Estonian-english parallel corpus, estengparcorp

1. BASIC INFORMATION

1.1 Corpus composition

The corpus contains:

a) Estonian laws and their translations into English;

b) EU legislation in English and their translations into Estonian.

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

The corpus is represented as flat files.

1.3 Character encoding

The characters are UTF8 encoded.

2. ADMINISTRATIVE INFORMATION

2.2 Contact person

Name: Kadri Muischnek,

e-mail: kadri.muischnek@ut.ee

2.2 Copyright statement and information on IPR

The resource is free for research purposes, local license.

3. TECHNICAL INFORMATION

3.1 Data structure of an entry

This is not relevant as the corpus is provided as a text file

3.2 Corpora size (nmb. of tokens)

The corpus contains about 153,500 parallel units (sentences or list items); 1.7 million tokens in Estonian, 2.9 million tokens in English.

4 CONTENT INFORMATION

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

This corpus is a semtence-aligned parallel corpus.

4.2 The natural language(s) of the corpus

The natural languages of the corpus are Estonian and English.

4. 3 Domain(s)/register(s) of the corpus

The corpus represents the legislative and bureaucratic language variety (eurospeak).

4.4 Annotations in the corpus (if an annotated corpus)

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

The corpus is annotated at sentence level.

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

The tags <eesti> and </eesti> delimit the Estonian part; <inglise> and </inglise> delimit the English part.

The subscripts and superscripts are tagged with <hi rend="sub"> and <hi rend="sup">. It often happens that the original or the translated unit contains one of them, but the corresponding parallel unit does not.

4.4.3Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)

The texts have been sentence-aligned. The items of lists are treated as equal to sentences. The Estonian and English sentences may be in 1-1, 1-2 or 2-1 alignments. There are no other alignments (like 1-0, 0-1, 2-2 etc) in this corpus.

The aligning was done using the Vanilla aligner (nl.ijs.si/telri/Vanilla). It is a language independent aligner, based on the algorithm from: Gale, W. A. and Church, K. W. (1993) Program for aligning sentences in bilingual corpora. Computational Linguistics 19, 75-102.

4.5 Intended application of the corpus

The corpus can be used for training a machine translation system, bilingual term extraction, bilingual multi-word unit extraction etc

4.6 Reliability of the annotations (automatically/manually assigned) – if any

Annotation and aligning have been done automatically and can contain mistakes.

3.6 *Database of estonian multi-word expressions, estmwe*

1. BASIC INFORMATION

1.1 *Lexicon type*:

lexicon of multi-word units

1.2 Representation of the lexicon:

flat file

1.3 Character encoding
The characters are UTF8 encoded.

2. ADMINISTRATIVE INFORMATION

2.1 *Contact person:*
Name: Kadri Muischnek;
e-mail: Kadri.Muischnek@ut.ee

2.2 Copyright statement and information on IPR
The resource is free for research purposes, local license

3. TECHNICAL INFORMATION

3.1 Data structure of an entry

One entry per line; two fields: 1) the multi-word unit itself and 2) its morphological type, i.e consisting of a case form of a noun and a verb or of an particle and a verb

3.2 Lexicon size

12505 entries

4. CONTENT INFORMATION

1.1 The natural language(s) of the lexicon

Estonian

1.2 Entry Type

All entries have the same entry type

4.3 *Coverage of the lexicon*

unknown

4.4 Intended application of the lexicon

It can be used for linguistic research, a gold standard for multi-word unit extraction task, can be used in lexicon-based tagging of multi-word items in text etc

4.5 Reliability (automatically/manually constructed)

Data has been automatically extracted from 6 dictionaries and from a text corpus

5 RELEVANT REFERENCES AND OTHER INFORMATION

Kaalep, H.-J.; Muischnek, K. Multi-Word Verbs of Estonian: a Database and a Corpus. In: Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions: Marrakech; Morocco; 1. juuni 2008. , 2008, 23 - 26