TRIS corpus

1. BASIC INFORMATION
   1.1. Corpus composition
       Collection of domain specific national technical regulations of the EU member states
   1.2. Representation of the corpora (flat files, database, markup)
       Right now we have paired-MS Word files and pdf scanned documents. For the 30th of
       November we will upload just around 25 files in tmx format (translation memory exchange).
   1.3. Character encoding
       UTF-8

2. ADMINISTRATIVE INFORMATION
   2.1. Contact person (name, e-mail)
       Carla Parra Escartín, carla.parra@uib.no
   2.2. Copyright statement and information on IPR
       The database will be public and the files will have a special license to avoid commercial usage.

3. TECHNICAL INFORMATION
   3.1. Data structure of an entry
       TMX standard tags.
   3.2. Corpora size (num. of tokens)
       n/a

4. CONTENT INFORMATION
   4.1. Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)
       Bilingual, parallel, aligned at sentence level, raw.
   4.2. The natural language(s) of the corpus
       German (Germany), German (Austria), Spanish (Spain)
   4.3. Domain(s)/register(s) of the corpus
       B00: CONSTRUCTION.
   4.3.1. Annotations in the corpus (if an annotated corpus)
       1.3.1.1. Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up,
              syntactic mark-up, semantic mark-up, discourse mark-up)
              none
   4.3.2. Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed)
              none
   4.3.3. Alignment information (if the corpus contain aligned documents: level of alignment,
              how it was achieved)
       Sentence alignment using the SDL Trados tool “Winalign”. Resulting files are
       converted to tmx format automatically with a script.
   4.4. Intended application of the corpus
       Research in word alignment and automatic dictionary extraction. It can be used to other
       purposes though.

2. RELEVANT REFERENCES AND OTHER INFORMATION
   For the moment we do not have any. In the future I hope we will have a paper describing the
   project and additional information.