УДК 004

SPELL CHECKING APPLICATION IN KAZAKH LANGUAGE

Aitimov A.K. *MSc student,* **Amirgaliyev Y. N.** *Prof*
Suleyman Demirel University

Түйін

Бұл мақаланың басты мақсаты Қазақ тіліне арналған сөз және сөйлемнің жазылуын тексеретін бағдарламасын құру. Мақалада бағдарламаның басты модульдері және олардың құрылыстары баяндалған. Сонымен қатар, тексеру көрсеткіштері де қамтылған.

Кілт сөздер: тексеру емлені , қазақ тілі, сөз тексеру, сөйлем тексеру , емлені тексерушіні қолдану.

Резюме

Эта статья предназначена для создания программы проверки правописания Казахского языка. Здесь рассмотрено главные модули и их архитектуры. А также включены результаты тестов.

Ключевые слова: проверка произношения, казахский язык, проверка слова, проверка предложения, запись приложения информатора.

Abstract. *In computing, a spell checker (or spell check) is an application program that flags words in a document that may not be spelled correctly. Spell checkers may be stand-alone, capable of operating on a block of text, or as part of a larger application, such as a word processor, email client, electronic dictionary, or search engine. However, in many languages such kind of application is not developed yet, and Kazakh language is one of such languages.*

This paper is aimed to create spell checker application both word and sentences. Also the paper includes description of main modules which were used in creating the application, and result of the application test.

Key words: *spell check, Kazakh language, word check, sentence check, spell checker application.*

Introduction

Kazakh language

Kazakh is generally verb-final, though various permutations on SOV (subject–object–verb) word order can be used [1]. Verbal and nominal morphology in Kazakh exists almost exclusively in the form of agglutinative suffixes.

Kazakh has 7 cases [2]. Case endings are applied only to the last element of a noun phrase—e.g., a noun, an abject, or a nominalised verb phrase. The endings are applied to a word ending in a front vowel, a word ending in a back vowel, a word ending in each of those with a voiced consonant, and a word ending with each of this and an unvoiced consonant [3]. Note that the suffixes for the instrumental case do not follow vowel harmony—the vowel is a front vowel regardless of the other vowels in the word.

Kazakh language may express different combinations of tense, aspect, and mood through the use of various verbal morphology or through a system of auxiliary

verbs, many of which might better be considered light verbs [4]. For example, the (imperfect) present tense in Kazakh bears different aspectual information depending on whether basic present-tense morphology is used, or one of (commonly) four verbs is used

Spell checker application

In computing, a spell checker (or spell check) is an application program that flags words in a document that may not be spelled correctly. Spell checkers may be stand-alone, capable of operating on a block of text, or as part of a larger application, such as a word processor, email client, electronic dictionary, or search engine.

The first spell checkers were widely available on mainframe computers in the late 1970s. A group of six linguists from Georgetown University developed the first spell-check system for the IBM corporation [5].

The first spell checkers for personal computers appeared for CP/M and TRS-80 computers in 1980, followed by packages for the IBM PC after it was introduced in 1981. Developers such as Maria Mariani [6], Random House[7], Soft-Art, Microlytics, Proximity, Circle Noetics, and Reference Software rushed OEM packages or end-user products into the rapidly expanding software market, primarily for the PC but also for Apple Macintosh, VAX, and Unix. On the PCs, these spell checkers were standalone programs, many of which could be run in TSR mode from within word-processing packages on PCs with sufficient memory.

A basic spell checker carries out the following processes:

- It scans the text and extracts the words contained in it
- It then compares each word with a known list of correctly spelled words (i.e. a dictionary). This might contain just a list of words, or it might also contain additional information, such as hyphenation points or lexical and grammatical attributes.
- An additional step is a language-dependent algorithm for handling morphology. Even for a lightly inflected language like English, the spell-checker will need to consider different forms of the same word, such as plurals, verbal forms, contractions, and possessives. For many other languages, such as those featuring agglutination and more complex declension and conjugation, this part of the process is more complicated.

It is unclear whether morphological analysis provides a significant benefit for English, though its benefits for highly synthetic languages such as German, Hungarian or Turkish are clear.

As an adjunct to these components, the program's user interface will allow users to approve or reject replacements and modify the program's operation.

An alternative type of spell checker uses solely statistical information, such as n-grams. This approach usually requires a lot of effort to obtain sufficient

statistical information and may require a lot more runtime storage. This method is not currently in general use.

In some cases spell checkers use a fixed list of misspellings and suggestions for those misspellings; this less flexible approach is often used in paper-based correction methods. Application such as Kazakh spell checking application is not developed. there is only few application and they work very primitive. Their database is very poor and checks only words without checking whole sentence. The best spell checker in Kazakh language is plug-in in Microsoft word application. It has good database of words but still does not check sentences. Because of that our project is aimed to fill the gap and create an application that will face all modern demands of Kazakh language.

Application

Every text consist of many sentences which include lots of words. To carry out spell checking there is need to check both words and sentences. Because of that the application consist of two main modules: word module and sentence module.

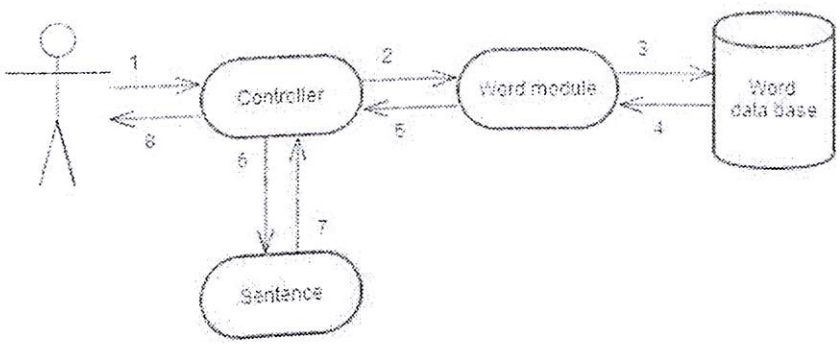

Figure 1. The application's structure

Word Module

According to figure 1 word module takes every word and checks for grammar mistakes. Word may consist of root, suffix, prefix and etc. Application takes one word look for same word in database. If chosen word is only root with no suffixes then application will find word in database and if there is no mistakes application show that there is no mistakes. But if word have mistake application will show it.

If word with suffixes is chosen then application will subtract every letter until there will be coincidence with data base and application will show that word have suffixes and spell checking will look if word is written by grammar rules, if yes then word is correct, if not then incorrect.

Sentence Module

After finishing word checking all text goes under sentence grammar checking. Application checks if words in sentence are on their grammatical places, if words on sentence are correctly connected with each other. In kazakh language there are four main types suffix: "септік", "тәуелдік", "жіктік", "көптік". Application looks only for this four types of suffix, other types of suffix have not implemented yet.

Test and evaluation To test the application 2685 words and 376 sentences was used. 731 out of 2691 words were complex (біріккен сөздер, қос сөздер, қысқарған

сөздер) and 1960 words were simple words. 234 sentences were complex sentences (салалас сөйлем, сабақтас сөйлем, аралас сөйлем) [7].

Result of the test was: 341 of 2685 words were incorrect and 2344 words were correct and 309 of 376 sentences were incorrect and 67 sentences were correct as shown in figure 2.

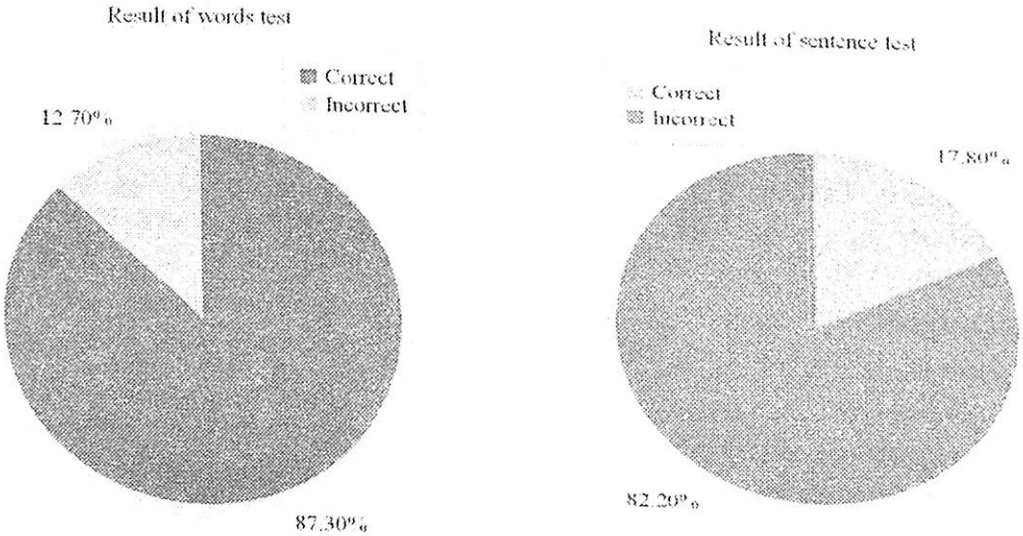

Figure 2. Test results

Conclusion In this paper we present application which checks texts for mistakes in Kazakh language. This application has two main module: words and sentence modules. Word module uses dictionary to check and sentence module uses different synthetic rules to check. The application was tested and results are: word module's efficiency is 87.3% and sentence module's efficiency is 17.8%.

References

[1]. Beltranslations.com

[2]. Базылхан, Б. 1977. Қазақ, Монгол Тілі Граммати асынан Қысқаша Салыстырмалы Белгілеулер. In: Қазақша-Монголша Сөздік. Уланбатыр.

[3]. Балакаев, М.Б., Н.А. Баскаков, С.К. Кенесбаев ed. 1962. Современный Казахский Язык. Фонетика и Морфология. Алма-ата.

[4]. Laude-Cirtautus, Ilse. 1974. The Past Tense in Kazakh and Usbek as a Means of Emphasizing Present and Future Actions. Central Asiatic Journal 18:149-158.

[5]. "Georgetown U Faculty & Staff: The Center for Language, Education & Development". Retrieved 2008-12-18., citation: "Maria Mariani... was one of a group of six linguists from Georgetown University who developed the first spell-check system for the IBM corporation."

[6]. Advertisement (November 1982). "The Spelling Bee Is Over". PC Magazine. p. 165. Retrieved 21 October 2013.

[7]. Серғалиев, Мырзатай. 1992. Қазақ Тілі. Методикалық оқу құралы. Алматы