UTX-Simple Specification Version 1.10

Table of Contents

0. Document information

Author/owner: AAMT Sharing/Standardization Working Group

http://www.aamt.info/english/utx/

Status: official release

Last updated: March 10, 2011

Document language: English

All Rights Reserved, Copyright (C) AAMT, 1996-2011

1. Introduction

1.1. Background

When using rule-based translation software in a CAT (computer-aided translation) workflow, specialized terminology, names of persons, and place names in the source document are often not included in basic system dictionaries, and they are not translated as well as one would expect. It is well established, however, that core terminological information (source and target terms) without additional details, if terms are well-chosen and appropriate for a specific domain, is sufficient to increase the adequacy and accuracy of machine translation. Unfortunately, user-created dictionaries (user dictionaries) are often incompatible across different MT systems, rendering the effort to create such dictionaries futile. To address this issue, AAMT (Asia-Pacific Association for Machine Translation) <www.aamt.info> has undertaken to establish a set of specifications for sharable dictionaries, which can be used across different MT systems. AAMT created its first version of specification, UPF (Universal PlatForm), with support from IPA (Information-technology Promotion Agency, an institute in Japan) in 1995. In 2006, AAMT started to create new specifications to reflect and incorporate the subsequent advancement of technology and the changing usage of MT. In 2007, the new format received a new name **"UTX,"** short for **universal terminology eXchange**. UTX is an open standard, and AAMT does not charge the use of its specification. In 2009, AAMT has established UTX-Simple, which is intended to the simple, tab-delimited version of UTX. As AAMT learned that more complex functions can be achieved by TBX and TBX-Basic, it has focused on developing UTX-Simple.

AAMT also produces and collects open user dictionary data for specialized domains. AAMT also hopes to create a user community for generating, sharing, and accumulating user dictionaries in a sustainable way.

1.2. Goal

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. UTX-S places emphasis on usability and simplicity over advanced manageability and lossless conversion.

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.

UTX-S is suitable to be used, for example, in the process of rapid compilation of glossaries from multiple resources, or distribution and reuse of glossaries across a wide range of applications. For a more complete, long-term terminological management, TBX may be suitable.

2. Intended 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, XML, or machine translation software etc. to create, edit, or use. UTX-S can be made from a minimum of data for both source and target language.

3. Target domains

UTX-S can be used in any domain of translation, 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. If a dictionary contains distinctly different domains, each domain should form a separate dictionary. By doing so, it is easier to manage dictionaries, and dictionaries can be easily reused by combining with other dictionaries.

UTX-S may not be suitable for translation of non-specialized, general contents. When used for general contents, the benefit of UTX-S is limited. The framework of UTX-S assumes that the target MT system already has a well-developed system dictionary. UTX-S can increase the adequacy of translation where the existing system dictionary cannot. Non-technical terms can only be included in a UTX dictionary when they have specific meanings and target terms within the domain.

4. Definitions

4.1. Dictionary administrator

While a dictionary may have multiple contributors, a dictionary administrator is ultimately responsible for a dictionary and defines the framework of a dictionary. The dictionary administrator would decide whether an entry is approved as an appropriate entry for the dictionary, as explained in "7.4.2 approved." If the dictionary is compiled by a single individual, the dictionary administrator is also the contributor.

4.2. UTX converter

UTX converter is a generic name for a tool that converts a UTX dictionary to/from other formats.

4.3. Translation direction

A UTX dictionary typically has a source language and a target language (monodirectional). A complete bidirectional dictionary is a dictionary where all terms are known to be usable for the reversed translation direction. In this case, a UTX converter may export the entire dictionary to a monodirectional dictionary whose translation direction is reversed. When only some entries have approved status as explained below, only these entries may be exported as a reversed dictionary to be used for the opposite translation direction.

Example

Consider a Japanese-English UTX dictionary. When some of its entries have approved status, these can be exported as an English-Japanese dictionary with an appropriate UTX converter.

4.4. Term

A term is a headword of either the source or target language. It should be the basic form of the word (lemma) such as a headword of a dictionary. See "9 UTX-Simple guidelines." In terms of UTX, a term is a technical term in a specific domain. A "word," in contrast, refers to a grammatical unit that represents a notion, which may or may not be in a UTX dictionary.

4.5. Entry

An entry in a UTX dictionary is a single logical line. A UTX dictionary should follow the "one term, one meaning" rule, that is, one term carries only one meaning in the specific domain of the dictionary. Thus, a source term ideally has a single target term, and together they form a single entry (single line). If a target term needs to have multiple target terms (thus multiple meanings), the rationale must be justified. This promotes an effort to compile a glossary in a systematic manner. Although exceptional, when multiple target terms are required, concept ID can be specified to indicate the same concept (see "7.5 Concept ID (optional)").

UTX is most effective when used in an organized authoring/translation project. It is assumed the author of the source document, such as a technical writer, is making an effort to use coherent terms following a style guide. When UTX is used in a less-organized project, its effectiveness may be limited.

5. 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 the following:

1. Dictionary information. See "6.1. Dictionary information."
2. Column definitions. See "6.2. Column definitions"
3. Body consisting of the entries (tab delimited text). See "7. Body."

1 and 2 form the header.

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

6. Header

A UTX-S header consists of two mandatory lines that begin with "#." As an option, one or more lines that have been commented out can be added between the two mandatory lines for a description of the dictionary.

6.1. Dictionary information

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>;
<creator>; <license>; <bidirectional (optional)>; <dictionary ID (optional)>;
<other optional fields>
```

- Source language/target language: ISO 639 and 3166 formats.
- Date created: ISO 8601 format

Example:

```
#UTX-S 1.10; en-US/ja-JP; 2009-08-10T14:28:00Z+09:00; comment: This is an
example of UTX-S header.
```

6.1.1. License (optional)

The license of dictionary can be declared in the form of Creative Commons, public domain, or other forms of license. It is strongly recommended to clarify how the dictionary can be shared and used.

6.1.2. Bidirectional (optional)

When all terms of a dictionary are known to be used in the opposite translation direction, the dictionary can have a "bidirectional" flag in its header. In this case, individual term status does not need to be indicated, because all terms are assumed to be approved.

Example:

```
#UTX-S 1.10; en-US/fr-FR; 2008-03-15T10:00:00Z+09:00; copyright: AAMT (2010);
license: CC-by 3.0; bidirectional
```

6.1.3. Dictionary ID (optional)

A dictionary ID is a unique identifier for a dictionary. It consists of four case-insensitive alphanumeric characters chosen by the dictionary administrator. It may be required to distinguish dictionaries when multiple dictionaries are merged. If two dictionaries happened to share entries with the same concept ID (see 7.5 Concept ID (optional)), unrelated entries could be grouped into one group of entries. Unique dictionary IDs can help avoid such a situation. A dictionary ID is not mandatory. It can be added when multiple dictionaries need to be merged.

Example:

When merged, unrelated entries may be grouped without IDs to distinguish a dictionary.

	Entries	Concept ID
Dictionary A:	outlet	76531
Dictionary B:	instantiate	76531

With dictionary IDs, entries happened to have the same concept ID can be distinguished even when the originating dictionaries are merged.

	Entries	Concept ID	Dictionary ID
Dictionary A:	outlet	76531	AD64
Dictionary B:	instantiate	76531	5d32

6.2. Column definitions

The second line, or if there are additional descriptive lines, the last line of UTX-S header (also begins with "#") includes a set of column definitions. Column definitions consist of three mandatory columns, followed by optional user-defined columns, separated by tab characters. The details are explained in "7. Body," as the column definitions and the body are closely related.

7. Body

The body of a UTX consists of entries in each logical line.

7.1. Source term (mandatory)

The first column `src` contains source terms, that is, terms in the source language.

7.2. Target term (mandatory)

The second column `tgt` contains target terms, that is, terms in the target language. The target language column is mandatory, however, in the case of monolingual dictionary, it can be left blank.

7.3. Part of speech (mandatory)

The third column `src:pos` contains parts of speech for the source term.

UTX-S has following parts of speech:

```
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 rather than a dictionary.

7.4. Term status (optional)

An entry can optionally have one of four term statuses (provisional, approved, non-standard, or forbidden) to indicate the terminological state of the term. Term status (`term status`) is only managed for the primary

translation direction of a dictionary. If term status needs to be managed for the reversed translation direction, a separate dictionary should be compiled.

7.4.1. provisional

The term status "provisional" means that an entry is entered into a dictionary by a contributor but not yet checked by the dictionary administrator. As a provisional status is temporary, the dictionary administrator is expected to promptly decide if the term should be any of "approved," "non-standard," or "forbidden," as explained below. The dictionary administrator may also choose to exclude (delete) the term from the dictionary.

7.4.2. approved

The term status "approved" means that an entry has been approved by the dictionary administrator. An approved status indicates that the term MUST be used. The rationale could vary, but usually because it is a technical term within a specific domain or it belongs to a glossary of an organization. If the word form of the term has variations, such as "plug-in" and "plugin," only the approved form should be used.

When there is a clear reason, a source term can have multiple target terms (thus multiple entries), but only in one entry is its term status approved.

If its term status is "approved," a term can be "reversed" to be used for the opposite of the translation direction that is defined in a dictionary (see 6.1.2 Bidirectional (optional)).

An approved term is always bidirectional. When there are multiple translations to a source term, and a user choose to use them for a reversed translation direction, an approved term will be the only valid term.

If the dictionary has a single contributor, the contributor (who is also the dictionary administrator) may choose to assign the "approved" status immediately after adding an entry to the dictionary. Alternatively, the contributor may also choose to leave the status blank or assign "provisional," until he or she can confirm that the new entry works fine in the translation project.

7.4.3. non-standard

The term status "non-standard" indicates one or more non-standard source terms. Non-standard terms are only permitted to accommodate variations of source terms. Non-standard terms should not be used as target terms. If a UTX dictionary is used as a glossary for authoring of documents (rather than translation), non-standard terms should not be used as terms, because they are entered in the dictionary so that a MT system can translate even if the author of the source document used improper words that are not approved.

7.4.4. forbidden

The term status "forbidden" means that an entry includes a target term which SHOULD NOT be used. Such words are explicitly forbidden from linguistic, social, terminological, branding, or other viewpoints. A target term may also need to be suppressed to avoid conflict with different domain-specific dictionaries, when a translation tool does not properly honor the priorities among multiple dictionaries.

Example

In the context of ICT, an English term "window" is very unlikely to be translated as, for example, a Japanese word "窓". This translation may need to be explicitly suppressed if the MT system cannot handle it appropriately.

A forbidden term could have an approved status in a reversed entry. Also, a non-standard term could have forbidden status in a reversed entry (for example, if the concept ID is the same).

It is preferable that a translation tool is capable of suppressing the use of forbidden terms. It is possible that a term is forbidden in one dictionary, but if the same term exists in another dictionary, it may not be forbidden. In other words, when using a set of dictionaries in a translation tool, the forbidden status may differ among the dictionaries. A translation tool or a UTX converter tool should preferably have a mechanism to detect such conflict.

Forbidden terms can be extracted to be used for terminological check outside of a translation tool.

7.5. Concept ID (optional)

When there are multiple entries with different term status, use a "concept ID" to indicate that they share the same concept. If a source term has only one target, concept ID is not required.

A concept ID consists of serial, unique numbers within a dictionary. A concept ID must be a numeric value of up to ten digits. When multiple dictionaries are merged, entries with the same concept ID can be distinguished by their dictionary IDs.

Example (English to Japanese):

(Term number)	src	tgt	term status	concept ID
1	outlet	コンセント	approved	73
2	outlet	アウトレット	forbidden	73
3	power point	コンセント	non-standard	73
4	PowerPoint	PowerPoint	approved	
5	outlet store	アウトレット ストア	approved	
6	plugin	プラグイン	approved	245
7	plug-in	プラグイン	non-standard	245

(Term numbers are given only for the sake of explanation. They do not exist in a UTX dictionary. Part of speech is mandatory, but not shown.)

In the above example, term numbers 1, 2, and 3 point the same concept. So do 6 and 7. In 4 and 5, concept ID is blank because they follow "one word, one meaning" principle, therefore, no other entries exist that need to be distinguished. Note that the word "アウトレット" is forbidden as the target term of "outlet," but it is approved as a part of the word "アウトレット ストア." Also note that where multiple entries share the same concept ID, there is only one approved term. All others are either forbidden or non-standard.

Example (Japanese to English)

src	tgt	term status	concept ID
コンセント	outlet	approved	73
アウトレット	outlet	approved	98
コンセント	power point	forbidden	73
PowerPoint	PowerPoint	approved	
アウトレット ストア	outlet store	approved	

7.6. Other optional columns

The fifth column and any other additional columns are optional; a user can define as much information as he/she wants to. Additional columns for conjugation and variation are described as the following:

- in the case of defining some form of the source term: **src:***some_information*
- in the case of defining some form of the target term: **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

A hyphen "-" is used to explicitly 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 of pre-defined symbols:

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

Example of actual content:

#UTX-S 1.10; en-US/ja-JP; 2010-03-15T10:00:00Z+09:00; copyright: AAMT (2010); license: CC-by 3.0				
#src	tgt	src:pos	term status	src:plural
early adopter	アーリー アドプター	noun	approved	early adopters
fast	高速な	adjective	provisional	
optional	省略可能な	adjective	approved	
optional	オプショナルな	adjective	forbidden	
save	保存する	verb	approved	

8. Preferred MT functions

UTX-S works better if MT systems have the following functions:

- A sound system dictionary for non-technical terms
- Capability of using a set of multiple user dictionaries
- Priority of longer compound words over shorter words
- Capability of suppressing the use of forbidden entry

9. UTX-Simple guidelines

9.1. 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 term 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 basic words in the system dictionary.
- Define the specific 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 term. 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 system.
- 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 "#."

9.2. Guidelines particular to the English language

- Always begin an entry with lowercase (except proper nouns).
- Do not include articles such as "a", "an" and "the," except in the special case that it is a part of a proper noun.

9.3. 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adjectiveとして示す。
- 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。

10. Version history

UTX 1.00 (November 10, 2009):

- Initial release.

UTX 1.10 (November 22, 2010):

- Overall revision.
- Added term status, dictionary administrator, bidirectional, concept ID, dictionary ID, and preferred MT functions.