

MONGOLIAN SPELL CHECKER

J.Purev and Ch. Odbayar

CRLP

Center for Research on Language Processing
National University of Mongolia

(NUM)

OUTLINE

- Introduction to NUMTeam project
- About Mongolian (in brief)
- Spell-checker Design
- Experiment and Corpus Correction
- Conclusion

NUM Team

- NUMTeam, Mongolia
 - At Center for Research on Language Processing (CRLP)
 - First center in Mongolia
 - Established in May 2007
 - Supported by National University of Mongolia and PANLocalization
 - Currently 7 researchers and staffs including computer scientists and linguists
 - Goal: *Mongolian Localization and Processing*
 - Our team joint in PANL10n Project in 2006
- Project Objectives:
 - 5 million tagged words corpus
 - Cleaning Tools and Spell Checker
 - POS Tagset
 - POS Tagger

MONGOLIAN

- Ancient language
- Spoken by around 8 million in Mongolia, Inner Mongolia, China, some parts of Russia
- Included in Altaic family languages
 - Highly agglutinative
- Two alphabets
 - Classical (old)
 - Cyrillic (daily used)



МИНИЙ
МОНГОЛ
БИЧИГ



ᠮᠢᠨᠢᠶᠢ
ᠮᠣᠩᠭᠣᠯ
ᠪᠢᠴᠢᠭ

SPELL CHECKER

- Spell checker
 - Dictionary based
 - Rule based
- At first, a corpus based lemmatizer for Mongolian has been developed
- Need more research work
- Have more complex rules and variations
- Need a large corpus

CORPUS BASED LEMMATIZER

Here is the Screenshot of the Lemmatizer based on the corpus

Tagged corpus

The screenshot shows the 'Dictionary builder' application window. On the left, a file explorer displays a folder named 'POS-III' containing a list of XML files from 16603.xml to 16675.xml. A red arrow points from the text 'Tagged corpus' to this folder. The main window has a menu bar (File, Edit, Tools, Help) and a toolbar. Below the menu bar are four tabs: 'Selecting word', 'Lemmatizing word', 'Dictionary', and 'Morph_generator'. The 'Lemmatizing word' tab is active, showing an 'Extract' table with the following data:

| Surface | Expression | Possible suffixes | Lemma |
|------------------|------------|--------------------|----------------|
| аажмаар | N+l | +аар/ээр/оор/ө... | аажим |
| аваар | N+N | +0 | аваар |
| авагчдын | N+PG | +чууд/чүүд/чуул... | авагч |
| аварга | N+C | +ыг/ийг/г/ | аварга |
| авианы | N+G | +ын/ийн/ы/ий/н/ | авиа |
| авлага | N+N | +0 | авлага |
| автомат | N+G | +ын/ийн/ы/ий/н/ | автомат |
| автоматжилт | N+C | +ыг/ийг/г/ | автоматжилт |
| автоматжуулал... | N+G | +ын/ийн/ы/ий/н/ | автоматжуулалт |
| автомашин | N+C | +ыг/ийг/г/ | автомашин |
| автомашины | N+G | +ын/ийн/ы/ий/н/ | автомашин |
| авьяас | N+S | +аар/ээр/оор/ө... | авьяас |
| авьяасыг | N+C | +ыг/ийг/г/ | авьяас |
| агаар | N+N | +0 | агаар |

CORPUS BASED LEMMATIZER

the POS tags of the words

the word list from the corpus

Lemmatization panel

Possible suffixes of the particular tag

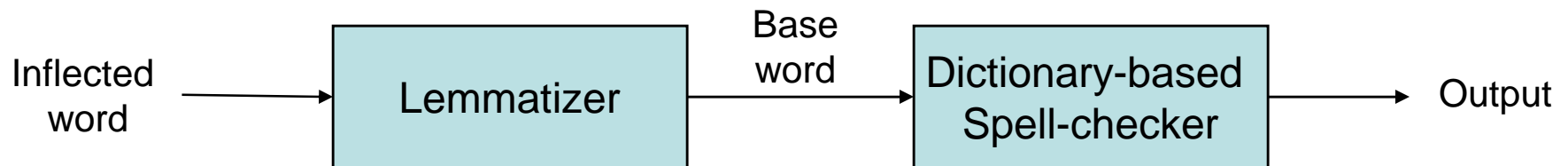
Lemma of the word

The screenshot shows the 'Dictionary Builder' application window. On the left, a file explorer displays a folder named 'POS-III' containing a list of XML files from 16603.xml to 16675.xml. The main window has a menu bar (File, Edit, Tools, Help) and a toolbar. Below the toolbar, there are tabs for 'Selecting word', 'Lemmatizing word', 'Dictionary', and 'Morph_generator'. The 'Lemmatizing word' tab is active, showing a table with the following columns: Surface, Expression, Possible suffixes, and Lemma. The table contains 16 rows of data, each representing a word and its corresponding POS tag, possible suffixes, and lemma.

| Surface | Expression | Possible suffixes | Lemma |
|------------------|------------|--------------------|----------------|
| ажмаар | N+N | +аар/ээр/оор/е... | ажим |
| аваар | N+N | +0 | аваар |
| авагчдын | N+PG | +чууд/чүүд/чуул... | авагч |
| аварга | N+C | +ыг/ийг/г/ | аварга |
| авианы | N+G | +ын/ийн/ы/ий/н/ | авиа |
| авлага | N+N | +0 | авлага |
| автомат | N+G | +ын/ийн/ы/ий/н/ | автомат |
| автоматжилт | N+C | +ыг/ийг/г/ | автоматжилт |
| автоматжуулал... | N+G | +ын/ийн/ы/ий/н/ | автоматжуулалт |
| автомашин | N+C | +ыг/ийг/г/ | автомашин |
| автомашины | N+G | +ын/ийн/ы/ий/н/ | автомашин |
| авьяас | N+IS | +аар/ээр/оор/е... | авьяас |
| авьяасыг | N+C | +ыг/ийг/г/ | авьяас |
| агаар | N+N | +0 | агаар |

RULE-BASED SPELL CHECKER

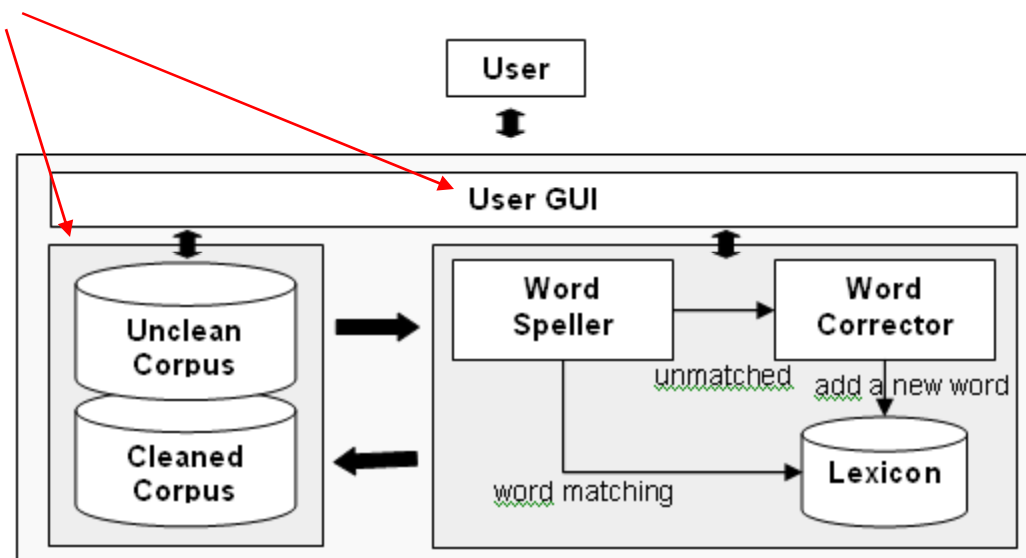
- After developing a Mongolian lemmatizer
- Combine it with the dictionary based spell checker



Currently, we developed dictionary based spell checker
For the rule based spell checker Mongolian lemmatizer is needed
So we are developing lemmatizer
But developing lemmatizer for agglutinative language needs more time

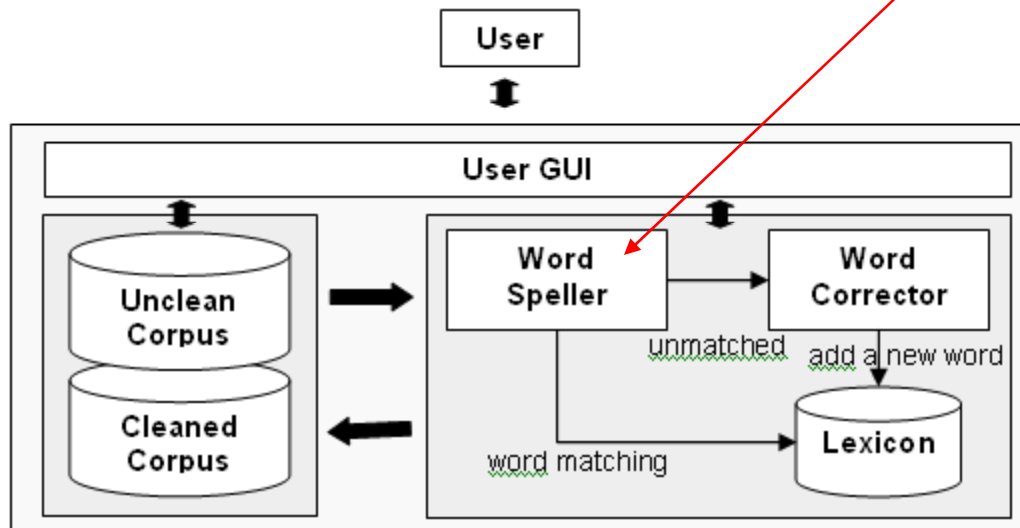
DICTIONARY BASED SPELL CHECKER DESIGN

GUI is designed for Corpus Spelling



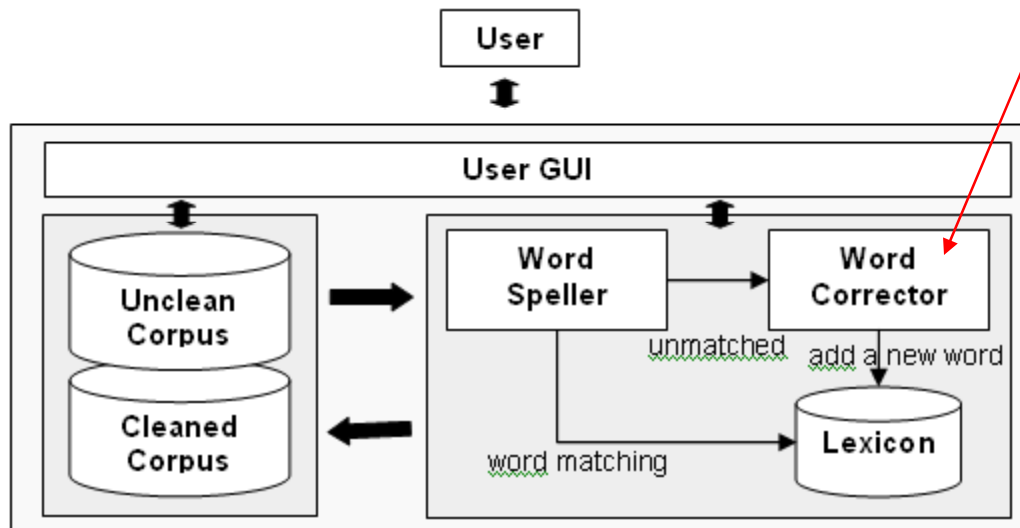
DICTIONARY BASED SPELL CHECKER DESIGN

Check an input word
by using lexicon

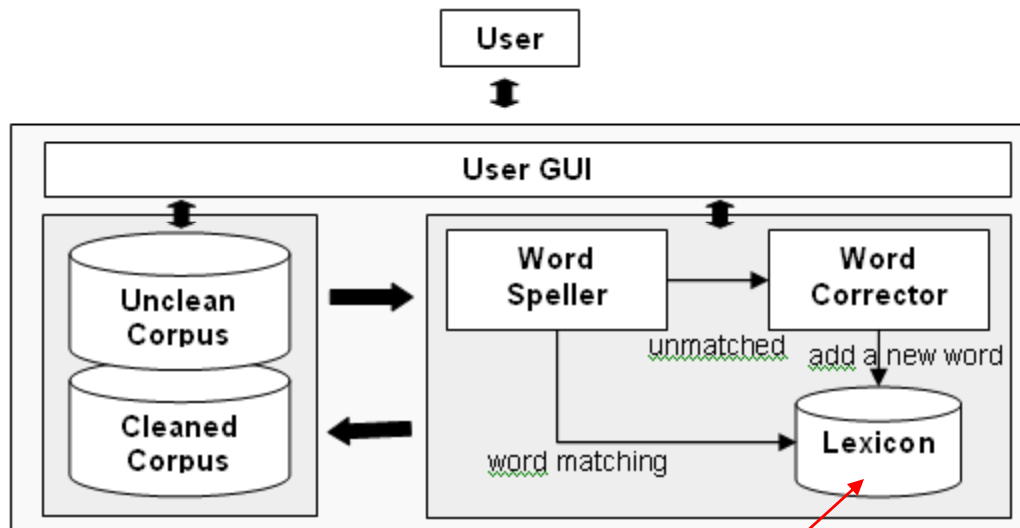


DICTIONARY BASED SPELL CHECKER DESIGN

If the input is incorrect, find its possible correct forms



DICTIONARY BASED SPELL CHECKER DESIGN



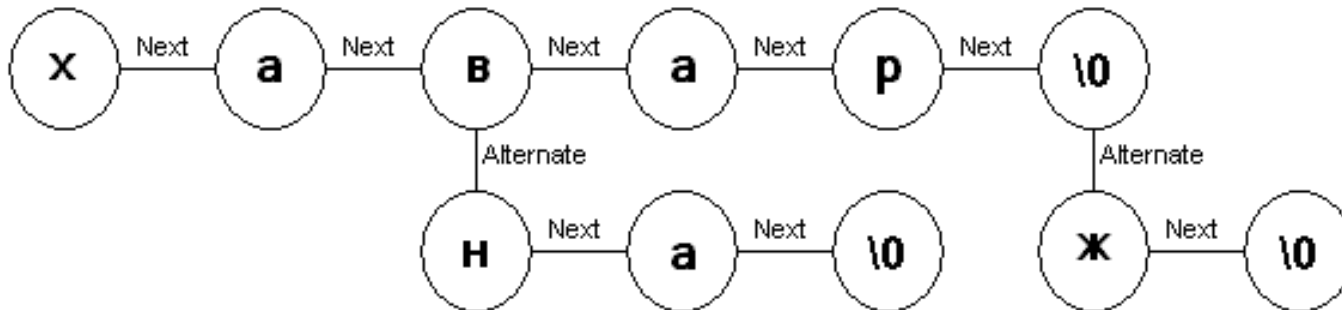
New words could be added
Lexicon is updated from the corpus

-Currently around 160 thousand words

Slide:12

DATA STRUCTURE

- Considered features
 - Fast
 - Less memory (for agglutinative)
 - Applicable for spell-checker
- Example
 - 3 words: *хавар* (spring), *хаварж* (to pass the spring) and *хана* (wall)



WORD CORRECTION [1]

- After analyzing the misspelling errors in the corpus
- Considered four types of errors

| No. | Mistyping Types | Incorrect Form | Correct Form |
|-----|----------------------|-----------------|--------------|
| 1. | Wrong character | <i>t</i> ather | father |
| 2. | Inserted character | father <i>r</i> | father |
| 3. | Missed character | fath <i>r</i> | father |
| 4. | Exchanged characters | fa <i>h</i> ter | father |

WORD CORRECTION [2]

- To correct the **wrong character** error
 - Replace each character in the word with *
 - Search the replaced word in the lexicon
 - All words found in the dictionary are suggested to correct the input word
- Example
 - *марчин* is a mistyped word
 - The correct form is *малчин* (*herder*)

| 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 |

WORD CORRECTION [2]

- To correct the **wrong character** error
 - Replace each character in the word with *
 - Search the replaced word in the lexicon
 - All words found in the dictionary are suggested to correct the input word
- Example
 - *марчин* is a mistyped word
 - The correct form is *малчин*

| 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 |

WORD CORRECTION [2]

- To correct the **wrong character** error
 - Replace each character in the word with *
 - Search the replaced word in the lexicon
 - All words found in the dictionary are suggested to correct the input word
- Example
 - *марчин* is a mistyped word
 - The correct form is *малчин*

| 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 |

WORD CORRECTION [3]

- To correct the **inserted character** error
 - Delete each character of a word one by one
 - Search the word in the lexicon
 - All words found in the dictionary are suggested to correct the input word
- Example
 - *малячин* is a mistyped word
 - The correct form is *малчин*

| 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 |

WORD CORRECTION [4]

- To correct the **missed character** error
 - Insert * before and after each character of a word
 - Search the word in the lexicon
 - All words found in the dictionary are suggested to correct the input word
- Example
 - *мачин* is a mistyped word
 - The correct form is *малчин*

| 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 |

WORD CORRECTION [5]

- To correct the **exchanged characters** error
 - Exchange each a pair of characters of a word
 - Search the word in the lexicon
 - All words found in the dictionary are suggested to correct the input word
- Example
 - *мачлин* is a mistyped word
 - The correct form is *малчин*

| 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 |

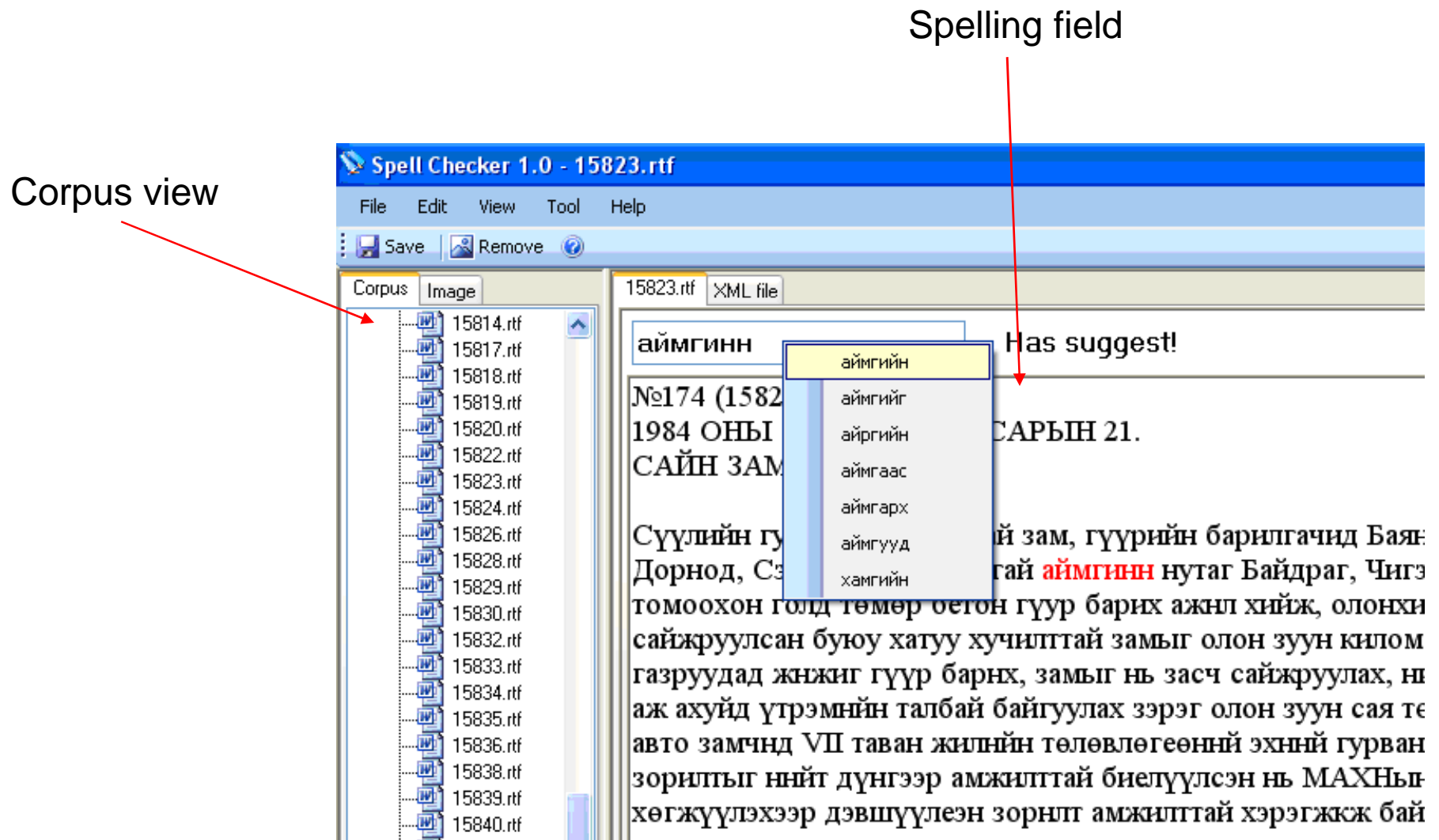
EXPERIMENT

| # | 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 (80 %) | 29 (20 %) |

EXAMPLE CORRECTION

| Two Misspelled Letters | | Over Two Misspelled Letters | |
|---|----------------|---|-------------------|
| Incorrect | Correct | Incorrect | Correct |
| х о дөлмөрийн | хөдөлмөрийн | о й о л ьгн | онолын |
| су ы н | сумын | хү и үү я лд й н | хүмүүжлийн |
| мачид | малчид | хү т аш й | хүчний |
| өн ө гийн | өнөөгийн | хэм г ээ п ий | хэмжээний |
| тех н к | техник | хэм я с э з н ий | хэмжээний |
| байгуулл з гын | байгууллагын | түм н н п х э | түмнийхээ |
| х ө дөлмөр | хөдөлмөр | тел ө злөл т инг | төлөвлөлтийг |
| үйлд й эрлэлийн | үйлдвэрлэлийн | дун д ь ш | дундын |
| меха к азм | механизм | це е к г уйг | цөөнгүйг |
| эмэгтэйчүү д п й н | эмэгтэйчүүдийн | д з в п ч л ттэй | дэвшилттэй |
| тэж э л я йн | тэжээлийн | у й д д вэрлэлийн х ээс | үйлдвэрлэлийнхээс |
| ү н лдвэрлэлин н | үйлдвэрлэлийн | үйлдвэрлэл ш п 1 | үйлдвэрлэлийг |
| х о нд н йрүүлдэг | хөндийрүүлдэг | к е е ц н й г | нөөцийг |
| д с р в н т ой | дорвитой | аш п ш т | ашгийг |
| ж н л н йн | жилийн | т в л в э | төлөө |
| шийд з эрт з й | шийдвэртэй | х з к л с э з н ий | хэмжээний |
| з и ч н лгээ | эмчилгээ | х н ш п | хийж |
| б з жүүлэх з д | бэхжүүлэхэд | х э д ел мер н ийн | хөдөлмөрийн |
| н э в ц инг | нөөцийг | к а л ы п | намын |
| бан л даа н ы | байлдааны | х э г ш эс э ц | хэмнэсэн |
| төл е ер | төлөөр | т э л э вл ө г ө өг | төлөвлөгөөг |

SPELL CHECKER GUI



CONCLUSION

- In this phase
 - Develop Mongolian Spell-Checker for Corpus Cleaning
 - Dictionary-based (around 160 thousand words)
 - Developing lemmatizer for a rule-based version
 - Currently noun lemmatization has developed
 - This noun lemmatization would be a model for verbs and other POSs
 - GUI is designed for the corpus cleaning
 - Accuracy
 - For spelling mistakes, more than 98%
 - For orthographic errors caused by OCR, 85%

THANK YOU