1

Complex Technology of Machine Translation Resources Extension for the Kazakh Language

The material is devoted to creating linguistic resources such as parallel corpora and dictionaries for machine translation for low resourced languages. We describe manual methods to building corpora, usage of Bitextor tool for mining parallel corpora from online texts, usage of dictionary enrichment methodology so that people without deep linguistic knowledge could improve word dictionaries. All describe methods were applied to Kazakh, Russian and English languages with a task of machine translation between these languages in mind.

1.1. INTRODUCTION

Linguistic resources are an important part of any linguistic study. While languages that have been subjects of computational studies for a long time have a lot of resources ready to be used, other languages have an urgent need to develop such resources. Linguistic resources such as monolingual and parallel corpora, electronic dictionaries, and rule dictionaries are very important both for statistical and rule-based language processing. Development of the resources requires a lot of effort and time. It is only logical that low resourced languages need to take all possible opportunities to make that process easier and faster. Here, we describe how to use on-line texts and specialized tools to build and improve linguistic resources. With tools and approaches described it is possible to create sufficient amount of different linguistic resources for low resources languages.

The contribution of this work consists of the following: it unites several technologies of linguistic resources extension for parallel corpora and word dictionaries; the combined technology is applied to Kazakh-English and Kazakh-Russian language pairs. The combination of the three technologies allows using their results together for the improvement of each other. Larger corpora help to increase coverage of dictionaries.

1.2. RELATED WORKS

Creation of linguistic resources that are being considered in this work has been an important task for all the languages. Techniques used in the work have been tried for different language pairs, but not for Kazakh-English or Kazakh-Russian.

Development of parallel corpora using Bitextor has been described in following works. [1] describes Bitextor and apply it for collecting Catalan–Spanish–English parallel corpora. [2] describes the creation of English-French parallel corpus. [3] is devoted to the Finnish-English parallel corpus. [4] deals with English-Croatian corpus. There are also similar works on Portuguese-English, Portuguese-Spanish, Slovene-English and Serbian-English language pairs.

Dictionary enrichment methodology for people without deep linguistic knowledge is described in [5] for Spanish and in [6] for Croatian. There are no similar works performed for Kazakh or Russian.

The complex technology that is described in the paper has not yet been used as such for one language pair. Only parts of it have been tested and applied to different languages.

1.3. MANUAL APPROACHES TO BUILDING CORPORA

In order to collect parallel text corpora we used three different approaches:

a) finding all significant ready to use aligned parallel texts;

b) using scrips for crawling texts from websites that contain same texts in several languages and using InterText tool with integrated hunalign tool for aligning them;

c) using bitextor tool for crawling websites that contain same texts in several languages and aligning them.

Approaches 2 and 3 seem to be similar to each other, but they have produced different amount of results which is described below.

There are not many places to find ready to use parallel text corpora that have Kazakh as one of the languages. In fact, there is one such place - the OPUS project (opus.lingfil.uu.se). There is some parallel Kazakh-English texts collected from Tatoeba and OpenSubtitles.

Another ready to use resource is the Bible. It has been repeatedly translated into many languages. The Kazakh is also among them. There were several translations prepared by several organizations. The most resent one is called "New World Translation of the Christian Greek Scriptures" published on different media by Jehovah's Witnesses. Despite the nature of the organization it is turned out to be a great parallel resource since the text of the book has strictly numbered chapters and verses across all translations.

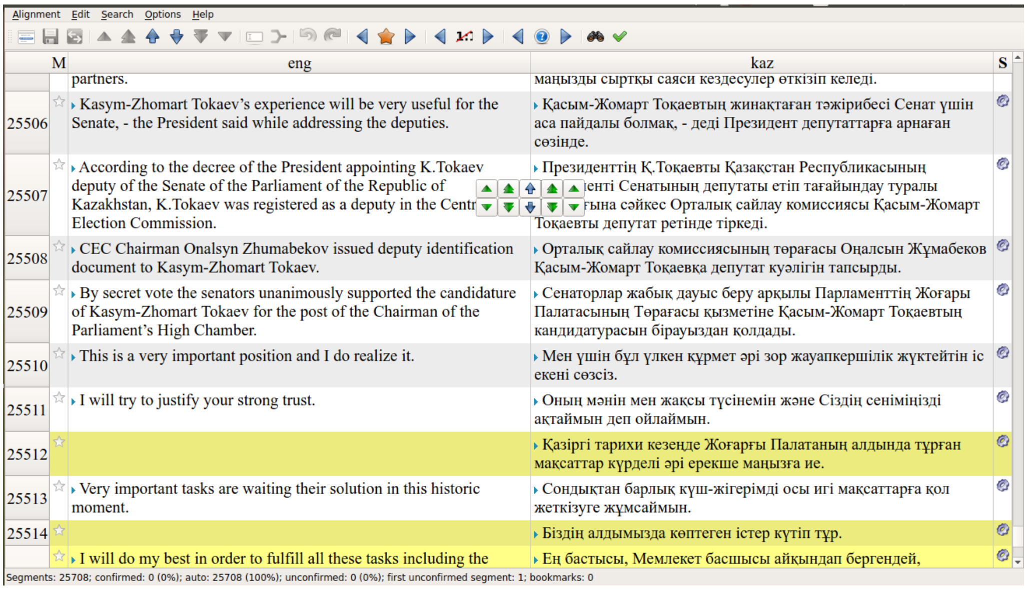
The second approach is partially automated but also involves manual checking of the results. It consists of following stages:

- crawling parallel texts on the internet;
- cleaning and formatting of gathered texts;
- sentence splitting;
- sentence alignment;
- manual checking.

All stages except the last one can be automated. But the quality of the parallel text will affect quality of the tasks to be solved with them. So in our opinion human involvement is mandatory.

As a source for parallel texts we used web-sites http://www.akorda.kz/ and https://www.ted.com/. Texts from the first one were collected using scripts links.pl and extract_text.pl, that are available in Apertium project's repository using the following link: https://sourceforge.net/p/apertium/svn/59905/tree/languages/apertium-kaz/texts/akorda/. Texts from the second site were collected manually.

After cleaning, formatting and sentence splitting we had two lists of sentences in two languages that were translations of each other but the sentences themselves were not aligned due to various translation reasons. To align them we used hunalign tool. Hunalign has remarkably high quality: we got 6-8% of incorrectly aligned sentences out of unaligned lists mentioned above. But low percentage still meant that we had 2000-3000 alignment mistakes. It is quite many and that is why manual checking was due. Parallel text alignment editor called InterText was used for that (Fig. 1.1).



	M	eng	kaz	S
		partners.	маңызды сыртқы саяси кездесулер өткізіп келеді.	
25506	☆	Kasym-Zhomart Tokaev's experience will be very useful for the Senate, - the President said while addressing the deputies.	Қасым-Жомарт Тоқаевтың жинақтаған тәжірибесі Сенат үшін аса пайдалы болмақ, - деді Президент депутаттарға арнаған сөзінде.	
25507	☆	According to the decree of the President appointing K.Tokaev deputy of the Senate of the Parliament of the Republic of Kazakhstan, K.Tokaev was registered as a deputy in the Centr Election Commission.	Президенттің Қ.Тоқаевты Қазақстан Республикасының енті Сенатының депутаты етіп тағайындау туралы ғына сәйкес Орталық сайлау комиссиясы Қасым-Жомарт Тоқаевты депутат ретінде тіркеді.	
25508	☆	CEC Chairman Onalsyn Zhumabekov issued deputy identification document to Kasym-Zhomart Tokaev.	Орталық сайлау комиссиясының төрағасы Оңалсын Жұмабеков Қасым-Жомарт Тоқаевқа депутат куәлігін тапсырды.	
25509	☆	By secret vote the senators unanimously supported the candidature of Kasym-Zhomart Tokaev for the post of the Chairman of the Parliament's High Chamber.	Сенаторлар жабық дауыс беру арқылы Парламенттің Жоғары Палатасының Төрағасы қызметіне Қасым-Жомарт Тоқаевтың кандидатурасын бірауыздан қолдады.	
25510	☆	This is a very important position and I do realize it.	Мен үшін бұл үлкен құрмет әрі зор жауапкершілік жүктейтін іс екені сөзсіз.	
25511	☆	I will try to justify your strong trust.	Оның мәнін мен жақсы түсінемін және Сіздің сеніміңізді ақтаймын деп ойлаймын.	
25512	☆		Қазіргі тарихи кезеңде Жоғарғы Палатаның алдында тұрған мақсаттар күрделі әрі ерекше маңызға ие.	
25513	☆	Very important tasks are waiting their solution in this historic moment.	Сондықтан барлық күш-жігерімді осы игі мақсаттарға қол жеткізуге жұмсаймын.	
25514	☆		Біздің алдымызда көптеген істер күтіп тұр.	
	☆	I will do my best in order to fulfill all these tasks including the	Ең бастысы, Мемлекет басшысы айқындап бергендей,	

Segments: 25708; confirmed: 0 (0%); auto: 25708 (100%); unconfirmed: 0 (0%); first unconfirmed segment: 1; bookmarks: 0

Fig. 1.1. InterText editor for aligned parallel texts

Source: own elaboration

Approach described above resulted in two text corpora:

1. Akorda - 24 148 aligned sentences.
2. TED - 6 120 aligned sentences.

1.4. BUILDING BILINGUAL LINGUISTIC CORPORA USING BITEXTOR

Creation of multilingual parallel corpora is one of the important tasks in the field of machine translation, especially for statistical machine translation. Today, the Internet can be considered a large multi-lingual corpus, because it contains a large number of websites with texts in different languages. Pages of the sites can be considered as parallel texts (bitexts). Bitextor is the tool that is used to collect and align parallel texts from websites.

Bitextor is a free open source application for collection of translation memories from multilingual websites. The application downloads all HTML files from a website then pre-processes them into a consistent format and applies a set of heuristics to select the file pairs that contain the same text in two different languages (bitexts). Using LibTagAligner library translation memories in TMX format are created from these parallel texts. The library uses HTML-tags and length of the text segments for alignment [1]. After cleaning the resulting translation memory from TMX format tags, we receive a parallel corpus with sentences in different languages aligned with each other.

Previous methods for aligning parallel texts were based on the length of sentences [7]. Systems based on these methods show good results for languages with a high correlation between the length of the sentences, but their disadvantage is that many texts are not translated sentence-to-sentence. One sentence in the source text may correspond to two or more sentences in the translation. Therefore, Bitextor uses two methods to determine parallel texts: structure of HTML tags and length of sentences.

A key element in Bitextor is the ability to compare file pairs and identify parallel texts in them. To do this, first of all, it uses file metrics (they can be called "fingerprints"), which are determined from numbered text segments. But before comparing file metrics, a set of heuristics is used. After applying heuristics, Bitextor does not need to process every pair of files to compare all of the metrics to each other. Metrics comparison is performed only if the file pair meets all the heuristics. List of heuristics:

1. Comparison of a language of the text: if two files are written in the same language, one cannot be a translation of the other.
2. Comparison of file extensions: if within the same site one file is a translation of another file, they usually have the same extension.
3. File size coefficient: this parameter is relative and used to filter a pair of files whose size is different from each other.
4. The total difference between lengths of the texts: this option has the same function as the previous one, but it measures the size of the plain text of every file in the symbols.

The process of creating corpora with Bitextor consists of several successive stages described below.

During download stage, website files are copied onto a computer using HTTrack application. This application downloads all HTML files from a multilingual website. Doing that it maintains directory tree structure.

During the next stage, all downloaded files are pre-processed in order to adapt them to the next stages. Bitextor uses LibTidy library to standardize possibly incorrect HTML files into valid XHTML files. It guarantees that tag structure within these files is proper. Original HTML file encoding is converted into UTF-8.

Once the files have been pre-processed, next step is to gather some information needed to compare files and generating the translation memory, such as name and file extension. The language of each text is determined using LibTextCat library. File metrics are also determined in this step.

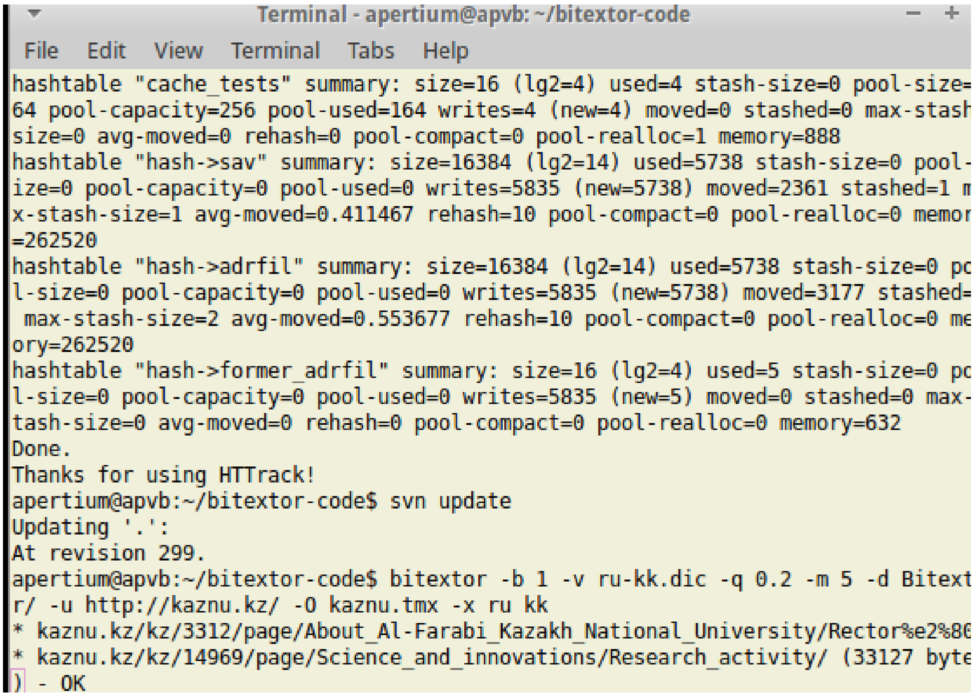
Information obtained from files is stored in a list, organized in accordance with a position of analyzed file in the directory tree. It makes access to information easier, as file comparison is done level by level.

On the stage of comparing files and translation memory generation, comparison of files begins with a comparison of the levels. The user can limit the difference in depth of the directory tree during the comparison. Parallel texts, as a rule, are at the same level in the tree or at very close levels, so there is no need to compare each file with all files at different levels.

Generation of translation memory in TMX format is done using LibTagAligner library. As with the metrics, Bitextor uses integer numbers for representing tags and text blocks in TagAligner.

1.5. RESULTS OF THE DEVELOPMENT OF BILINGUAL PARALLEL CORPORA FOR KAZAKH-ENGLISH AND KAZAKH-RUSSIAN LANGUAGE PAIRS

Bitextor has beer run for following websites: http://www.kaznu.kz (Fig. 1.2), http://www.bolashak.gov.kz, http://www.enu.kz, http://egov.kz, http://www.kazpost.kz, http://www.archeolog.kz, http://e-history.kz, http://inform.kz, http://primeminister.kz, http://tengrinews.kz and other, about 20 in total.

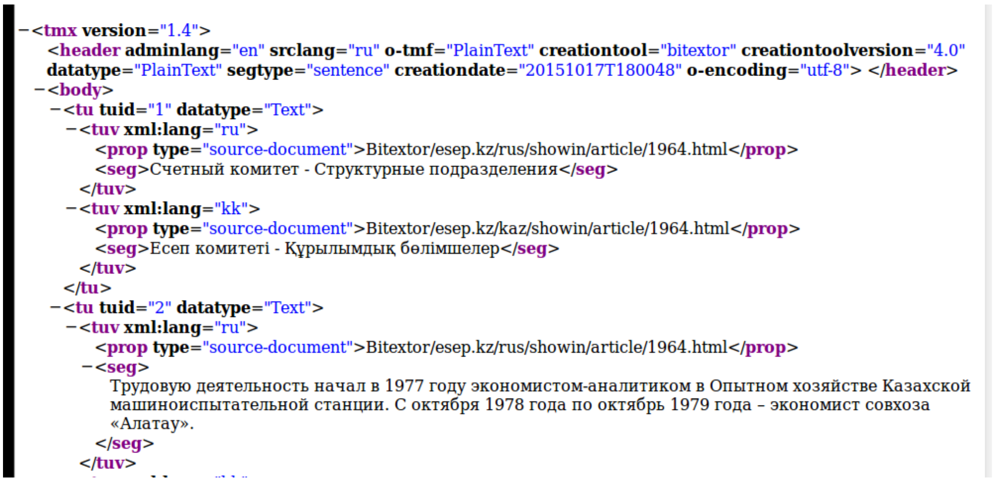

Terminal - apertium@apvb: ~/bitextor-code
File Edit View Terminal Tabs Help

```
hashtable "cache_tests" summary: size=16 (lg2=4) used=4 stash-size=0 pool-size=
64 pool-capacity=256 pool-used=164 writes=4 (new=4) moved=0 stashed=0 max-stash
size=0 avg-moved=0 rehash=0 pool-compact=0 pool-realloc=1 memory=888
hashtable "hash->sav" summary: size=16384 (lg2=14) used=5738 stash-size=0 pool-
ize=0 pool-capacity=0 pool-used=0 writes=5835 (new=5738) moved=2361 stashed=1 m
x-stash-size=1 avg-moved=0.411467 rehash=10 pool-compact=0 pool-realloc=0 memor
=262520
hashtable "hash->adrfil" summary: size=16384 (lg2=14) used=5738 stash-size=0 po
l-size=0 pool-capacity=0 pool-used=0 writes=5835 (new=5738) moved=3177 stashed=
 max-stash-size=2 avg-moved=0.553677 rehash=10 pool-compact=0 pool-realloc=0 me
ory=262520
hashtable "hash->former_adrfil" summary: size=16 (lg2=4) used=5 stash-size=0 po
l-size=0 pool-capacity=0 pool-used=0 writes=5835 (new=5) moved=0 stashed=0 max-
tash-size=0 avg-moved=0 rehash=0 pool-compact=0 pool-realloc=0 memory=632
Done.
Thanks for using HTTrack!
apertium@apvb:~/bitextor-code$ svn update
Updating '.':
At revision 299.
apertium@apvb:~/bitextor-code$ bitextor -b 1 -v ru-kk.dic -q 0.2 -m 5 -d Bitext
r/ -u http://kaznu.kz/ -O kaznu.tmx -x ru kk
* kaznu.kz/kz/3312/page/About_Al-Farabi_Kazakh_National_University/Rector%e2%80
* kaznu.kz/kz/14969/page/Science_and_innovations/Research_activity/ (33127 byte
) - OK
```

Fig. 1.2. An example of running Bitextor for www.kaznu.kz

Source: own elaboration

As a result of Bitextor's work from each site we obtained *.tmx file with the format presented in Fig. 1.3.



```
−<tmx version="1.4">
    <header adminlang="en" srclang="ru" o-tmf="PlainText" creationtool="bitextor" creationtoolversion="4.0"
    datatype="PlainText" segtype="sentence" creationdate="20151017T180048" o-encoding="utf-8"> </header>
  −<body>
    −<tu tuid="1" datatype="Text">
      −<tuv xml:lang="ru">
          <prop type="source-document">Bitextor/esep.kz/rus/showin/article/1964.html</prop>
          <seg>Счетный комитет - Структурные подразделения</seg>
        </tuv>
      −<tuv xml:lang="kk">
          <prop type="source-document">Bitextor/esep.kz/kaz/showin/article/1964.html</prop>
          <seg>Есеп комитеті - Құрылымдық бөлімшелер</seg>
        </tuv>
      </tu>
    −<tu tuid="2" datatype="Text">
      −<tuv xml:lang="ru">
          <prop type="source-document">Bitextor/esep.kz/rus/showin/article/1964.html</prop>
        −<seg>
           Трудовую деятельность начал в 1977 году экономистом-аналитиком в Опытном хозяйстве Казахской
           машиноиспытательной станции. С октября 1978 года по октябрь 1979 года – экономист совхоза
           «Алатау».
          </seg>
        </tuv>
```

Fig. 1.3. A format of obtained parallel corpus for Kazakh-Russian language pair

Source: own elaboration

In this format (Fig. 1.3), tag <tu> includes a pair of aligned segments (in this case - sentences); tag <tuv> - separate sentences in two languages; tag <prop> - HTML file addresses from which these sentences have been extracted; tag <seg> - sentences themselves. In such *.tmx file sentence in one language corresponds to the sentence in another language. It should be noted that comparison quality depends on the website. Thus, we receive a file with parallel texts.

As it can be seen Bitextor allows saving human and time resources and obtaining parallel aligned corpora from multilingual websites. The corpora then can be used for ensuring dictionary coverage.

Bitextor is not the only way to gather parallel corpora. There are some other approaches that also can be used. Two of them are:

a) finding all significant ready to use aligned parallel texts;
b) manually gather texts from websites that contain same materials in several languages and using InterText tool with an integrated hunalign tool for aligning them.

All mentioned and described approaches can help to produce a significant amount of corpora. All the raw text corpora described in this section are available at https://drive.google.com/drive/folders/0B3f-xwS1hRdDM2VpZXRVblRRUmM.
Information about all the text corpora that we have gathered is provided in Tab. 1.1.

Table 1.1. Kazakh-English and Kazakh-Russian corpora stats.

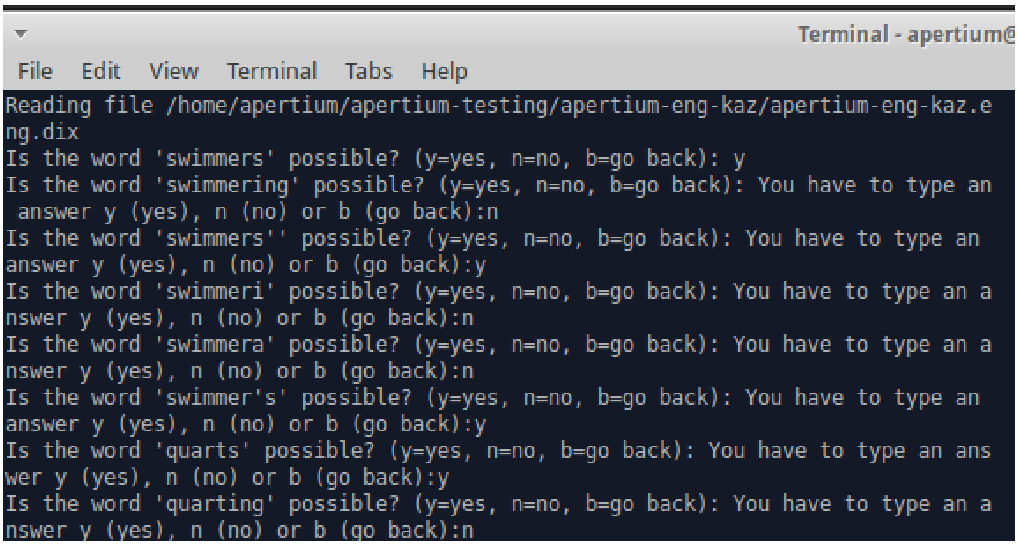
#	Corpus	Languages involved	# of sentence pairs	# of words kaz	# of words eng / rus
1	OPUS	Kazakh-English	4 480	19 892	27 839
2	New World Bible	Kazakh-English	32 358	548 258	824 398
3	Lab IIS	Kazakh-English	5 925	112 658	157 313
4	Akorda	Kazakh-English	24 148	341 154	456 689
5	TED	Kazakh-English	6 120	54 965	79 320
6	Kaz-rus	Kazakh-Russian	14 290	235 189	243 398
		TOTAL:	87 321	1 312 116	1 788 957

1.6. AUTOMATED ENRICHMENT OF MACHINE TRANSLATION SYSTEM DICTIONARIES

Dictionaries are necessary for translation of texts from one language to another. There are thousands of translation dictionaries between hundreds of languages and each of them can contain many thousands of words. Usually, paper version of a dictionary is a book of hundreds of pages for which a search for the right word is a fairly long and laborious process. Dictionaries used in machine translation may contain translations into different languages of hundreds of thousands of words and phrases, as well as provide users with additional features. Such as giving a user an ability to select the languages and translation direction, provide a quick search for words, ability to enter phrases, etc.

Today there are many methods of expanding dictionaries. We used method realized in Apertium by Miquel Esplà-Gomis. We used the tool to fill dictionaries for English-Kazakh, Kazakh-English language pairs in the free/open-source Apertium machine translation system. English-Kazakh MT system has three types of dictionaries: English monolingual, Kazakh monolingual and English-Kazakh bilingual dictionary. All dictionaries, except Kazakh monolingual, have XML format, each word has tag showing which part-of-speech it is [8].

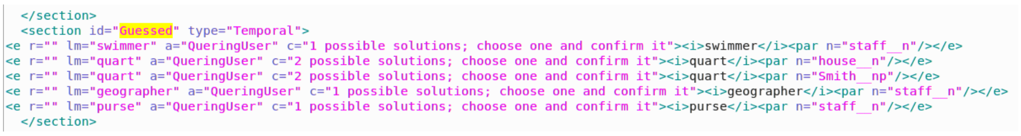
The method is used to assign stems and inflectional paradigms to unknown words if unknown word's paradigm (word pattern) does not appear in dictionaries. The tool needs a file with a list of unknown words that will be added into monolingual dictionary, monolingual dictionary, new dictionary that will be created with the new words added to special section marked “Guessed”, information about a number of questions to be asked. For Kazakh language, it also needs the automation of second-level rules for MT system. The list of unknown words has to be pre-processed with the collection of scripts provided with the tool. After the tool is launched a user can choose among different combination of candidate stems and paradigms correct ones by answering questions asked by system. When a user confirms that the words have been detected correctly they get moved to appropriate dictionary section. In case when the system finds more than one solution for a word all possible options are written to the dictionary along with the number of found possible options.



```
Reading file /home/apertium/apertium-testing/apertium-eng-kaz/apertium-eng-kaz.e
ng.dix
Is the word 'swimmers' possible? (y=yes, n=no, b=go back): y
Is the word 'swimmering' possible? (y=yes, n=no, b=go back): You have to type an
 answer y (yes), n (no) or b (go back):n
Is the word 'swimmers'' possible? (y=yes, n=no, b=go back): You have to type an
answer y (yes), n (no) or b (go back):y
Is the word 'swimmeri' possible? (y=yes, n=no, b=go back): You have to type an a
nswer y (yes), n (no) or b (go back):n
Is the word 'swimmera' possible? (y=yes, n=no, b=go back): You have to type an a
nswer y (yes), n (no) or b (go back):n
Is the word 'swimmer's' possible? (y=yes, n=no, b=go back): You have to type an
answer y (yes), n (no) or b (go back):y
Is the word 'quarts' possible? (y=yes, n=no, b=go back): You have to type an ans
wer y (yes), n (no) or b (go back):y
Is the word 'quarting' possible? (y=yes, n=no, b=go back): You have to type an a
nswer y (yes), n (no) or b (go back):n
```

Fig. 1.4. Example of using the method for adding words to English monolingual dictionary

Source: own elaboration



```
  </section>
  <section id="Guessed" type="Temporal">
<e r="" lm="swimmer" a="QueringUser" c="1 possible solutions; choose one and confirm it"><i>swimmer</i><par n="staff__n"/></e>
<e r="" lm="quart" a="QueringUser" c="2 possible solutions; choose one and confirm it"><i>quart</i><par n="house__n"/></e>
<e r="" lm="quart" a="QueringUser" c="2 possible solutions; choose one and confirm it"><i>quart</i><par n="Smith__np"/></e>
<e r="" lm="geographer" a="QueringUser" c="1 possible solutions; choose one and confirm it"><i>geographer</i><par n="staff__n"/></e>
<e r="" lm="purse" a="QueringUser" c="1 possible solutions; choose one and confirm it"><i>purse</i><par n="staff__n"/></e>
  </section>
```

Fig. 1.5. Generated dictionary entries

Source: own elaboration

We have been using this methodology to extend a number of words in dictionaries, which mainly effects to the quality of the translation in machine translation. The technology allows non-expert users who do not have a deep knowledge in computational representation of morphology but understand the language being developed participate in building dictionaries. That means more people can add dictionary entries creating larger dictionaries in less time. As for the moment there are 33 174 entries in Apertium's English-Kazakh dictionary and 31 189 entries in Apertium's Kazakh-Russian dictionary.

1.7. EXPERIMENT RESULTS

After implementing the technologies described in the paper experiments on machine translation quality for the language pairs have been conducted. Collected resources were incorporated in Apertium machine translation platform. After that translation quality was compared with Sanasoft and Google Translate – both machine translation applications that support Kazakh-English and Kazakh-Russian language pairs. The results of experiments are shown in Tab. 1.2-1.5.

Table 1.2. BLEU scores for Russian-Kazakh translation

MT Application	Unigrams %	Bigrams %	Trigrams %	Total %
Google Translate	18,1176	6,7296	2,6470	9,1648
Sanasoft	11,5294	0,8403	0	4,1232
Apertium	18,9411	4,8011	1,7352	8,4925

Table 1.3. BLEU scores for Russian-Kazakh translation

MT Application	Unigrams %	Bigrams %	Trigrams %	Total %
Google Translate	11,4375	3,6458	2,9220	6,0018
Sanasoft	17,8125	4,3229	2,1306	8,0886
Apertium	12.6875	4.3020	0	5.6631

Table 1.4. BLEU scores for English-Kazakh translation

MT Application	Bigrams %	Trigrams %	Total %
Google Translate	11,49	4,9	20,57
Sanasoft	6,298223	0,9127744	15,74
Apertium	43,87676	30,24746	58,97

Table 1.5. BLEU scores for Kazakh-English translation

MT Application	Bigrams %	Trigrams %	Total %
Google Translate	23,3475	17,52087	33,08
Sanasoft	4,072844	0,4729024	13,97
Apertium	18,077	5,6	34,5

1.8. Conclusion

We have described complex technology of building linguistic resources for low-resourced languages. We have shown how to create parallel corpora manually and using Bitextor and a method of enriching dictionaries with new words without much of linguistic knowledge. The results of applying described methods for Kazakh, Russian and English languages show that they allow producing sufficient amount of linguistic resources within a time period of several months and thus helping to support research work in the field of natural language processing.

The work contributes to saving human and time resources when creating linguistic resources for machine translation and has been applied to Kazakh-English and Kazakh-Russian language pairs.

References

[1] Esplà-Gomis M. (2009) *Bitextor: a Free/Open-source Software to Harvest Translation Memories from Multilingual Websites*. Proceedings of MT Summit XII, Ottawa, Canada, Association for Machine Translation in the Americas.

[2] Esplà-Gomis M. and Forcada M. (2010) *Combining content-based and URL-based heuristics to harvest aligned bitexts from multilingual sites with Bitextor*. The Prague Bulletin of Mathematical Linguistics, 93, pp.77-86.

[3] Rubino R., Pirinen T., Espla-Gomis M., Ljubešic N., Ortiz Rojas S., Papavassiliou V., Prokopidis P., and Toral, A. (2015) *Abu-MaTran at WMT 2015 Translation Task: Morphological Segmentation and Web* Crawling. In Proceedings of the Tenth Workshop on Statistical Machine Translation, pp. 184-191.

[4] Esplà-Gomis M., Klubicka F., Ljubesic N., Ortiz-Rojas S., Papavassiliou V., and Prokopidis P. (2014) May. *Comparing two acquisition systems for automatically building an English-Croatian parallel corpus from multilingual websites*. In LREC, pp. 1252-1258.

[5] Espla-Gomis M., Carrasco R.C., Sánchez-Cartagena V.M., Forcada M.L., Sánchez-Martınez F., and Pérez-Ortiz J.A. (2014) *An efficient method to assist non-expert users in extending dictionaries by assigning stems and inflectional paradigms to unknown words*. In Proceedings of the 17th Annual Conference of the European Association for Machine Translation, pp. 19-26.

[6] Ljubešic N., Espla-Gomis M., Klubicka F., and Preradovic, N.M. (2015) *Predicting Inflectional Paradigms and Lemmata of Unknown Words for Semi-automatic Expansion of Morphological Lexicons*. In Proceeding of International Conference Recent Advances in Natural Language, pp. 379-387.

[7] Brown P., Lai J., and Mercer R. (1991) *Aligning sentences in parallel corpora.* In Proceedings of the 29th annual meeting on Association for Computational Linguistics., Association for Computational Linguistics Morristown, NJ, USA, pp. 169–176.

[8] Forcada M. L., Ginestí-Rosell M., Nordfalk J., O'Regan J., Ortiz Rojas S., Pérez Ortiz J. A., Sánchez Martínez F., Ramírez Sánchez G., and Tyers F. M. (2011)

Apertium: a free/open source platform for rule based machine translation. Machine translation, 25(2), pp. 127-144.

[9] Karlsson F., Voutilainen A., Heikkilä J., and Anttila, A. (1995) *Constraint Grammar: A language independent system for parsing unrestricted text.* Mouton de Gruyter.

[10] Och F.J. and Ney H. (2003) *A systematic comparison of various statistical alignment models.* Computational Linguistics. 29(1), pp. 19–51.