Abstract. The goal of the paper is to present different problems related to the building of Parallel Corpus for two small languages, namely, Latvian and Lithuanian. The Lithuanian-Latvian-Lithuania Parallel Corpus (LILA) will contain 8 million running words; will be bidirectional, aligned on the sentence level. The problems include identifying, acquiring, preparing, and aligning parallel texts.

Keywords. Lithuanian, Latvian, parallel corpus

Introduction

Nowadays most of the parallel corpora are build for designing and training MT systems, however this is not the case with the Lithuanian-Latvian-Lithuanian Parallel corpus (LILA). The building of the corpus is one of the goals in the project “Development of Research Infrastructure for Education in the Humanities in Eastern Latvia and Lithuania” and its’ initial purpose is the compilation of a unique linguistic resource for scientists, language students, translation students, teachers, translators, and lexicographers. This is the first parallel corpus of such kind between small Baltic languages. Typically Lithuanian and Latvian texts are included in parallel corpora with other more widely used languages (e.g. INTERCORP2, The Multilingual Translation Memory of the Acquis Communautaire3).

The goal of the paper is to present different problems related to the building of parallel corpus for two small languages, namely, Latvian and Lithuanian. LILA corpus will contain approx. 8 million running words; it will be bidirectional, aligned on the sentence level. The parallel corpus will be freely available on the web4. Users of the...
corpus will be able to select a language of search (Lithuanian or Latvian) and to obtain bilingual concordance lines.

The building of parallel corpus is not an easy task even for dominant and well-resourced languages, but the task of building a parallel corpus for two small languages is especially challenging. The problems include identifying, acquiring, preparing, and aligning parallel texts for the two languages [1]:

- corpus representativeness and balance criteria that work for the dominant languages might be hard to fulfill in minor languages,
- limited availability of mutual translations,
- the lack of translated texts in specific genres,
- software is usually developed for dominant languages, therefore in the process of text preparation quite significant amount of work have to be done manually.

1. Selection Criteria

Typically, compilation of the parallel corpus begins with carefully chosen selection criteria. Initially, we have set up the following text selection criteria:

- time – texts published from 1991,
- originals – language source should be originally written in Latvian or Lithuanian with the respective target translation,
- media – both printed and electronic texts (texts on the Web) will be included in the corpus.

It must be said that there exist many more text selection criteria (e.g. popularity, geography, variation of authors, size, etc.) for a corpus, as often there are many more texts available than could possibly be included in a corpus. However, this is not the case with the parallel corpus of two small languages. Already before starting the collection of parallel texts, we have already known that we will be challenged by the lack of parallel texts. Thus our decision was to limit the criteria to the three.

2. Acquiring of Parallel Texts

Based on these criteria, we have started identifying and collecting necessary texts from different media (printed and electronic texts) and different genres (fiction, legal, etc.).

2.1. Printed Sources

It turned out that for the two languages there are not many parallel resources. For instance, based on (Rodiklis [2]), ~ 20 fiction sources (~ 670 000 running words) for LV-LT direction and ~ 32 fiction sources (~ 1 700 000 running words) for LT-LV direction. The amount of legal, news, and scientific texts is not very large either. Besides, legal texts usually are not freely available, except of EU legislation documents.
2.2. Web Sources

For the web sources it has been observed that:

- Most of news and reviews are not translated directly, as typically news is localized for end-users;
- When taking Internet texts, it is not always clear, which text is original.
- Often texts from the web are not stable, as they can be changed or deleted.

3. Preparation of Parallel Texts

As the parallel aligner (see Chapter 5) used for the project works only with plain text files (character encoding Windows 1257), all collected texts have had to be converted accordingly. For each text type (books, periodicals, web-pages) sometimes different preparation has been needed: although some books could be received in electronic form, but others had to be scanned or converted from PDF or MS Word formats into plain text format. In all cases the correction of errors and checking for correct character encoding is a time-consuming task. Downloaded texts from internet and periodicals should be checked for coding and format as well.

4. Indirect Translations of Acquis Communautaire

The lack of direct translations between the two languages has made us to look for different solutions and to review some of the initial selection criteria. The possible solutions are: indirect translations (with an intermediary language, e.g. Latvian ->English->Lithuanian) or parallel translation (when translation of a given text is produced into both the languages, e.g. Latvian <- English->Lithuanian). While the former is marginal, the latter is quite common. Thus, for the parallel corpus we have considered the Multilingual Translation Memory of the Acquis Communautaire developed by the EU. This database offers about one million aligned sentences for both Baltic languages. Original data in TMX format were converted by a Haskel function into the XML format of the aligner.

5. Aligning

The aligning of parallel texts is done by Aligner 2.0.6.7 [3], a semi-automatic language-independent sentence aligning tool. The tool makes use of so called Gale and Church algorithm [4], which calculates links between parallel segments based on the number of characters within a given segment. Table 1 presents the main features of the tool.
Table 1. Features of the aligning tool Aligner 2.0.6.7.

<table>
<thead>
<tr>
<th>Features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>Plain text</td>
</tr>
<tr>
<td>Output</td>
<td>XML</td>
</tr>
<tr>
<td>Formatting</td>
<td>No formatting is preserved</td>
</tr>
<tr>
<td>Anchors</td>
<td>Paragraph and sentence boundaries</td>
</tr>
<tr>
<td>User interface</td>
<td>Plain, tag-based</td>
</tr>
</tbody>
</table>

As neither paragraph alignment, nor sentence alignment is fully automatic, some subjective decisions by human aligners must occasionally be taken. During the alignment process the manual corrections are needed, and it is done by removing erroneous tags or adding missing tags, and then realigning the text. Typical errors include: untranslated sentences or paragraphs, different division into sentences by a translator, creative translation decisions, different order of translated sentences, alignment of direct speech.

Conclusions

- Translations in both directions (LV-LT and LT-LV) are limited in respect to genres and amounts.
- Fiction after 1991 can cover only limited part of the parallel corpus.
- Available data from the Web have to be included in the parallel corpus, because of lack of the other data. But data from the Web are very important because of text types (and respectively the vocabulary) they represent.
- Parallel data from other sources (for instance, translation agencies) would be valuable.
- Asymmetry: more translations from LT into LV to the opposite direction.
- As to build the parallel corpus of two small languages is a very challenging endeavor, some pragmatic solutions have to be applied in order to build it. In general, initial criteria had to be reviewed to meet practical needs:
  - Including texts of earlier periods;
  - Using Internet texts, when it is not always clear, which text is original;
  - Using EU documents, with other original languages.

Acknowledgements

The compilation of the LILA corpus has been financed by the Latvian-Lithuanian Cross Border Cooperation program (2007-2013) within the project “Development of Research Infrastructure for Education in the Humanities in Eastern Latvia and Lithuania” (HipiLatLit).
References


