Corpus of Latvian literature

1 BASIC INFORMATION

1.1 Corpus composition

The Corpus of Latvian literature contains literary works of Latvian authors which are not protected by copyright low. It contains works of 21 authors – poems, stories, novels and other literary works, 69 in total which correspond to 15 000 printed pages .

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

The corpus is stored in XML files.

1.3 Character encoding – UTF-8

1. ADMINISTRATIVE INFORMATION

1.1. Contact person (name, e-mail)

For further information, please, contact Roberts Rozis (Roberts.rozis@tilde.lv) or Anita Vasiļjeva (anita.vasiljeva@tilde.lv)

1.2. Copyright statement and information on IPR

The corpus is freely available for browsing from http://www.letonika.lv/literatura/

2. TECHNICAL INFORMATION

2.1. Data structure of an entry

Corpus is available in proprietary format.

2.2. Corpora size

15 000 printed pages

3. CONTENT INFORMATION

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

Monolingual

3.2. The natural language(s) of the corpus

Latvian

3.3. Domain(s)/register(s) of the corpus

Fiction.

3.4. 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)

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

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

3.7. Intended application of the corpus

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

Content of corpus is manually checked.

4. RELEVANT REFERENCES AND OTHER INFORMATION

Available from the Web for browsing: http://www.letonika.lv/literatura/.