**GLOTTAL (GLOssary Tool To Alleviate Lethargy)**

**Functional Specification.**

1. **Purpose**

To generate glossary for GLPC

1. **Supported input formats.**

Minimal set of supported input formats include B4X and TT files.

The B4X is XML based content file format that holds cards information for card oriented client software e.g. Byki, CW1 unit player etc.

The TT is XML based content file format that holds title text information corresponding to the video lesson content. The TT files always come in pairs with suffixes \_known\_tt.xml and \_learning\_tt.xml.

1. **Glossary.**

Glossary must contain information necessary to build glossary presentation according to the glossary specification created by the content development department (<http://confluence.transparent.com:8080/download/attachments/19007890/Glossary+Model+for+GLPC+080409.docx>).

Each glossary entry is based on the first letter of the content that goes in to this entry. Each entry lists words and phrases separately. Each entry may include arbitrary number of annotations. Depending on the sort order o the glossary the key will be in L1 or L2 language and translation in L2 or L1 correspondingly. Annotations are usually in L1.

It is a responsibility of software that uses glossary XML to present glossary in the form required for that software. Glossary XML must provide enough information to do so but it is up do the particular user of the glossary XML what and how to render in L1 and L2 based glossaries.

1. **Functionality**

The tool must be able to process all of the files of the supported types under given root folder and produce two glossary XML output files, one sorted by L1 language and another by L2. The name and location of the output files must be customizable.

The tool must do the following processing of the collected content before it generates the glossary:

* For all of the sort fields for a given language remove stop words if the list of stop words is defined for this language;
* Remove all of the duplicates. The two content elements are considered duplicates if both L1 and L2 fields of these elements exactly matching each other;
* Merge content of every two related TT files together to form complete content element (one that contains both L1 and L2 parts);
* Sort all collected content by L1 and L2 sort fields;

For B4X content the following fields must be included: L1, L2, card level comment and hint. Other fields can be added to this list in the future if needed.

For TT files the only content that can be included is the phrase.

The tool must be flexible enough to accommodate wide range of input formats as long as they are XML based.