PAN LOCALIZATION PROJECT

MONGOLIAN SPELL CHECKER

Phase 2.1

October 11, 2008

CENTER FOR RESEARCH ON LANGUAGE PROCESSING
NATIONAL UNIVERSITY OF MONGOLIA, ULAANBAATAR, MONGOLIA

Table of Contents

List of Figures and Tables

Abstract

This document provides an overview of a dictionary-based spell checker for Mongolian. The spell checker is discussed focusing on its design, spell-checking the corpus texts and its further development.

1. Introduction

In third phase of the project for building a 5 million words corpus, we have developed a dictionary based spell-checker named Moozuur, and spell-checked the words of the corpus. Around 30 thousand words [3, 4] are included in the dictionary of the spell-checker at first. Currently, the dictionary contains around 100 thousand words by obtaining during the corpus texts spelling and checking. After that, the ability of Moozuur to correct mistyped words is increased and the time spending on the corpus cleaning is reduced.

Because some part of the corpus has been collected by scanning text sources, spelling and correcting the words of the corpus has took much more time. For the scanned texts, to note, the spell checker needs some special algorithms because the errors made by OCR are different from human typing errors.

The Mongolian spell checker is divided into two main parts, spell-checking an input word and correcting the input if it is incorrect. To check that the input word is typed correctly, the spell-checking part compares it with the words of a dictionary in a tree structure. If the input is incorrect, the correcting part calculates the correct forms for it by using the dictionary, and shows possible correct forms. If any possible correct form of the input is not found, the program considers the input as a new word to its dictionary, or asks the user to correct it by typing. The spell-checker stores the input into the dictionary in both of the above cases.

Moozuur checks and corrects words more than 90 percent of accuracy. In the future, we develop a rule based module on the spell-checker.

2. Spell Checker

High-Level Architecture

There are two general kinds of spell-checker: one is dictionary-based, and another is rule-based [1, 2]. We have developed a dictionary-based spell-checker for cleaning a Mongolian corpus. The overall system is comprised of a user GUI, a word speller, a word corrector, a lexicon and corpora as shown in Figure 1.

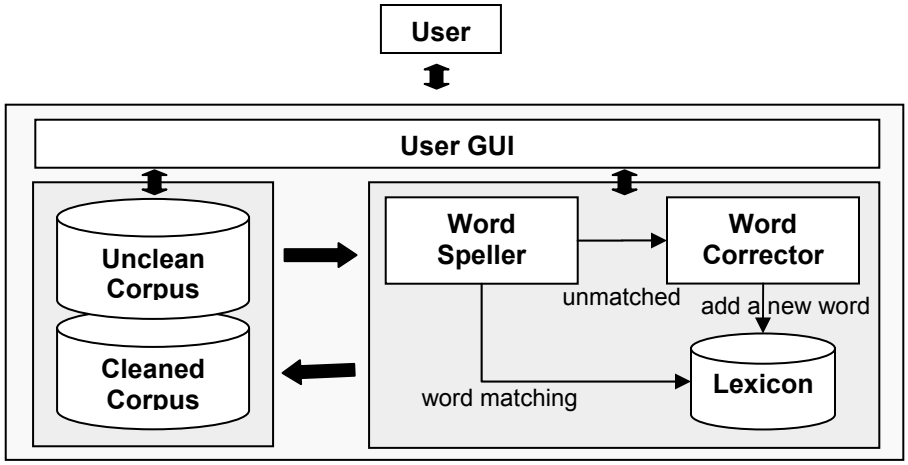

Figure 1. The high-level architecture of Mongolian Spell-checker

Low-Level Architecture

Word speller

Word speller is an important part of the system. It will perform an operation of checking an input word that is given by a user. For doing that, the input word is compared to the words in a dictionary the speller contains. The dictionary is loaded into memory from a lexicon, currently containing around 100 thousand words. Its data structure is a tree shown in Figure 2.

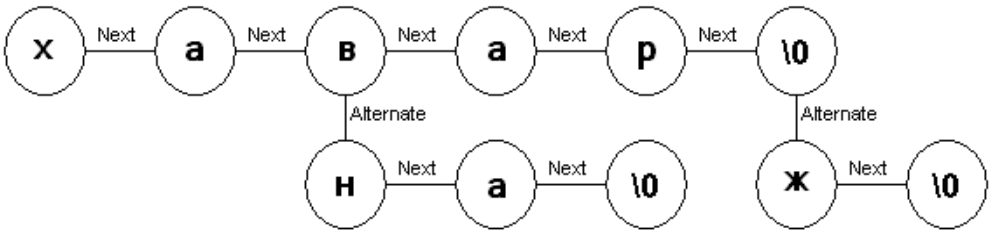

Figure 2. Tree Structure of Dictionary

As shown in Figure 2, three words that are *хавар* (spring), *хаваржс* (to pass the spring) and *хана* (wall), are stored in a tree structure. The number of characters of the strings used in the above example is 15, while that of the tree nodes is 11 including three ends of the tree, '\0' character. As largely increasing the number of the words in dictionary, the number of nodes in the tree is increased less.

Here is an experiment that shows the effective memory usage and data structure. For loading 24385 headwords whose characters are 184777 in total into the memory, the number of tree

nodes allocated into memory is 89250, or two times less than that of the actual words characters. Thus, this kind of tree structure is suitable to storing a large amount of words and retrieving a word from them in a short time.

If the input word is not retrieved from the dictionary, the word is considered a misspelled word and passed into a correcting part of the system, explained in the next part.

Word corrector

Word corrector is a part to correct an input word if it is not found in the dictionary, or could be misspelled. This module of the system finds possible correct forms of the input word by matching it with the dictionary used in the word speller. A correcting algorithm considers four kinds of misspelling errors. First, one of the letters in a word is mistyped, for example, a word *father* can be mistyped as *tather*. Second, a letter is mistakenly inserted into a word, for example, *father* can be mistyped as *fatherr*. The third error type is that one letter of a word is missed, for example, *father* can be mistyped as *fathr*. Last, two letters of a word can be exchanged, for example, *father* can be mistyped as *fahter* (see Table 1). The algorithm is potential to correct a wrong word whose up to three letters are misspelled.

Table 1. Spelling Error Types

No.	Mistyping Types	Incorrect Form	Correct Form
1.	Wrong character	*tather*	*father*
2.	Inserted character	*fatherr*	*father*
3.	Missed character	*fathr*	*father*
4.	Exchanged characters	*fahter*	*father*

For correcting the error types shown in Table 1, following algorithms are used in the word corrector.

For the first error type, each letter of the mistyped word is replaced with a special character, and the model words whose number is the same to the letters of the mistyped word are created (see Table 2).

Table 2. Algorithm correcting a wrong character error type

Mistyped Word	Model Word	Possible Word
марчин	1. *арчин	Not match
	2. м*рчин	мөрчин
	3. ма*чин	махчин, малчин
	4. мар*ин	Not match
	5. марч*н	Not match
	6. марчи*	Not match

As shown in Table 2, *марчин* is used as an example to show an algorithm correcting a wrong character error type. Its correct form is *малчин*, a herder in English. Letter *л* is mistyped as *р*. To correct that mistyped word, each letter from the beginning of the word is replaced with *, a special character. The number of the created model words is 6 according to the number of *марчин*'s letters. After matching the model words with the dictionary, two model words could be a correct form of the mistyped word. One model word has a form, *мөрчин* (tracker), and another has two forms, *махчин* (carnivorous) and *малчин* (herder), for correcting the mistyped word, respectively. As a result of the algorithm, three possible words are found to be a correct form of the mistyped word, and they are added into the list that is shown to the user who chooses one of the suggestions to correct it.

For the second error type shown in Table 1, the model words are created by deleting each letter of the mistyped word. For example, a word *малячин*, an input word, is mistyped in the inserted error type. Its correct form is *малчин*. There is an inserted letter *я*. According to the algorithm, the

input word consists of seven letters. Thus, there must be six model words that could be possible correct forms of the input word after deleting each letter of it one by one (see Table 3).

Table 3. Algorithm correcting an inserted character error type

Mistyped Word	Model Word	Possible Word
малячин	1. алячин	Not matched
	2. млячин	Not matched
	3. маячин	Not matched
	4. малчин	Matched
	5. маляин	Not matched
	6. малячи	Not matched

As shown in Table 3, there is only one possible word that matches one of the dictionary words, so the possible word is added into a suggestion list that will be shown to a user, and the user choose which possible word is to correct the input word.

For the third misspelling type shown in Table 1, the model words are created by inserting a special character before and after each letter of the input word (see Table 4).

Table 4. Algorithm correcting a missed character error type

Mistyped Word	Model Word	Possible Word
мачин	1. *мачин	Not matched
	2. м*ачин	Not matched
	3. ма*чин	махчин, малчин
	4. мач*ин	Not matched
	5. мачи*н	Not matched
	6. мачин*	Not matched

In Table 4, *мачин* is a mistyped word, and its correct spelling is *малчин*. There are two possible correct word forms of a model word *ма*чин*, and they are added into the suggestion list because they can be a correct form of the input word. Other model words do not match with any word of the dictionary.

For the fourth misspelling type shown in Table 1, the model words are created by exchanging each two letters from the beginning of the word (see Table 5).

Table 5. Algorithm correcting an exchanged error type

Mistyped Word	Model Word	Possible Word
мачлин	1. амчлин	Not matched
	2. мчалин	Not matched
	3. малчин	Matched
	4. мачилн	Not matched
	5. мачлни	Not matched

As shown in Table 5, third model word, *малчин*, is matched with one of the dictionary words, and added into the suggestion list. Other model words are not matched with any of the dictionary words.

Summarizing four kinds of algorithms mentioned above for correcting the input word, the input word is considered as mistyped in all four misspelling types (see Figure 3).

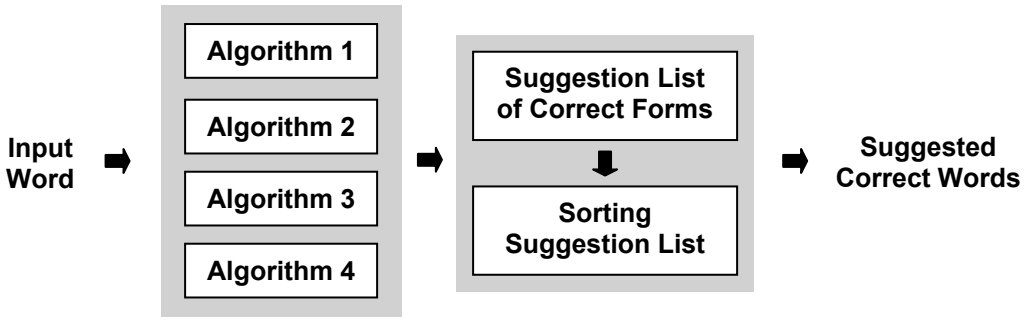

Figure 3. Structure of Word Corrector Module

As shown in Figure 3, the possible correct forms of an input word are calculated using four kinds of algorithms explained in Table 2, 3, 4 and 5, respectively. If some possible correct forms are found, they are placed into the suggestion list that is shown to the user. Before the suggestion list is shown, it is sorted in the order of which form is most probably correct.

If the module does not make any suggestion form to correct the input word, the user is asked to add the word as a new one to the dictionary, or to correct it by typing. The word corrected by typing is also stored into the dictionary if it is not in it.

Lexicon
Lexicon contains words that are loaded into the dictionary of the word speller. Its word entities are enriched new words that occurred during the corpus spelling. Currently, the lexicon contains around 100 thousand word entities.

Unclean corpus
Unclean corpus is a corpus whose texts are just collected and not cleaned for research use. The words in that corpus are not checked and may be misspelled, so they have to be checked and corrected if some of them are incorrect.

Cleaned corpus
Cleaned corpus is a corpus prepared for processing by cleaning and spell-checking its words. It is separately created from the unclean corpus so that two corpora can be compared for statistics of incorrect words and other studies such as typing error features.

User GUI
The main and working views of the spell-checker's user GUI are shown in Figure 4 and 5.

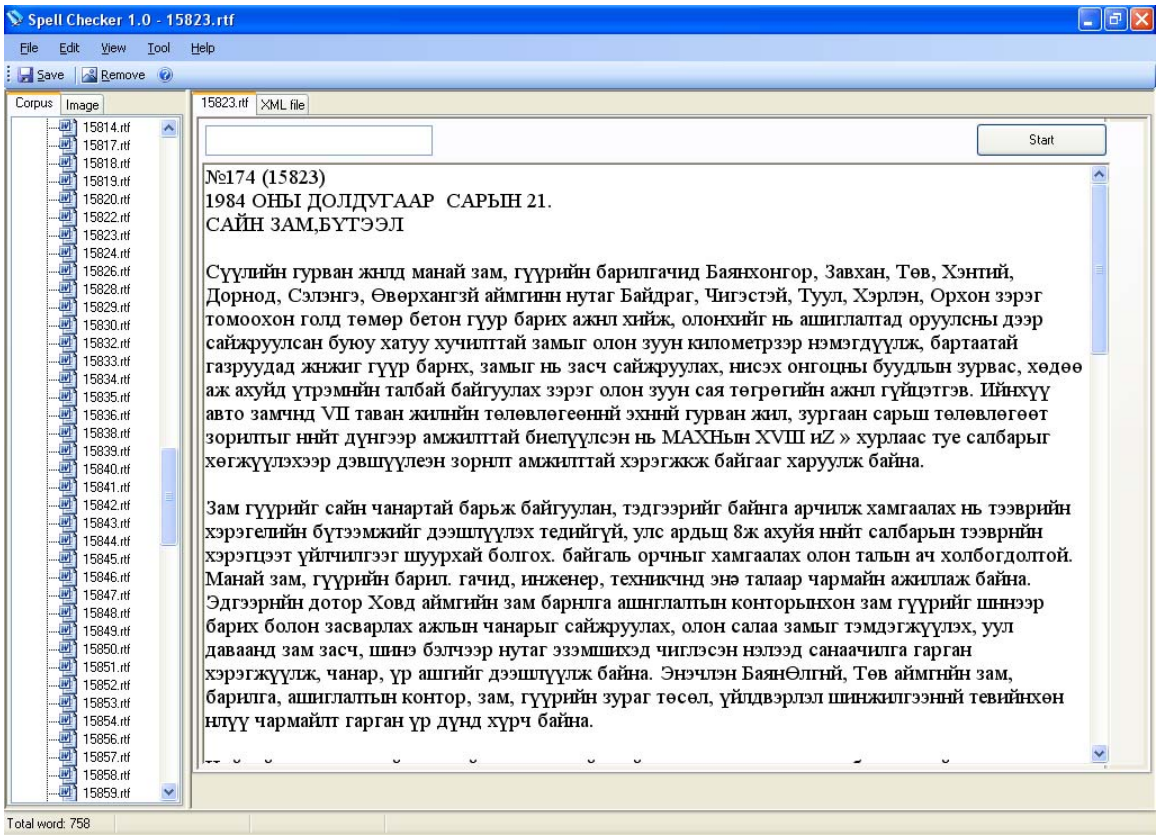

Figure 4. Main View of Spell-Checker

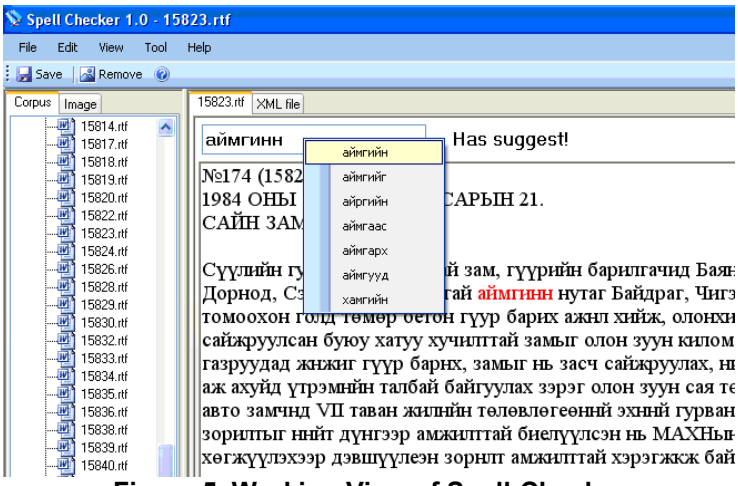

Figure 5. Working View of Spell-Checker

As shown in the above figures, the GUI is divided into two main views that are a corpus view and a text view. The corpus view has two sub views. One is a text view, and it shows the texts of a corpus in a tree view. The user can load the corpus from any directory, and selects to open a text in the corpus. Another sub view shows images of the texts so that the user can compare the words in a text that are not identified with its original image that can be identified if the texts are OCRed.

The text view is a working field of the spell-checker. The user checks and corrects the words on this field after opening a text from the corpus view. As clicking *start* button, it is started an

operation of checking the opened text. Then, the text is tokenized into words, and they are fetched to the word speller mentioned earlier one by one as an input word.

The incorrect words are highlighted with a red color in the text field one by one. The incorrect word is also located in a text box above the text field allowing the user to correct it by typing. The message showing a state whether the suggestions of the correct forms are available or not is appeared on the right side of the text box.

After spelling and correcting a text, the user is asked to add new words found in the spelled text into the dictionary of the spell checker (see Figure 5). The user can review and edit the new words in the list view.

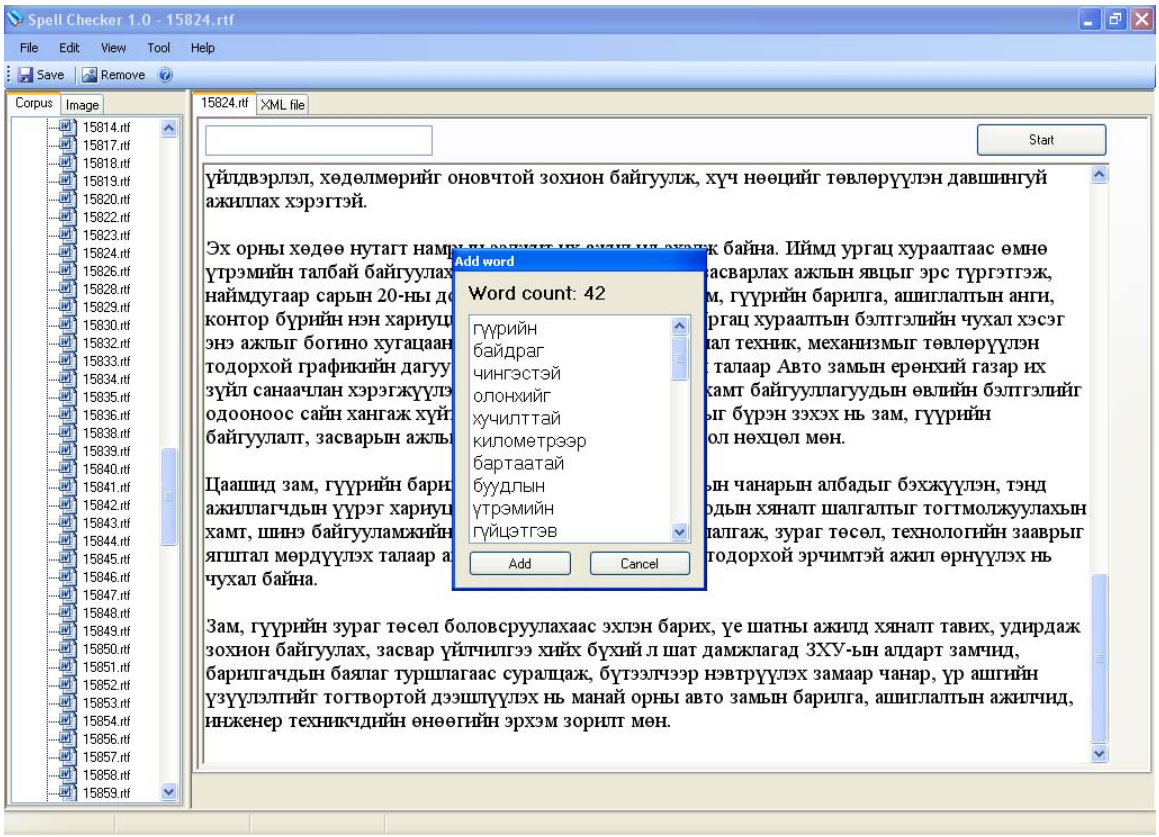

Figure 6. Add Window of Spell Checker

3. Corpus Cleaning with Spell-Checker

The corpus for Mongolian has been collected from both of electronic and hard sources. Thus, the texts in the corpus have words spelled incorrectly. For cleaning the corpus, the spell checker explained in the previous section is used as one of the corpus cleaning tools.

As using the spell-checker, the time to spend on spell-checking texts by hand is reduced ten times. The rate how the spell-checker works is shown in Table 6.

Table 6. Rate of the Moozuur ability to correct the misspelled words

#	Words in text	Incorrect words	Corrected words by Moozuur	Corrected words by hand
1.	683	83	63	20
2.	725	114	95	19
3.	726	257	189	68
4.	805	85	74	11
5.	772	104	86	18
6.	939	208	181	27
7.	728	118	99	19
8.	810	243	177	66
9.	770	122	99	23
10.	729	200	174	26
	768 (100%)	**153 (19.9%)**	**124 (16.1%)**	**29 (3.8%)**

As shown in Table 6, the number of incorrect words is decreased after correcting each text because Moozuur is based on the dictionary. Consequently, some correct words not in the dictionary are considered incorrect. However, these words are added into the dictionary at the first time as occurring. Considering that, Moozuur ability is 95 per cent of accuracy to correct the misspelled words in a typed or electronic text and 80 per cent of accuracy to correct the ones in a text recognized by OCR application, if the dictionary is increased enough.

Some of the incorrect word forms are shown in Table 7.

Table 7. Some of the Errors Occurred in the Text Cleaning

Two Misspelled Letters		Over Two Misspelled Letters	
Incorrect	**Correct**	**Incorrect**	**Correct**
х**о**дөлмөрийн	хөдөлмөрийн	о**й**ол**ь**гн	онолын
суын	су**м**ын	хү**и**үү**я**л**д**йн	хүмүүжлийн
мачид	ма**л**чид	хү**таш**й	хүчний
өн**е**өгийн	өнөөгийн	хэм**гк**ээ**п**ий	хэмжээний
техн**н**к	техник	хэм**яс**э**з**ний	хэмжээний
байгуулл**з**гын	байгууллагын	түмн**нп**х**з**э	түмнийхээ
х**е**дөлмөр	хөдөлмөр	т**е**лө**з**лөлти**н**г	төлөвлөлтийг
үйлд**й**эрлэлийн	үйлдвэрлэлийн	дунд**ь**ш	дундын
меха**ка**зм	механизм	ц**еек**г**у**йг	цөөнгүйг
эмэгтэйчү**у**д**п**йн	эмэгтэйчүүдийн	д**з**в**пчг**лттэй	дэвшилттэй
тэж**з**эл**я**йн	тэжээлийн	**у**й**д**двэрлэлий**и**хээс	үйлдвэрлэлийнхээс
үн**н**лдвэрлэли**н**н	үйлдвэрлэлийн	үйлдвэрлэл**шп1**	үйлдвэрлэлийг
х**о**нд**н**йрүүлдэг	хөндийрүүлдэг	**кее**ц**н**йг	нөөцийг
д**с**рв**н**той	дорвитой	аш**пшт**	ашгийг
ж**н**л**н**йн	жилийн	т**в**л**вэ**	төлөө
шийд**з**эрт**з**й	шийдвэртэй	х**зклсз**эн**н**й	хэмжээний
э**и**ч**н**лгээ	эмчилгээ	х**нш**п	хийж

бзхжүүлэхзд	бэхжүүлэхэд	хэделмернйн	хөдөлмөрийн
нәөцинг	нөөцийг	калып	намын
банлдааиы	байлдааны	хэгшэсэц	хэмнэсэн
төлеер	төлөөр	тэлэвлегеег	төлөвлөгөөг

In Table 7, the misspelled words are shown by classifying according to the number of misspelled letters in a word, less than or equal to two in the first column and more than two in the second column. Moozuur can correct misspelled words like ones shown in the first column, but the misspelled words in the second column are corrected around 80 per cent because they are not real mistakes, instead they are orthographic errors caused by OCR. To improve the correction ability of Moozuur, we are analyzing the information about the misspelled words, and it is possible that Moozuur can correct misspelled words till 90 per cent in case of ORCing.

4. Conclusion

In this project phase, we have developed a dictionary based spell-checker for the Mongolian corpus cleaning. The corpus texts have been collected by using two methods, scanning and electronic sources. The words of the texts that are scanned are around 1 million, and the rest of the words, counted 4 million, in the corpus are collected from electronic sources. Moreover, the scanned texts take more time to spell and correct their words because of the errors caused by OCR application. 20 percent of the scanned texts words are incorrect words and most of the incorrect words are corrected by hand.

Moozuur considers four types of mistyping errors: wrong character, inserted character, deleted character and exchanged character, respectively. For words mistyped in the errors, Moozuur's ability to correct them is more than 90 percent.

In further, the dictionary of the spell checker is extended to be rule-based and updated from the corpus automatically.

Reference

[1] Speech and Language Processing
Daniel Jurafsky, James H. Martin, 2000, Singapore

[2] What is Morphology
Mark Aronoff, Kirsten Fudeman, 2005, United States

[3] Монгол хэлний дүрмийн толь (Mongolian Spelling Dictionary)
Ц.Дамдинсүрэн, Б.Осор, Улаанбаатар, 1983 он

[4] Монгол хэлний хэлзүйн толь бичиг I, II, III, IV боть (Mongolian Dictionary, I-IV Volumes)
Ш.Чоймаа, М.Баярсайхан, Э.Мөнх-Учрал, С.Батхишиг, Улаанбаатар, 2005 он