data can be used, modified or redistributed under this license and that all users of the data must accept the terms of it.

At the time of writing of this article (March 2012), we have obtained positive feedbacks from more 50 different sources including important ones such as www.tolafghan.com, www.rohi.af or www.benawa.com.

3.3 Corpus collection

The corpus collection consists mainly in crawling identified and negotiated sources, cleaning and formatting the data. Several difficulties arise when computing with Pashto as explained in section 2. In addition to the specificity of Pashto as a language, we have to face encoding character problems or the use of non Pashto text in web pages.

Once we have obtained the right to use the data from a specific website we apply a fully automatic process to get the useful texts, clean the data and format them in XML. We developed a fully automatic process for collecting the data from a specific website. This process can be summarized by the following sequential steps:

1. Crawl the entire website.
2. Extract the content of each crawled web page and convert it into text.
3. Remove empty files and duplicated files.
4. Detect Pashto text files by using n-gram language models of words and remove non-pashto text files.
5. Convert each Pashto text file into XML and identify Pashto sentences within one file using n-gram models of characters.

The resulting corpus is made of XML documents which follow a specific Definition Type Document (DTD).

We use following DTD for the monolingual corpus:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!ELEMENT CORPUS (FILE)*>
<!ELEMENT FILE (DATE, TITLE?, INFO, BODY)>
<!ELEMENT DATE (#PCDATA)>
<!ELEMENT TITLE (#PCDATA)>
<!ELEMENT INFO (#PCDATA)>
<!ELEMENT BODY (p*)>
<!ELEMENT p (#PCDATA | NON-PS)*>
<!ELEMENT NON-PS (#PCDATA)>
<!ATTLIST NON-PS lang (too_short | albanian |
arabic | azeri | basque | bengali | bulgarian |
chinese | cebuano | croatian | czech | danish |
dutch | english | estonian | farsi | finnish |
french | german | greek | hausa | hawaiian | hindi
| hungarian | icelandic | indonesian |
italian | kazakh | klingon | kyrgyz | latin |
latvian | lithuanian | macedonian | mongolian |
nepali | norwegian | pidgin | pig_latin | polish
 | portuguese | romanian | russian | serbian |
slovak | slovene | somali | spanish | swahili |
swedish | tagalog | turkish | ukrainian | ukran
ian | urdu | uzbek | vietnamese | welsh | unknown ) "unknown" >
```

Figure 1 Document Type Definition (DTD) of the XML format for the monolingual corpus

where:

- DATE is the crawling date,
- TITLE is the title of a document,
- INFO is the exact URL of the crawled document,
- BODY contains the cleaned texts.
- P delimits the physical sentences, e.g. the texts between two carriage returns (‘\n’).
- NON-PS identifies non-pashto sentences and gives the language of the sentence.

So far we have crawled and cleaned more than 60 million words and we do expect to reach the 100 million words by September 2012.

3.4 Validation of the monolingual corpus

Since the procedure is fully automatic, the validation aims at checking the procedure used for crawling, cleaning and formatting the data. For this purpose a Pashto native speaker checked manually 1000 physical paragraphs for a total of 150k words.

The validation consisted of checking the following items:

- Is there any character encoding problem?
- Is there any difference with the corresponding text of original web document?
- Is the language identification tag correct?

No character encoding errors were found.

No differences were found between the cleaned paragraphs and the original ones.

Out of the 1000 samples, two items were wrongly tagged as Arabic instead of Pashto. These two sentences were very short ones and were using only Arabic letters of the Pashto alphabet which confused our language identification tool based on n-gram models of characters.

4. Speech corpus

The speech corpus is made of broadcast news recordings transcribed orthographically. The target size of the corpus is 100 hours with recordings coming from different sources to cover different dialects.

4 http://www.voanews.com/english/news/69075687.html

5 http://creativecommons.org/licenses/by-sa/3.0/

As for the monolingual text corpus, international broadcasters (VOA, BBC, DW, CRI, VOR, TRT) and local broadcasters from Pakistan and Afghanistan are selected. The same tasks of identification and negotiation have been carried out. The corpus collection consists of recording broadcast news shows from TV and radio channels through the digital satellite or by capturing data on the web. Once the data have been recorded they are audited to validate the quality and the content of the recordings. Once successfully validated broadcast news shows can be transcribed orthographically.

For transcription, we use Xtrans[6] a transcription tool that supports Pashto natively.

The transcription conventions are adapted for Pashto but follows standard transcription guidelines for broadcast news speech.

4.1 Transcription issues

Transcription involves many theoretical issues and may not be regarded just a process of writing down what we hear. According to scholars at Linguistic Society of America (LSA) on transcription (Bucholtz et al. 2006) "Transcription systems differ in the specific sets of analytic choices they offer for representing the spoken language data, and the set of available transcription choices may lead different analysts to see different things in the same recorded data."

Pashto transcription also presents multiple choices, with at least two major transcription variations, attributed as Afghani and Peshawari (Yousufzai) styles. There are also variations, which arbitrarily combine different features of these two different styles. This variation presents a challenge for transcribing Pashto. The situation is further aggravated by the fact that the mutually agreed transcription standard between Pakistani and Afghani Pashto scholars, which was finalized in a meeting in Bara Gali Pakistan 1990, (Kakakhel, 2012) is not propagated and completely adopted by community, and the language authorities in both Afghanistan (originally Pakhto Tolana in Kabul) and Pakistan (Pashto Academy in Peshawar) have not managed to address this issue.

These differences include lexicalized word forms possibly due to dialectal differences based on pronunciation and are reflected in the writing styles. Some variations sound predictable as in the following examples of Table 3 where the presence of a long vowel in the middle of the word makes the difference between the Afghani and Peshawari writing styles.

English	Yousufzaï	Afghani
It can be	کيدي شي	کيداي شي
He can do	کړي شي	کړاي شي
He can understand	پوهيدي شي	پوهيداي شي
He can do	کولي شي	کولاي شي

Table 3 Lexical variations between YousufzaÏ and Afghani Pashto

Foreign loan words are spelt differently according to the way they are accented. Some examples are given in Table 4

English	Yousufzaï	Afghani
Parliament	پارليمان	پارلمان
Process	پروسيسه	پروسه
Conference	کانفرنس	کنفرانس

Table 4 Lexical variations for foreign loan words

Difference in the spelling of foreign proper nouns has been noted as depicted in Table 5

English	Yousufzaï	Afghani
Chicago	شکاګو	شيکاګو
George Bush	جورج بش	جورج بوش
Tunisia	تيونس	تونس

Table 5 Lexical variations for proper names

The way how the words are joined is also significant in the writing styles. It is interesting to note that in Afghani writing style there is a tendency of writing the words jointly but in Yousufzaï the case is reverse as showed in Table 6

English	Yousufzaï	Afghani
signature	لاس ليك	لاسليك
victory	بريالي توب	بريالیتوب
welcome	پخير	په خير

[6] http://www.ldc.upenn.edu/tools/XTrans

capital	پای تخت	پایتخت

Table 6 Examples of word joining confusions

Of all orthographic issues, the use of multiple forms of ی (**'**Yay**'**) by Pashto writers has been noted as the greatest source of spelling variation, especially among the writers of Peshawari Pashto writing style. Formerly, writers of both Afghani and Peshawari writing styles used only two forms of this character, that is, 'ی' and 'ے' . Later on, Afghani alphabet included five forms of this character (ي, ې, ۍ, ى, and ئ) and the character 'ے' replaced with that is 'ی' word-finally and two vertical dots below the normal 'ی' word-medially. Afterwards, writers in Peshawari style of writing also followed this set of five forms of ی but didn't abandoned 'ے' and continued to use this character according to the decisions made at Baragali conference (Kakakhel, 2012). Table 7 gives some examples of words using different forms for the letter 'Yay'.

English	**Yousufzaï**	**Afghani**
Village	کلے	کلی
Lion	زمرے	زمری
Man	سړے	سړی
Place	ځائے	ځای
God	خدائے	خدای

Table 7 Lexical variations for the use of the lettrer 'Yay'

4.2 Transcription process and validation

One transcription team is used to audit and transcribe the data.

The transcription team is made of:

- Several transcribers who are native speaker of Pashto
- One supervisor who manages the pool of transcriber and support them

Since the dialectal variations of Pashto are quite important, the transcription team is made of both mother-tongue speakers of Afghanistan and Pakistan.

The transcriptions are firstly transcribed by one transcriber. Once the file is totally transcribed the transcriber revises it entirely. Then the file is passed to another transcriber who revises it entirely and corrects any error he/she might found (cross validation process). The transcription team makes use of lexicons and dictionaries to correct any misspelling and error.

Finally the file is passed to ELDA who conducts a formal validation of the transcription file.

The validation procedure includes the following steps:

- Pashto native speakers are asked to control and assess the quality of the transcriptions according to the transcription manual and the validation criteria
- For each delivery, 3 to 5% of the data are checked randomly. Ten hours of transcription represent around 100k words and a sample of 3% (3k words) is enough for a confidence interval of 95% and a word error rate of 4%.
- The word error rate must be below 5%. To calculate this word error rate, we ask our Pashto experts to correct the samples and we automatically align the original samples with the corrected ones and compute the word error rate as described in (Hunt, 1990).
- The same assessment grid is used to validate each batch of transcriptions
- After applying the assessment grid, if the sample contains more errors than the threshold of 5%, the entire batch of transcriptions is rejected and the transcriptions must be revised.
- If a batch of transcription is rejected and sent back to transcribers, the corrected transcriptions will follow a new phase of validation.
- A first hour of transcription is validated at the start of the collection (prevalidation phase).
- Then the validation process is applied every 10 hours of transcription.

5. Parallel text corpus.

In addition to the monolingual text corpus and the speech database the project is producing a parallel Pashto-French corpus of around 2 million words. For this, 200 hours of transcription of Pashto recordings is translated into French by professional translators. The source texts are made of the transcription of the previously described corpus in section 4 and an existing corpus of 100 hours of transcription of conversational speech in Pashto. When producing translations from language A to language B, it is usually required that translators are native speakers of the target language B. But it is very unlikely that we will be able to find sufficient Pashto-to-French translators who are native speakers of French. Therefore we decided to use the services of native Pashto translators who are fluent in French and then revise the produced translations by few native French speakers fluent in Pashto. The source and translated files are formatted in XML and follow an adapted DTD derived from NIST MT evaluations[7].

The process of translation has started with a first batch of 10 hours to be translated.

The translations are validated at ELDA. An automatic validation is applied to check the format of the files, duplicated sentences and spelling errors.

Then a manual validation is applied.

- For each delivery, a 1200 words sample of the data is checked randomly by a Pashto/French bilingual professional translator at ELDA.
- Every couple of sentences (Pashto/French) is checked and the type of error is indicated and the corrected translation is produced.
- To ensure consistency from one review to another, the following system has been adopted for judging translations.
-

[7] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.3.dtd

Error	Penalty points
Syntactic 4 points	4 points
Deviation from guidelines	3 points
Lexical	2 points
Poor French usage	1 point

Table 8 Penatly points for the validation of translations

"Poor usage" reveals a non-nativeness style like "Je vous souhaite une *journée bonne*" à la place de "je vous souhaite une *bonne journée*".

6. Conclusion

This paper reports on the production of new language resources for Pashto: a monolingual text corpus of 100 million words, a speech database of 100 hours of transcribed broadcast news and a parallel Pashto-to-French text corpus of around 2 million words.
For the monolingual corpus, we developed a fully automatic procedure that can be reused to develop monolingual corpora for any language.
Out of our three tasks, transcribing Pashto recordings is the more challenging one since Pashto is an oral language and due to the lack of orthography conventions. The bilingual corpus of Pashto-French data raises also some difficulties due to the low number of available professional translators for this translation direction.
Nevertheless, at the time of writing of this article, we have collected, cleaned and formatted more than 60 million words for the monolingual corpus. For the transcription task more 30 hours have been transcribed and for the parallel corpus, the translation work has just started.
We expect to finish the collection of the monolingual and audio corpora by September 2012 and the parallel Pashto-French corpora by the end of 2012. These resources will be publicly available for research purposes.

7. Acknowledgements

This material is based on the work in the scope of the PEA TRAD project supported by the Direction Générale de l'Armement (DGA).

8. References

Akram M. and Hussain S (2010). Word segmentation for urdu ocr system. In *Proceedings of the 8th Workshop on Asian Language Resources*, COLING2010, Beijing,

Bucholtz, M. and Du Bois, J. W. (2006), in the Session Report on "Transcription Issues in Current Linguistic Research," *80th Annual Meeting, Linguistic Society of America,* New Mexico, USA

Durrani N. and Hussain S. (2010). Urdu word segmentation. In *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*, Los Angeles, US, 1-6 June.

Hunt, M. J. (1990) Figures of Merit for Assessing Connected Word Recognisers. In *Speech Communication*, 9, 1990, pp 239-336.

Kakakhel, S. T. U. and Khattak, R. S. (2012). Pakhto Liklar, *Pashto Academy*, University of Peshawar, Peshawar, Pakistan.

Kathol A., Precoda, K., Vergyri D.,Wang W., and Riehemann S.. (2005). Speech translation for low-resource languages: The case of Pashto. In *Proceedings of Interspeech*, Lisbon, Portugal, September 4-8 2005.

C. A. Kopris. 2005. Computing in pashto: An overview of a major language in Afghanistan and Pakistan. In *Multilingual Magazine*, Volume 16, pages 27–30, March.