UTX-Simple Specification Version 1.00

0. Document information

●Owner: AAMT Sharing/Standardization Working Group
http://www.aamt.info/english/utx/
●Status: preliminary
●Last updated: November 10, 2009

1. Introduction

The goal of UTX-Simple (UTX-S) is to create from a user's viewpoint a simple, easy to make, easy to use dictionary that can be used by machine translation systems. The same UTX-S dictionary can be used by different manufacturers' translation software. In addition, a UTX-S dictionary is human-readable, and can be used as a glossary that does not involve translation software at all.

When a user of translation software makes the effort to prepare user dictionaries, they are fragmented and dispersed, and thus not effective. Also, even a simple plain text file is difficult to share or to reuse, unless its format is standardized. However, if a single UTX-S standard is adopted, shared dictionaries can be used widely across various tools, such as translation software from different manufacturers, and are also highly reusable.

1.1. Target users

UTX-S is specifically designed to be used by end users of translation software, or translators. UTX-S does not require any advanced technical knowledge of linguistics, grammar, or machine translation software etc. to create or use. UTX-S can be made from a minimum of data for both source and target language.

1.2. Target domains

UTX-S can be used in any domain, but should be specific to some subject or topic, such as ICT, medicine, law, or engineering. Ideally, a domain should be highly specific, such as “Ruby (scripting language),” “cardiovascular surgery,” etc. UTX-S may not be suitable for translation of non-specialized, general contents.

2. Specification

2.1. UTX-S file

A UTX-S file should be encoded in UTF-8 (without BOM). Its new line code is “\r\n” (CR+LF). The file extension is “.utx.”

A UTX-S file consists of a header and a body:
1. A descriptive header (lines 1 & 2)
2. The actual entries (tab delimited text)

Entries can be commented out by placing “#” at line start.

2.2. UTX-S header

A UTX-S header consists of two lines that begin with “#.”

2.2.1. First line of a UTX-S header

The first line consists of necessary information about the UTX-S file, delimited by semicolons. It is specified as follows:

#UTX-S <version>; < source language >/< target language>; <date created>; <optional fields (creator, license, etc.)>
- UTX-S version is currently 1.00.
- Source language/target language: ISO 639, 3166 formats.

In the case of monolingual dictionary, target language should be omitted.
- Date created: ISO 8601 format
- Optional fields (creator, license, etc.)

Example 1:

```
#UTX-S 1.00; en-US/ja-JP; 2009-08-10T14:28:00Z+09:00; comment: This is an example of UTX-S header.
```

Example 2:

```
#UTX-S 1.00; en-US/fr-FR; 2008-03-15T10:00:00Z+09:00; copyright: AAMT (2008); license: CC-by 3.0
```

2.2.2. Second line of a UTX-S header

The second line of UTX-S header consists of three mandatory columns, followed by optional user-defined columns, separated by tabs:

2.2.2.1. Mandatory columns

- First column: **#src**: contains words in the source language
- Second column: **tgt**: contains words in the target language
- Third column: **src:pos**: contains parts of speech for words in the source language.

UTX-S has following parts of speech:
noun / properNoun / verb / adjective / adverb / sentence

Refer to the section “2.4 Part of speech” for details.

2.2.2.2. Optional columns

The fourth column and all additional columns are optional; a user can define as much information as he/she wants to. Additional columns can be described as the following:

- in the case of defining some form of the source language word: **src:***some_information*
- in the case of defining some form of the target language word: **tgt:***some_information*

For English, the following symbols denoting *some_information* are pre-defined:
plural: plural form
3sp: third-person singular form
past: past tense form
presp: present participle form
pastp: past participle form
comparative: comparative form
superlative: superlative form

In the case of monolingual dictionary, information on the target language can be left blank.
A hyphen "-" is used to indicate that a property does not exist for the word in a particular language. For instance, as the English word "information" has no plural form, you may specifically use "-" to indicate it.

Example:

```
#src tgt src:pos src:plural src:3sp src:past src:pastp src:presp
src:comparative src:superlative
```

2.3. UTX-S body

A UTX-S body consists of one or more entries. An entry is in a tab-delimited format. The first, second, and third columns are mandatory fields. A UTX-S body can also contain one or more comments, followed by an initial "#."

2.3.1. First column (mandatory)

A word in the source language.

2.3.2. Second column (mandatory)

A word in the target language.

2.3.3. Third column (mandatory)

The part of speech of a source language word.

2.3.4. Fourth and the following columns (optional)

User defined attributes.

2.3.5. Comments

A comment line begins with "#."

2.4. Part of speech

Only the following parts-of-speech should be used:

```
noun
properNoun
verb
adjective
adverb
sentence
```

If the part of speech is unknown then leave it blank: " ".
`sentence` should only be used when necessary; entries of pairs of translated sentences should be stored in a translation memory.

3. UTX-Simple guidelines

3.1. General guidelines

In general, a UTX dictionary should only contain technical terms of a specific domain. In most cases, entries are nouns, especially compound nouns. Translation accuracy can be improved by collecting, sharing, and reusing the data of fine-tuned bilingual translations which are not included in translation software out-of-box. Sentences should not usually be included in a UTX dictionary, except when it is appropriate to treat them as a kind of "words." As a rule, UTX should be separated from translation memory, which is a bilingual database of sentences, rather than words.

For example, a word like "XML declaration" can be correctly translated into its Japanese equivalent, "XML 宣言" by just registering it in a user dictionary. Basic vocabulary like "window" should not be included, because such a word is already contained in the system dictionaries of translation software.

- Add only one translation for each entry.
- Avoid words in the system dictionary.
- Define the domain of the dictionary clearly.
- The basic form of word should be entered (singular form for a noun, root form for a verb - as you would see in a commercial dictionary).
- Any comments should be noted separately in the comment field, not as a part of the entry.
- Choose only the single, most appropriate translation corresponding to a source word. If it has multiple DISTINCTLY different meanings, they can be treated as separate entries.
- Do not add words that are dependent on a specific MT.
- Alphabets and numbers should be written in single-byte characters, not multi-byte characters.
- Do not use ellipsis (…) to indicate a variable within an entry.
- Do not add any comments directly in a mandatory field; add a comment by either adding a comment column in the dictionary table or by adding a comment line that begins with “#.”

3.2. Guidelines particular to the English language

- Always begin an entry with a small letter (except proper nouns).
- Do not contain articles such as “a”, “an” and “the,” except in the special case that it is a part of a proper noun.

3.3. Guidelines particular to the Japanese language

- 半角カタカナなど機種依存文字は使用しない。
- サ変動詞は「する」で終わる。例：強調する
- 形容動詞は「な」で終わる。例：静かな
- 音引きは省略しない。例：ユーザー、セキュリティー、コミュニティー
- 中黒(middle dot)は省略しない。もしくは半角スペースで代用する。

すでに定着しているものを除き、カタカナ語、和製英語は避ける。