Legislation Corpus of the Republic of Latvia

1. BASIC INFORMATION

1.1 Corpus composition

Latvian-English legislation corpus of Republic of Latvia is composed from public legal documents of the Republic of Latvia available in Latvian to English. It contains the Laws of the Republic of Latvia and Cabinet Regulations in the period of 2000-2010. It contains text from the total of 1275 documents: 270 Laws, 2 Cabinet Instructions and 1003 Cabinet Regulations.

1.2 Representation of the corpora (flat files, database, markup)

The corpus is represented in a single TMX[1] standard file and the documents data has been aligned at a sentence level.

1.3 Character encoding – UTF-8

2. ADMINISTRATIVE INFORMATION

2.1. Contact person (name, e-mail)

For further information, please, contact Roberts Rozis (Roberts.rozis@tilde.lv)

2.2. Copyright statement and information on IPR

MSC_BYNCND

3. TECHNICAL INFORMATION

3.1 Data structure of an entry

The corpus is provided in a singlefile in TMX format, metadata data information is encoded is document header.

3.2 Corpora size (num. of tokens)

The corpus contains about 1 660 000 tokens: 1 million tokens in English, 660 000 tokens in Latvian

3. CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

Parallel

4.2.The natural language(s) of the corpus

Latvian, English

4.3 Domain(s)/register(s) of the corpus

Legal. Legislation.

4.4 Annotations in the corpus (if an annotated corpus)

4.5 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

The corpus is aligned in TMX format, sentence level mark-up.

4.6 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),

4.7 Alignment information (if the corpus contain aligned documents: level of alignment, how it was achieved)

The corpus is aligned at sentence level, alignment has been performed by using Hunalign alignment tools

4.8. Intended application of the corpus

NLP application: training of MT systems.

Human use: Analysis of legal language.

4.9. Reliability of the annotations (automatically/manually assigned)

Alignment is done automatically.

---

[1] http://en.wikipedia.org/wiki/Translation_Memory_eXchange

http://web.archive.org/web/20110102010600/http://www.lisa.org/Translation-Memory-e.34.0.html

4. RELEVANT REFERENCES AND OTHER INFORMATION

All the source documents among many other documents available from the Web for browsing: http://www.likumi.lv/