# ENGLISH-KAZAKH PARALLEL CORPUS FOR STATISTICAL MACHINE TRANSLATION

Ayana Kuandykova[1], Amandyk Kartbayev[2] and Tannur Kaldybekov[3]

School of Mechanics and Mathematica, al-Farabi Kazakh National University, Almaty, Kazakhstan[1,2]

Department of Information Systems, NRU ITMO, St. Petersburg, Russia[3]

## ABSTRACT

*This paper presents problems and solutions in developing English-Kazakh parallel corpus at the School of Mechanics and Mathematics of the al-Farabi Kazakh National University. The research project included constructing a 1,000,000-word English-Kazakh parallel corpus of legal texts, developing an English-Kazakh translation memory of legal texts from the corpus and building a statistical machine translation system. The project aims at collecting more than ten million words. The paper further elaborates on the procedures followed to construct the corpus and develop the other products of the research project. Methods used for collecting data and the results are discussed, errors during the process of collecting data and how to handle these errors will be described.*

## KEYWORDS

*Parallel corpus, statistical machine translation, collecting data, legal texts, English-Kazakh corpus*

## 1. INTRODUCTION

The Kazakh language is a majority language within Kazakhstan Republic. While 11 million claim fluency in the language (of an overall population of about 17 million), it is estimated that the number who use Kazakh as their daily language of communication is in the region of 7-8 million. The areas in which Kazakh is a daily community language are geographically located in southeast of the land to the southwest and south, with one exception in the Almaty city. Others who use Kazakh on a daily basis include village dwellers, some of which work for local government structures or are involved in Kazakh education. While the conditions for the spoken language have been poor over the years, resources relating to the written language have been in development with considerable dedication and enthusiasm. Much activity has focussed on digitization of sources over the past ten years, with the result that we now have several sources.

There are many resources that contain sentences and words aligned in two languages. One of the most well known parallel corpora is Europarl (Koehn, 2002), which is a collection of material including 11 European languages taken from the proceedings of the European Parliament. Another often-used parallel corpus is the JRC-Acquis Multilingual Parallel Corpus (Steinberger et al., 2006). The corpus consists of documents of legislative text, covering a variety of domains for above 20 languages. The Bible translated to a large number of languages and collected and annotated by Resnik et al. (1999) is another example of a freely available parallel language resource.

Kazakh language corpus belongs to a new type of corpus resources; consequently we find a small

10.5121/ijnlc.2014.3306

amount of available resources. Kazakh National corpus (website: til.gov.kz) is one of the earliest corpuses, but it is not annotated and very small. Kazakh Language Corpus (Makhambetov et al., 2013) is another resource with linguistic annotation and which is rapidly developing by researchers from Nazarbayev University. We also read about a Kazakh corpus that has been developed at Xinjiang University and used in research by Altenbek and Xiao-long, 2010. But couldn’t find enough information about a corpus, as we can’t find any published projects in the research done by Mukan, 2012. Baisa et al., 2012, built also the new small corpus as a part of Turkic languages corpus. We think these Kazakh corpuses will have an own impact and it will be useful tool in the analysis of Kazakh.

The corpora mentioned below are Legal Text Corpora developed for lexicographical and research purposes. That the corpora cover the span of legal texts freely available from official web site: www.adilet.zan.kz. This is a parallel English-Kazakh corpus containing about 46,000 sentences from articles of the legal codes.

By “parallel corpus”, we mean a text which is available in English and Kazakh languages: it is an original text and its translation, also there are texts which has been written by a consortium of authors in a variety of languages (e.g. UN conventions), and then published in various language versions. Building a parallel corpus of high quality from that kind of raw data is not straightforward because of low initial precision, frequent embedding of nonparallel fragments in parallel texts, and low-quality documents. The rest of the paper is organized as follows: next section provides an overview of the system architecture and addresses specific problems at the preparatory stage. Section 3 describes the sentence-alignment algorithm and procedure. Section 4 we evaluate the quality of the final parallel corpus and provide some statistical information about English-Kazakh language pair. We conclude in final section with short summary remarks.

Our final goal is to build a big corpus of parallel sentences good enough for training a statistical machine translation system.

## 2. SYSTEM OVERVIEW

The first idea is to find similar candidate pairs of sentences using parallel tags, often accompanied by anchors, or pairs of filenames which differ only in the identification of a language, e.g. with alternative directories in the paths, or suffixes such as «en» and «kz».

These candidates are then evaluated by comparing, in a very simplistic manner, their content: since they are usually HTML documents, it is usually quite easy to align the HTML markup (heading and paragraph, for example), and to compare the amount of text between each anchor. In this way, we get a rough map of the structures of the two documents. These can then be compared using variety techniques, which may include the kinds of linguistic methods used in the alignment of known parallel sentences. For example, next paragraph shows part of parallel English and Kazakh pages with minor differences in markup and content:

```
</div><div class="container_omega text text_new"><div class="gs_12">
<article><p id="z2">TEXT</p></article>
</div></div></div>
```

In this paper we address the tasks of extraction of the best parallel fragments. Mining parallel texts from a big document collection usually involves detecting a set of potential parallel document pairs with low-precision algorithms, filtering of unwanted texts, e.g. footnotes. In our experiments, we used a sentence-alignment algorithm similar to (Brown et al., 1991; Gale and

Church, 1993), it is mainly aimed at achieving high precision rather than high recall. Finding parallel texts in web documents is a task that has methods mainly based on the analysis of HTML markup and HTML parsing tools, e.g. BeautifullSoup.

We designed a crawler based on BeautifulSoup library, that depends from a specific structure of the source document. BeautifulSoup is a Python library for pulling data out of HTML and XML files. It provides idiomatic ways of navigating, searching, and modifying the parse tree. So, the corpus building procedure includes several steps presented in Fig. 1. Our main sources of documents are web pages from web site database with their textual contents already extracted and sentence boundaries detected. Nevertheless documents often include sentences that are site-specific and carry some meta-information, advertising, or just some noise. When often repeated such sentences may confuse training process, so we removed subsequent sentences that have been encountered recently.

After cleaning up the raw data received from the html files, the text data is processed automatically by using tools for tokenizing, segmenting, the tokens are manually light reviewed. Extracted raw sentences are aligned automatically, and English words are linked to each other in Kazakh words. We use standard techniques for the establishment of links between source and target language segments. Paragraphs and sentences are aligned by using the length-based approach developed by Gale and Church (1993). Once the sentences are aligned in the source and target language, we send it for manual correction to students who speak both languages.

The results show that between 70% and 87% of the sentences were correctly aligned by the automatic aligner depending on the text quality and dictionary. Also this is a project for building corpus for language pairs dissimilar in language structure.  We partly build the corpus by using a Hunalign tool for automatic alignment. Therefore, efforts are put on developing a general method and using tools that can be applied to similar resources.

## 3. ALIGNMENT AND DATA SELECTION

Standard In order to use for various purposes our parallel text, at first necessary to align the two texts at paragraph or sentence level. By align we mean the association of chunks of text in the one document with their translation or equivalent text in the other document. Part of approaches to alignment use of some type of traditional analysis of the texts (e.g. parsing etc.), while others take an entirely automatic approach. For our project, i.e. extraction of parallel sentences depends on precisely the kind of information we are trying to extract. There is a lot of paper on this subject (Antonova et al., 2011) and the approaches that follow are not intended as an entire review, we did more detailed review that we used.

Gale & Church and Brown et al. both developed alignment programs based on the simple assumption that there is a significant correlation in the relative length of texts, which are translations of each other.

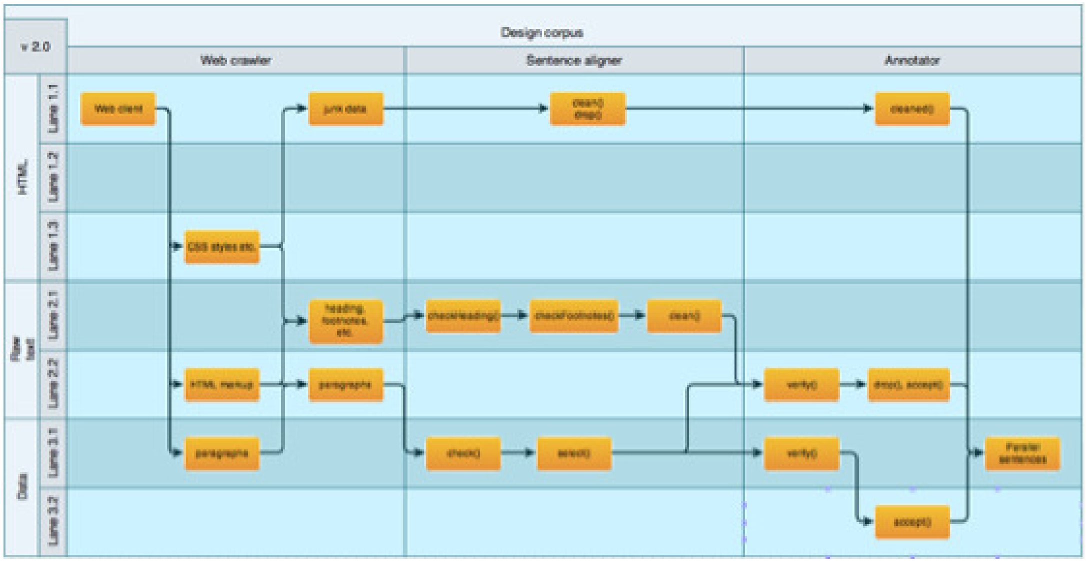

Figure 1. Corpus building procedure

The former measured length in characters, the latter in words. Simard et al. suggested some improvements. So, Gale & Church took the output of their alignment program and used it to identify correspondences at the word level.

Much of the early success of all these approaches was no doubt due to the fact that they used Canadian Hansard corpus was very good in that most of the sentences and paragraphs lined up nicely, and also syntactically and lexically French and English are quite similar.

If we try to illustrate it in another way, consider two texts with different segments length. At first sight, the most intuitive alignment<A> pairs up the segments of nearly equal length, as indicated by the arrows. But a more likely alignment, which accounts for more of the text addressed the problem of «noisy» texts<N> by trying to align on the basis of similar short stretches of characters. So we looked at texts<T> that had been extracted from document copy, and so contained misalignment problems<P> caused by different pagination, e.g. a footnote suddenly appearing in the middle of a paragraph, or figure headings out of sequence. Most of the approaches have in common a technique, which involves identification of anchor<a> points and verification of the comparability of the textual material between the anchors. These anchors can, in the simplest case, be structural, as in early work by Gale & Church, where sentence boundaries<b> are taken to make an initial segmentation. So we can say, the alignment depends from following situation: □□□̣□̱□□□□□□□□̤̃□□□□□□̤.

That we first identify potential anchor points throughout the text, and then pick those that are closest to the ideal alignment, which is a horizontal line. These then define sub regions in which the process can be iterated. “Smoothing” techniques can be used to reduce the search space even further.

Apart from automatic estimation of translation pairs, a number of sentence alignment algorithms rely on dictionaries as a method for finding lexical anchor points. This technique of course relies on the availability of a suitable dictionary, not to mention the need for efficient lemmatization in the case of highly inflected language as Kazakh.

Improving of the alignment timely depends on extracted vocabulary; so aligned parallel corpora can be used for the extraction not of everyday vocabulary, just specific lexical pairings, notably:

- Novel terminology.
- Proper names.
- Abbreviations etc.

Method was used for dictionary extraction is a hybrid of sentence and word-alignment. The approach is to find word pairs, which are most probably align-able on the basis of similar distribution. This distribution is defined in terms of text sectors, and Dice's coefficient is used to quantify the probability. Dice's coefficient is a simple calculation which compares c, the number of times the two candidate words occur in the same sector with a and b, the number of times the source or target words occur independently.

The algorithm is iterative in that the sentences containing high-scoring word pairs are established as anchors, which allow the text to be split up into smaller segments, affording more and more results.

To select the data we used a method based on Bayesian classifier, which we have applied as follows. Consider, we use some bag of words W= $\{w_1, w_2, ..., w_m\}$, each of word has a certain set of characteristics from the set of features F = $\{f_1, f_2, ..., f_q\}$ and has one tag from a set of tags T = $\{t_1, t_2, ..., t_k\}$. Our task is to identify the most likely class of the words W, relying on the set of its features $F_s$= $\{fw_1, fw_2, ..., fw_n\}$. In other words, we need to compute a value of the random variable T, at which a posteriori maximum (1) is reached. We decompose (1) on the Bayes formula (2), consider, we are looking for an argument, maximizing the likelihood function, the denominator is a constant and not depends on this variable, may remove (3) a value of the total probability P(w), because the logarithm is monotonically increasing for any argument, then the maximum of any function will be equal to the maximum (4), we need it for do not work with numbers close to zero on programing.

$$t_{map} = \arg\max_{t\in T} P(t \mid w) = \tag{1}$$

$$= \arg\max_{t\in T} P(w|t)P(t)/P(w) = \tag{2}$$

$$= \arg\max_{t\in T} P(w|t)P(t) = \tag{3}$$

$$= \arg\max_{t\in T} ln(P(w|t)P(t)) \tag{4}$$

Naive Bayesian classifier model takes two assumptions: does not matter the order of the features; Characteristics probabilities are independent on each other within the class (5).

$$P(f_i \cap f_j|t) = P(f_i|t)P(f_j|t) \tag{5}$$

Considering the above-stated assumptions, we extend the derivation of (6-9):

$$t_{map} = \arg\max_{t\in T} ln(P(w|t)P(t)) \tag{6}$$

$$t_{map} = \arg\max_{t\in T} lnP(f_1, f_2, ... f_n|t)P(t) \tag{7}$$

$$t_{map} = \arg\max_{t\in T} lnP(t)\prod_{i=1}^{n} P(f_i|t) \tag{8}$$

$$t_{map} = \arg\max_{t\in T} lnP(t) + \sum_{i=1}^{n} lnP(f_i|t) \tag{9}$$

Using the algorithms based on these formulas, we have identified areas of unnecessary text.

## 4. Quality Estimation

We evaluate corpus quality in two ways:

- Selecting each 100-sentence pair from the corpus and manually checking the sentences is parallel or not;
- Training a statistical machine translation system on the corpus and testing its output with BLEU metric.

Most of the algorithms make certain assumptions about the nature of parallel corpora:

- Words have one sense per domain;
- There are no missing translations in both languages;
- The frequencies the sentence translations are comparable;
- The positions of words are comparable.

Word have one sense per domain is underlying the approach to natural language processing. It is often true, especially for words, which have terminological status; but homonymy is not avoidable, even in narrow domains.

Translation uniqueness is undermined further by the fact that local syntactic conditions might result in inflectional morphology in one language but not the other: in particular, the distribution of singular and plural can differ widely between otherwise closely related languages, without even considering grammatical case and gender. This can be overcome by subjecting the corpora to a process of lemmatization.

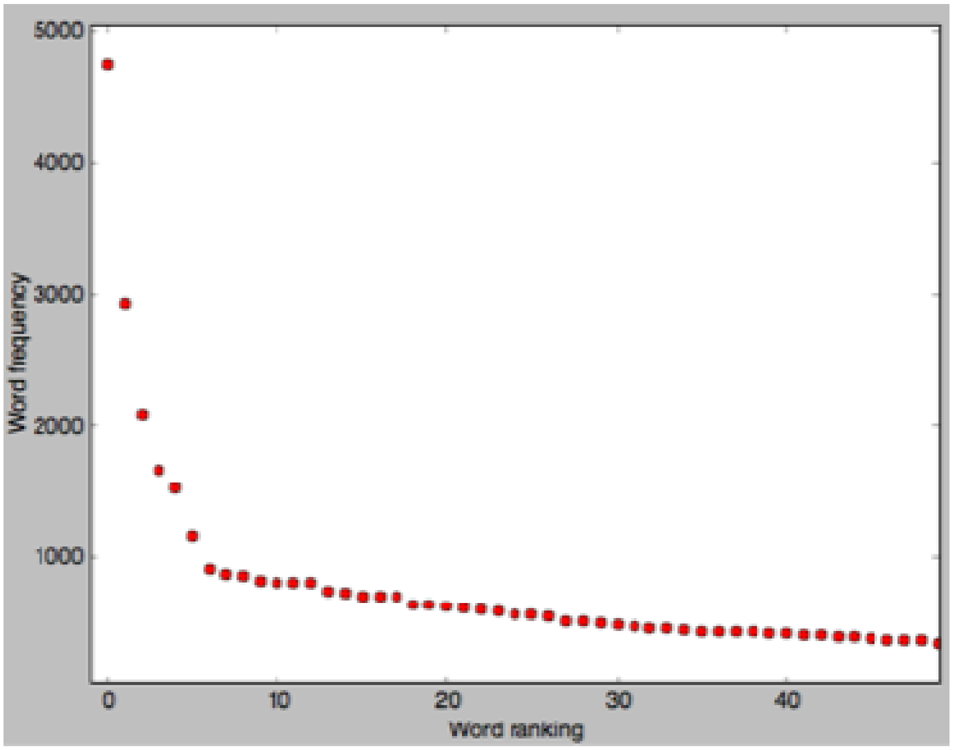

Figure 2. Word frequency diagram shows no critical differences

Another problem is that multi-word compounds in one language may correspond to what are typographically single words in another.

Word frequency difference is again the fact that a single word in one language can have a variety of translations in the other just because of grammatical inflection(See Fig.3). Word-order is a fundamental difference between many languages like English(SVO) and Kazakh(SOV).

We tested Kazakh-to-English translation systems on 4000 test sentences varying the language model order from trigram to 5-gram. BLEU measured about 10.03. The system performance can be improved by training a bigger language model, so our goal is to show the corpus is suitable for training statistical machine translation system.

## 5. CONCLUSIONS

We have described our approaches to problems for building a parallel English-Kazakh corpus from the Web. We have proposed a method of automatically alignment of texts. It allowed us to build sentence pairs that extracts from documents with more than 70% .

The approach relies on general properties of the state-of-the-art both languages and therefore is applicable to many other language pairs.

See project files in the our repository: bitbucket.org/kzmt/public.

We also would like to sincerely thank our validators and annotators. The annotation and validation work they have done helped a great deal in designing corpus. This work would not be possible without their contribution.

We presented results of evaluation of the English-Kazakh parallel corpus. We are sure that English-Kazakh corpus of parallel sentences used in this paper is a useful resource for machine translation research and machine translation contests.

## REFERENCES

[1] Altenbek G. and WANG Xiao-long. 2010. Kazakh segmentation system of inflectional affixes. *In Joint Conference on Chinese Language Processing*, CIPS- SIGHAN, p. 183–190.

[2] Antonov A., Misureyev A., Building a Web-based parallel corpus and filtering out machine-translated text. 2011. *Proceedings of the 4th Workshop on Building and Using Comparable Corpora, 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oregon, p. 136–144.

[3] Brown, P.F., Lai, J.C., Mercer, R.L. 1991. Aligning Sentences in Parallel Corpora. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California , p.169–176.

[4] Baisa and Suchomel. 2012. Large corpora for turkic languages and unsupervised morphological analysis. *In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12),* Istanbul, Turkey. European Language Resources Association (ELRA), p. 28–32.

[5] Gale, W. A., & Church, K. W. 1993. A program for aligning sentences in bilingual corpora. Computational Linguistics, 19(3), p. 75-102.

[6] Makhambetov O., Makazhanov A., Yesssenbayev ZH., Matkarimov B., Sabyrgaliev I., Sharafudinov A. 2013. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA, p. 1022–1031

[7] Mukan A. 2012. A Learner's Dictionary of Kazakh Idioms. Georgetown University Press.

[8] Philip Koehn. 2002. Europarl: A multilingual corpus for evaluation of machine translation. Technical report, Information Sciences Institute, University of Souther California.

[9] Resnik Ph., Olsen M, Diab M. 1999. The bible as a parallel corpus: Annotating the book of 2000 tongues. Computers and the Humanitie, 33(1- 2): p. 129–153.

[10] Steinberger R., Pouliquen B., Widiger A., Ignat C., Erjavec T., Tufis D., Varga D. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006).*

[11] Simard, M., G. Foster, P. Isabelle. 1992: Using cognates to align sentences in bilingual corpora. Quatrième colloque international sur les aspects théoriques et méthodologiques de la traduction automatique, *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-92),* Montréal, Canada, p. 67–82.