# Towards 100M Morphologically Annotated Corpus of Tajik

Gulshan Dovudov, Vít Suchomel, Pavel Šmerk

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{xdovudov,xsuchom2,xsmerk}@fi.muni.cz

**Abstract.**
The paper presents a work in progress: building morphologically annotated corpus of Tajik language of the size more than 100 million tokens. The corpus is and will be by far the largest available computer corpus of Tajik: even its current size is almost 85 million tokens. Because the available text sources are rather scarce, to achieve the goal also the texts of a lower quality have to be included. This short paper briefly reviews the current state of the corpus and analyzer, discusses problems with either "normalization" or at least categorization of low quality texts and finally also the perspectives for the nearest future.

**Key words:** web corpora, Tajik

## 1 Introduction

The Tajik language is a variant of the Persian language spoken mainly in Tajikistan and also in some few other parts of Central Asia. Unlike closely related Iranian Persian which uses Arabic script, Tajik is written in Cyrillic.

Since the Tajik language internet society (and consequently the potential market) is rather small and Tajikistan itself is ranked among developing countries, available tools and resources for computational processing of Tajik as well as publication in the field are rather scarce. Availability of Tajik corpus data does not seem to change during the last year: aside from our corpus, the biggest freely available corpus is still the one within the Leipzig Corpora Collection project [5] (ca. 100 000 sentences, 1.8 million words, huge amount of errors), the biggest planned corpus is still the one prepared by the Tajik Academy of Sciences (target size 10 million, no visible changes in the last year)[1]. For further details and for information on other minor corpus projects see our previous work [1].

In this paper we present our corpus of contemporary Tajik language of more than 85 million tokens. After a brief review of its current state we will discuss problems with low quality texts. Finally we will debate possible improvements in the nearest future.

[1] http://www.termcom.tj/index.php?menu=bases&page=index3&lang=eng (in Russian)

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2012, pp. 91–94, 2012. 

## 2 The Current State of the Corpus

Our corpus is built only from online sources (or, to be precise, only from sources which were online at some time, as the accessibility of some data changes rapidly). We use two different approaches to obtain the data. For the details refer to our previous papers [1] and [2].

The main part of the corpus was collected by crawling several portals, mostly news portals, in Tajik language.[2] Each portal is processed separately to get the maximum of relevant (meta)information, i.e. correct headings, publication date, rubric etc. The data for the second part of the corpus was obtained with SpiderLing, a general web crawler for text corpora [6], which automatically walks through the internet and searches for texts of a particular language. The crawling process started with 2570 seed URLs (from 475 distinct domains) collected with Corpus Factory [3]. The obtained data was uniformely tokenized and then deduplicated using Onion[3] with moderately strict settings[4].

The actual size of the corpus is 84,557,502 tokens and 70,665,499 words i.e. tokens which contain only Cyrillic characters (the rest is punctuation, numbers, Latin words etc.). The semi-automatically crawled part has 57,636,441 tokens, which means that the contribution of the automatically crawled part is 26,921,061 tokens. Our morphological analyzer recognizes 92.5 % of words (i.e. of the above mentioned 70 million tokens).

The corpus is not freely available for a download at the moment, but eventual interested researchers can access it through a very powerful corpus manager the Sketch Engine[5] [4].

### 2.1 Dealing with texts of lower quality

As was mentioned in the Introduction, Tajik uses Cyrillic script. Unfortunately, the Tajik alphabet contains six letters which are missing in the probably most widespread codepage in the post-Soviet world, i.e. cp1251. As there is or has been almost no support for Tajik in Windows and also in other major OSs, in many occasions people writing Tajik texts were not able to write proper Tajik-specific characters and were about to use some replacements. The most frequently used replacement sets can be seen in the table.[6]

Note that letters Љљ, Њњ, Її, and Ўў are shared by the Replacement sets 1 and 2, but except for Ўў the replacements have different meanings. Thus it would not be correct to directly replace all these replacement characters, but

[2] Paradoxically, the two largest Tajik news portals are not located in Tajikistan, but in the Czech Republic (ozodi.org, Tajik version of Radio Free Europe/Radio Liberty) and the United Kingdom (bbc.co.uk, Tajik version of BBC).

[3] http://code.google.com/p/onion/

[4] Paragraphs with more than 50 % of duplicate 7-tuples of words were removed.

[5] http://ske.fi.muni.cz/open/

[6] Note that Ss is not Latin S, but "Cyrillic Dze", Ii is not Latin I, but "Cyrillic Byelorussian-Ukrainian I" and the accents above ЌќЃѓ should be acute, not grave (probably wrong glyph in LaTeXfont).

**Table 1.** Replacement sets for Tajik.

| Char | RS1 | RS2 | RS3 | RS4 |
|---|---|---|---|---|
| Ғғ | Ѓѓ | Љљ | Uu | Гг |
| Ӣӣ | Її | Њњ | Bb | Ии |
| Ҷҷ | Љљ | Її | Хх | Чч |
| Ҳҳ | Њњ | Іі | {[ | Хх |
| Ққ | Ќќ | Ss | Rr | Кк |
| Ӯӯ | Ўў | Ўў | Ee | Уу |

one have to guess the whole replacement set which the particular document uses. Moreover, sometimes people write й instead of -и and bigrams х, к, or ч, (i.e. letter and comma) instead of proper letters ҳқҷ.

Out of 192,664 documents (web pages, newspapers articles etc.), 169,233 need not any change (87.8 %), 21,524 needs some replacements, in 1323 documents the "diacritics" was restored (RS4), and finally for 584 there was need both for replacements and the diacritics restoration.

2391 documents use RS1, 778 RS3 and 113 documents use RS2. The bigrams were used in 2453 words and й instead of -и in 77812 words. In total, 859641 words was somehow modified which is more than 1 % of all words in the corpus.

Numbers of particular changes and other useful information are described in each document's metadata which allows users to create specific subcorpora, e.g. subcorpus of texts without any changes (probably the most quality ones).

## 3 Future Work

The semi-automatic crawling was run a year ago, then after four months and then now. The automatic crawling was run a year ago, then after four months, after six months, and finally now. From the respective tables it can be seen, that the growth is not strictly linear, but the achievement of 100 million tokens seems to be possible during the next year.

**Table 2.** Growth of the semi-automatically crawled part.

| date | tokens | increment | per month |
|---|---|---|---|
| 11/2011 | 40.6 M | — | — |
| 03/2012 | 43.3 M | +6.7 % | 1.7 % |
| 11/2012 | 47.2 M | +9.1 % | 1.1 % |

In the nearest future we want to further improve the analyzer to achieve better coverage and also we want to employ some kind of morphological guessing of unknown words. Then we will be able to develop a tagger and some tools for complex verbs and noun phrases detection which all will allow

**Table 3.** Growth of the automatically crawled part.

| date | tokens | increment | per month |
|---|---|---|---|
| 11/2011 | 34.6 M | — | — |
| 03/2012 | 41.1 M | +18.6 % | 4.7 % |
| 05/2012 | 44.7 M | +8.6 % | 4.3 % |
| 11/2012 | 54.8 M | +22.6 % | 3.8 % |

us to create word sketches [4] for Tajik words. That is why we need to have the corpus as big as possible: 100 million words is considered as a minimum for word sketches to work reasonably.

## Acknowledgements

This work has been partly supported by Erasmus Mundus Action II lot 9: Partnerships with Third Country higher education institutions and scholarships for mobility, and by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013.

## References

1. Dovudov, G., Pomikálek, J., Suchomel, V., Šmerk, P.: Building a 50M Corpus of Tajik Language. In: Proceedings of the Fifth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2011. Masaryk University, Brno (2011)
2. Dovudov, G., Suchomel, V., Šmerk, P.: POS Annotated 50M Corpus of Tajik Language. In: Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL 8/AfLaT 2012). European Language Resources Association (ELRA), Istanbul (2012)
3. Kilgarriff, A., Reddy, S., Pomikálek, J., PVS, A.: A Corpus Factory for Many Languages. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010. Valleta, Malta (2010)
4. Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: The Sketch Engine. In: Proceedings of EURALEX. pp. 105–116 (2004)
5. Quasthoff, U., Richter, M., Biemann, C.: Corpus Portal for Search in Monolingual Corpora. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006. Genoa (2006)
6. Suchomel, V., Pomikálek, J.: Practical Web Crawling for Text Corpora. In: Proceedings of the Fifth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2011. Masaryk University, Brno (2011)