OLANTO

myCAT

Specifications for Corpus Documents

May 2012

myCAT is a Computer-Assisted Translation (CAT) tool, and more specifically a concordancer, *i.e.* a full-text bitext-based search engine with an automated alignment feature.

The corpus to be used by myCAT should have the following structure:

- The documents should be stored in a root folder called “docs”
- They can be organized in sub-folders such as “Budget”, “Studies”, etc. and sub-sub-folders such as “2011”, “2012”, etc. The maximum depth of folders should be 3 or 4 (beyond that it becomes difficult to use the collection box in the GUI)
- Supported file formats are MS-Word (doc and docx files), Excel (xls but not xlsx at this stage), PowerPoint (ppt but not pptx at this stage), PDF (indexable, not graphic format), HTML and TXT

The documents should meet the following naming convention:

- File names should have only letters and numbers, separated only by the underscore character
- All language versions of a same documents must have the same name, except for the language suffix
- The language suffix should be compliant with the ISO 639-1 standard:
  - EN = English
  - FR = French
  - ES = Spanish
  - AR = Arabic
  - RU = Russian
  - ZH = Chinese
  - etc...

Example:

- Budget_2011_Annex1_EN.doc
- Budget_2011_Annex1_FR.doc
- Budget_2011_Annex1_ES.doc