Reversing the Palladius Mapping of Chinese Names in Russian Text

31 October 2012

Dr. Katherine M. Young

N-Space Analysis, LLC

Human Trust & Interaction Branch (711HPW/RHXS)

711th Human Performance Wing

Air Force Research Laboratory

Wright-Patterson AFB

Integrity ★ Service ★ Excellence

DISTRIBUTION A: Approved for public release; distribution unlimited
(Date cleared: 24 September 2012; Case Number: 88ABW-2012-5094)

Acknowledgments

Co-authors:

Jeremy Gwinnup
Joshua Reinhart
SRA International, Inc.

Thanks to:

Dr. Ray Slyh, AFRL
Dr. Tim Anderson, AFRL
Mr. Bill McIntyre, USAF

Overview

- The Problem
- Workflow
- Interface Design
  - Human-in-the-loop
  - Automatically-generated options
  - Saving a dictionary
- Applying the Dictionary to New Documents
- Future Possibilities

The Problem

When using Systran to support Russian to English translation:

- Systran translates known words, and sounds out unknown words
- Chinese names follow a separate Palladius sound mapping

Meaning	“British expert Phillip Clark”
Cyrillic	Британский эксперт Филлип Кларк
Systran	The British expert Of fillip Clark

Meaning	"Zhai Zhigang was the youngest of six children"
Cyrillic	Чжай Чжиган был младшим из шести детей
Systran	Chzhay Of chzhigan was young of six children

Palladius Sound Mappings

Pinyin	Cyrillic	typical sound	
r	ж	[ž]	Palladius maps initial zh to чж
ch	ч	[č]	
zh	чж	[č ž]	
n__	н__	[n]	Palladius maps word-final ng to н
__n	__нь	[n’]	
__ng	__н	[n]	
xi	си	[ɕi]	Palladius maps x to с
sa	са	[sa]	

Reverse Palladius (RevP) Algorithm

1. User selects candidate words

Чжай Чжиган был младшим из шести детей

RevP

2. Syllabification

#Чжай# #чжи#ган# был младшим из шести детей

(C)(G)V(N)

C=consonant
G=glide {i, u}
N=coda{n,ng,i,u,r}

3. Reverse Palladius mapping

Zhai zhigang был младшим из шести детей

RevP Rules
чж → zh
н# → ng

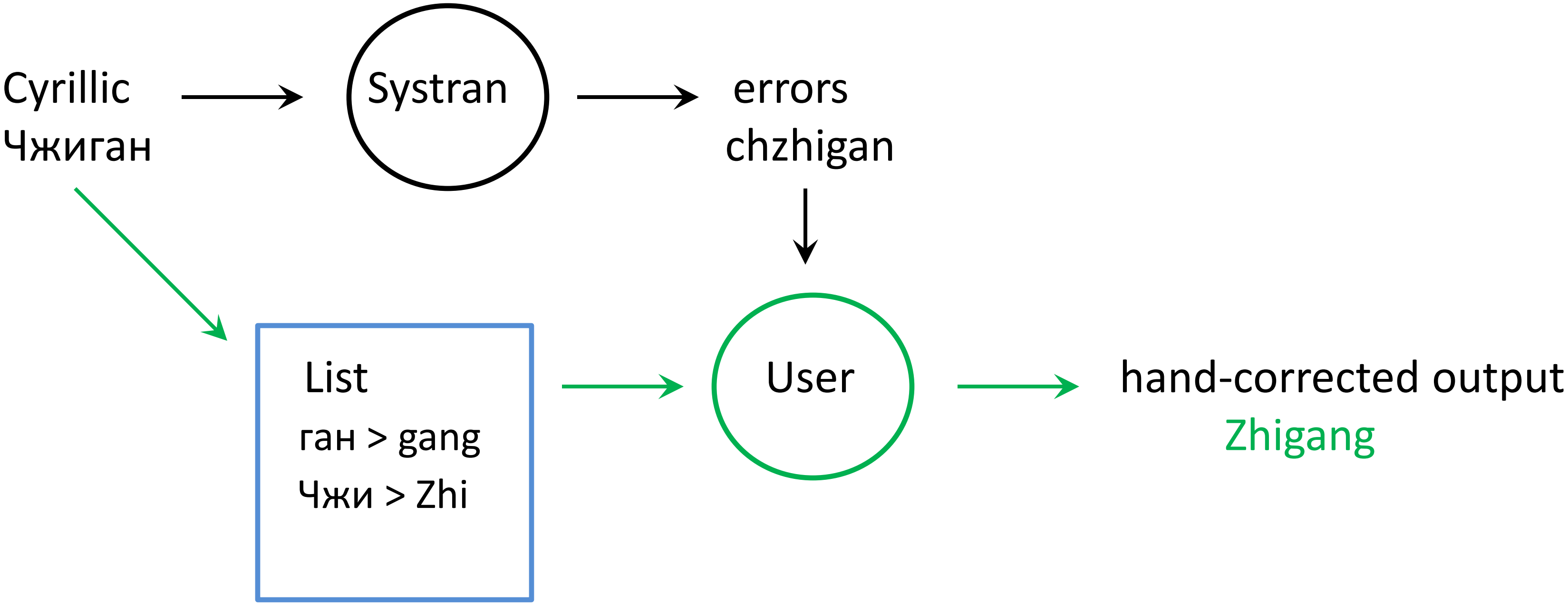
4. Systran

Zhai zhigang was young of six children

Distribution A: Approved for public release; distribution unlimited

Workflow -- Original

The user compiles a list of Chinese/Russian syllable correspondences and uses this list to hand-edit the Systran errors.

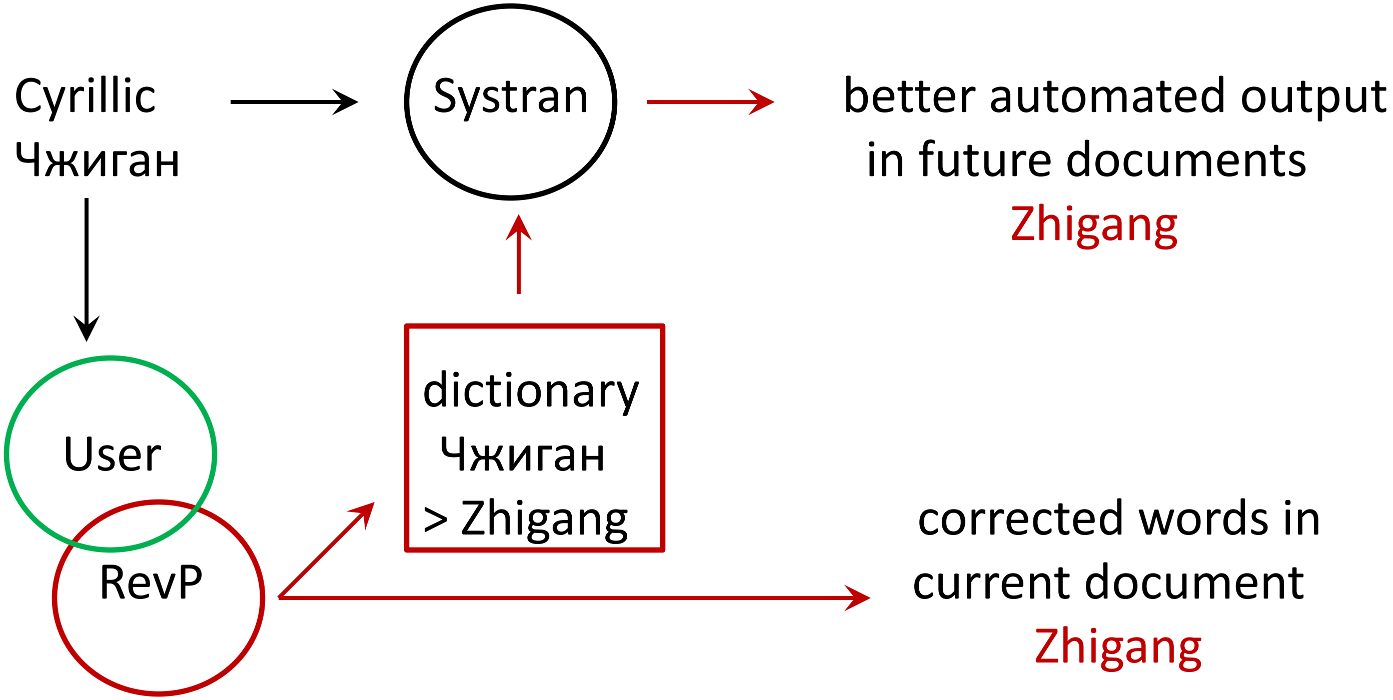


- The user has to look back at the original spelling of the name.
- Each instance in the text has to be separately corrected.
- These corrections must be repeated in each new document.

Distribution A: Approved for public release; distribution unlimited

Workflow -- with RevP Program

- The user selects a word and the program provides the Reverse Palladius form
- The new mappings are saved in a dictionary for future use, either with Systran or as a stand-alone process.



Cyrillic
Чжиган

Systran

better automated output
in future documents
Zhigang

User

RevP

dictionary
Чжиган
> Zhigang

corrected words in
current document
Zhigang

Interface Design

- We use a human-in-the-loop to identify the candidate words
- We provide alternative mappings
- We store the user’s choices in a dictionary

- The dictionary can be applied to new documents within Systran
- The dictionary can be applied independently

Human-in-the-loop: Identifying the Chinese Words

We don’t want to apply the Reverse Palladius mapping to Russian words, or to words borrowed from other languages

Cyrillic	Лисов	
Traditional Sound Mapping	Lisov	<< correct form
Reverse Palladius Mapping	Lisuow	

Cyrillic	Малинди	
Traditional Sound Mapping	Malindi	<< correct form
Reverse Palladius Mapping	Malingdi	

Human-in-the-loop: Identifying Russian Inflection

If the name occurs as the object of a verb or preposition, it may appear with Russian case endings.

Cyrillic	ReversePalladius	pinyin	
Бомин	Boming	Boming	
Бомина	Boming*a*	Boming	(genitive)
Бомином	Boming*om*	Boming	(instrumental)

We optionally apply a stemming program to remove inflections before the sound mapping:

meaning:	" ... together with Liu Boming"
Cyrillic:	... вместе с Лю Бомином
Reverse Palladius:	Liu Bomingom
Stemming + Reverse Palladius:	Liu Boming

Distribution A: Approved for public release; distribution unlimited

Human-in-the-loop: Identifying Russian Inflection

We cannot apply the stemmer in all instances.

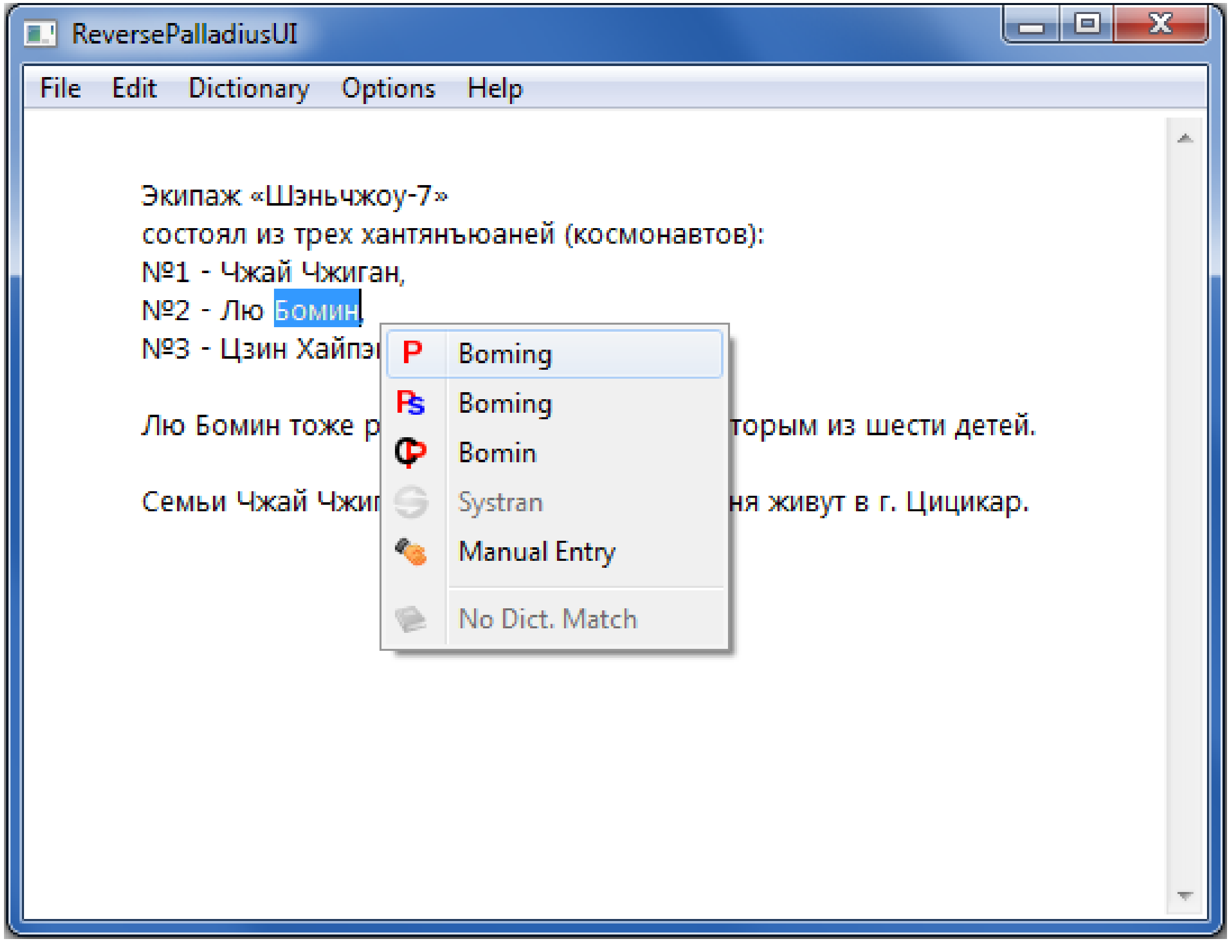
For example, the Chinese name Baohua ends in /a/, which is a potential Russian inflectional ending.

Stemming of Баохуа creates an error:

Cyrillic	Stemmed	RevP	Correct pinyin
Баохуа	Баоху	Baohu	Baohua

The RevP program generates the various possible forms, and we rely on the user to select the correct alternative.

Interface: User Selection



Экипаж «Шэньчжоу-7»
состоял из трех хантяньюаней (космонавтов):
№1 - Чжай Чжиган,
№2 - Лю Бомин,
№3 - Цзин Хайпэ

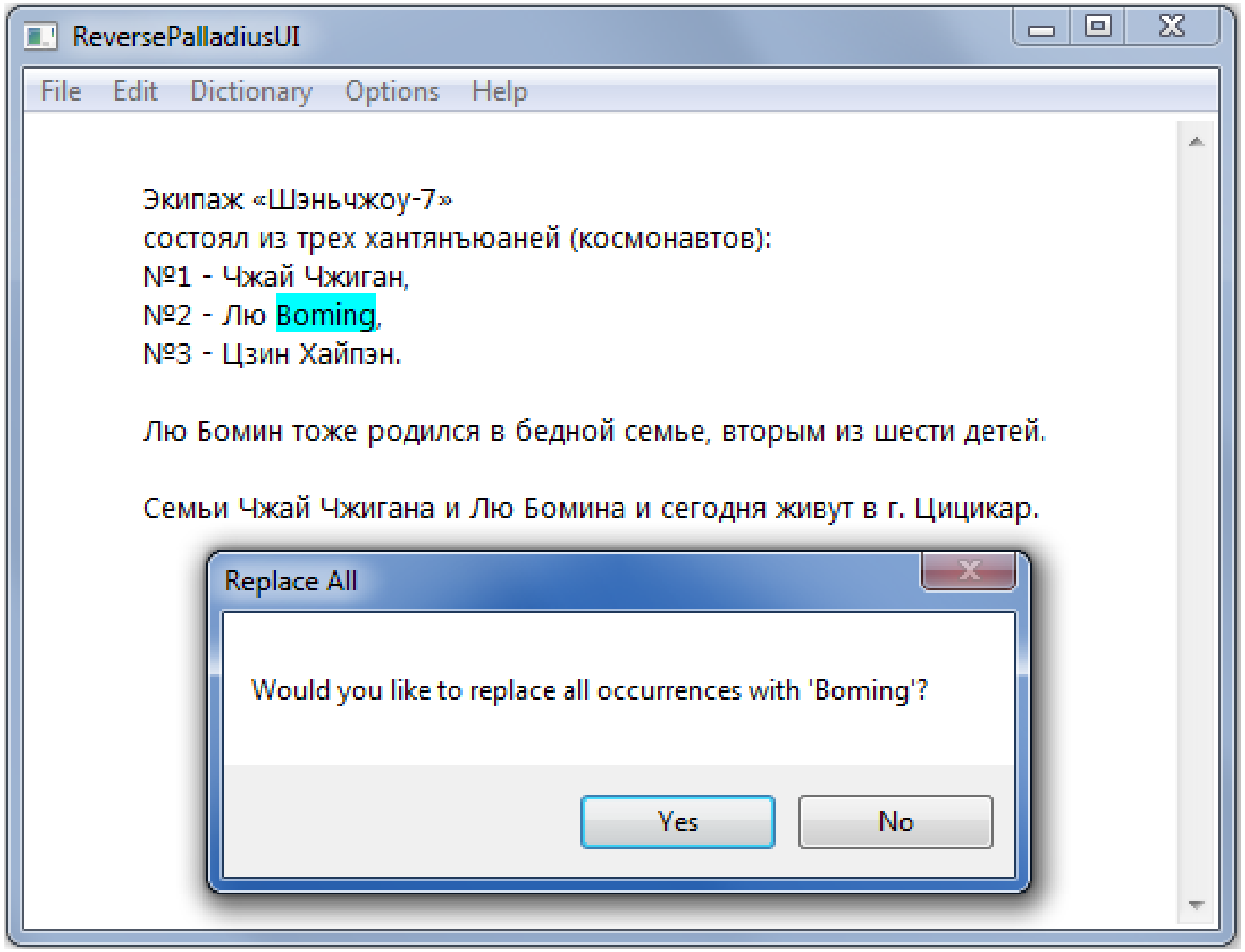
Лю Бомин тоже р торым из шести детей.

Семьи Чжай Чжи ня живут в г. Цицикар.

The user has selected the name, Бомин.

The program presents the user with the various options for the selected word. The first option is the Reverse Palladius Mapping.

Interface: Replacing the Word



Экипаж «Шэньчжоу-7»
состоял из трех хантянъюаней (космонавтов):
№1 - Чжай Чжиган,
№2 - Лю Boming,
№3 - Цзин Хайпэн.

Лю Бомин тоже родился в бедной семье, вторым из шести детей.

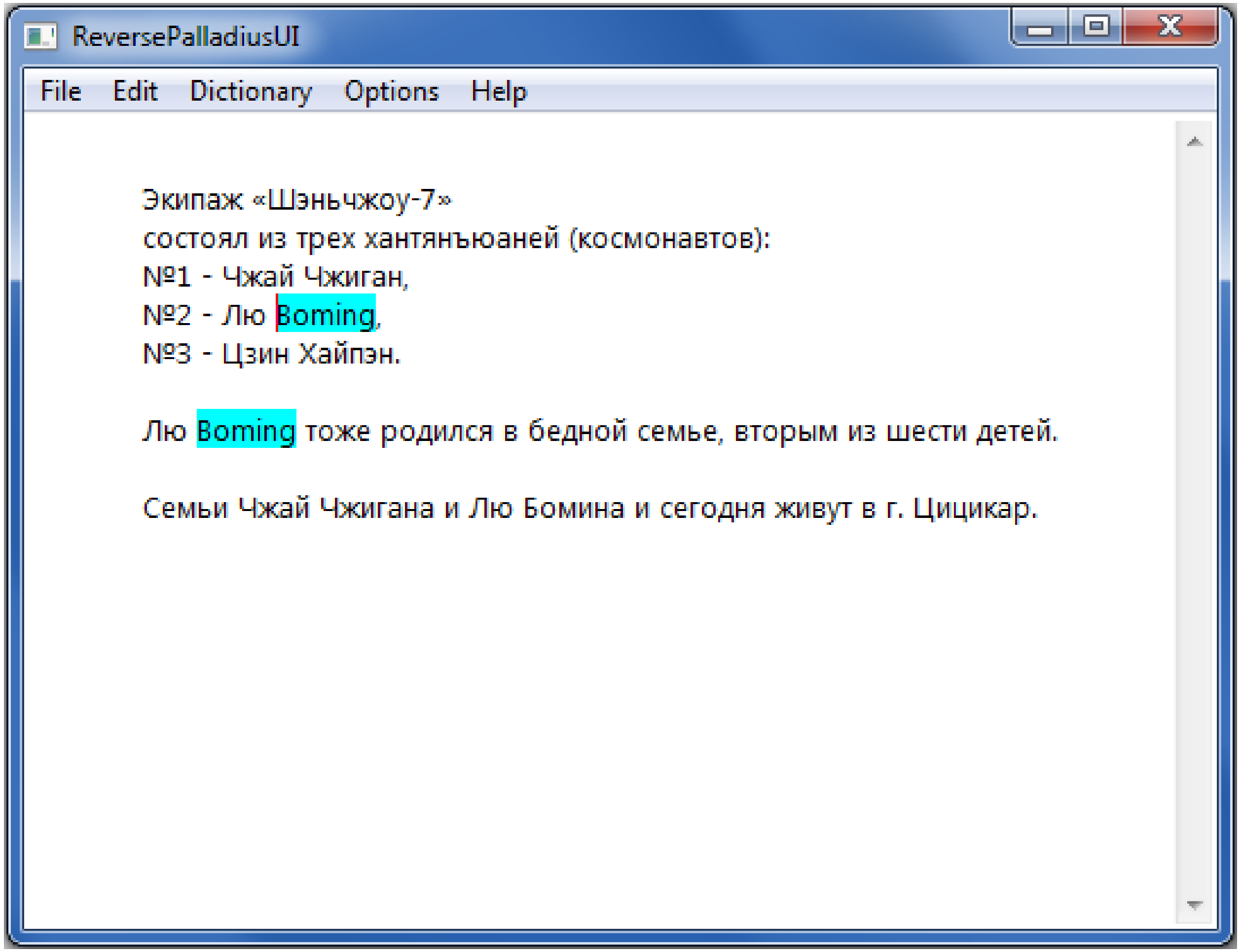
Семьи Чжай Чжигана и Лю Бомина и сегодня живут в г. Цицикар.

Replace All

Would you like to replace all occurrences with 'Boming'?

The selected form replaces the original, and the program offers to replace all instances of this word.

Interface: Replacing the Word



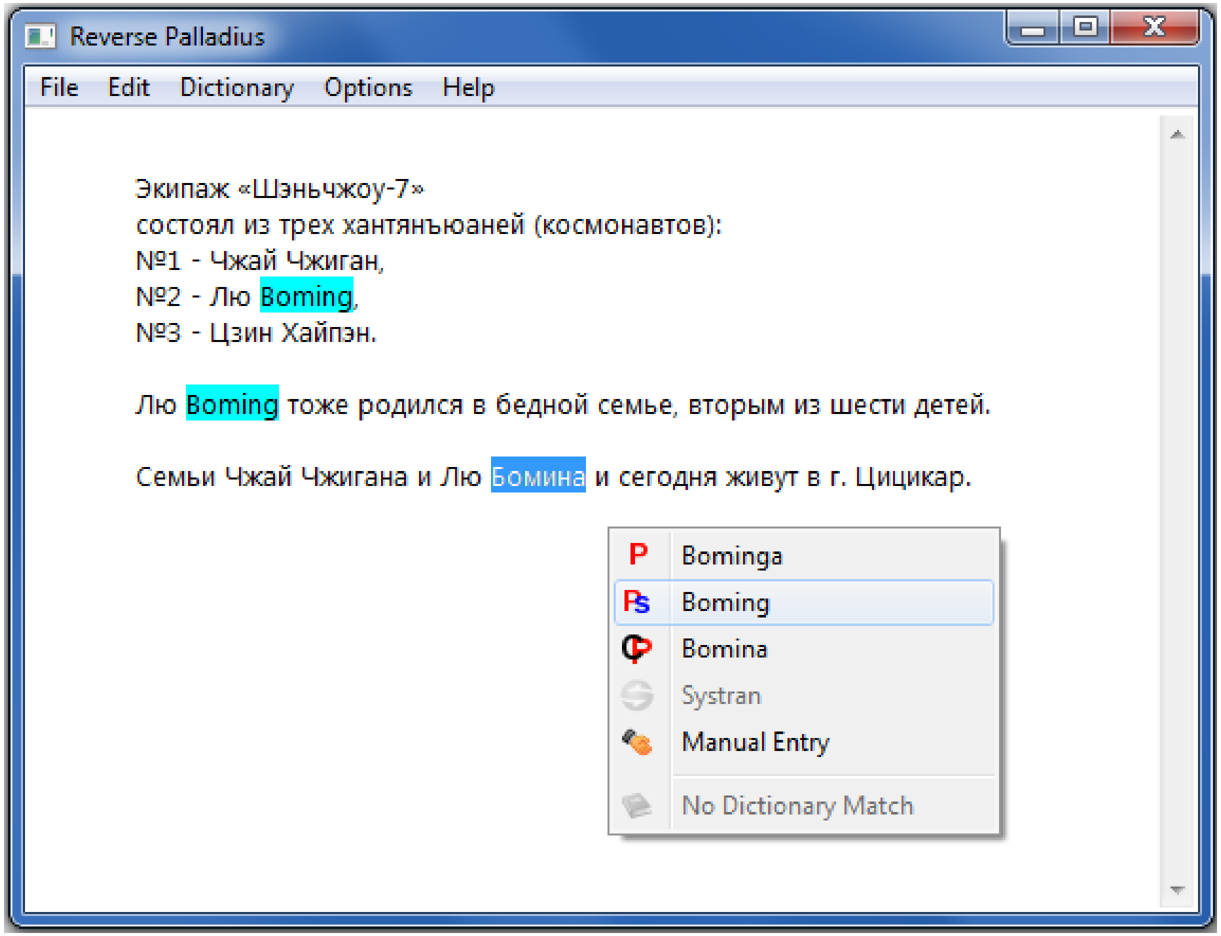
Экипаж «Шэньчжоу-7»
состоял из трех хантянъюаней (космонавтов):
№1 - Чжай Чжиган,
№2 - Лю Boming,
№3 - Цзин Хайпэн.

Лю Boming тоже родился в бедной семье, вторым из шести детей.

Семьи Чжай Чжигана и Лю Бомина и сегодня живут в г. Цицикар.

Distribution A: Approved for public release; distribution unlimited

Interface: Stemmed Form



Экипаж «Шэньчжоу-7»
состоял из трех хантянъюаней (космонавтов):
№1 - Чжай Чжиган,
№2 - Лю Boming,
№3 - Цзин Хайпэн.

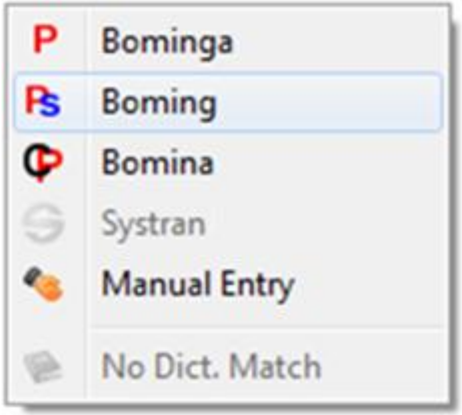
Лю Boming тоже родился в бедной семье, вторым из шести детей.

Семьи Чжай Чжигана и Лю Бомина и сегодня живут в г. Цицикар.

Here, the user has selected the word Бомина.
Now we need the stemmed form, which is the second option.

Interface: Other Options

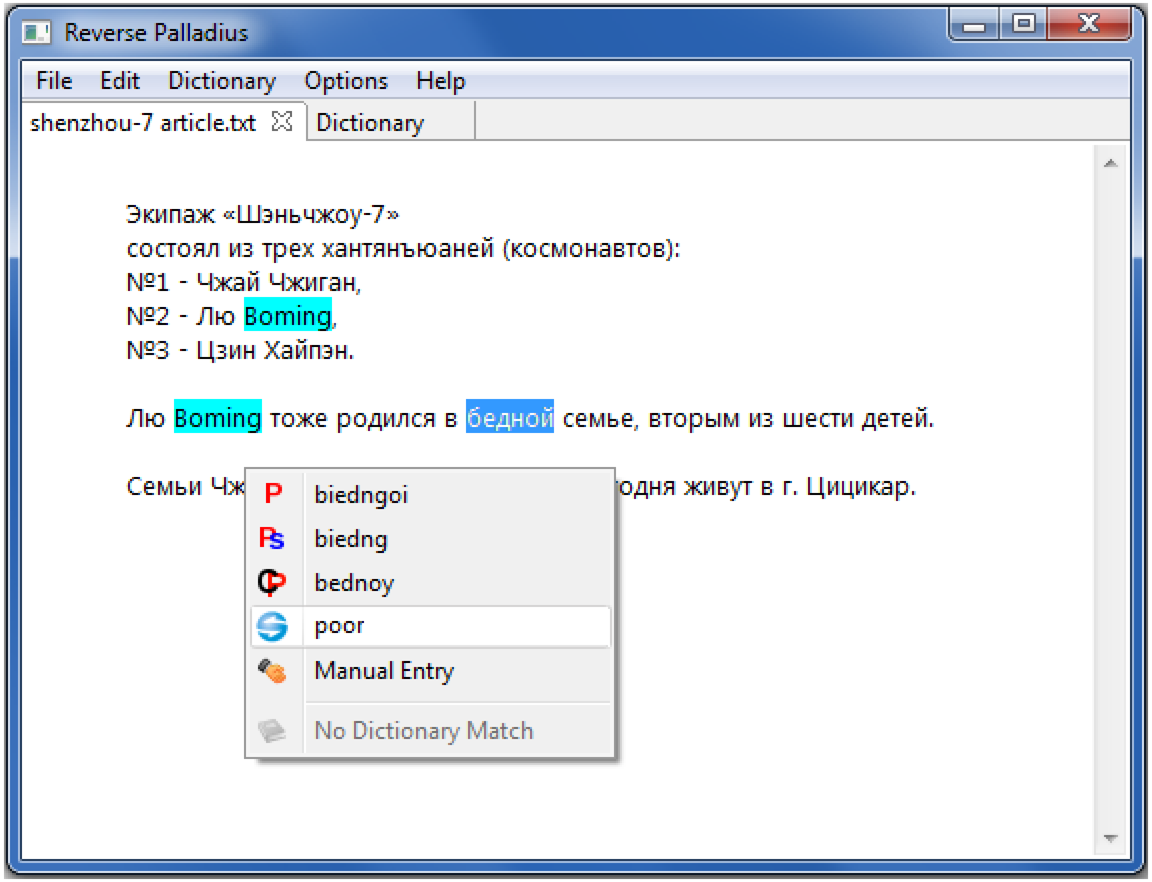
The RevP options:



Symbol	Source	Details
P	Palladius	apply reverse Palladius mapping
Ps	stemmed Palladius	remove Russian inflectional endings, then apply reverse Palladius mapping
Cp	Cyrillic phonetics	apply Cyrillic phonetic mapping
S	Systran translation	use the Systran translation for this word (if you have Systran enabled)
hand	Manual Entry	type in the correct spelling
book	Dictionary Match	use the previous dictionary choice for this word (if any)

Distribution A: Approved for public release; distribution unlimited

Interface: Systran Preview



Экипаж «Шэньчжоу-7»
состоял из трех хантянъюаней (космонавтов):
№1 - Чжай Чжиган,
№2 - Лю Boming,
№3 - Цзин Хайпэн.

Лю Boming тоже родился в бедной семье, вторым из шести детей.

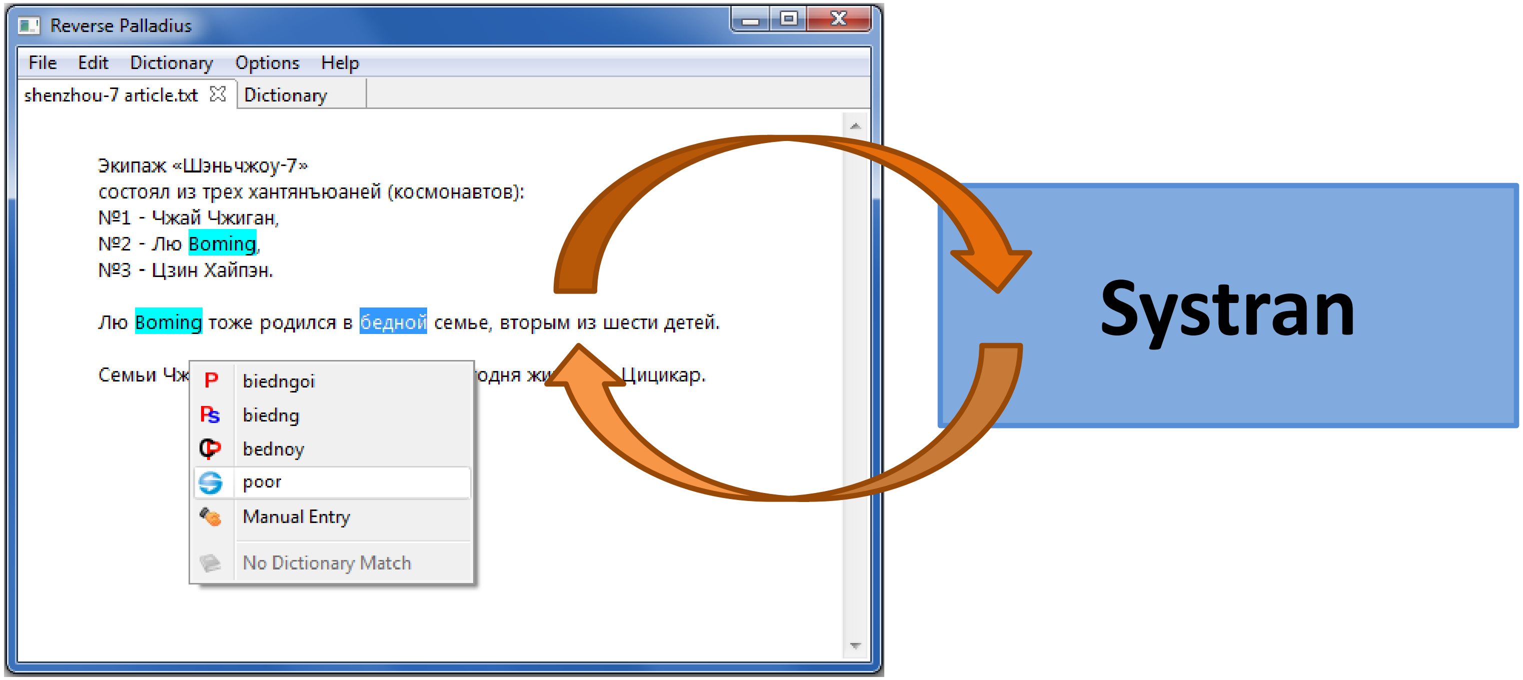
Семьи Чж… …одня живут в г. Цицикар.

Systran translation: a preview of how Systran would translate this word or phrase

Distribution A: Approved for public release; distribution unlimited

Interface: Connection to Systran

- Systran 7 SOAP API – use code generated from WSDL to allow RevP to send highlighted text to a Systran server for immediate translation



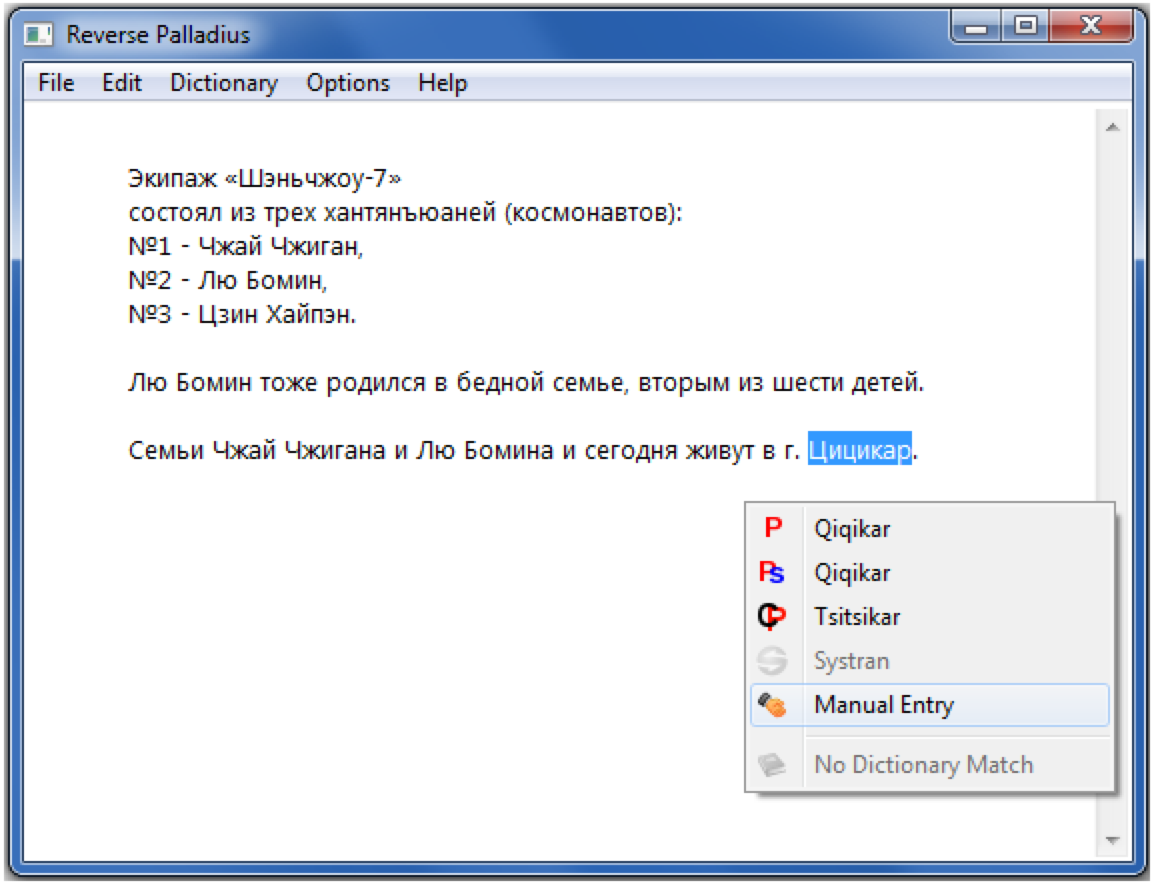
Экипаж «Шэньчжоу-7»
состоял из трех хантяньюаней (космонавтов):
№1 - Чжай Чжиган,
№2 - Лю Boming,
№3 - Цзин Хайпэн.

Лю Boming тоже родился в бедной семье, вторым из шести детей.

Семьи Чж одня жи Цицикар.

Systran

Interface: Manual Entry



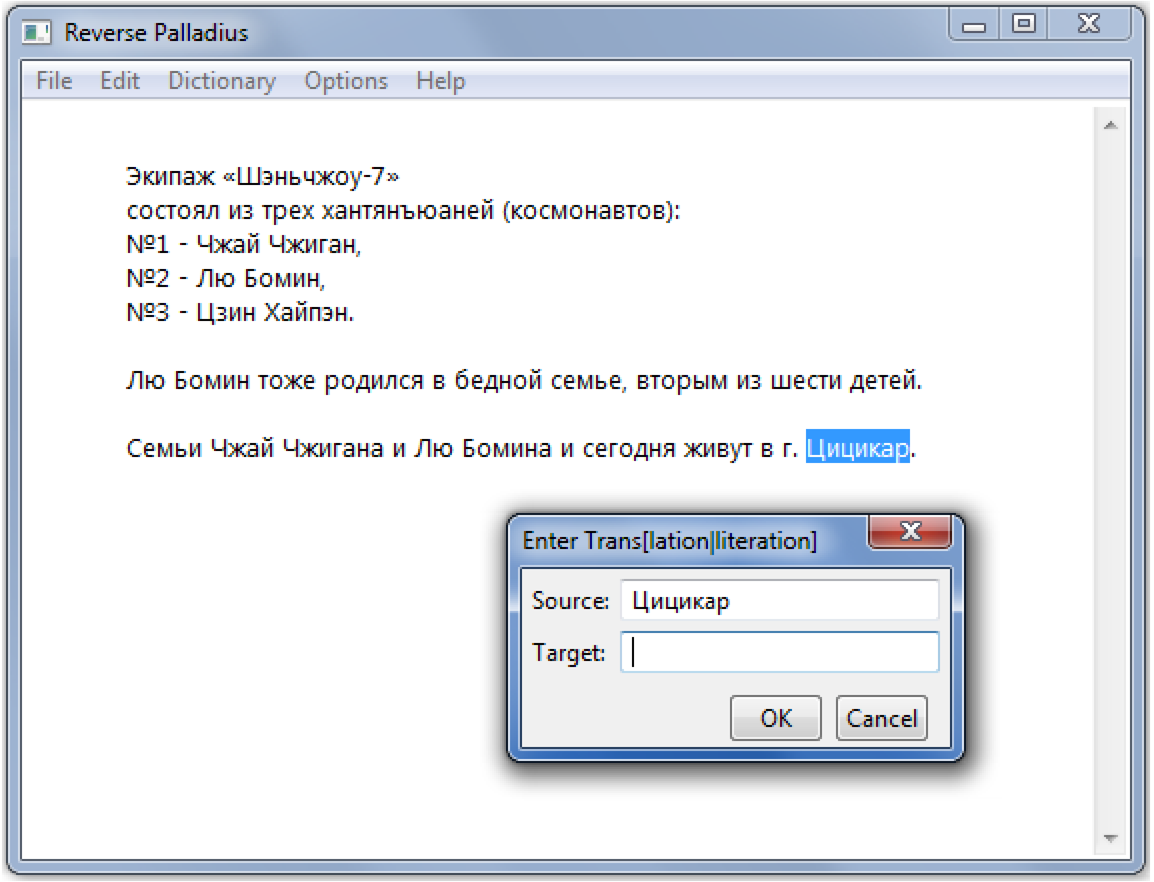
Экипаж «Шэньчжоу-7»
состоял из трех хантяньюаней (космонавтов):
№1 - Чжай Чжиган,
№2 - Лю Бомин,
№3 - Цзин Хайпэн.

Лю Бомин тоже родился в бедной семье, вторым из шести детей.

Семьи Чжай Чжигана и Лю Бомина и сегодня живут в г. Цицикар.

Manual Entry: If none of the provided forms are correct, there is an option to type in the correct form.

Interface: Manual Entry



Экипаж «Шэньчжоу-7»
состоял из трех хантяньюаней (космонавтов):
№1 - Чжай Чжиган,
№2 - Лю Бомин,
№3 - Цзин Хайпэн.

Лю Бомин тоже родился в бедной семье, вторым из шести детей.

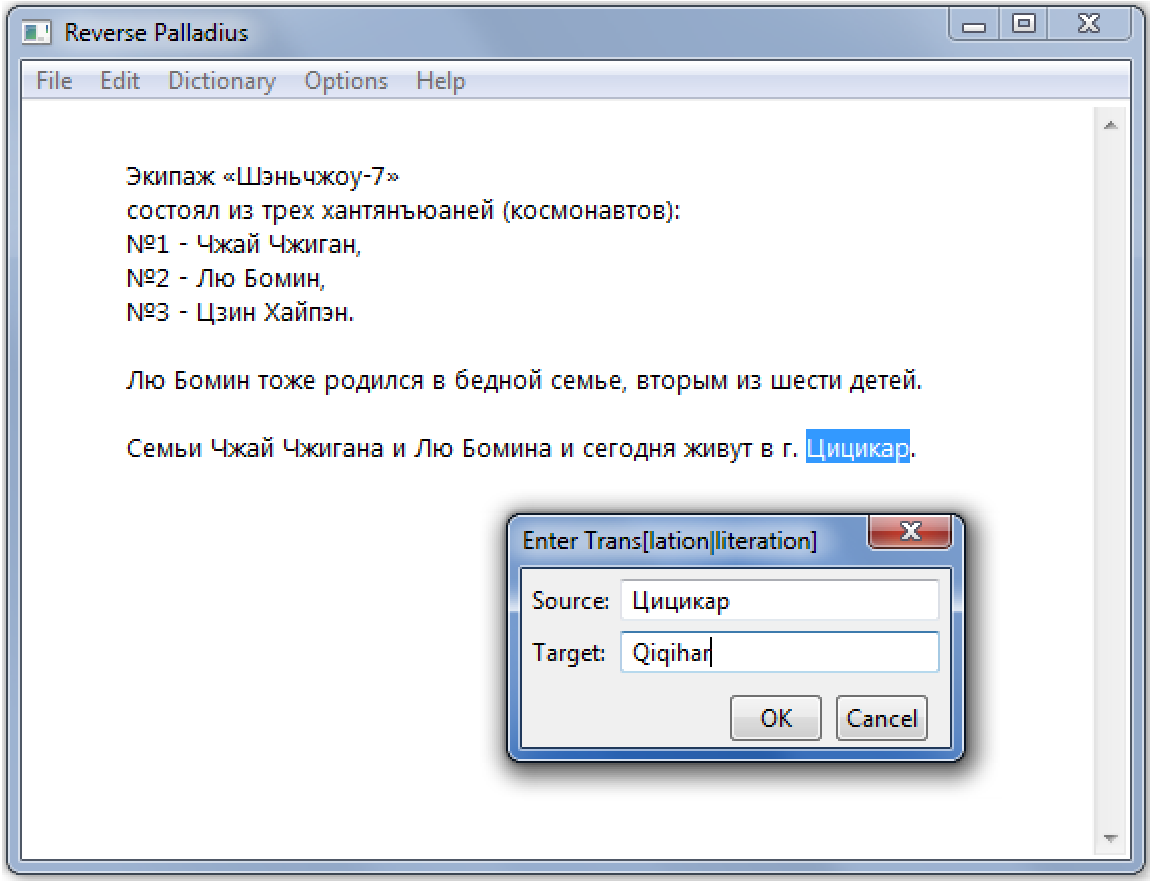
Семьи Чжай Чжигана и Лю Бомина и сегодня живут в г. Цицикар.

Source: Цицикар

Target:

Distribution A: Approved for public release; distribution unlimited

Interface: Manual Entry



Экипаж «Шэньчжоу-7»
состоял из трех хантянъюаней (космонавтов):
№1 - Чжай Чжиган,
№2 - Лю Бомин,
№3 - Цзин Хайпэн.

Лю Бомин тоже родился в бедной семье, вторым из шести детей.

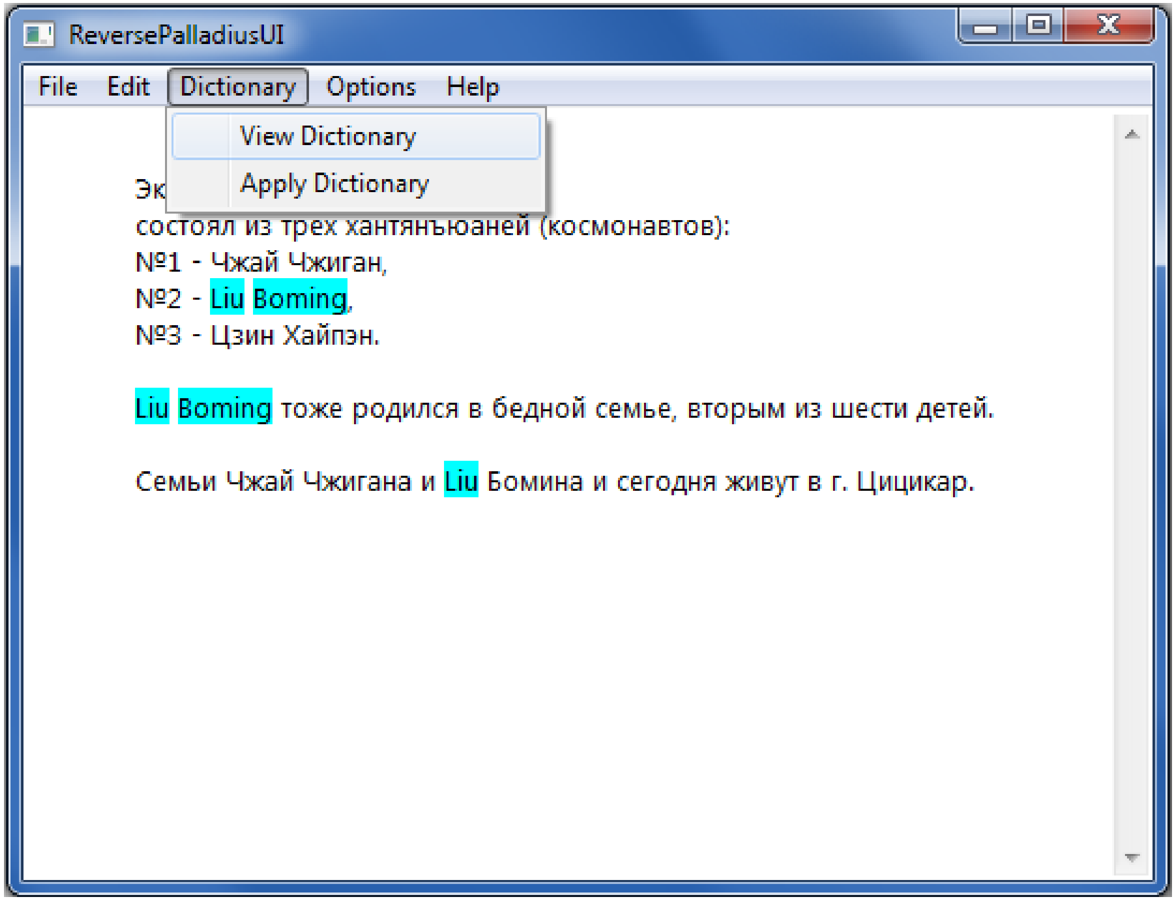
Семьи Чжай Чжигана и Лю Бомина и сегодня живут в г. Цицикар.

Enter Trans[lation|literation]

Source: Цицикар

Target: Qiqihar

Saving the Dictionary



Эк
состоял из трех хантянъюаней (космонавтов):
№1 - Чжай Чжиган,
№2 - Liu Boming,
№3 - Цзин Хайпэн.

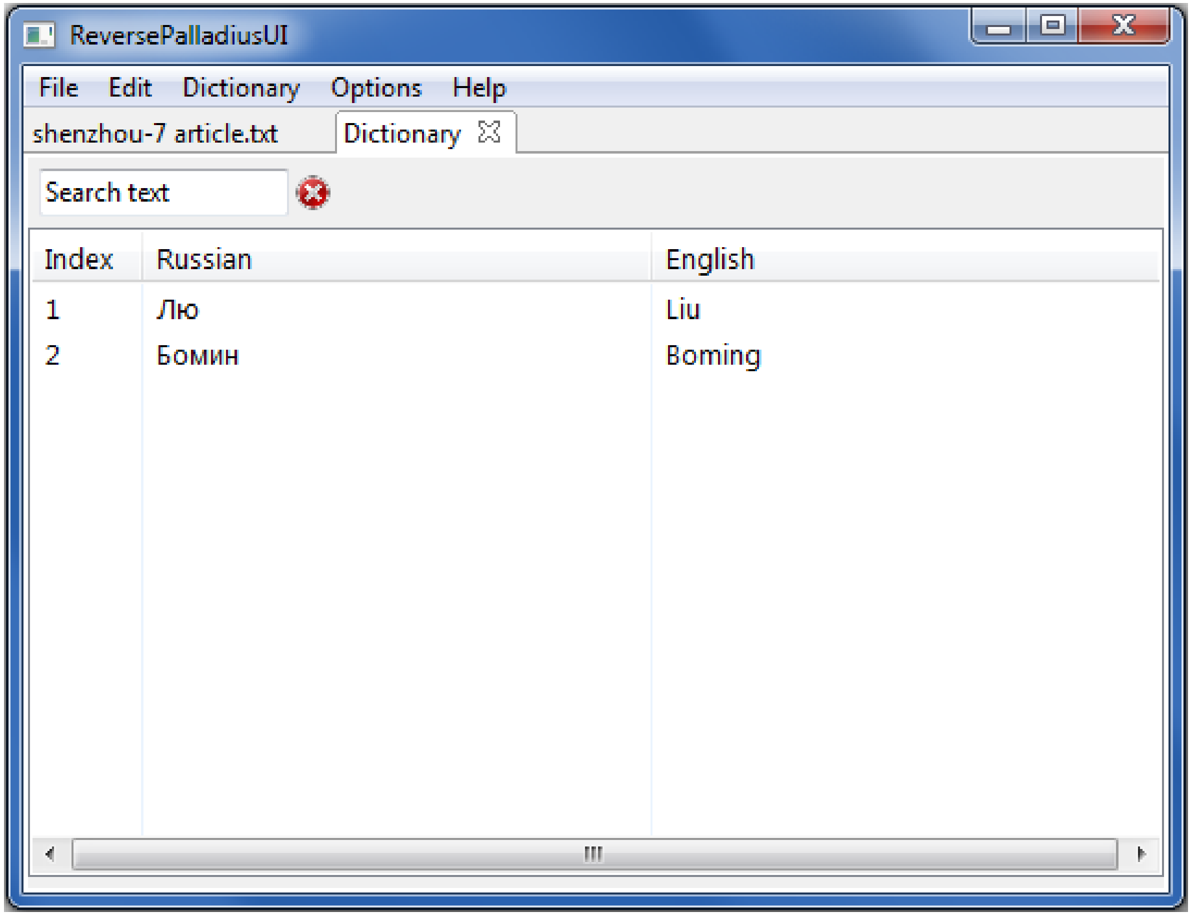
Liu Boming тоже родился в бедной семье, вторым из шести детей.

Семьи Чжай Чжигана и Liu Бомина и сегодня живут в г. Цицикар.

In this example, the user has processed the names Liu and Boming. We select the “View Dictionary” option…

Distribution A: Approved for public release; distribution unlimited

Saving the Dictionary

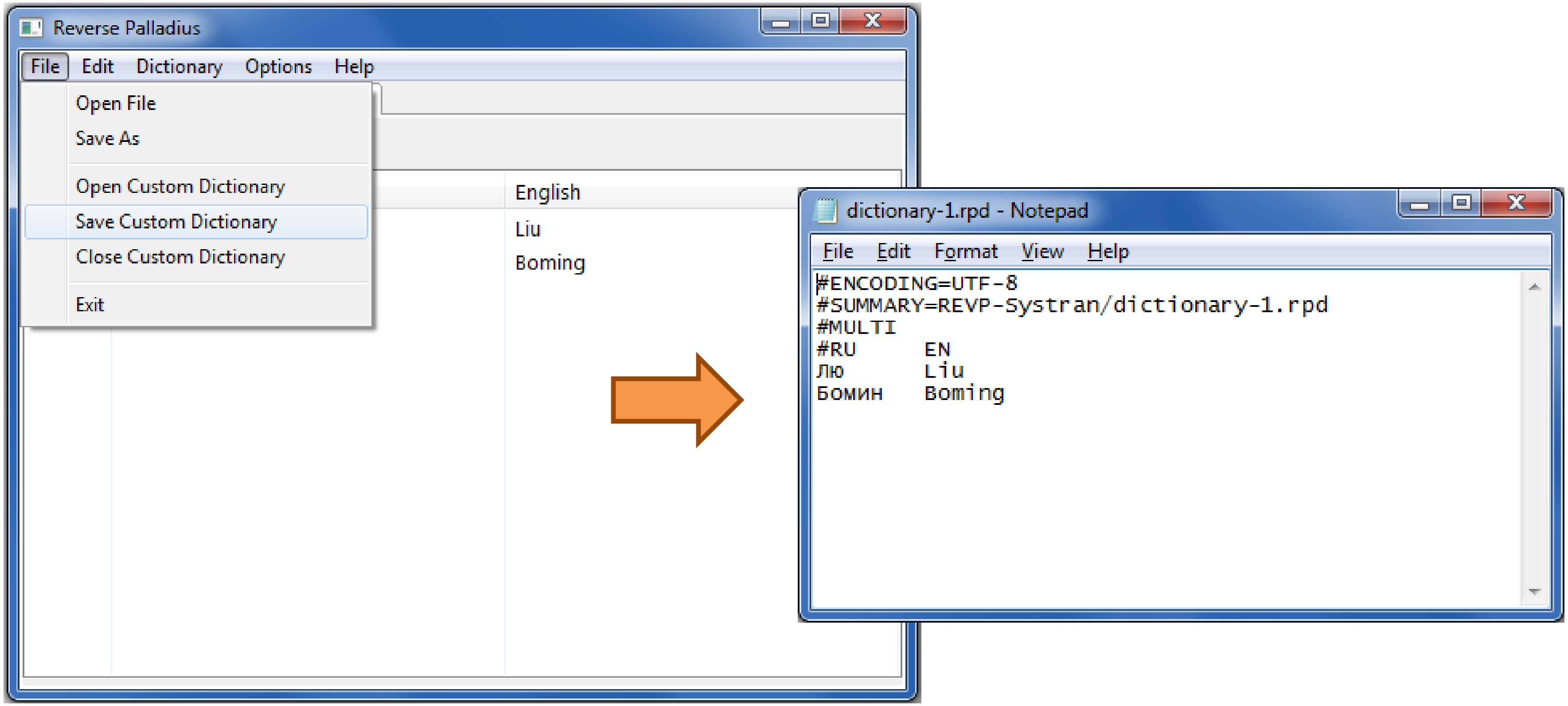


Index	Russian	English
1	Лю	Liu
2	Бомин	Boming

… and see the two entries that RevP has created.

Distribution A: Approved for public release; distribution unlimited

Saving the Dictionary



```
#ENCODING=UTF-8
#SUMMARY=REVP-Systran/dictionary-1.rpd
#MULTI
#RU     EN
Лю      Liu
Бомин   Boming
```

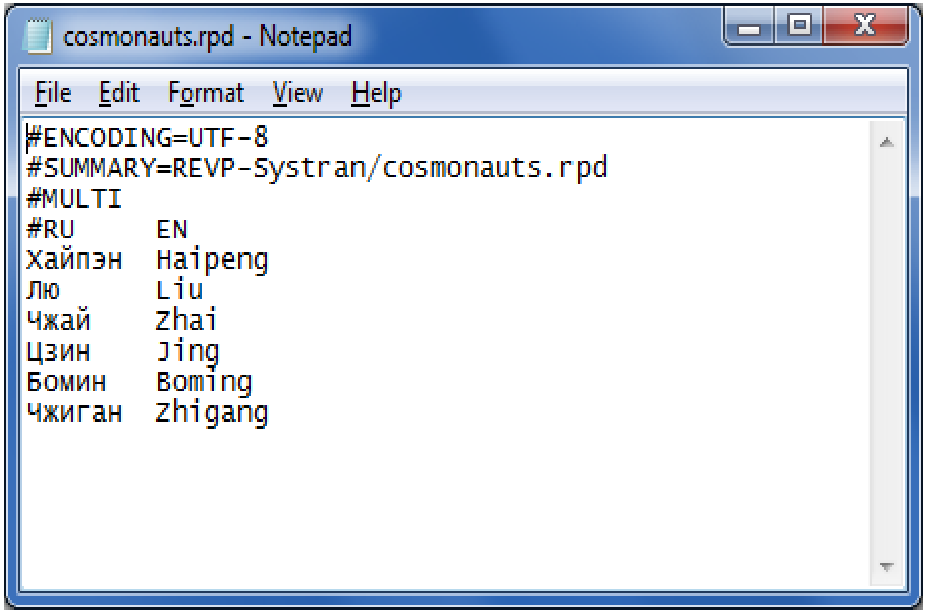
These entries can be stored to a file by using the menu option, File/Save Custom Dictionary.

The dictionary is saved in a Systran-readable format.

Applying the Dictionary to New Documents Using RevP

We can use the RevP program to apply a saved dictionary to a new document.

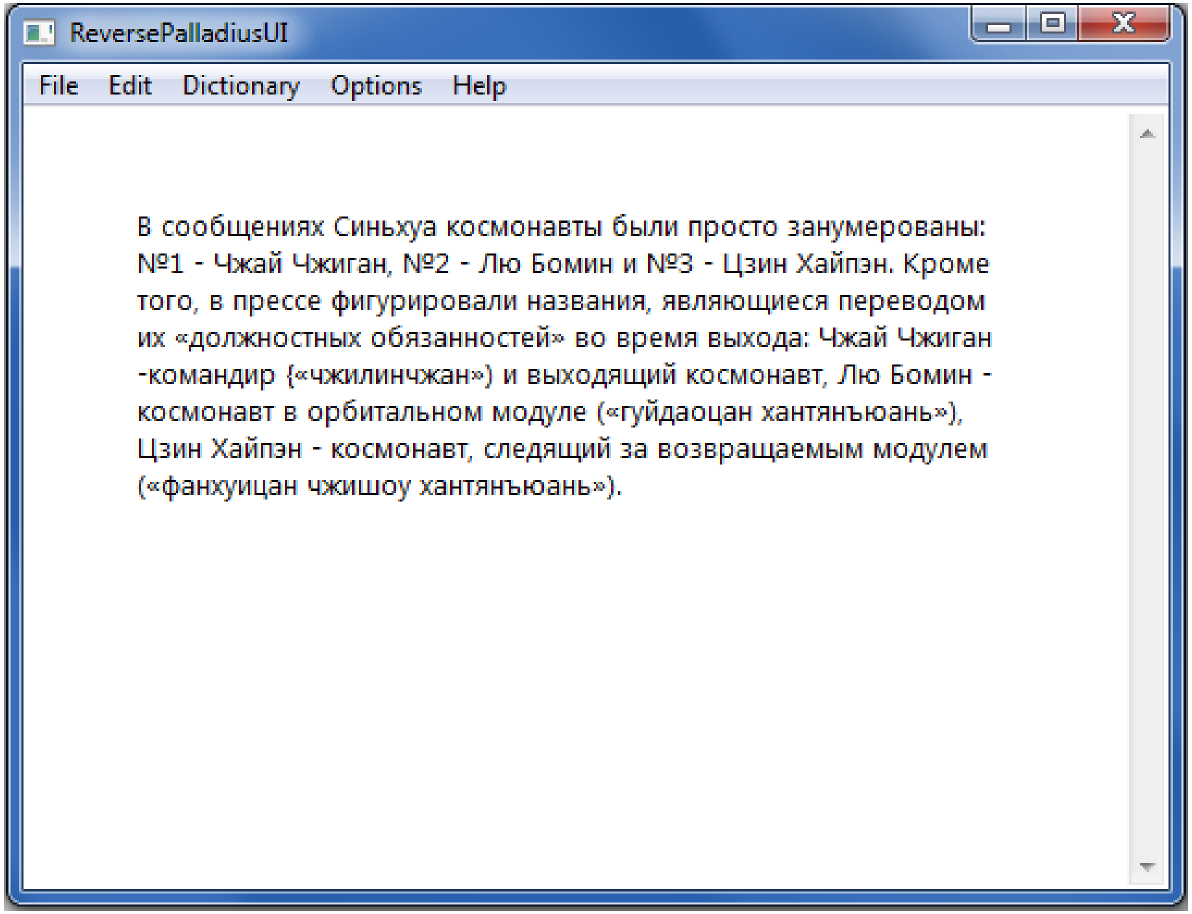
In this example, we have created a dictionary with the names of three cosmonauts.



```
#ENCODING=UTF-8
#SUMMARY=REVP-Systran/cosmonauts.rpd
#MULTI
#RU     EN
Хайпэн  Haipeng
Лю      Liu
Чжай    Zhai
Цзин    Jing
Бомин   Boming
Чжиган  Zhigang
```

Distribution A: Approved for public release; distribution unlimited

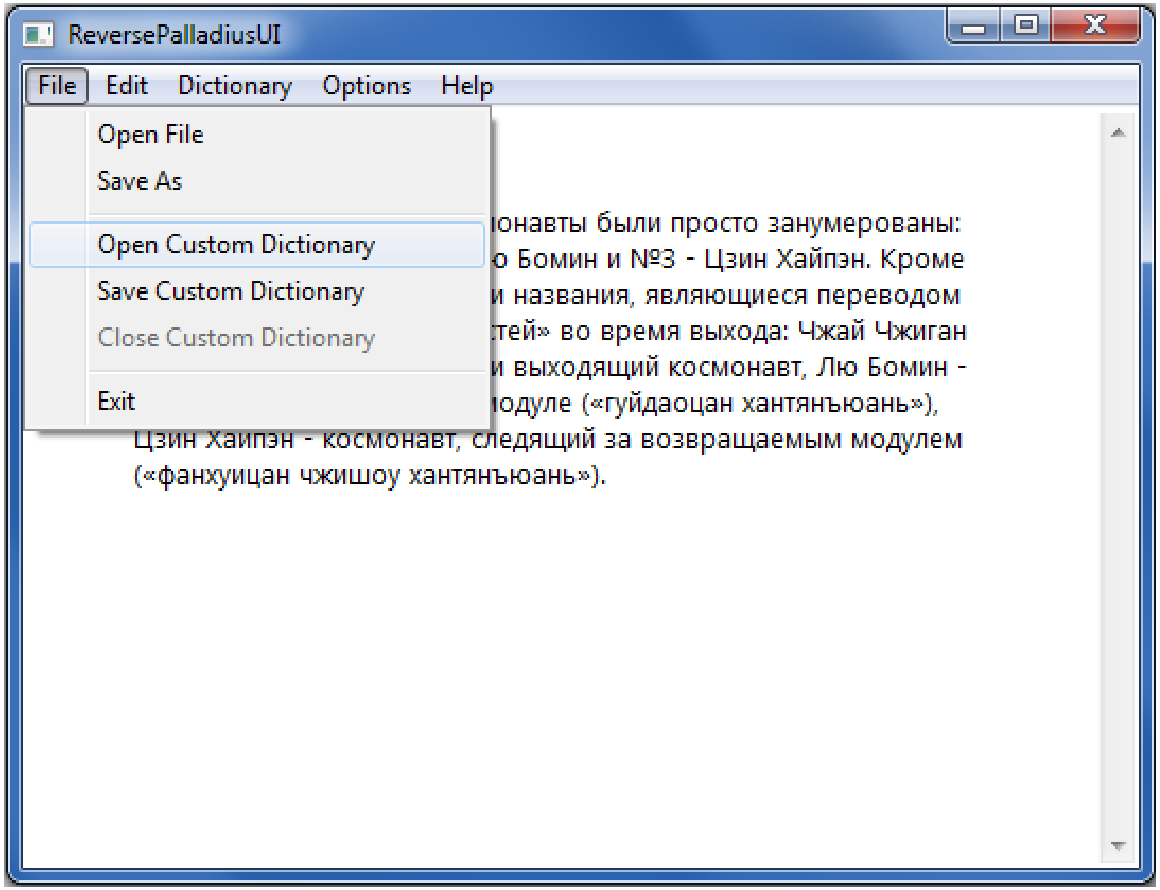
Applying the Dictionary to New Documents Using RevP



В сообщениях Синьхуа космонавты были просто занумерованы: №1 - Чжай Чжиган, №2 - Лю Бомин и №3 - Цзин Хайпэн. Кроме того, в прессе фигурировали названия, являющиеся переводом их «должностных обязанностей» во время выхода: Чжай Чжиган -командир {«чжилинчжан») и выходящий космонавт, Лю Бомин - космонавт в орбитальном модуле («гуйдаоцан хантянъюань»), Цзин Хайпэн - космонавт, следящий за возвращаемым модулем («фанхуицан чжишоу хантянъюань»).

We start the RevP program and open a new document.

Applying the Dictionary to New Documents Using RevP

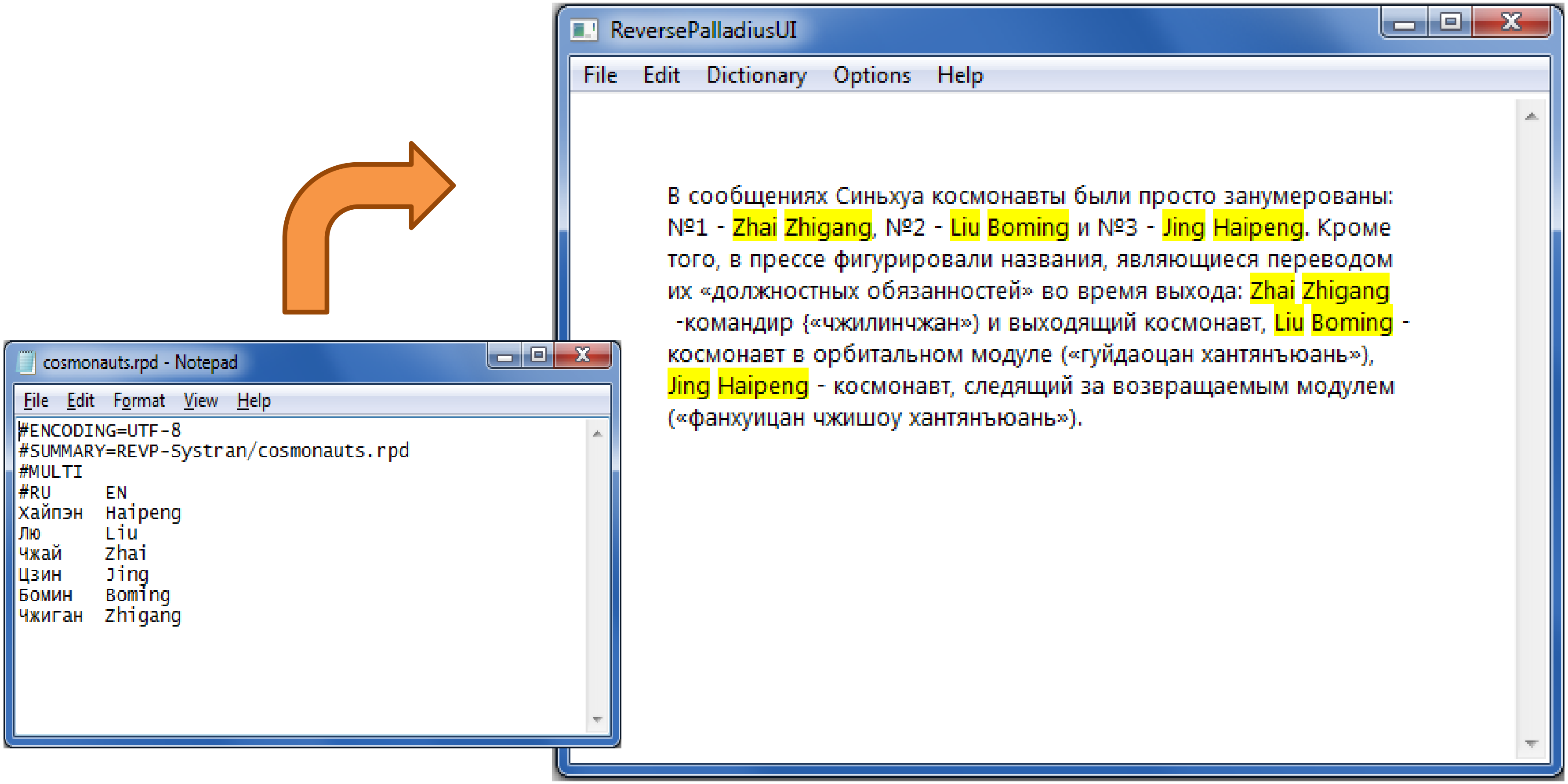


онавты были просто занумерованы:
о Бомин и №3 - Цзин Хайпэн. Кроме
и названия, являющиеся переводом
тей» во время выхода: Чжай Чжиган
и выходящий космонавт, Лю Бомин -
одуле («гуйдаоцан хантянъюань»),
Цзин Хайпэн - космонавт, следящий за возвращаемым модулем
(«фанхуицан чжишоу хантянъюань»).

We select the menu option, File/Open Custom Dictionary, and load our cosmonaut dictionary.

Distribution A: Approved for public release; distribution unlimited

Applying the Dictionary to New Documents Using RevP



```
#ENCODING=UTF-8
#SUMMARY=REVP-Systran/cosmonauts.rpd
#MULTI
#RU     EN
Хайпэн  Haipeng
Лю      Liu
Чжай    Zhai
Цзин    Jing
Бомин   Boming
Чжиган  Zhigang
```

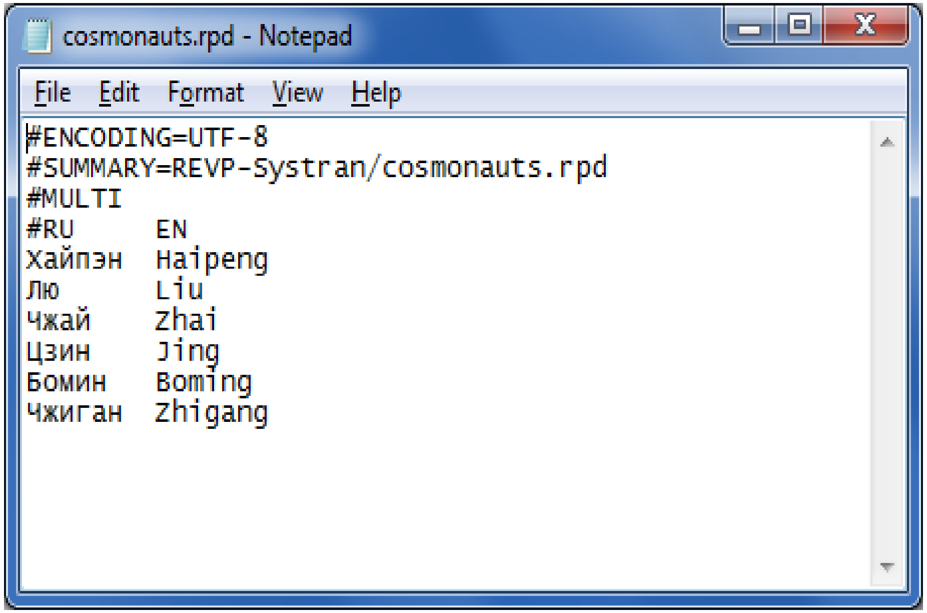
В сообщениях Синьхуа космонавты были просто занумерованы: №1 - Zhai Zhigang, №2 - Liu Boming и №3 - Jing Haipeng. Кроме того, в прессе фигурировали названия, являющиеся переводом их «должностных обязанностей» во время выхода: Zhai Zhigang -командир {«чжилинчжан») и выходящий космонавт, Liu Boming - космонавт в орбитальном модуле («гуйдаоцан хантянъюань»), Jing Haipeng - космонавт, следящий за возвращаемым модулем («фанхуицан чжишоу хантянъюань»).

The program pre-translates all the matching names in the document.

Distribution A: Approved for public release; distribution unlimited

Applying the Dictionary to New Documents Using Systran

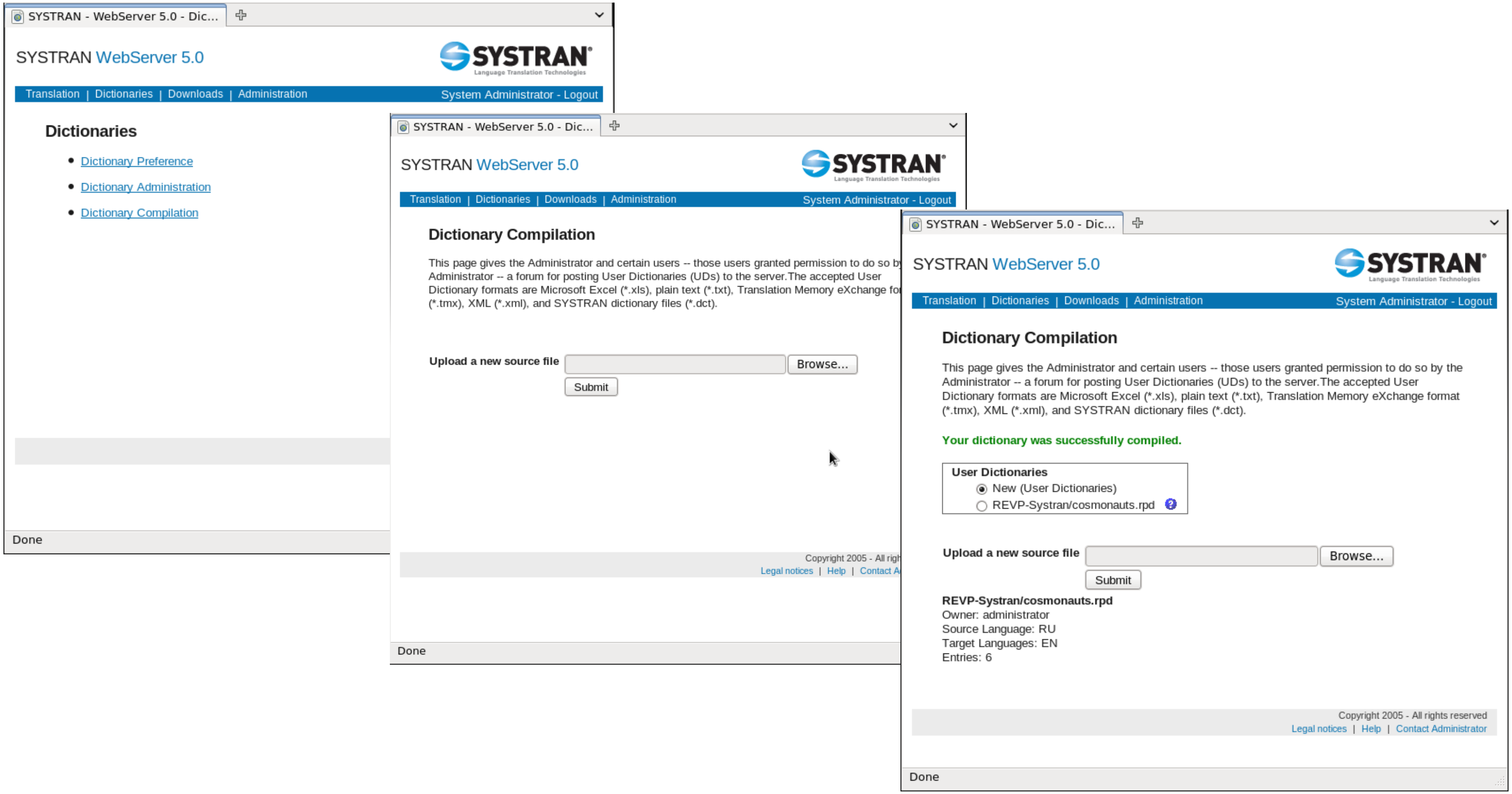
We can also compile our cosmonaut dictionary as a Systran user dictionary.



```
#ENCODING=UTF-8
#SUMMARY=REVP-Systran/cosmonauts.rpd
#MULTI
#RU     EN
Хайпэн  Haipeng
Лю      Liu
Чжай    Zhai
Цзин    Jing
Бомин   Boming
Чжиган  Zhigang
```

Distribution A: Approved for public release; distribution unlimited

Applying the Dictionary to New Documents Using Systran



SYSTRAN - WebServer 5.0 - Dic...

SYSTRAN WebServer 5.0

Translation | Dictionaries | Downloads | Administration
System Administrator - Logout

Dictionaries

- Dictionary Preference
- Dictionary Administration
- Dictionary Compilation

Done

SYSTRAN - WebServer 5.0 - Dic...

SYSTRAN WebServer 5.0

Translation | Dictionaries | Downloads | Administration
System Administrator - Logout

Dictionary Compilation

This page gives the Administrator and certain users -- those users granted permission to do so by Administrator -- a forum for posting User Dictionaries (UDs) to the server.The accepted User Dictionary formats are Microsoft Excel (*.xls), plain text (*.txt), Translation Memory eXchange for (*.tmx), XML (*.xml), and SYSTRAN dictionary files (*.dct).

SYSTRAN - WebServer 5.0 - Dic...

SYSTRAN WebServer 5.0

Translation | Dictionaries | Downloads | Administration
System Administrator - Logout

Dictionary Compilation

This page gives the Administrator and certain users -- those users granted permission to do so by the Administrator -- a forum for posting User Dictionaries (UDs) to the server.The accepted User Dictionary formats are Microsoft Excel (*.xls), plain text (*.txt), Translation Memory eXchange format (*.tmx), XML (*.xml), and SYSTRAN dictionary files (*.dct).

Your dictionary was successfully compiled.

User Dictionaries

● New (User Dictionaries)

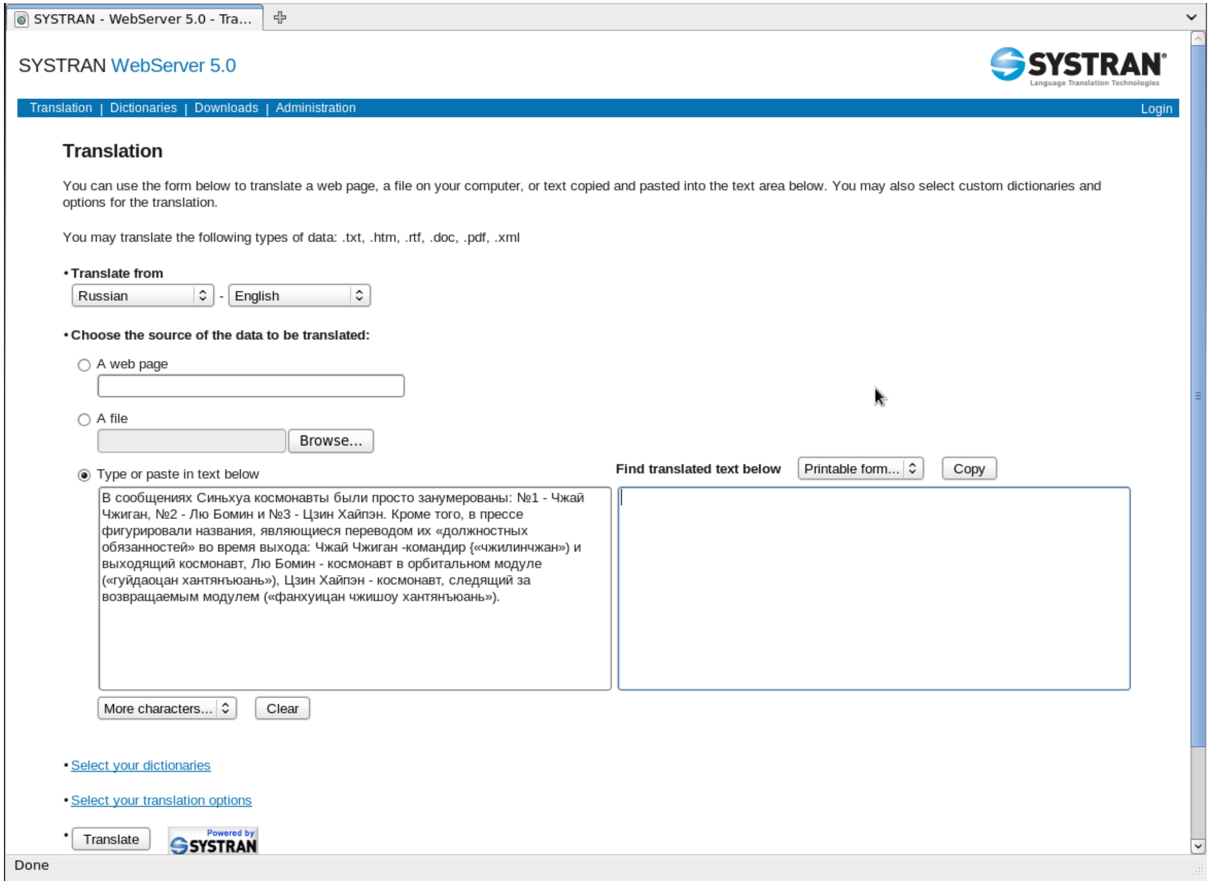
○ REVP-Systran/cosmonauts.rpd

REVP-Systran/cosmonauts.rpd
Owner: administrator
Source Language: RU
Target Languages: EN
Entries: 6

Copyright 2005 - All rights reserved
Legal notices | Help | Contact Administrator
Done

Select the dictionary tab, select dictionary compilation, and compile the dictionary file.

Applying the Dictionary to New Documents Using Systran


SYSTRAN - WebServer 5.0 - Tra...
SYSTRAN WebServer 5.0
SYSTRAN
Language Translation Technologies
Translation | Dictionaries | Downloads | Administration
Login
Browse...
Printable form...
Copy
More characters...
Clear
Translate
Powered by SYSTRAN
Done

Translation

You can use the form below to translate a web page, a file on your computer, or text copied and pasted into the text area below. You may also select custom dictionaries and options for the translation.

You may translate the following types of data: .txt, .htm, .rtf, .doc, .pdf, .xml

- **Translate from**
  Russian - English
- **Choose the source of the data to be translated:**
  ○ A web page
  ○ A file
  ● Type or paste in text below

В сообщениях Синьхуа космонавты были просто занумерованы: №1 - Чжай Чжиган, №2 - Лю Бомин и №3 - Цзин Хайпэн. Кроме того, в прессе фигурировали названия, являющиеся переводом их «должностных обязанностей» во время выхода: Чжай Чжиган -командир {«чжилинчжан») и выходящий космонавт, Лю Бомин - космонавт в орбитальном модуле («гуйдаоцан хантянъюань»), Цзин Хайпэн - космонавт, следящий за возвращаемым модулем («фанхуицан чжишоу хантянъюань»).

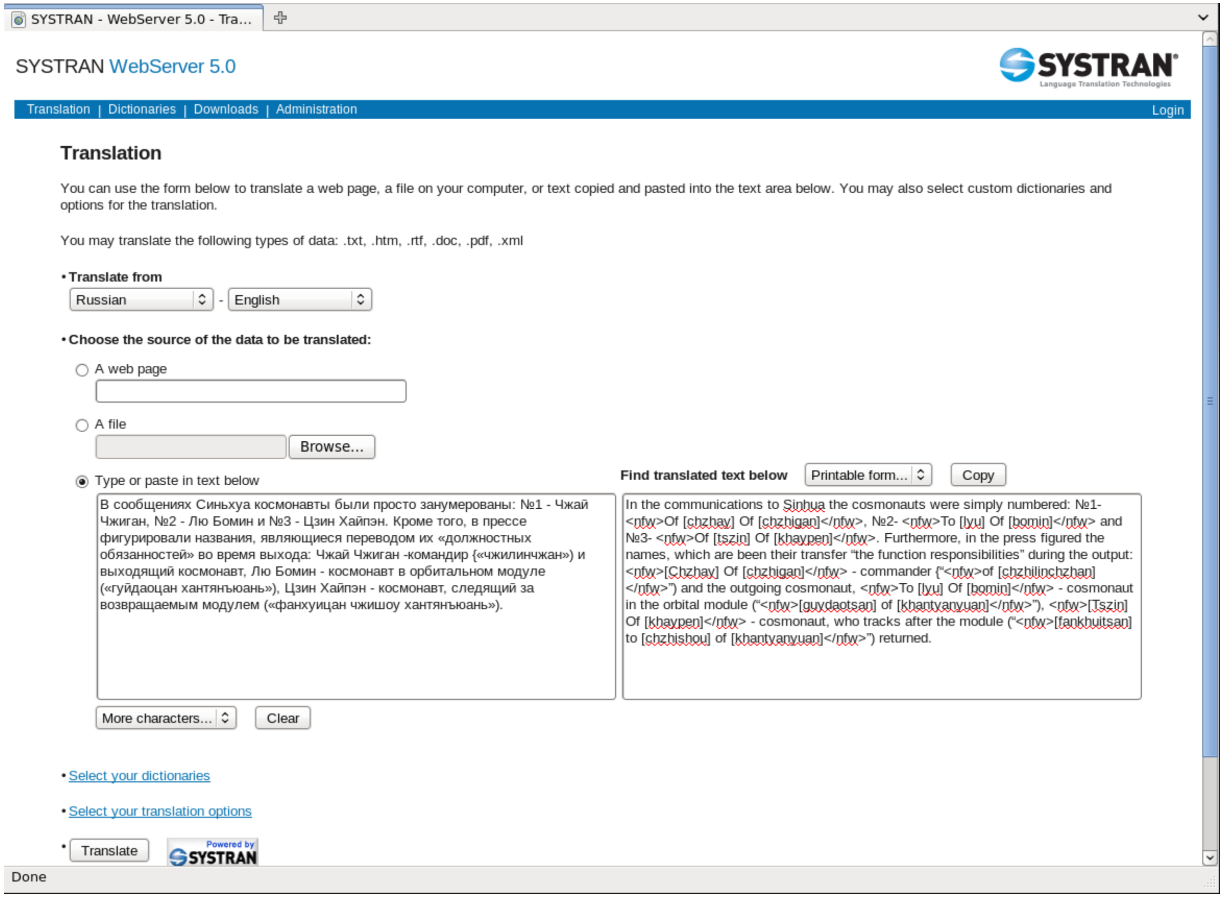
Find translated text below

- Select your dictionaries
- Select your translation options

Select the translation tab and enter the Russian text (or upload a file).

Distribution A: Approved for public release; distribution unlimited

Applying the Dictionary to New Documents Using Systran


SYSTRAN - WebServer 5.0 - Tra...
SYSTRAN WebServer 5.0
SYSTRAN Language Translation Technologies
Translation | Dictionaries | Downloads | Administration
Login
Browse...
Printable form...
Copy
More characters...
Clear
Translate
Powered by SYSTRAN
Done

Translation

You can use the form below to translate a web page, a file on your computer, or text copied and pasted into the text area below. You may also select custom dictionaries and options for the translation.

You may translate the following types of data: .txt, .htm, .rtf, .doc, .pdf, .xml

- **Translate from**

  Russian - English

- **Choose the source of the data to be translated:**

  ○ A web page

  ○ A file

  ● Type or paste in text below

В сообщениях Синьхуа космонавты были просто занумерованы: №1 - Чжай Чжиган, №2 - Лю Бомин и №3 - Цзин Хайпэн. Кроме того, в прессе фигурировали названия, являющиеся переводом их «должностных обязанностей» во время выхода: Чжай Чжиган -командир {«чжилинчжан») и выходящий космонавт, Лю Бомин - космонавт в орбитальном модуле («гуйдаоцан хантянъюань»), Цзин Хайпэн - космонавт, следящий за возвращаемым модулем («фанхуицан чжишоу хантянъюань»).

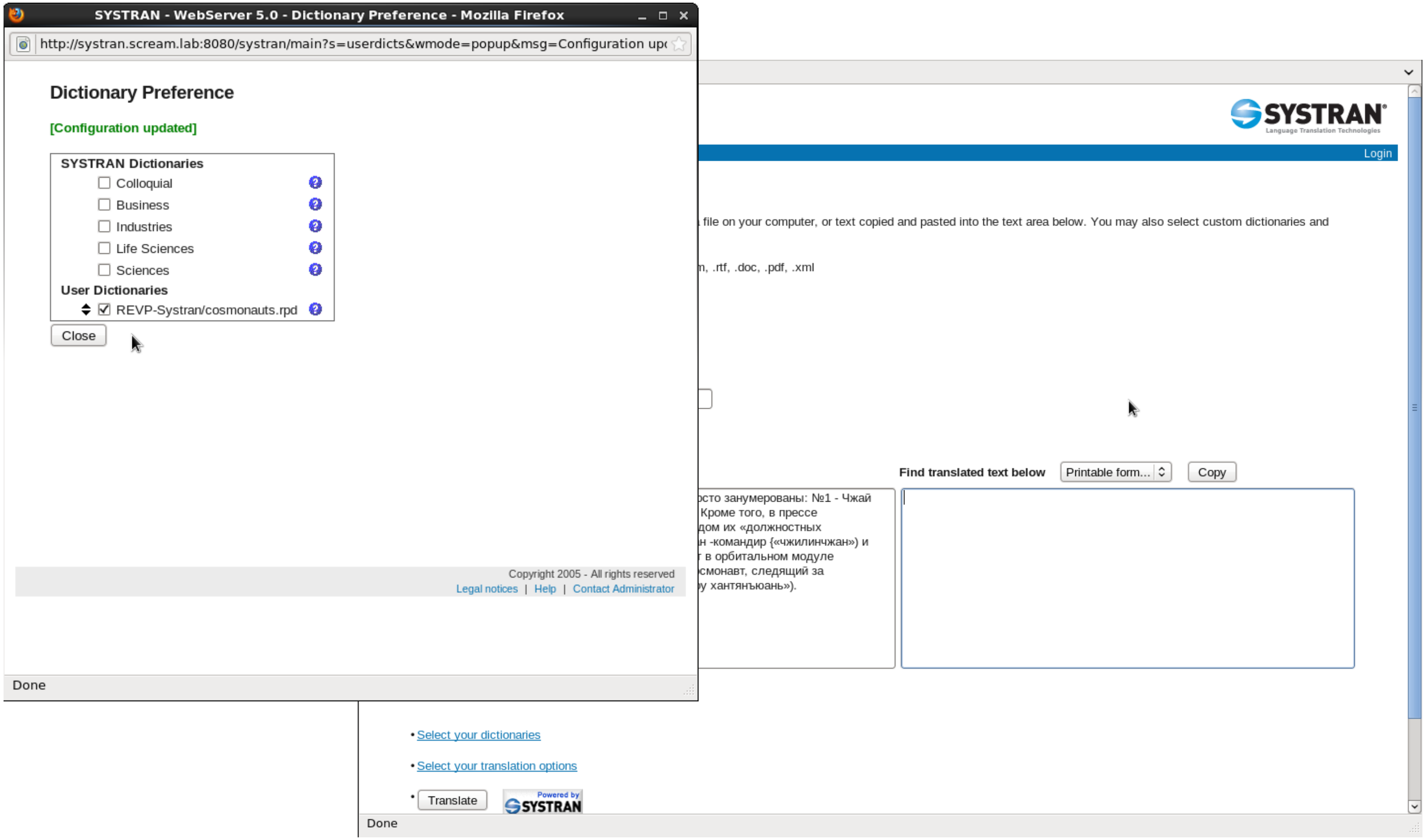
Find translated text below

In the communications to Sinhua the cosmonauts were simply numbered: №1- <nfw>Of [chzhay] Of [chzhigan]</nfw>, №2- <nfw>To [lyu] Of [bomin]</nfw> and №3- <nfw>Of [tszin] Of [khaypen]</nfw>. Furthermore, in the press figured the names, which are been their transfer “the function responsibilities” during the output: <nfw>[Chzhay] Of [chzhigan]</nfw> - commander {“<nfw>of [chzhilinchzhan]</nfw>”) and the outgoing cosmonaut, <nfw>To [lyu] Of [bomin]</nfw> - cosmonaut in the orbital module (“<nfw>[guydaotsan] of [khantyanyuan]</nfw>”), <nfw>[Tszin] Of [khaypen]</nfw> - cosmonaut, who tracks after the module (“<nfw>[fankhuitsan] to [chzhishou] of [khantyanyuan]</nfw>”) returned.

- Select your dictionaries
- Select your translation options

Before applying our user dictionary, we get transliterated versions of the Chinese names: chzhay chzhigan, tszin khaypen, lyu bomin

Applying the Dictionary to New Documents Using Systran



SYSTRAN - WebServer 5.0 - Dictionary Preference - Mozilla Firefox

http://systran.scream.lab:8080/systran/main?s=userdicts&wmode=popup&msg=Configuration upc

Dictionary Preference

[Configuration updated]

SYSTRAN Dictionaries

- ☐ Colloquial
- ☐ Business
- ☐ Industries
- ☐ Life Sciences
- ☐ Sciences

User Dictionaries

- ☑ REVP-Systran/cosmonauts.rpd

Close

Copyright 2005 - All rights reserved

Legal notices | Help | Contact Administrator

Done

file on your computer, or text copied and pasted into the text area below. You may also select custom dictionaries and

n, .rtf, .doc, .pdf, .xml

Find translated text below Printable form... Copy

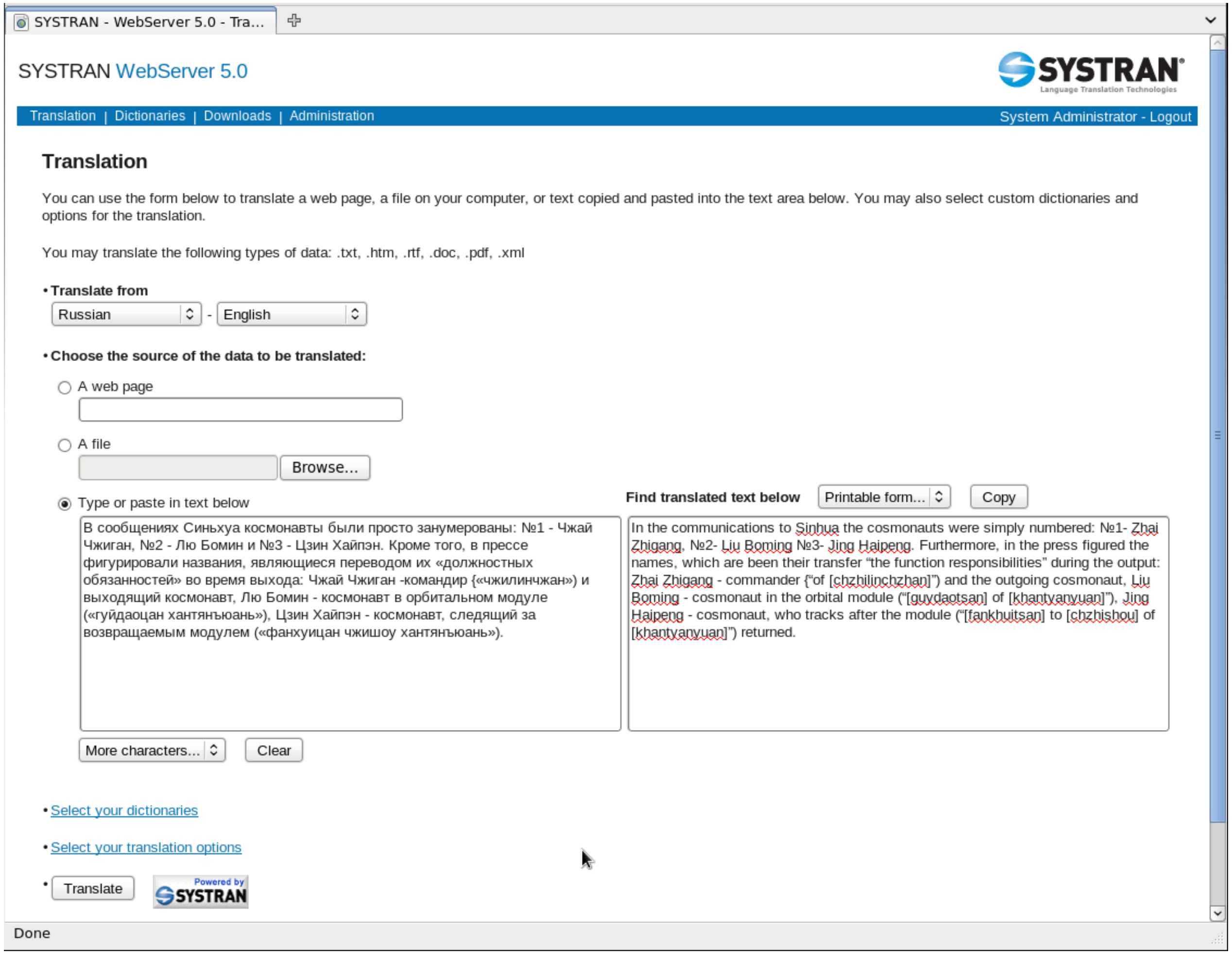
осто занумерованы: №1 - Чжай
Кроме того, в прессе
дом их «должностных
н -командир {«чжилинчжан») и
т в орбитальном модуле
смонавт, следящий за
у хантянъюань»).

- Select your dictionaries
- Select your translation options
- Translate

We click "Select your dictionaries" and check the RevP dictionary.

Distribution A: Approved for public release; distribution unlimited

Applying the Dictionary to New Documents Using Systran



SYSTRAN - WebServer 5.0 - Tra...
SYSTRAN WebServer 5.0
SYSTRAN
Language Translation Technologies
Translation | Dictionaries | Downloads | Administration
System Administrator - Logout

Translation

You can use the form below to translate a web page, a file on your computer, or text copied and pasted into the text area below. You may also select custom dictionaries and options for the translation.

You may translate the following types of data: .txt, .htm, .rtf, .doc, .pdf, .xml

- **Translate from**

Russian - English

- **Choose the source of the data to be translated:**

○ A web page

○ A file

● Type or paste in text below

В сообщениях Синьхуа космонавты были просто занумерованы: №1 - Чжай Чжиган, №2 - Лю Бомин и №3 - Цзин Хайпэн. Кроме того, в прессе фигурировали названия, являющиеся переводом их «должностных обязанностей» во время выхода: Чжай Чжиган -командир {«чжилинчжан») и выходящий космонавт, Лю Бомин - космонавт в орбитальном модуле («гуйдаоцан хантянъюань»), Цзин Хайпэн - космонавт, следящий за возвращаемым модулем («фанхуицан чжишоу хантянъюань»).

Find translated text below

In the communications to Sinhua the cosmonauts were simply numbered: №1- Zhai Zhigang, №2- Liu Boming №3- Jing Haipeng. Furthermore, in the press figured the names, which are been their transfer “the function responsibilities” during the output: Zhai Zhigang - commander {“of [chzhilinchzhan]”) and the outgoing cosmonaut, Liu Boming - cosmonaut in the orbital module (“[guydaotsan] of [khantyanyuan]”), Jing Haipeng - cosmonaut, who tracks after the module (“[fankhuitsan] to [chzhishou] of [khantyanyuan]”) returned.

- Select your dictionaries
- Select your translation options
- Translate

Now we get the correct forms: Zhai Zhigang, Jing Haipeng, and Liu Boming.

Distribution A: Approved for public release; distribution unlimited

Future Possibilities

- Integrate a named entity tagger to help select the Chinese names
- Include an option for Tongyong pinyin

  (used in Taiwan)
- Tag the corrected names as do-not-translate items going into Systran
- Adapt the program to other languages, to assist in information retrieval

Questions?

Katherine M. Young, PhD
N-Space Analysis, LLC
(937) 475-3637
nspaceanalysis@earthlink.net