UCT JOURNAL OF SOCIAL SCIENCE AND HUMANITIES RESEARCH 2018(02)

Available online at http://journals.researchub.org

Personal names spell-checking – a study related to Uzbek

Isroilov Jasur* , Abdurakhmonova Nilufar

Faculty member of Islamic studies department, Iran-Zahedan University of Medical Science.

ARTICLE INFO

Article history:
Received 02 Dec 2017
Received in revised form 04 Jan 2018
Accepted 20 Feb 2018

Keywords:
Database,
Uzbek names,
Suffix,
Prefix,
Dictionary.

ABSTRACT

Objective: In the paper we describe the development process of the dictionary of Uzbek names and surnames. **Methodology**: The dictionary is created to support the identification of personal names in Uzbek texts, and to aid the spell-checking of texts written in Uzbek. **Results**: Apart from discussing the development process, we also evaluate the dictionary by performing a set of experiments. **Conclusion**: We verify whether the information collected in the dictionary can be successfully used to find and, if needed, correct the misspelled names and surnames.

1. Introduction

In today's world, we are surrounded by information coming from many different sources. These sources include verbal and non-verbal communications, as well as various textual forms. Take short messages, emails, twits or newspapers as some of the many text-based communication examples (Waterman, 2002). What is more, we are both the recipients, and the producers of these pieces of information, a big part of which is generated through social media. However, even in the shortest messages we produce, we are prone to making spelling, punctuation or grammar mistakes. Some of them can seem unimportant. But when it comes to writing someone's name incorrectly, the matter becomes much more serious (Kapinus et al., 2008). The motivation behind our current research stems from the need for new spell-checking methods and tools. In particular, we search for the tools that can also support languages with fewer number of linguistic resources, such as the Uzbek language. Uzbek belongs to the family of Turkic languages. It is written using Latin and Cyrillic scripts, which in itself poses problems during transliteration as stated in (Shaalan et al., 2012). The literature studies of Uzbek are mainly related to machine translation (Davrondjon & Janowski, 2002) or corpora alignment tasks. Although many existing texts editors support spellchecking, none of them supports Uzbek. As the response to this problem, we present a dictionary of names and surnames used in Uzbek. The dictionary has been developed manually based on (Bahodir o'g'li, 2018). We present the process of the development of this dictionary, pointing out some of the features that are typical to Uzbek language. We also discuss some statistics describing dictionary's size and content. Finally, to show that the dictionary can support spell-checking we perform a set of experiments. In these experiments, we evaluate whether the dictionary can aid the tasks of personal names identification and correction. The paper is divided into 5 sections. In Section 2. we review the literature related to spell-checking methods and tools. In Section 3. we describe the development process of the dictionary and some statistics about it. In Section 4., we present the results of conducted experiments. Section 5. contains conclusions and future research perspectives.

1.1 Related Works

* *Corresponding author:* Jasur@uzhk.ru
DOI: https://doi.org/10.24200/jsshr.vol6iss02pp1-6

There are two types of spelling mistakes: non-word mistakes (Doğruöz, 2010) and real-word mistakes. The non-word mistakes are the errors, when a word does not belong to the language. The real-word mistakes are those in which a word belongs to the language, but is not properly used in the given context.
There are various methods for error detection and error correction, for both types of errors. Among the methods for error detection of non-word errors one can find dictionary lookup methods and N-gram based methods (Trevisani, 2010). Statistical and machine learning methods, N-gram based algorithms and noisy channel models have been used for real-word spellchecking (Tantug, 2010). The error correction methods include methods based on the edit distance or language grammar rules (Akın, 2007).
Due to space constraints we have to refer the reader to (Atkinson, 2006) and their references for further reading on the existing spellchecking methods.
Our current research is closely related to the non-word error detection by means of dictionary lookup. In particular, we focus on the detection and correction of errors related to Uzbek names and surnames.

2. Materials and methods

2.1 Dictionary Development

To facilitate the spell-checking of Uzbek texts we decided to build a complete dictionary of names and surnames that are used in Uzbekistan and hence can appear in texts written in Uzbek. Currently, the dictionary is available as a spreadsheet, although porting it to some Relational Database Management System is possible without much additional effort (Akhatov, 2012).
The development of the dictionary has been divided into the following stages:

- ✓ development of the part related to names
- ✓ development of the part related to surnames
- ✓ development of the part related to both names and surnames.

2.2 Development of the name part

The development process regarding the part of the dictionary related to names, has been divided into the following steps:

1. The dictionary was filled in with the names that are listed as being currently used in Uzbekistan. The list of used names was obtained from (Küster, 1999). The list contained over 15 000 names.
2. The list of names was supplemented by the names produced according to the following rules:
   (a) Uzbek names consist of the base part, followed by an additional suffix. Therefore, for each name from the initial list we created "an extended version", which consisted of the name followed by one of the predefined suffixes.
   (b) Uzbek names can be also formed by preceding the base part of the name with an additional pre- fix. Following this rule, the previously created set of names was complemented with the names formed by prefixing. However, rules 2a and 2b are mutually exclusive. As a consequence, we had to remove from the list of the names formed according to rule 2a, those names that are subject to rule 2b.
   (c) Finally, if the base part of the name already contains a particular suffix then the additional suffix cannot be added. Therefore, from the list of names produced in the previous steps we had to remove the names violating this rule.

By following the above rules, the number of names became greater than 35 000. The suffixes and prefixes used in the process of extending the initial set of names are gathered in Tab. 1 and 2.

Table 1. Suffixes used in rule 2a to form the Uzbek names

Name-forming suffixes
ali, bakir, begim, bek, beka, berdi, bergan, bibi, bonu, boy, gul, g'ozi, hoji, jahon, jamol, jon, mirza, oy, poshsha, qul, sher, sho, shoh, to'ra, xo'ja, xon

To give the reader some better understanding of how the words shown in Tab. 1 and 2 were used with the rules 2a-2c let us consider the following examples:
Example 1 Let the base name be Jasur. This name is subject to the suffixing rule (i.e. rule 2a). Hence, we can create the following names out of this base name:
Jasurbek, Jasurjon, Jasurxon.
Example 2 Let the base name be Zohid. This name in turn is subject to the prefixing rule (i.e. rule 2b). So, we can create the following names out of this base name: Abduzohid, Mirzohid.
Example 3 Let the base name be Otabek. This name is subject to the last rule (i.e. rule 2c), since it already contains the suffix bek. Therefore, this base name can neither form additional names by prefixing nor by suffixing.

2.3 Development of the surname part

We extended the dictionary described in the previous subsection by adding the information on the surnames that can appear in Uzbek texts. To this aim we have observed the following rules:

1. In the former Soviet Socialist Republic (SSR) surnames were formed by adding certain suffixes, such as: *-ev, -yev, -ov, -v,* to the names of the grandfathers. However, by analogy to rule 2c, this rule could be applied only to surnames, which did not already have such an ending.

2. To distinguish between surnames of men and women, an additional suffix *-a* should be added to form a female surname.

Table 2. Prefixes used in rule 2b to form the Uzbek names

Name-forming prefixes
abdu, abu, ali, amir, anna, aziz, baxt, bayram, bek, berdi, besh, bibi, bobo, bo'l, bolta, bo'ron, boy, chaqqon, chin, davlat, dil, din, do'st, dur, egam, eshon, eson, fayzi, fozil, g'affor, g'ani, g'iyos, g'ozi, g'ulom, habib, hakim, hamid, hayit, hazrat, hoji, ibni, imom, iris, iso, jahon, jon, jo'ra, kumush, mehmon, mir, mirza, muhammad, mulla, murod, muso, nabi, nafas, nazar, niyoz, nor, nur, ochil, omon, oq, oraz, o'rin, o'roz, ortiq, oxun, ozod, polvon, qilich, qodir, qora, qori, qo'sh, qo'zi, qul, qurbom, qutli, rahim, rahmon, rajab, rasul, ro'zi, safar, sahad, said, salom, sari, sayid, soat, sohib, sulton, temir, tilla, to'g'ay, toji, to'ra, tosh, to'xta, to'y, turdi, tursun, tuvoq, ubay, ulug, umar, umr, usmon, usta, uzoq, vali, xalil, xayri, xidir, xo'ja, xol, xon, xudoy, yaxshi, yazdon, yo'l, yor, yoz, yusuf, zamon, zay, zikr, ziyo, zohid, zokir, zul

By applying the two rules described above we have inserted over 46 000 surnames to the dictionary.

Let us note, that the choice of an ending for a particular base surname is governed by a set of simple rules, which are summarized in Tab. 3. These rules state that, if the surname ends with one of the letters listed in the second column of Tab. 3, then we should add the suffix given in the first column. Note however, that letter *f* is marked with an asterisk ('*'). This is to show that if the surname ends with *f*, then we need to replace this letter with the suffix *-pov*, instead of just appending the suffix.

Table 3. Surname forming rules

Suffix	Suffix Last letter of the surname
-ev	h, y
-ov	b, d, g, g', j, k, l, m, n, p, r, q, s, t, x, z
-pov	f*
-yev	a, e, i, o', u
-v	o

Let us now provide a few examples showing how the aforementioned rules were applied in practice (for more examples see (Lawson et al., 1969)).

Example 4. Let the base part of the surname (i.e. grandfather's name) be Isroil*ov*. Then by rule 1 the new male surname is Isroilov (see the second row in Tab. 3). By rule 2 the female surname becomes Isroilov*a*.

Example 5. Let the base part of the surname be Orif. Then taking into account the exceptional case shown in the third row of Tab. 3, the male surname is Ori*pov*, while the female surname is Oripov*a*.

Example 6. Let the base part of the surname be Bektil*ov*. Since it already ends with suffix *-ov*, only the female surname Bektilov*a* will be created.

2.4 Development of the combined part

At the final stage we followed the Uzbek grammar rules, which say that inflected names and surnames acquire additional suffixes. As the result, the size of the dictionary grew up to over 106 names and surnames.

The summary of suffixes that may be combined with the names and surnames is shown in Tab. 4. In case of suffixes *-ng* and *-ngiz*, depending on the name or surname to which the suffix is applied, it is sometimes necessary to prepend the suffix with letter i (see Examples 7 and 8).

Table 4. Inflection-related suffixes for names and surnames

Suffix	Suffix Last letter of the surname
names	*-da, -dan, -dek, -ga, -gacha, -ng/-ing, -ngiz/-ingiz, -ni, -niki, -ning, -siz*
surnames	*-da, -dan, -dek, -ga, -gacha, -lar, -ni, -niki, -ning, -siz*

Example 7. Let us consider the name Asror. Since it is a name, then the following inflected forms are all valid, and thus can appear in Uzbek texts: Asror*da*, Asror*dan*, Asror*dek*, Asror*ga*, Asror*gacha*, Asror*ing*, Asror*ingiz*, Asror*ni*, Asror*niki*, Asror*ning*, and Asror*siz*.

Example 8. Let us consider the name Anora. The following inflected forms are valid forms for this name: Anora*da*, Anora*dan*, Anora*dek*, Anora*ga*, Anora*gacha*, Anora*ng*, Anora*ngiz*, Anora*ni*, Anora*niki*, Anora*ning*, and Anora*siz*. Note that the suffixes *-n*g and *-ngiz* were used here, instead of *-ing* and *-ingiz*.

Example 9. Let us then consider the surname Aliyev. Then the following inflected forms are allowed: Aliyev*da*, Aliyev*dan*, Aliyev*dek*, Aliyev*ga*, Aliyev*gacha*, Aliyev*lar*, Aliyev*ni*, Aliyev*niki*, Aliyev*ning*, and Aliyev*siz*.

Latin alphabet (excluding letters C and W), and 4 additional symbols: O', G', Sh and Ch. The distributions of male names, female names and surnames among all 28 symbols of the alphabet are shown as histograms in Fig. 2, 3 and 4.

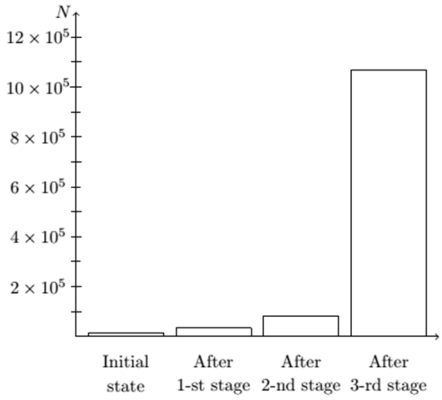

Figure 1. Size expansion of the dictionary of Uzbek names and surnames (N denotes number of entries)

To sum up, in Fig. 1 we show how the three stages contributed to the expansion of the dictionary. We have also analyzed the distribution of male names, female names and surnames among the letters of Uzbek alphabet. This way we have gained some insight into the most popular names and surnames. The Uzbek alphabet consists of 24 letters of the Latin alphabet (excluding letters C and W), and 4 additional symbols: *O‘, G‘, Sh* and *Ch*. The distributions of male names, female names and surnames among all 28 symbols of the alphabet are shown as histograms in Fig. 2, 3 and 4, respectively. The histograms do not include the inflected forms.

From the analysis of the histograms it follows that most names and surnames start with letter *M*. In particular, it is the initial letter of 15% of names and surnames. On the other end we find letter *L*, which appears in less than 1% of names, as well as letters *U, V, O‘, G‘* and *Ch*, which are the initial letters of around 1% of names and surnames.

Figure 2: Male names distribution among the letters of Uzbek alphabet (*N* denotes number of entries)

Figure 3: Female names distribution among the letters of Uzbek alphabet (*N* denotes number of entries)

Let us also observe that the more male names we have for a particular initial letter, the more surnames we have for this letter as well. This follows directly from the way the surnames are generated. However, when we take into account the combined number of male and female names starting with some letter, then the above statement no longer holds.

3. Discussion and results

3.1 Experiments

We conducted the experiments using a simple program written in Java, which allowed to load the dictionary into memory, parse input text files and identify the correctly and incorrectly spelled names and surnames. The dictionary loading process has been realized using Apache POI library version 3.16 (The Apache Software Foundation, 2017). Due to memory constraints we had to divide the dictionary into two separate spreadsheets, containing names and surnames. It took around 57 seconds to load the names part and approximately 63 seconds to load the surnames part.

The parsed text files contained short stories in Uzbek. We found the stories in (To’rayev and Pakhunov, 2016). For the purpose of the experimental evaluation we used five short stories: *Hur qiz* (Hello girl), *“Zingerli boy”* (Rich with Zinger), *O’tmishdan ertaklar* (Fairy tailes), *Hasan bilan Husan* (Hasan and Husan) and *Qushcha* (Bird).

Based on the original texts, we generated five modified versions for each text. In some words, we changed the initial letter into uppercase one (with the probability 0.33). We also made some of the words lowercase (again with the probability 0.33). This way we have obtained 30 test files. In the sequel we will refer to the groups of input files resulting from each story by *S1, S2, . . ., S5*.

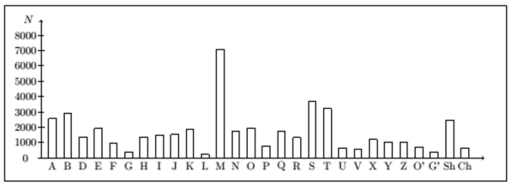

Figure 4. Surnames distribution among the letters of Uzbek alphabet (*N* denotes number of entries)

For each of the input files we have determined the total number of words N, the total number of unique words N', the total number of words starting with a capital letter M and the total number of unique words starting with a capital letter M'. The average values and standard deviations for the four parameters, computed separately for each input set are gathered in Tab. 5.

Table 5: Word statistics for the input sets (avg. – average value, st. dev. – standard deviation)

Statistic	S_1	S_2	S_3	S_4	S_5
avg. *N*	1966	1009	448	410	533
st. dev. *N*	0	0	0	0	0
avg. *N'*	1343	761	384	270	400
st. dev. *N'*	46	25	9	11	9
avg. *M*	664	337	151	146	167
st. dev. *M*	182	94	45	32	46
avg. *M'*	474	263	134	97	138
st. dev. *M'*	145	81	41	29	43

The identification of properly and improperly spelled names and surnames consisted in comparing the words starting with an uppercase letter to the dictionary contents, and comparing the words starting with a lowercase letter to the dictionary contents. We performed the second step to detect potentially misspelled names and/or surnames.

The summary of the experimental results is shown in Fig. 5. The figure contains the information on the number of properly identified names and surnames (true positives, TP), the number of words improperly identified as names or surnames (false positives, FP) and the number of corrected words. On the X axis of each subplot, label S_{ij} corresponds to the j-th member of the i-th set of texts, where $i = 1, 2, \ldots, 5$ and $j = 0, 1, \ldots, 5$. The S_{i0}, for $i = 1, 2, \ldots 5$, corresponds to the original story in each set. Hence there are no corrected words for these elements.

From the results shown in Fig. 5 it follows that using the dictionary we were able to correct all misspelled words in all cases. On the other hand, it is also worth noticing that we were also suggested to correct some properly spelled words due to their polysemy (see the black bars in Fig. 5). Therefore, we conclude that the results of the comparison with the dictionary contents should always be verified by the user. Otherwise, apart from correcting the misspelled words, we can also introduce some new errors into our texts.

4. Conclusion

In the paper we have discussed the development process of the dictionary of Uzbek names and surnames. The dictionary has been developed as a spreadsheet. The data included in the dictionary covers the names currently used in Uzbekistan, involving also the names resulting from inflection.

The dictionary has been evaluated by a set of experiments. We have verified whether the information contained in the dictionary can be used to identify the names and surnames appearing in Uzbek texts. We have also checked whether the dictionary can support the task of spell-checking of the texts written in Uzbek, with respect to personal names. The experiments have shown that the dictionary can successfully aid both tasks.

In the future we plan to perform the experiments on larger sets of texts obtained from different sources. We also intend to check whether applying measures of text similarity, such as the edit distance, can help to improve the spell-checking process even further.

REFERENCES

Akhatov, A. R. 2012. Methods for controlling the authenticity of textual information transfer on the basis of statistical and structural redundancy. International Journal of Automation and Computing, 9(5), 518-529.

Akın, A. A., & Akın, M. D. 2007. Zemberek, an open source nlp framework for turkic languages. Structure, 10, 1-5.

Atkinson, K. 2006. Gnu aspell 0.60. 4.

Bahodir o'g'li, I. J. 2018. PERSONAL NAMES SPELL-CHECKING-A STUDY RELATED TO UZBEK. Министерство высшего и среднего специального образования Республики Узбекистан Национальный офис Erasmus+ в Узбекистане Национальная команда экспертов в области высшего образования, 33.

Davrondjon, G., & Janowski, T. 2002, October. Developing a Spell-Checker for Tajik using RAISE. In International Conference on Formal Engineering Methods (401-405). Springer, Berlin, Heidelberg.

Doğruöz, A. S. 2010. Analyzing language change in syntax and multiword expressions: A case study of Turkish Spoken in the Netherlands. In First Workshop on Language Resources and Technologies for Turkic Languages (20).

Kapinus, O., Michailov, I., & Yurin, I. 2008. Russian IM-applications:«Mail. ru Agent».

Küster, M. W. 1999. Multilingual Ordering—the European Ordering Rules. Multilinguale Corpora: Codierung, Strukturierung, Analyse: 11. Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung, 21-33.

Shaalan, K. F., Attia, M., Pecina, P., Samih, Y., & van Genabith, J. 2012, May. Arabic Word Generation and Modelling for Spell Checking. In LREC (719-725).

Tantug, A. C. 2010. A probabilistic mobile text entry system for agglutinative languages. IEEE Transactions on Consumer Electronics, 56(2), 1018-1024.

Trevisani, T. 2010. Land and power in Khorezm: Farmers, communities, and the state in Uzbekistan's decollectivisation (23). LIT Verlag Münster.

Waterman, S. 2002. Scholar, manager, mentor, mensch: Saul B. Cohen. Political Geography, 21(5), 557-572.

How to Cite this Article:

Jasur I., Nilufar A., Personal names spell-checking – a study related to Uzbek, UCT Journal of Social Sciences and Humanities Research 6(2) (2018) 1–6.