POS Annotated 50M Corpus of Tajik Language

Gulshan Dovudov, Vít Suchomel, Pavel Šmerk

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
zarif_dovudov@mail.ru,xsuchom2@fi.muni.cz,smerk@mail.muni.cz

Abstract

Paper presents by far the largest available computer corpus of Tajik language of the size of more than 50 million words. To obtain the texts for the corpus two different approaches were used and the paper offers a description of both of them. Then the paper describes a newly developed morphological analyzer of Tajik and presents some statistics of its application on the corpus.

Keywords: Tajik language, Tajik corpus, morphological analysis of Tajik

1. Introduction

1.1. Tajik Language

The Tajik language is a variant of the Persian language spoken mainly in Tajikistan, where it plays a role of the national and official language. Tajik is spoken also in some few other parts of Central Asia, among which neighboring Uzbekistan is the most notable, because the biggest group of Tajik native speakers outside Tajikistan resides there.

Unlike closely related and mutually intelligible Iranian Persian (Farsi) and Afghan Persian (Dari), which are written in the Arabic script, Tajik is written mostly in the Cyrillic alphabet which is the official script.

According to its grammatical structure the Tajik language inflectionally belongs to analytical type of languages. Although the nominal morphology itself is rather poor—Tajik has neither gender nor case, only pluralization suffixes and specificity/undefiniteness marker—nouns, adjectives and participles can appear in many different forms thanks to the direct object marker, possesive enclitics (*dust-am: friend your*), copulas (*dust-am: friend [I] am*) and ezafe marker which are all written together with the word in the Cyrillic script and some of them even can combine together. Tajik verbal morphology is much richer: besides many compound verbal forms Tajik has affixes (mostly suffixes) for expressing person, number, tense, mood and voice. Together with participles, the number of distinct forms which can be generated from a single verbal root can easily exceed 1000 items.

1.2. Existing Tools and Resources

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 publications in the field are rather scarce. Moreover, many of the publications were published only in Russian.

For Tajik there exist several online[1] or offline Tajik–Russian and Tajik–English dictionaries, from which the biggest (Usmanov et al., 2007) and (Usmanov et al., 2008) covers ca. 120,000 Russian words and phrases. The corresponding numbers for English are rather lower, around 35,000 words and phrases at most.

Megerdoomian and Parvaz (2008; 2009) aim to transliterate Tajik from Cyrillic alphabet to Arabic script, which would allow to employ tools developed for closely related Iranian Persian. More elaborated transliteration was offered by Usmanov et al. (2009). Usmanov et al. also created a morphological analyzer (2011) and a spellchecker (2012) for OpenOffice.org suite. Another spellchecker was made by Davrondjon and Janowski (2002). There is also available a Tajik text-to-speech system built by Khudoiberdiev (2009). If we add Tajik extension of the multilingual information extraction system ZENON (Hecking and Sarmina-Baneviciene, 2010), the list of all more interesting Tajik language processing tools and non-corpora resources we are aware of is complete.

The computer corpora of Tajik language are either small or even still only in the stage of planning or development. The University of Helsinki offers a very small corpus of 87 654 words.[2] Megerdoomian and Parvaz (2008; 2009) mention a test corpus of approximately 500 000 words, and the biggest and the only freely available corpus is offered within the Leipzig Corpora Collection project (Quasthoff et al., 2006) and consists of 100 000 Tajik sentences which equals to almost 1.8 million words.[3] Iranian linguist Hamid Hassani is said to be preparing a 1 million words Tajik corpus[4] and Tajik Academy of Sciences prepares a corpus of 10 million words[5]. Unfortunately, at least by now the latter is not a corpus of contemporary Tajik, but rather a collection of works—and moreover mainly a poetry—of a few

[1] E.g. http://lugat.tj, http://sahifa.tj, http://termcon.tj or http://www.slovar.kob.tj.

[2] http://www.ling.helsinki.fi/uhlcs/readme-all/README-indo-european-lgs.html

[3] Unfortunately, the encoding and/or transliteration vary greatly: more than 5 % of sentences are in Latin script, almost 10 % of sentences seem to use Russian characters instead of Tajik specific characters (e.g. х instead of Tajik ҳ, which sound/letter does not exist in Russian) and more than 1 % of sentences uses non-Russian substitutes for Tajik specific characters (e.g. Belarussian ў instead of proper Tajik ӯ) — and only the last case is easy to repair automatically.

[4] http://en.wikipedia.org/wiki/Hamid_Hassani, http://www.tajikistan.orexca.com/tajik_language.shtml

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

notable Tajik writers (one of them is even from the 13th century).

1.3. Structure of the Paper

In this paper we present a newly built corpus of the contemporary Tajik language of more than 50 million words. All texts were taken from the internet. We used two different approaches to obtain the data for the corpus and we describe these methods and their results in the following two sections. In the Section 3 we present a new morphological analyzer of Tajik and the results of annotating the corpus with this analyzer. Finally, in the last section we discuss some planned future improvements.

2. Bulding the Corpus

As it was said in the Introduction, the new[6] corpus consists of two parts. The first one is supposed to contain data of a higher quality, the second one contains texts which we are able to download and process only in a less controlled way.

2.1. Semi-automatically Crawled Part

The first part of the corpus was collected by crawling several news portals in Tajik language.[7] If the articles of a particular portal were numbered, we tried to download all possible numbers, otherwise we got a list of articles either directly from the portal, or from the Google cache. Each single portal was processed separately to get the maximum of relevant (meta)information, i.e. correct headings, publication date, rubric etc.

In the Table 1 we present some statistics of the obtained data. Docs is a number of documents downloaded from the given source, pars is a number of paragraphs (including headings), words is a number of tokens which contain only characters from Tajik alphabet, w/doc is a words / document ratio (i.e. average length of possibly continuous texts), tokens is a number of all tokens (words, interpunction etc.) and MB is the size in megabytes of the data in vertical corpus format (i.e. plain text). The table is sorted by number of words. From the electronic library on `gazeta.tj` we choose only prose and omit all more structured texts as poetry, drama or e.g. computer manual. The articles in `gazeta.tj` archive are joined in one file on a weekly basis and that is why the words / document ratio is so high.

On almost all websites, alongside the articles in Tajik there were also many articles in Russian. Because both alphabets, Tajik and Russian, contain characters which do not occur in the other alphabet, it is easy to distinguish between the two languages and discard the Russian articles even without any further language analysis.

2.2. Automatically Crawled Part

SpiderLing, a web crawler for text corpora (Suchomel and Pomikálek, 2011), was used to automatically download documents in Tajik from the web. We started the crawl using 2570 seed URLs (from 475 distinct domains) collected with Corpus Factory (Kilgarriff et al., 2010). The crawler combines character encoding detection[8], general language detection based on a trigram language model trained on Wikipedia articles, and a heuristic based boilerplate removal tool jusText[9] which removes content such as navigation links, advertisements, headers and footers etc. and preserves only paragraphs (preferrably continuous groups of paragraphs) containing full sentences.

The crawler downloaded 9.5 GB of HTML data in ca. 300,000 documents over three days. That is not much compared to crawling documents in other languages by the same tool. For example the newly built web corpus of Czech, which has roughly only two times more native speakers compared to Tajik, has more than 5 billion words, of course, not obtained in three days. We conclude that the available online resources in Tajik are very scarce indeed. An overview of internet top level domains of URLs of the documents obtained can be found in Table 2.

TLD	docs downloaded	docs accepted
tj	55.0 %	51.7 %
com	23.0 %	28.1 %
uk	8.9 %	7.2 %
org	6.8 %	7.7 %
ru	2.6 %	1.4 %
ir	1.6 %	2.4 %
other	2.0 %	1.5 %

Table 2: Number of documents by internet top level domain

Since Russian is widely used in government and business in Tajikistan (and other language texts may appear), 33 % of the downloaded HTML pages were removed by the SpiderLing's inbuilt language filter. The obtained data was tokenized and deduplicated using Onion[10] with moderately strict settings[11]. Some statistics of the automatically crawled part of the corpus are in the Table 3 (only the ten most productive sources of data are detailed).

2.3. Corpus of Tajik Language

The two partial corpora were joined together and the result was deduplicated using the tool Onion. We obtained a corpus of more than 50 million words, i.e. corpus positions which consists solely of Tajik characters, and more than 60 million tokens, i.e. words, interpunction, numbers etc. Detailed numbers follow in the Table 4.

[6] It should be noted that this is not the very first information about the new corpus. We presented it at a local event few months ago (Dovudov et al., 2011). Since that time, as the following numbers and text show, we extended the first part of the corpus and repaired some errors in the tools for automatic crawling and then processed the previously crawled data again, which positively affected the second part of the corpus and among others it enabled its much better deduplication.

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

[8] `http://code.google.com/p/chared/`

[9] `http://code.google.com/p/justext/`

[10] `http://code.google.com/p/onion/`

[11] removing paragraphs with more than 50 % of duplicate 7-tuples of words

source	docs	pars	words	w/doc	tokens	MB
ozodi.org	58228	348583	12597156	216	14738546	178
gazeta.tj archive	480	163565	5006469	10430	6032000	67
co.uk	9240	164252	4124668	446	4767707	57
jumhuriyat.tj	8104	104331	3703806	457	4397596	53
tojnews.org	9598	72407	2529370	264	3073951	37
khovar.tj	16860	67325	2502295	148	3055984	39
millat.tj	2540	47144	2131354	839	2511507	29
gazeta.tj	1940	37460	1352150	697	1622370	18
gazeta.tj library	130	98690	1053206	8102	1358510	15
ruzgor.tj	1753	26417	1046598	597	1241107	14
			…			
all	**115364**	**1200645**	**38269686**	**332**	**45473773**	**537**

Table 1: Statistics of the semi-automatically crawled part of the corpus.

source	docs	pars	words	w/doc	tokens	MB
*.wordpress.com	3298	85923	3683551	1117	4489937	50
bbc.co.uk	5194	94441	2480519	478	2857831	34
ozodi.org	5930	81147	2017199	340	2399314	28
khovar.tj	10345	41604	1599389	155	1953702	25
millat.tj	1349	21274	1080308	801	1268268	14
gazeta.tj	1704	27631	1040471	611	1244311	14
*.blogspot.com	793	21849	854055	1077	1060820	12
firdavsi.com	559	24315	804196	1439	966421	11
pressa.tj	2317	17130	637781	275	771169	9
ruzgor.tj	889	10408	489143	550	573800	7
			…			
all	**53424**	**675366**	**24723099**	**463**	**29709116**	**346**

Table 3: Statistics of the automatically crawled part of the corpus.

It was rather surprising for us that the fully automated crawling yielded even smaller data than the semi-automated approach. It has to be said that at least 25 % of semi-automatically crawled data were inaccessible to the general crawler, as it cannot extract texts from RAR-compressed archives (gazeta.tj archive and library) and because there does not seem to exist any link to the bigger part of older BBC articles although they remained on the server (we exploited Google cache to get the links). It is highly probable that also the other sites contain articles unreachable by any link chain and thus inaccessible for the general crawler. But even if we discount these data, the automated crawling did not outperform the semi-automated one in such an extent that we expected and which is common for many other languages. As we remarked in the previous subsection, we attribute it to the scarceness of the online texts in Tajik language. It means that we probably reach or almost reach the overall potential of internet resources, i.e. even if we somehow get all Tajik online texts, the corpus might be bigger by half, but surely not, for example, ten times or even three times.

3. Morphological Analysis of Tajik and Annotation of the Corpus

As it was mentioned in the Introduction, a morphological analyzer for the Tajik language already exists (Usmanov et al., 2011). Unfortunately, for the annotation of corpus data it showed up quite unusable, mainly for the following reasons:

- the program is too slow, it processes less than 2 words per second;
- its code is written in MS Visual Basic 6.0 and the executable can run only on MS Windows, but we process corpus data on Linux servers (and may be there are ways to compile such MS VB code under Linux, but we consider it too complicated);
- the program splits the input word form to morphs (among others), but it cannot offer the lemma (base form, citation form) of the word.

3.1. New Morphological Analyzer

For the creation of the new morphological analyzer we wanted to use information about Tajik morphs and their possible combinations from the current analyzer. It was not a simple task, because the information on morph combining is not described in some external file but “encoded” directly in the source code, so it is still a work in progress.

We use an approach of Jan Daciuk (Daciuk, 1998), who invented an algorithm for both space and time efficient building of minimal deterministic acyclic finite state automata (DAFSA). On his pages he offers source codes of tools

source	**docs**	**pars**	**words**	**w/doc**	**tokens**	**MB**
ozodi.org	59943	384932	13426445	224	15738683	189
gazeta.tj archive	480	163555	5006432	10430	6031951	67
co.uk	9288	164436	4129179	445	4772807	57
jumhuriyat.tj	8106	104776	3703685	457	4397650	53
*.wordpress.com	3080	74652	3235436	1050	3946319	44
tojnews.org	9653	72575	2532572	262	3077917	37
khovar.tj	17079	68022	2512232	147	3082293	39
millat.tj	2803	49846	2268000	809	2673004	30
gazeta.tj	2209	38060	1389318	629	1665672	19
kemyaesaadat.com	1863	33859	1182353	635	1404072	16
			...			
all	**138701**	**1541470**	**51722009**	**373**	**61837585**	**723**

Table 4: Statistics of the resulting corpus.

which allows to convert morphological data into such an automaton and to use it for morphological analysis.

The data for the analyzer are triplets *word:lemma:tag*, e.g. `kardem:kardan:tag` where the lemma is, in fact, not present in a full form, but encoded in the following way `kardem:Can:tag`, which means "to obtain a lemma, delete 2 (`A` = 0, `B` = 1, `C` = 2, ...) letters from the end of the wordform and add the string `an`".[12] The list of such triplets can be viewed as a finite formal language and for such languages always exists the minimal DAFSA. If we generate such triplets for all word forms the analyzer should know, the data will be quite big, but also very redundant: the conversion to an automaton serves as a very good compression.

The data for the analyzer are generated from our dictionary of Tajik base forms (lemmata), where each lemma has an information about its POS and optionally some additional information (e.g. that pluralization suffix can be not only common *-ho*, but also *-on*). The possible word forms are generated according to a rather simple description (80 lines) of possible suffixes and their allowed combinations. The process also uses an information about possible phonological or ortographical changes at morpheme boundaries.

The whole analysis is then just effortless travelling through the automaton: the analyzer simply (the automaton is deterministic) follows the path which corresponds to the analyzed word form and the delimiter, e.g. `kardem:`. Then the labels of each possible path to the final state represents one of the possible analyses. The new analyzer is therefore very simple: the whole source code has only around 450 lines in C++ (we have simplified Daciuk's code considerably). The new analyzer is also much faster then the current one: at the moment it processes around million words per second.[13] This number will get lower in the future as the size of the data will increase, but surely it always will remain very fast.

Because many words of our corpus were unknown to the current analyzer, we needed to enrich the lexicon. We took all unknown word forms from the corpus and for each such word form we generated possible lemmata. Then we manually evaluated such lemma candidates from the most frequent ones. In principle, our approach was very similar to (Sagot, 2005), but we avoided all the maths which Sagot uses to rank lemma candidates: the manual decision process was so fast that there was no need for some tiny optimization. But unlike Sagot, we did not use whole morphology at once, but we started with searching for possible proper nouns, then common nouns, then adjectives, which can express degree etc.—it simplifies the work of the annotator, because deciding only e.g. proper nouns is faster and less erroneous than deciding possible lemmata of all kinds in one pass. Before searching for lemma candidates we also tried to detect and lowercase words which started with capital letter yet not being a proper noun: we selected words whose lowercase form was more frequent than the form with the first letter capitalized. We obtained more than 7000 lemmata and almost 5000 of them were proper nouns. Of course, it is an ongoing process, we have evaluated only the most frequent candidates so far.

count of word+lemma+tag triplets	8,476,108
size of input data in bytes	175,845,264
size of automaton in bytes	1,138,480
bytes per line of input data	0.13
count of lemmata in dictionary	14934
average number of triplets per lemma	568

Table 5: Some statistics of the new analyzer data.

3.2. Annotation of the Corpus

We decided to use only a lemma and POS for the annotation of the corpus data. The information which is represented by the rest of the morphological tag is currently not in a fully consistent state[14] and also the current format is a subject to change, because the current analyzer uses tags which are too long and thus hard to read. The Table 6 shows the meaning of tags and counts of word forms for each POS.

[12] This example would work only for suffixes. To handle also the prefixes, it is possible to employ the same principle again, for example: `namekardem:ECan:tag`, where the first `E` denotes that to get the correct lemma kardan the first four letters are to be deleted (and nothing is to be added).

[13] The speed of analyzers was not compared on a same computer, but the difference is obvious.

[14] At least in our new analyzer—and it is not possible to directly use the "knowledge" of the current analyzer, as we have mentioned it in the previous subsection.

Meaning	Tag	# of forms
nouns	01	6267182
adjectives	02	941209
numerals	03	25572
pronouns	04	52
verbs	05	372778
infinitives	06	646500
adjectival participles	07	217273
adverbial participles	08	5253
adverbs	09	86
prepositions	10	44
postpositions	11	3
conjunctions	12	52
particles	13	35
interjections	14	64
onomatopoeia	15	0
numeratives	16	5

Table 6: Meaning of the tags and numbers of forms with the given tag in the data.

Our new analyzer is able to annotate 87.2 % of the 51,722,009 words from the corpus[15] (and 20.5 % of the wordlist). 25.6 % of known words are ambiguous (13.9 % of the wordlist) and for known words the analyzer offers 1.33 lemma+POS combinations in average. As this is the very first work in this field, we cannot offer any comparison with existing results. We also do not have any manually tagged data and thus we cannot calculate the standard measures like coverage, accuracy or F-measure.

dar	dar:01;dar:05;dar:10	1626855
ba	ba:10	1572867
va	va:12	1417227
ki	ki:04;ki:12	1226404
az	az:10	1173474
in	in:04;in:14	773985
bo	bo:10	513154
ast	ast:05	347578
on	on:04	301493
Tojikiston	Tojikiston:01	281627

Table 7: The top ten most frequent words, their analyses and frequency in corpus.

The annotated corpus is not freely available for a download at the moment, but eventual interested researchers can access it through the Sketch Engine on `http://ske.fi.muni.cz/open/`. This web interface displays concordances from the corpus for a given query. The program is very powerful with a wide variety of query types and many different ways of displaying and organising the results.

4. Future Work

The Table 8 shows statistics of the texts which were new in the automatically crawled part compared to the semi-automatically crawled data. It is worth mentioning the difference between the two parts of the corpus: the analyzer knows 89.1 % of words from the first part, but only 81.8 % words from the second part. Our interpretation is that data in the first part display a higher quality, higher regularity and we expect them to be less noisy than the data from the second part of the corpus. The numbers in the table indicate that there is still some room for an extension of the semi-automated part. We will prepare specialized scripts for the most productive portals to download their data in a some more controlled way. We would like to extend the first part of the corpus to at least 50 million words.

It is worth clarifying the cases of ozodi.org and kemyaesaadat.com, as data from these sites were downloaded also semi-automatically. The general crawler tries to get all reasonable texts on the page, which, on the news portals, may include the readers' comments. On the other hand, because the comments may contain a substandard language features, they were omitted during the semi-automated crawling. Thus the 1715 documents from ozodi.org and 346 from kemyaesaadat.com were not some newly added ones, but they were results of the deduplication which discarded the article itself and leaved only the comments as it processed the corpus by single paragraphs. This is also one of the reasons why we prefer the semi-automated crawling when it is possible: in the future we want to mark these comments to allow a creation of a subcorpus of the (presumably standard) language of articles as well as a subcorpus of the (potentially substandard) language of comments.

Another problem with the comments—but not only with them—is a common absence of Tajik-specific characters. The language model for the general crawler was trained using Tajik Wikipedia[16] so the crawler searches for texts in language, which looks like the language of Tajik Wikipedia. Unfortunately, in many Wikipedia articles the Tajik-specific characters are replaced by some other characters. The unambiguous replacements were trivially repaired in the whole corpus, but e.g. Cyrillic х can sometimes stand either for the Tajik-specific ҳ or also for х itself. On the one hand we plan to tag such texts to allow a creation of subcorpora with or without them, on the other hand we want to either develop a program which would be able to repair them or use (Usmanov and Evazov, 2011). We will also train the language model with another sets of texts to see how it will affect the crawled data.

Unlike the corpus—which is still "work in progress", but the progress is already rather moderated—the work on the analyzer and the morphological data is still at the beginning. It is neccessary to properly describe the "morphological" system of Tajik, design the tagset, enlarge the lexicon (preferably exploiting word derivation) and start working on at least partial disambiguation.

[15] Word is here a token which contain only characters from Tajik alphabet, because analyzer cannot analyze anything else. More than 10,000,000 of corpus tokens are numbers, interpunction etc., but also e.g. words in Latin alphabet.

[16] The use of Wikipedia to train the language model is a part of default settings or a default scenario of the process of building corpora for new languages without any other utilizable resources. Of course we have better Tajik texts at hand, but the automatically crawled part of our corpus had also to act as a test of a general suitability of our technologies for the case of building corpora for low-density languages.

source	**docs**	**pars**	**words**	**w/doc**	**tokens**	**MB**
*.wordpress.com	3080	74652	3235436	1050	3946319	44
ozodi.org	1715	36386	829816	484	1000827	11
*.blogspot.com	723	17901	690797	955	865401	10
firdavsi.com	428	19861	614062	1435	737959	8
abdulov.tj	99	9392	403686	4078	488243	6
nahzat.tj	2406	7096	394437	164	463538	6
sahifa.tj	56	478	390618	6975	480398	5
ozodagon.com	1425	6352	371081	260	446249	5
kemyaesaadat.com	346	8526	342394	990	411893	4
bayynattj.com	298	4810	293010	983	340423	4
			…			
all	**23337**	**340950**	**13454032**	**577**	**16365975**	**186**

Table 8: The contribution of automatically crawled part.

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 Center.

5. References

Jan Daciuk. 1998. *Incremental Construction of Finite-State Automata and Transducers, and their Use in the Natural Language Processing*. Ph.D. thesis, Technical University of Gdańsk, Gdańsk.

Gafurov Davrondjon and Tomasz Janowski. 2002. Developing a Spell-Checker for Tajik using RAISE. In *Proceedings of the 4th International Conference on Formal Engineering Methods: Formal Methods and Software Engineering*. Springer Verlag.

Gulshan Dovudov, Jan Pomikálek, Vít Suchomel, and Pavel Šmerk. 2011. Building a 50M Corpus of Tajik Language. In *Proceedings of the Fifth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2011*, Brno. Masaryk University.

Leonid A. Grashenko, Zafar D. Usmanov, and Aleksey Y. Fomin. 2009. Tajik-Persian converter of the graphic writing systems. National patent 091TJ, National Patent Information Centre, Republic of Tajikistan. (In Russian).

Matthias Hecking and Tatiana Sarmina-Baneviciene. 2010. A Tajik Extension of the Multilingual Information Extraction System ZENON. In *Proceedings of the 15th International Command and Control Research and Technolgy Symposium (ICCRTS)*, Santa Monica, CA.

Khurshed A. Khudoiberdiev. 2009. *Complex Program of Tajik Text-to-Speech Synthesis*. Ph.D. thesis, Khujand Polytechnical Institute of Tajik Technical University, Dushanbe. (In Russian).

Adam Kilgarriff, Siva Reddy, Jan Pomikálek, and Avinesh PVS. 2010. A Corpus Factory for Many Languages. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, Valleta, Malta.

Karine Megerdoomian and Dan Parvaz. 2008. Low-density Language Bootstrapping: The Case of Tajiki Persian. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech, Morocco.

Karine Megerdoomian, 2009. *Language Engineering for Lesser-Studied Languages*, chapter Low-density Language Strategies for Persian and Armenian, pages 291–312. IOS Press, Amsterdam.

Uwe Quasthoff, Matthias Richter, and Christian Biemann. 2006. Corpus Portal for Search in Monolingual Corpora. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006*, Genoa.

Benoît Sagot. 2005. Automatic Acquisition of a Slovak Lexicon from a Raw Corpus. In *Text, Speech and Dialogue*, volume 3658 of *Lecture Notes in Computer Science*, pages 156–163. Springer.

Vít Suchomel and Jan Pomikálek. 2011. Practical Web Crawling for Text Corpora. In *Proceedings of the Fifth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2011*, Brno. Masaryk University.

Zafar D. Usmanov and Khisrav A. Evazov. 2011. Computer correction of Tajik text typed without using special characters. *Reports of Academy of Sciences, Republic of Tajikistan*, 54(1):23–26. (In Russian).

Zafar D. Usmanov, Sanovbar D. Kholmatova, and Odilchodja M. Soliev. 2007. Tajik–Russian computer dictionary. National patent 025TJ, National Patent Information Centre, Republic of Tajikistan. (In Russian).

Zafar D. Usmanov, Odilchodja M. Soliev, and Khurshed A. Khudoiberdiev. 2008. Russian–Tajik computer dictionary. National patent 054TJ, National Patent Information Centre, Republic of Tajikistan. (In Russian).

Zafar D. Usmanov, Gulshan M. Dovudov, and Odilkhodja M. Soliev. 2011. Tajik computer morfoanalyzer. National patent ZI-03.2.220, National Patent Information Centre, Republic of Tajikistan. (In Russian).

Zafar D. Usmanov, Odilchodja M. Soliev, and Gulshan M. Dovudov. 2012. Tajik language package for system OpenOffice.Org. National patent ZI-03.2.222, National Patent Information Centre, Republic of Tajikistan. (In Russian).