Compiling Japanese-Russian Aligned Parallel Corpus

Alexandr Pershin[1], Satoru Fujitani[2], Kanji Akahori[1]
[1]*Department of Human System Science, Graduate School of Decision Science and Technology, Tokyo Institute of Technology, 2-12-1, O-okayama, Meguro, Tokyo 152-8552 Japan.*
[2]*Mejiro University College, 4-31-1, Naka-ochiai, Shinjuku, Tokyo 161-8539 Japan.*
E-mail: sasha@ak.cradle.titech.ac.jp, fujitani@mejiro.ac.jp, akahori@ak.cradle.titech.ac.jp

Abstract

This paper describes the process of semi-automatic compilation of a Japanese-Russian parallel corpus for language learning. Aligning software was developed for this research. The TMX XML standard for storing aligned text has been applied. The features of aligning for structurally different languages will be discussed.

1. Introduction

In the last few years, the parallel corpus has been steadily gaining importance in language teaching. The latest researches and case studies have showed that parallel corpora and concordancers can be successfully used in the classroom. Such famous software application as "ParaConc" of Barlow [1] and "WordSmith", and many others have been created. Note that the main research to date has been concentrated in creating parallel corpora for European language pairs. The number of parallel corpora for others languages pairs is scant. We tried to fill this gap with our research.

2. Corpus Aligning Method

Most known aligning methods are oriented on the European language pairs, or languages which belong to similar language families. Japanese and Russian languages belong to different language families and are different in these respects [6]:

- Positional characteristics of the Japanese and Russian units
- Lexical-grammatical characteristics of Japanese and Russian languages
- Lexical features of Japanese and Russian languages

Russian texts can be translated in different ways into Japanese, because Japanese has three alphabets (hiragana, katakana and kanji). For example, the Russian phrase

" Ya(I) polozhil(put) karandash(pencil) na(on) stol(table)" may be translated as

" Watashi(I, not always used) wa(subject marker) enpitsu(pencil) wo(object marker) tsukue(table) no(preposition) ue(on) ni(not always translated as on) okimashita(put, not always translated as okimashita)" .

In aligning languages with different structure, Daille [3] found that Mutual Information (MI) worked better only joint frequency of the words.

MI is defined by such formula:

$$MI=\log_2 [n.A/(A+B)(A+C)] \quad (1) ,$$

where n-total number of segments in aligning texts, **A** – number of aligned regions in which the Russian word **R** and the Japanese word **J** both occur, **B** is the number of regions where **R** occurs but **J** does not occur, **C** is the number of regions where **J** occurs but **R** does not occur.

As far as we know there is no method of aligning Russian-Japanese language pair, so we have applied Haruno [4] algorithm, adjusted for structurally different language pair. We used two types of word correspondences:

- Machine translation systems (Japanese-English and English-Russian machine translation systems)
- Word correspondences statistically acquired in the text alignment process.

3. Sources for the Corpus

Tribble [5] emphasized that a language teacher should operate a few small or middle-sized corpora, which are adjusted to the current curriculum. Based on this, we took into account the opinion of our learners when we compiled our corpus. In addition, we selected materials for the intermediate or high-level proficiency Russian learners in Japan. Moreover, the materials we chose were reliable and interesting for the learner. The term 'reliable materials' [2] was defined as:

- Printed in multiple copies for distribution
- Copyrighted registered or recorded by a major indexing service.

At the moment, our corpus contains about 100,000 Russian words and their Japanese equivalents. We are planning to increase this number.

4. Compiling Japanese-Russian Parallel Corpus

As mentioned above, the materials used should be reliable, authentic and interesting for the learner with intermediate or high Russian language proficiency. We are planning to use our parallel corpus in the Russian classroom, where learners have interest in the following:

- Science
- Literature
- News

The opinion of our learners played a crucial role in the process of corpus compilation.

Genre	Science	Literature	News
%	45	20	35

Table 1.Distribution of sources in corpus.

As you can see on the Table.1, our corpus consists of three components (science, literature, news).

Corpus has been semi-automatically compiled by the software, which we created and then manually cleared from misalignments.

5. Corpus Encoding

We encoded our corpus in Translation Memory Exchange (TMX) format. The TMX- format was developed within scope of the Localization Industry Standards Association (LISA, www.lisa.org)

Below is one of the examples of TMX translation units:

```
<tu tuid="0001" srclang="*all*">
  <prop type="Domain">News</prop>
  <tuv lang="JP"><seg>小泉首相</seg></tuv>
  <tuv lang="RU"><seg>Премьер-министр Коидзуми</seg></tuv>
</tu>
```

Fig.1 Example of the TMX translation unit.

6. Conclusion

This paper presents the process of compiling Japanese-Russian parallel corpus.

Our compiling of the parallel corpus of a Japanese-Russian parallel corpus using the software we created was a success. In the future, we are planning to enhance our corpus and experiment with it in Russian language classroom in Japan.

7. References

[1] Barlow, M. (1995), A Guide to ParaConc, Athelsan, Houston.

[2] Biber, D. (1993), "Representativeness in Corpus Design", Literary and Linguistic Computing 8, pp.243-57

[3] Daille, B. (1995), "Combined Approach For Terminology Extraction:"Lexical Statistics And Linguistic Filtering", COLLING 94,pp.515-521

[4] Haruno, M., Ikehara, S., & Yamazaki, T. (1996), "High Performance Bilingual Text Alignment Using Statistical And Dictionary Information", *Proceedings of COLING-96*, University of Copenhagen, Denmark, pp.525-530.

[5] Tribble, C. (1997) Improvising corpora for ELT: quick-and-dirty ways of developing corpora for language teaching , Proceedings of Practical Applications in Language Corpora , University of Lodz, Poland

[6] Modina-Shalyapina,(1995) — Modina L.S.,Shalyapina Z.M. Principi organizacii linguisticheskih znanij v ob'ektno-orientirovannoj modeli leksiko-morfologicheskoj sistemi yaponskogo yazika. — V kn.: DIALOG '95. Trudi Mezhdunarodnogo seminara po kompyuternoj lingvistike I ee prilozheniyam.-Kazan, s. 198-205.