Progress on TBX-Min

Overview

TBX-Min is a minimal TBX dialect designed to contain roughly the same information as a UTX file. For more information on UTX format, see the UTX homepage here: http://www.aamt.info/english/utx/. This document will introduce you to the available tools that support TBX-Min. These tools, and indeed the TBX-Min dialect itself, are a work in progress, and feedback will be much appreciated.

The intro package you have received contains the following files and directories:

- `samples`- UTX and TBX-Min files. The UTX is from the UTX homepage, though edited to make the file size manageable for this demo. `basic-min.tbx` is similar to the sample originally provided to us, but has been changed as TBX-Min has changed. `basic-min-bad.tbx` is an invalid file for testing the schemas.
- `output`- if this demo does not work for you, then you may inspect the conversion output provided in this folder.
- `schemas`- contains the TBX-Min RNG and XSD schemas

The Standard

Here is a small sample of TBX-Min (it might look familiar to you):

```
<TBX dialect="TBX-Min">
    <header>
        <id>D001</id>
        <description>A TBX-Min example containing two concept entries with multiple terms in two languages (English and Spanish)</description>
        <languages source="en" target="de"/>
        <creator>BYU TRG</creator>
        <directionality>bidirectional</directionality>
        <license>CC BY license can be freely copied and modified</license>
    </header>
    <body>
        <termEntry id="C001"><!--terminological entry-->
```

```
<subjectField>Currency</subjectField>
<langSet xml:lang="en">
    <tig><!--terminological information group-->
        <term>money</term>
        <partOfSpeech>noun</partOfSpeech>
        <termStatus>preferred</termStatus>
        <noteGrp>
            <note>
                <noteValue>may refer to physical or banked currency</noteValue>
            </note>
        </noteGrp>
        <customer>TRG</customer>
    </tig>
</langSet>
<langSet xml:lang="de">
    <tig><!--terminological information group-->
        <term>Geld</term>
        <partOfSpeech>noun</partOfSpeech>
        <termStatus>preferred</termStatus>
        <noteGrp>
            <note>
                <noteValue>may refer to physical or banked currency</noteValue>
            </note>
        </noteGrp>
        <customer>TRG</customer>
```

```
                </tig>
            </langSet>
        </termEntry>
        <termEntry id="C002">
            <subjectField>Music</subjectField>
            <langSet xml:lang="en">
                <tig>
                    <term>music</term>
                    <partOfSpeech>noun</partOfSpeech>
                    <termStatus>preferred</termStatus>
                    <noteGrp>
                        <note>
                            <noteValue>may mean music in
general or may mean an instance of music ("Let's play music")</noteValue>
                        </note>
                    </noteGrp>
                    <customer>TRG</customer>
                </tig>
            </langSet>
            <langSet xml:lang="de">
                <tig><!--terminological information group-->
                    <term>Musik</term>
                    <partOfSpeech>noun</partOfSpeech>
                    <termStatus>preferred</termStatus>
                    <customer>TRG</customer>
                </tig>
            </langSet>
```

```
        </termEntry>
    </body>
</TBX>
```

The included features were chosen so that UTX could be converted with little loss, and the structure is DCT-style TBX. Most of the information in the document is element text, and a few fields (language, id and dialect) are XML attributes.

The header contains the following:

- a unique ID to identify the document
- the creator name
- prose description of the term base
- the directionality (monodirectional or bidirectional)
- the source and target languages (only two languages may be used)
- the termbase license
- the date that the termbase was created (in ISO 8601 format)

The body contains the usual nested termEntry, langSet and termEntry.

In UTX, there is no specific location for subject field, and a document is assumed to only contain terms of one subject field. This was not acceptable for a TBX dialect, and as a result conversion of TBX-Min to UTX may require the creation of multiple files if many subject fields are used.

The tig element contains the following:

- the term text (the only required field)
- a noteGrp (which contains notes)
    - Each note contains a noteValue and a noteKey(optional). The noteValue stores the note, and the noteKey stores the type of note)
- the part of speech
- the name of the applicable customer
- the term status

The term status and part of speech are only allowed to have certain values. These values were taken from the TBX-Basic standard rather than from the UTX standard.

Converting between TBX-Min and UTX

The converter has been placed on the gevterm server and may be accessed here: http://tbxconvert.gevterm.net/tbx-min/.

Validation

Using the provided RNG and XSD files in the schemas directory, you can validate your TBX-Min files. We have set up a web page for you to upload your TBX-Min files for validation using one of these schemas. Try validating `samples/basic-min.tbx` using the tool at this URI: http://tbxconvert.gevterm.net/tbx-min/validate/validate.php.

One invalid sample file, `samples/basic-min-bad.tbx`, has been provided so that you can see what a validation error looks like. The errors emitted by the RNG validator are rather cryptic, but those given by the XSD validator are pretty useful. The error we inserted into the file is an incorrect `termStatus` value: `deprecated` should be `obsolete`.

Code Status

For TBX-Min support, we currently have the following (see http://tbxconvert.gevterm.net/tbx-min):

- TBX::Min, a Perl module for reading, writing, and editing TBX-Min files.
- TBX-Min Viewer (http://viewer.tbxinfo.net/min/)
- A bidirectional converter for UTX and TBX-Min files.
- Converter to and from TBX-Basic
- A converter from a spreadsheet Glossary (.xlsx, .xls, or .csv) into TBX-Min
- RNG and XSD schemas.

XLIFF:doc conversion support was put on hold because of some of its limitations. We will continue to support and polish these tools, and will also update them with any desired changes to the TBX-Min dialect.

The code is hosted on GitHub. The converter is here (https://github.com/byutrg/p5-Convert-TBX-UTX) and TBX::Min and the schemas are here (https://github.com/byutrg/p5-TBX-Min).

James Hayes has been working hard and learning a lot about Perl, XML, TBX and more. He will continue to work with the Translation Research Group and contribute to the development of TBX-Min tools.