UTX 1.20 Specification

Table of Contents

Document information

Authors: AAMT Sharing/Standardization Working Group: YAMAMOTO Yuji (CosmosHouse), MURATA Toshiki (Oki Electric Industry Co., Ltd.), Francis Bond (Nanyang Technological University), OKURA Seiji (Fujitsu Laboratories Limited), Michael Konin Kato (Japanese Greats Co., Ltd.), AKIMOTO Kei (Kotobaya Inc.), TAKAHASHI Hiroyuki (Cross Language Inc.), KAMEYA Hiroshi (SunFlare Co., Ltd.)

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

Status: official release

Last updated: February 16, 2018

Document language: English

0. Introduction

0.1 Background

UTX was first established to be used as a standardized user dictionary format for rule-based translation software. More recently, UTX has been reconceived as a glossary format that can be used in more areas, including CAT (computer-aided translation) and natural language processing.

Documents to be translated (source documents) frequently include specialized terminology, names of persons, and place names. This type of information is often insufficient, missing, or inappropriate in basic system dictionaries of rule-based translation software. In this case, the machine translation system can only produce unsatisfactory results. Also, the structure and file formats of user-created dictionaries are often incompatible across different MT (machine translation) systems, making any effort to share and reuse such dictionaries difficult. It is well established, however, that core terminological information (source and target terms) is sufficient to increase the adequacy and accuracy of MT, if terms are well-chosen and appropriate for a specific domain. To address the need for a standardized glossary format, AAMT (Asia-Pacific Association for Machine Translation) <http://www.aamt.info> undertook the establishment of a set of specifications for sharable dictionaries, which can be used across different MT systems. AAMT created its first version of the specification, UPF (Universal PlatForm), with support from IPA (Information-technology Promotion Agency, an institute in Japan) in 1995.

In 2006, AAMT initiated the development of new specifications to reflect and incorporate the subsequent advances in technology and the changing usage of MT. In 2007, the new format was given the name **"UTX,"** short for **Universal Terminology eXchange**. In 2009, AAMT established UTX-Simple, which is a simple, tab-delimited version of UTX. Having learned that more complex functions can be achieved by TBX and TBX-Basic, AAMT shifted its focus to developing UTX-Simple. In April 2011, UTX-Simple changed its name to "UTX," dropping "-Simple."

0.2 Goal

The goal of UTX is to provide a set of rules for creating a simple, easy-to-make, easy-to-use glossary to a wide range of users including non-experts. UTX places emphasis on usability and simplicity over advanced manageability and lossless conversion. A UTX glossary is human-readable, and can be used as a glossary that is independent of translation software.

When an individual user of translation software makes a user dictionary, considerable time and effort is often required to make it effective. Also, even a simple plain text file is difficult to share or reuse, unless its format is standardized. However, if multiple users create user dictionaries following a single standard such as UTX, they can share their dictionaries more easily. They can use these dictionaries across various tools, such as translation software from different manufacturers.

A usage scenario of UTX includes, for example, a quick compilation of glossaries from multiple resources in daily translation work to share translation knowledge. UTX is also useful in distributing, sharing, and reusing glossaries across a wide range of applications.

UTX's simple structure is ideal for reducing the complexity of terminology management. The tabular structure of UTX allows editing in a spreadsheet application. For more complicated terminology

management that requires many languages and many term fields (attributes), TBX may be suitable.

0.3 Applications of UTX

The UTX format is designed for end users of translation software, or translators. To create, edit, or use a UTX glossary does not require any advanced technical knowledge of linguistics, grammar, XML, MT software, etc. A UTX glossary can be made from a minimum of term data.

UTX can be used in any domain of translation, but the glossary should be specific to some subject or topic, such as ICT, medicine, law, or engineering. Ideally, a domain should be highly specific, such as "Perl (scripting language) in the ICT domain," or "cardiovascular surgery in the medical domain."

> Note: If a translation project involves multiple domains, it is recommended to create an individual glossary for each domain. By changing the combination of glossaries, they can be reused and repurposed more effectively.

A UTX glossary can be directly viewed and edited with text editors and spreadsheet applications. It can also be used within terminology tools for terminology checking. A UTX glossary must be compiled with terminological consistency because its consistency determines the consistency of the documents based on the glossary.

UTX is not suitable for translations of non-specialized, general content. The UTX framework assumes that the target MT system already has a well-developed system dictionary. Non-technical words should only be included in a UTX glossary if they have specific meanings and target terms within the domain.

0.4 References to international standards

This document refers to the following international standards.

Language tags: IETF BCP 47, *Tags for Identifying Languages*

Date/time format: ISO 8601:2004, *Data elements and interchange formats – Information interchange – Representation of dates and times*

File encoding: ISO/IEC 10646:2014, *Information technology -- Universal Coded Character Set (UCS)*

Terminology format: ISO 30042:2008, *Systems to manage terminology, knowledge and content -- TermBase eXchange (TBX)*

Translation memory format: *TMX 1.4b Specification*

Requirement Levels: S. Bradner, *Key words for use in RFCs to Indicate Requirement Levels*, *http://www.ietf.org/rfc/rfc2119.txt* IETF (Internet Engineering Task Force) RFC 2119, March 1997.

0.5 Version history

UTX-Simple 1.00 (November 10, 2009)

- Initial release.

UTX-Simple 1.10 (November 22, 2010)

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

UTX 1.11 (May 25, 2011)

- Renamed UTX-Simple to UTX (without "-Simple").
- Clarification of some topics.

UTX 1.20 (August 3, 2016)

- A multilingual glossary can be created.
- Wording: "dictionary" is replaced with "glossary."
- The `dictionary ID` is changed to `glossary ID` and now allows text strings.
- A glossary may now include sub-glossaries.
- BOM must be added to a UTX file.
- Per-language term status is added.
- Term status: blank is now the same as "approved."
- Term status: rejected/obsolete term statuses are added.
- Header: `date created` is changed to `creation date`. The property now explicitly accepts the simple format: YYYY-MM-DD.
- Header: All glossary properties are now optional except UTX version.
- Glossary properties added: `last modified date`, `glossary administrator`, `domain`, `copyright`, `directionality`, `sortable`, `glossary version`.
- The `bidirectional` property is removed and superseded by the `directionality` property.
- `pos` field items added: `vt`, `vi`, `prenominal`.
- Field (column) order is no longer fixed.
- concept ID is no longer restricted to numbers.

UTX 1.20 (September 20, 2016)

- Linguistic corrections.

UTX 1.20 (May 24, 2017)

- Member changes.

UTX 1.20 (February 16, 2018)

- The license is changed to Creative Commons 4.0 BY.
- The disclaimer is changed.

1. Terms and definitions

This section explains key terms and definitions related to UTX.

1.1 Types of UTX glossaries

A **UTX glossary** is a collection of terms within a particular domain. It is designed to be both human and machine-readable.

A UTX glossary serves as a data dictionary for machine translation (MT) to improve accuracy and fluency. Some MT systems use two types of dictionary. A **user dictionary** is a dictionary created by a user, whereas a **system dictionary** is a dictionary built into an MT system.

With respect to the number of languages it contains, a UTX glossary can be monolingual, bilingual, or multilingual.

A **monolingual glossary** contains terms in only one language. It may be used for terminological checking, or standardizing variants of expressions or technological terms, for example. A monolingual glossary containing pronunciation information can be used by voice recognition and text-to-speech applications. For some languages, such information can be used for input method editors.

A **bilingual glossary** contains terms in two languages. A UTX glossary is most often bilingual. If not stated otherwise, the current specification assumes a UTX glossary is bilingual.

A **multilingual glossary** contains terms in three or more languages.

Note: UTX 1.11 and earlier recognize only bilingual and monolingual glossaries. UTX 1.20 or later allows the user to create multilingual glossaries (see "6. Multilingual glossary").

1.2 Term

A **term** is a headword of either the source or target language(s). A term in a UTX glossary should be in the basic form of the word such as a headword in a dictionary. See also "7. Appendix A: UTX content guidelines."

Note: Term definitions are optional in a UTX glossary.

1.3 Entry

An **entry** in a UTX glossary is a unit consisting of one or more terms and additional information.

An entry corresponds to a line. See "4. Field definitions and body" for details.

Note: If a UTX glossary is intended for rule-based machine translation, it should follow the **"one term, one meaning" rule** which states that one term corresponds to only one meaning (or, one concept) in a glossary in the specific domain.

1.4 Property, property value, and property item

A **property** is a characteristic of an object. Unless otherwise stated, a property in UTX refers to a glossary property.

A **property value** is the value assigned to a property. For example, a user determines the property value of the `glossary administrator` property.

A **property item** is a pre-defined pick-list item for a property value. For example, `uni`, `bi`, and `multi` are property items for the `directionality` property value.

1.5 Field, field value, and field item

A **field** is a characteristic (such as part of speech) of an entry or a term. In the tabular format, a field corresponds to a column.

A **field value** is the value set to a field. For example, a user would determine the field value of the `glossary ID` field.

A **field item** is a pre-defined pick-list item for a field value. For example, `noun` is a field item for the `pos` (part of speech) field value.

1.6 Synonyms

Synonyms are terms that share the same concept, or meaning. Each synonym forms an independent entry, the entries then being grouped under the same concept ID. Alternative spellings are treated in the same way as synonyms. Term status can be used to distinguish different statuses for synonyms and alternative spellings. For details, see "5.2 concept ID field."

1.7 User roles

There are three user roles in relation to a UTX glossary. These roles define what a user can do to a glossary.

A **glossary user** is an individual who uses the glossary. A glossary user may have knowledge of the domain of the glossary, and a minimal understanding of UTX. Glossary users cannot add, delete, or change terms, but they may express opinions and preferences about terms via comments and other means.

A **glossary contributor** is an individual who proposes the addition of new entries. A glossary contributor should have a good knowledge of the domain of the glossary, and a basic understanding of UTX. A glossary contributor subsumes the role of glossary user.

A **glossary administrator** is an individual who is responsible for a glossary. A glossary administrator should have expert knowledge of the domain of the glossary, and an excellent understanding of UTX. A glossary administrator decides whether an entry is approved as an appropriate entry for the glossary or removed, if necessary (See "4.5.2 approved"). A glossary administrator may also assign term statuses, or change one status to another. A glossary administrator subsumes the role of glossary contributor. One or more **delegates** can perform the function of glossary administrator.

1.8 Translation direction

The direction of translation from one language to another (translation direction) can be unidirectional, bidirectional, or multidirectional. This information can be specified with "3.2.14 `directionality` property."

A **unidirectional glossary** is a glossary whose translation direction is primarily one-way, i.e. from

the source language to the target language.

Example: unidirectional bilingual Japanese-English UTX glossary

Source language: Japanese, target language: English

Primary translation direction: Japanese to English

Note: Some terms in a unidirectional glossary may be exported and used in the reverse direction in an ad-hoc manner. This operation is called **reverse-exporting**. In this case, the source language becomes the target language, and vice versa. A reverse-exported unidirectional glossary may contain problems because the consequence of the reversal may not be thoroughly examined when compared with a full bidirectional glossary.

A **bidirectional glossary** is a glossary that is designed to be used in two-way translation. Terms in one language can be translated into another, and vice versa.

Example: bidirectional bilingual Japanese-English UTX glossary

Language 1: Japanese, language 2: English

Translation direction: Japanese ⇔ English

Example: bidirectional multilingual English-French-German glossary

Language 1: English, language 2: French, language 3: German

Translation direction: English ⇔ French, English ⇔ German (but not French ⇔ German)

A **multidirectional glossary** is a type of multilingual glossary that is designed to be used in any combination of languages in the glossary.

Example: multidirectional multilingual English-Japanese-Chinese glossary

Language 1: English, language 2: Japanese, language 3: Chinese

Translation direction: any combination of the above

1.9 UTX-enabled application and UTX converter

A **UTX-enabled application** is an application that is capable of reading and writing glossaries in the UTX format. Examples of such applications are translation software, a glossary tool, and a UTX editor.

Among UTX-enabled applications, a **UTX converter** is a tool that converts a UTX glossary to/from other formats.

2. The structure of a UTX file

The file extension of a UTX file is ".utx."

A UTX file has the following structure:

Header

The header consists of the following:

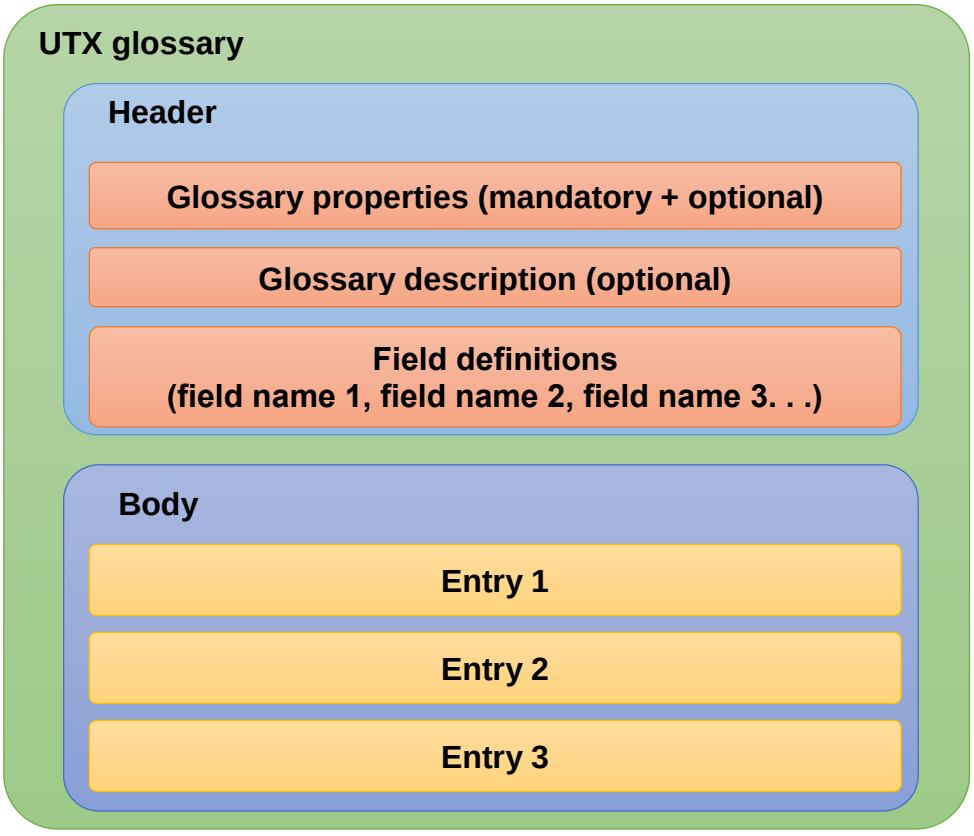
1. Glossary properties. See "3.2.2 Mandatory glossary property" and "3.2.3 List of optional glossary properties."
2. Glossary description (optional). See "3.3 Glossary description."
3. Field definitions. See "3.4 Field definitions."

Body

The body consists of the entries (tab-delimited text). See "4. Field definitions and body."

The following diagram illustrates the structure of a UTX glossary. The components are detailed in later sections.

Figure 1 UTX glossary structure



2.1 Core UTX example

The following is an example of a complete UTX glossary with the minimum required information. The elements are explained in the following chapters.

#UTX 1.20	
#term:en	term:ja
test	テスト

2.2 Character encoding

The character encoding of a UTX file should be UTF-8 with BOM (byte-order mark).

> Note: This is a change from UTX 1.11, in which BOM is omitted. BOM is a special character that is used to identify UTF-8 encoding in certain operating systems such as Microsoft Windows.

2.3 Line

The new line code used for a UTX file is "\r\n" (CR+LF). A new line code is a set of special characters that represent the end of a line. Blank lines are not allowed.

> Note: In a Microsoft Windows environment, CR+LF is often the default setting for text editors and no special attention is required. In an environment where LF is commonly used, such as UNIX or Mac OS, the creator of a UTX file should make sure to change the new line code to CR+LF.

2.4 Line comment

A line comment is any line that starts with # (hash symbol), which is treated as a comment. Existing entries can be commented out by placing "#" at the line start. For example, entries can be commented out to isolate potential technical problems in glossary conversion.

> Note: A UTX converter or UTX-enabled application should exclude line comments when exporting to a format that does not accept line comments.
>
> A UTX glossary that includes line comments should set a `sortable` property to `false` in the header to prevent unintentional sorting (see "3.2.15 `sortable` property").

3. Header

3.1 Language tags

Use IETF's BCP 47 language tags to indicate languages in any part of a UTX glossary. These are the same as the language tags used in HTML and XML. See <http://www.w3.org/International/articles/language-tags/Overview.en.php>. A language tag consists of a subtag to indicate language followed by a regional or script subtag if a distinction is required. If the regional difference is not important, the language should be specified without the regional subtag.

In addition to the language declaration, language tags are used to indicate that certain fields pertain to a particular language. See "4.3 List of Fields."

Examples: Languages tags

Language	Language tag	Language tag with regional subtag	Language tag with script subtag
Chinese (simplified, PRC)	`zh`	`zh-CN`	`zh-Hans`
Chinese (traditional, Taiwan)	`zh`	`zh-TW`	`zh-Hant`
English (US)	`en`	`en-US`	
English (UK)	`en`	`en-GB`	
Japanese	`ja`	`ja-JP`	
Korean	`ko`	`ko-KR`	

3.2 UTX header structure

A **UTX header** is a section that provides information about the entire glossary. All lines in a UTX header begin with "#," meaning that they are treated as commented-out lines.

A UTX header includes at least two mandatory lines, namely, glossary properties and field definitions. The glossary properties can have two or more lines.

One or more lines can be added between the two mandatory lines for a description of the glossary.

Example: UTX header

```
# <Glossary properties>      mandatory + optional
# <Additional glossary properties>    optional
(# <Glossary description >  optional)
(#<Continuing glossary description> optional)
. . .
# <Field definitions>          mandatory
```

3.2.1 Glossary properties

Glossary properties describes various properties of a glossary, including languages in the glossary, creation date, license, and so on.

A glossary property consists of a property name, a colon, a space, and a property value.

Syntax	Example
<property name>: <property value>	`copyright: AAMT (2016)`

Each glossary property is delimited by a semicolon and a single space that follows.

Example: A header for a bilingual UTX file

#UTX 1.20; lang: en/ja; creation date: 2016-04-15; copyright: AAMT (2016)

If a property value is not determined but is to be determined soon, the property item "`undetermined`" can be used.

Example

```
license: undetermined
```

3.2.2 Mandatory glossary property

Name	Syntax	Example
UTX version	`#UTX <version>`	`#UTX 1.20`

3.2.3 List of optional glossary properties

Name	Syntax	Example
lang (language declaration)	`lang: <language 1>/<language 2> or <src>:<source language>/ <tgt>:<target language> etc.`	`lang: en/ja or lang: src:en/tgt:ja or lang: src:en/tgt:ja/tgt:fr`
creation date	`creation date: YYYY-MM-DD`	`creation date: 2016-04-10`
	`creation date: YYYY-MM-DDThh:mm:ssTZD`	`creation date: 2016-04-10T12:34:56Z`
last modified date	`last modified date: YYYY-MM-DD`	`last modified date: 2016-05-10`
	`last modified date: YYYY-MM-DDThh:mm:ssTZD`	`last modified date: 2016-05-10T12:34:56Z`
glossary ID	`glossary ID: <string>`	`glossary ID: Brain surgery`
domain	`domain: <string>`	`domain: Aerospace`

creator	creator: <string>	creator: Yamada Tarou
glossary administrator	glossary administrator: <string>	glossary administrator: Yamada Hanako
copyright	copyright: <string> (<year>)	copyright: AAMT (2016)
license	license: <string>	license: CC BY 4.0
directionality	directionality: <uni/bi/multi>	directionality: bi
sortable	sortable: <true/false>	sortable: true
glossary version	glossary version: <number>	glossary version: 1.00

3.2.4 UTX version property

The `UTX version` property indicates the UTX version of the glossary. It is specified with "UTX", a space, and the version number "1.20." The `UTX version` is the only mandatory glossary property.

3.2.5 `lang` property (language declaration)

The `lang` property, also called **language declaration**, indicates the language(s) of the terms contained in a glossary.

The language declaration of a bilingual glossary is specified in the following format:

Syntax	Example
lang: <language 1>/<language 2>	lang: en/ja

The role of each language can be clarified using `src` (source) and `tgt` (target).

Syntax	Example
lang: <src>:<source language>/ <tgt>:<target language>	lang: src:en/tgt:ja

Note: The language declaration is optional. The languages contained in a glossary are also indicated in term (src/tgt) fields (see 4.2 `term (src/tgt)` fields).

3.2.6 `creation date` property

The `creation date` property indicates the date (and the time) when the glossary is first created. Use ISO 8601 format.

The local time can be indicated by adding the time difference to UTC (coordinated universal time).

Example: creation date property

14:28, April 10, 2016 in Japan Standard Time (UTC plus 9 hours) is represented as one of the following formats.

Type	Syntax	Example
Local date only (no time)	`YYYY-MM-DD`	`creation date: 2016-04-10`

Date plus hour	`YYYY-MM-DDThh:mm:ssTZD`	`creation date: 2016-04-10T05:28:00Z`
		or
		`creation date: 2016-04-10T14:28:00+09:00`

where:

YYYY = four-digit year

MM = two-digit month (01=January etc.)

DD = two-digit day of month (01 through 31)

hh = two digits of hour (00 through 23) (am/pm not allowed)

mm = two digits of minute (00 through 59)

ss = two digits of second (00 through 59)

TZD = time zone designator ("Z" when using UTC, or +hh:mm or -hh:mm when using local time)

See also: <http://www.w3.org/TR/NOTE-datetime>.

3.2.7 `last modified date` property

The `last modified date` property indicates the time and date when the glossary was last modified. Use ISO 8601 format. The local time can be indicated by adding the time difference to UTC.

3.2.8 `glossary ID` property

The `glossary ID` property is a text string that is used as a unique identifier for the glossary.

Note: The `glossary ID` property can be a "name" of the glossary that typically represents the domain of a glossary.

The `glossary ID` property is used only when a glossary contains terms from a single domain. If a glossary contains terms from multiple domains, the `glossary ID` "field" must be assigned for each domain instead of the `glossary ID` property. See "5.3 `glossary ID` field."

Example

```
glossary ID: Rocket engine
```

3.2.9 `domain` property

The `domain` property is a text string that indicates the domain of the glossary. This property is used when you need to group multiple glossaries into a domain. If you use glossary IDs as domain names, the `domain` property is not necessary.

Example

```
domain: Aerospace
```

3.2.10 `creator` property

The `creator` property indicates the name of the person who created the glossary. This is normally the individual responsible for the creation of the glossary. A personal name is preferable, but a name of a department or group may be used.

Example

```
creator: Yamada Tarou
```

3.2.11 `glossary administrator` property

The `glossary administrator` property indicates the name of the person who is responsible for the authorization of the terms. A personal name is preferable to clarify the responsibility. When there are multiple glossary contributors to a glossary, it is recommended to decide who the glossary administrator is. If the creator of the glossary is the sole glossary contributor, `glossary administrator` is not required.

Example

```
glossary administrator: Yamada Hanako
```

3.2.12 `copyright` property

The `copyright` property indicates the name of a person or an entity who holds the copyright over the entire glossary contents, and the year when the copyright took effect. Although this property is optional, it is recommended to clarify the copyright.

Example

```
copyright: AAMT (2013)
```

3.2.13 `license` property

The `license` property indicates the license, or how users can use the glossary. It can be declared in the form of Creative Commons <https://creativecommons.org>, public domain, or other forms of license. Although this property is optional, it is recommended to clarify the license.

In the following example, the license is Creative Commons 4.0 Attribution.

Example

```
license: CC BY 4.0
```

3.2.14 `directionality` property

The `directionality` property indicates the directionality of a glossary (see "1.8 Translation direction"). The property item `uni` indicates that the entries of the glossary are intended for a single translation direction. `bi` indicates that entries in the glossary can be used in both translation directions when the term status permits. `multi` indicates that entries in the glossary can be used in any translation direction when the term status permits. Only a multilingual glossary can have the property item `multi`.

Type of glossary	Property items of `directionality` property
Monolingual glossary	N/A
Bilingual glossary	`uni` or `bi`
Multilingual glossary	`uni`, `bi`, or `multi`

Note: In UTX 1.11, the `bidirectional` flag was used to indicate directionality. This is deprecated by the introduction of multi directionality in UTX 1.20.

3.2.15 `sortable` property

The `sortable` property indicates that whether a glossary's entries can be sorted using one or more fields as a key. The property value is Boolean (true/false). To keep the order of entries, set its property value to `false`.

Example

```
sortable: false
```

Note: By specifying the `sortable` property as `false`, users can loosely track the chronological order of entries when they append new entries at the end of a glossary. "`sortable: false`" can also be used to maintain the position of certain entries in the glossary, even if they have been commented out.

3.2.16 `glossary version` property

The `glossary version` property indicates the version of a glossary. It tracks the revisions of a glossary. This should not be confused with the version of the UTX format.

Example

```
glossary version: 1.02
```

3.3 Glossary description

A **glossary description** is one or more informative text line(s) that may be used for the description

of a glossary, legal notices, disclaimer, etc. A glossary description is preceded by "#."

Example

This is a glossary description.

This is "readme" information regarding the glossary.

This is a disclaimer.

…

3.4 Field definitions

The last line of the UTX header (also begins with "#") includes a set of field (or column) definitions. The details are explained in "4. Field definitions and body" because the field definitions and the body are closely related.

4. Field definitions and body

The body of a UTX glossary consists of entries in each line. Within a line, elements ("cells" in a spreadsheet) are separated by tab characters.

4.1 Language tags for fields

Fields that are specific to a particular language are indicated by the respective language tag. Language tags do not have a space after the colon.

Example: Language tags for fields

Language-specific fields	Example
English term (source language)	`src:en`
Japanese term (target language)	`tgt:ja`
Japanese term status	`term status:ja`
Chinese (simplified) terms (target language)	`tgt:zh-CN`
Chinese (simplified) term status	`term status:zh-CN`
Comments on Japanese terms (a user-defined field for Japanese entries)	`x-comment:ja`

Note: Some fields, such as a user defined `x-comment` field, do not require a language tag if it is relevant to an entire entry. However, a user defined field pertaining to a term in a specific language may require a language-specific field. For example, when there are a number of comments that are only relevant to Japanese target terms, `x-comment:ja` should be created rather than `x-comment` without a language tag.

The language used for the field value may or may not be the language indicated by the language tag. For example, it is possible to write a comment in `x-comment:ja` field in English.

4.2 `term (src/tgt)` fields

Use either `term`, `src`, or `tgt` field to indicate a field of a term. The languages of these fields are indicated by language tags. These fields must match with the language declaration of the glossary.

Use `term` field when the glossary is monolingual, or the roles of source and target languages need not be distinguished.

Use `src` field when you specify one or more source languages.

Use `tgt` field when you specify one or more target languages.

See the following table for examples. Leave no space after the colon before a language tag.

Examples: Glossary types and translation direction notation

For examples of a multilingual glossary, see "6. Multilingual glossary."

Types of glossary	Examples of translation direction notation	Description
Monolingual	`term:en`	English is the only language in this glossary.
Bilingual	`src:en | tgt:ja`	The source language is English and the target language is Japanese.
Bilingual	`term:en | term:ja`	A source/target distinction is not required.

4.3 List of Fields

Field name	Syntax/field item/field value	Example
term (src/tgt)	term:<language> src:<language> tgt:<language> etc.	`src:en | tgt:ja`
pos	noun/properNoun/verb/vt/vi/prenominal/ adjective/adverb/sentence	noun
concept ID	<number>	45
term status	blank/provisional/approved/ non-standard/forbidden/rejected/obsolete	approved
glossary ID	<string>	Brain surgery

4.4 pos field

4.4.1 pos field and its field items

The pos field indicates the part of speech. If the field has no language tag, the part of speech is assumed to be applicable to terms of all languages in an entry. It is also possible to indicate the part of speech for a particular language with the notation: pos:<language tag> (no space after the colon). The pos field is optional (a change from UTX 1.11). The pos field value can be left blank.

The following pos field items are defined:

pos field item	Description
noun	Noun
properNoun	Proper noun
verb	Verb

vt	Transitive verb
vi	Intransitive verb
adjective	Adjective
prenominal	Prenominal
adverb	Adverb
sentence	Sentence

`noun` and `properNoun` specify a noun and a proper noun respectively. Proper nouns include personal and place names.

`verb` specifies a verb. `vt` specifies a transitive verb. `vi` specifies an intransitive verb. `vt` and `vi` can be specified when the distinction is required. `vt`, `vi`, and `verb` can be mixed.

Example: verb, vt, and vi

term:en	**term:ja**	**pos**
receive	受信する	verb
acquire	取得する	vt
listen	リッスン状態になる	vi

`adjective` specifies an adjective. `prenominal` specifies a prenominal modifier.

> Note: A prenominal modifier is similar to an adjective but treated differently. For example, prenominal adjective (*rentaishi*) in Japanese ("円錐形の" for example) is a prenominal modifier that has no inflection. In some MT systems, it may be necessary to distinguish prenominal adjectives from adjectives.

`adverb` specifies an adverb. It may include an adverbial phrase.

4.4.2 `sentence` and special characters

`sentence` is a special `pos` field item that indicates that the "term" is a sentence.

> Note: `sentence` should only be used when necessary. `sentence` would be used for a user interface message in the form of a sentence, for example. Entries of pairs of translated sentences should be stored in a translation memory format (such as TMX) rather than a glossary. When a UTX glossary is exported for an MT system that does not treat sentence as a type of part of speech, `sentence` entries can be treated as nouns.

Some special characters can be used within a `sentence` entry. These are used only for a `sentence` entry.

Character	Notation
Tab character	`\t`
New line character	`\n`
Literal escape character (literal backslash)	`\\`

4.5 Term status field

The **term status** field indicates the status of a term. There are 7 statuses: blank, provisional, approved, non-standard, forbidden, rejected, or obsolete. Only a glossary administrator and a delegate can change the value of a term status.

Example: Term status field

#src:ja	tgt:en	term status:ja	term status:en
プラグイン	plug-in	approved	approved
プラグイン	plugin		non-standard
アドオン	add-on	provisional	

Note: If a glossary does not have a term status field, all entries are considered to be approved.

4.5.1 provisional

The term status "provisional" indicates that a target term is proposed by a contributor but not yet authorized by the glossary administrator. As provisional status is temporary, the glossary administrator should promptly decide the term status such as "approved."

Note: The glossary administrator may also choose to exclude (delete) the term from the glossary, or move it to another glossary.

4.5.2 approved

The term status "approved" indicates that an entry has been approved for the particular glossary (domain) by the glossary administrator. An approved status indicates that the term must be used with the highest priority, whenever applicable. If a term has synonyms or alternative spellings, such as "plug-in" and "plugin," only one of these should have approved status.

An approved term in one language is paired with another approved term in another language. If the parts of speech of these multiple entries are different, then they are different terms. For example, "plot" can be a noun and a verb, and each can have approved status.

4.5.3 Blank term status

If the term status is left blank, it is considered as approved (a change from UTX 1.11). The term status of a term paired with a non-standard, forbidden, rejected, and obsolete term (explained later) can also be blank, which implicitly indicates approved status. See the example at the beginning of "4.5 Term status field."

4.5.4 non-standard

The term status "non-standard" indicates one or more terms that are less-preferred within a group of synonyms or alternative spellings.

Note: The glossary administrator decides whether the term is less-preferred or not for a particular glossary. Therefore, this status could vary in different glossaries, or with a different glossary administrator.

4.5.5 forbidden

The term status "forbidden" indicates that a term must not be used. A term is marked as forbidden not only for being inappropriate as a translation, but also if it is inappropriate within the context of the end-result document.

A forbidden term, unlike a non-standard term, should not be provided as a translation candidate.

Note: A term is “forbidden” because it is inappropriate from linguistic, social, terminological, branding, or other viewpoints.

Up to UTX 1.11, only a target term could be indicated as forbidden. UTX 1.20 allows any term (including a source term) to be indicated as forbidden.

Forbidden terms can be exported from a UTX glossary for terminological checking. Based on this information, a function of a translation tool or a dedicated terminological checker can ascertain whether translation files contain any undesirable terms.

4.5.6 rejected

The term status "rejected" indicates that a term is not appropriate for inclusion in a glossary. Rejected terms can be kept in the glossary for record keeping, moved into a separate list, or deleted at a later time.

4.5.7 obsolete

The term status "obsolete" indicates that a term was previously used, but should no longer be used. Obsolete terms can be kept in the glossary for record keeping, moved into a separate list, or deleted at a later time.

4.6 User-defined fields

Any number of user-defined fields and their field items can be added to a UTX glossary.

For a language-specific field, use a language tag to indicate the language. Language tags do not have a space after the colon.

Syntax	Example
User-defined field:<language>	`x-termUsage:en`

5. Advanced concepts

5.1 Single term status and per-language term status

There are two methods of applying term status: single term status and per-language term status. A glossary can use either of these to indicate the term status.

5.1.1 Single term status

Term statuses can be specified by a single field. The term status noted in this way is called **single term status**. The single term status was the only type of term status up to UTX 1.11.

A single term status within an entry implicitly refers to the source term, target term, or both, depending on the kind of term status. The referent (the term that is referred to by the field) for each term status is shown in the following example.

Example: Single term status of an English/Japanese glossary:

src:en	tgt:ja	term status	(The referent of the term status)
outlet	コンセント	approved	Both source and target terms
outlet	アウトレット	forbidden	Target term
plugin	プラグイン	provisional	Both source and target terms
power point	コンセント	non-standard	Source term

The single term status is determined by the relation between the source and target terms rather than the individual status of source and target terms.

For the single term status, the following rules apply:

- "approved" always applies to a pair of source and target terms.
- "non-standard" applies to less-preferred synonyms or variations of source terms. (Therefore, when reverse-exported, non-standard terms must not be used as target terms.)
- "forbidden" applies to target terms only.
- "provisional" always applies to a pair of source and target terms.

Advantages: The term status is represented by a simple single field. It is suitable if the glossary is unidirectional, and there are no plans for using it in the reverse direction.

Disadvantages: The referents of the term status may be confusing. Additionally, single term status cannot adequately represent how the terms should be used in the reverse translation direction.

5.1.2 Per-language term status

Per-language term status specifies the term status of a term for a particular language rather than a pair of two languages. For example, if it is used in a bilingual unidirectional glossary, it requires two

term status columns, one for the source language, and one for the target language.

Syntax	Example
`term status:<language tag>`	`term status:ja`

If a language tag is not specified, the term status is treated as single term status (UTX 1.11 style).

Note: Per-language term status is introduced in UTX 1.20 to handle bilingual bidirectional glossaries and multilingual glossaries.

5.1.3 Term status behaviors for an MT dictionary

When exporting a dictionary for MT from a UTX glossary, several rules are applied based on per-language term status. A UTX glossary can be mono/bi/multilingual and uni/bi/multidirectional. However, an MT dictionary is usually bilingual and unidirectional. The following examples show how a bilingual/bidirectional UTX is exported to Japanese-to-English and English-to-Japanese MT dictionaries.

The following is a bidirectional UTX glossary. In this glossary, "操作" as source term has a priority over "アクション" because it is an approved term.

Example 1: UTX glossary with a non-standard term

#src:ja	tgt:en	term status:ja	term status:en
操作	action	approved	approved
アクション	action	non-standard	

When a Japanese-to-English MT dictionary is exported from this glossary, both "操作" and "アクション" are exported because both must be translated as "action." See the example below.

Example 2: Exported MT dictionary (Japanese to English) from Example 1

Japanese	English	Priority
操作	action	n/a
アクション	action	n/a

The priority is irrelevant because there is only one target term "action."

When an English-to-Japanese MT dictionary is reverse-exported from this glossary, the case is different from the previous example. The term pair "action/操作" has the higher priority, because "操作" was approved in the original UTX. The term pair "action/アクション" has a lower priority because "アクション" was non-standard in the original UTX. The MT system is expected to honor this priority. If the MT system cannot distinguish terms with a priority, "action/アクション" should not be exported.

Example 3: Reverse-exported MT dictionary (English to Japanese) from Example 1

English	Japanese	Priority
action	操作	high
action	アクション	low

The next example is a UTX glossary that has a forbidden term.

Example 4: UTX glossary with a forbidden term

#src:en	tgt:ja	term status:en	term status:ja
configuration	構成	approved	approved
configuration	コンフィグレーション		forbidden

Consider the case of the exporting an English-to-Japanese MT dictionary from the above UTX glossary. See below.

Example 5: Exported MT dictionary (English to Japanese) from Example 4

English	Japanese	Priority
configuration	構成	n/a

As "コンフィグレーション" is forbidden in the original UTX, it is not included in the exported dictionary.

Now consider the case of reverse-exporting a Japanese-to-English MT dictionary from the same glossary. See below.

Example 6: Exported MT dictionary (Japanese to English) from Example 4:

Japanese	English	Priority
構成	configuration	n/a
コンフィグレーション	configuration	n/a

In the same way as the example of non-standard status, both "構成" and "コンフィグレーション" are exported, because both must be translated as "configuration." The priority is irrelevant because there is only one target term "configuration."

A spelling error can be intentionally included as a source term in a glossary so that it can be translated. In this case, forbidden status must be used instead of non-standard so that it will not be used as a target term when reverse-exported.

Example 7: UTX glossary containing a deliberate spelling error

#src:en	tgt:ja	term status:en	term status:ja
configuration	構成	approved	approved
configulation	構成	forbidden	

Rejected or obsolete term statuses are treated in the same way as if they were forbidden. The decision to include or exclude provisional terms in or from an exported MT dictionary depends on the policy of the glossary administrator.

5.2 concept ID field

A "**concept ID**" can be assigned to multiple synonyms and alternative spellings to indicate that they belong to the same group. A group of such entries is called a "**concept group**." If a source term corresponds to only one target term, a concept ID for this term is not required.

A concept ID can be a number or a text string. When multiple glossaries are merged, entries with the same concept ID can be distinguished by their glossary IDs (see "5.3 glossary ID field").

Up to UTX 1.11, a concept ID was typically required in the case of entries with multiple source terms and a single common target term (n-to-1 relation). In addition to this scenario, UTX 1.20 allows entries containing a single common source term with multiple target terms (1-to-n relation), and multiple source terms with multiple target terms (n-to-n relation).

The example below explains the relationship between concept IDs, concept groups, and term statuses.

Table 1: Example (English to Japanese)

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

8	AAMT	AAMT	approved	3
9	Asia-Pacific Association for Machine Translation	アジア太平洋機械翻訳協会	approved	3

(Serial numbers are given only for the sake of explanation.)

The terms with serial numbers 1, 2, and 3 share the same concept. So do terms 5 and 6, as well as 8 and 9. Therefore, they belong, respectively, to three concept groups (enclosed in red lines). A concept group can be distinguished by its concept ID. The concept group containing the entries with concept ID 3 can be called "concept group 3."

One typical use of concept IDs is to bind an acronym and its spelled-out version as a concept group. In Table 1, term 8 "AAMT" is an acronym and term 9 "Asia-Pacific Association for Machine Translation" is its spelled-out version. They mean the same thing, therefore both have the same concept ID, 3. Entries that have the same concept ID often share at least one source term or target term. However, it is possible that entries have no common term even though the concept is shared, as in the case of "AAMT" and "Asia-Pacific Association for Machine Translation."

Within a concept group, if a term corresponds to multiple terms in another language, only one of these terms has an approved term status. Therefore, concept groups 1 and 2 have one and only one approved term each.

If entries in a concept group do not share a common source and target term, each of the term within these entries can have approved status. For example, concept group 3 contains two approved terms.

If multiple terms have different parts of speech, they have different concept IDs even if they have a common source/target term.

5.3 glossary ID field

A glossary ID field (previously called "dictionary ID" in UTX 1.11 and earlier) is a unique identifier for a glossary. It consists of a text string. A glossary ID field may be required to distinguish certain terms when multiple glossaries are merged.

UTX 1.20 allows a glossary to contain multiple "sub-glossaries" within a glossary (In UTX 1.11, "merging" means to leave no domain distinctions within a glossary). Thus, it is possible to have sub-glossaries with different domains in a single UTX glossary. The glossary ID is considered to be the name of a sub-glossary within a glossary.

When you merge two glossaries, for example, it is possible that two entries originating from different glossaries share the same concept ID (see "5.2 concept ID field"). In this case, different concepts may be accidentally grouped under the same concept ID, as shown below.

Origins of glossaries	Entries	concept ID
Glossary A	outlet	76
Glossary B	instantiate	76

In this case, entries with the same concept ID can be distinguished by using glossary IDs.

Origins of glossaries	Entries	concept ID	glossary ID
Glossary A	outlet	76	Electronics
Glossary B	instantiate	76	IT

5.4 Language-specific fields

Additional fields for properties of a language, such as conjugation, are specified as the following:

<field name>:<language tag> (no space after the colon)

The following fields are pre-defined:

Pre-defined field names for English

Field	Description	Example
plural	Plural form	instances
3sp	Third-person singular form	transmit
past	Past tense form	transmitted
presp	Present participle form	transmitting
pastp	Past participle form	transmitted
comparative	Comparative form	more opaque
superlative	Superlative form	most opaque

Example: Pre-defined fields for verb conjugation in English

#UTX 1.20; lang: en/ja; creation date: 2016-04-15T10:00:00+09:00; copyright: AAMT (2016); license: CC BY 4.0					
#src:ja	tgt:en	pos:en	plural:en	past:en	superlative:en
アーリーアドプター	early adopter	noun	early adopters		
手段	means	noun	-		
白濁した	opaque	adjective			most opaque
保存する	keep	verb		kept	

The field value should always indicate the full form to avoid confusion. For example, write "oranges" instead of "s" to indicate a plural form. The field value is optional, and does not have to be exhaustively filled. Any number of entries can be left blank.

A hyphen "-" is used to explicitly indicate that a field item is not applicable. For instance, as the English word "information" has no plural form, you may specifically use "-" in `plural:en` to indicate this.

6. Multilingual glossary

As of UTX 1.20, a multilingual UTX glossary is supported.

Example of a header for a multilingual UTX file

```
#UTX 1.20; lang: en-US/ja/fr-FR/ko; creation date: 2016-04-15; copyright: AAMT (2016);
license: CC-by 4.0; glossary ID: Rocket engine
```

6.1 lang property (language declaration)

In the language declaration in the header, use language tags to specify one or more languages, as well as their roles as a source or target language (if necessary) in a glossary.

For a bilingual glossary and multilingual glossary, the source language(s) and the target language(s) can be specified with `src` and `tgt`. If `src` and `tgt` are not specified, the languages are treated equally, without designation as a source or target language.

Note: This is a change from UTX 1.11. In UTX 1.11, the first language specified is the source language, and the second language is the target language.

Example: Language declaration

Type of glossary	Language(s)	Example
Monolingual glossary	(single language)	`lang: en`
Bilingual glossary 1	<source language>/<target language>	`lang: src:en/tgt:ja`
Bilingual glossary 2	<language 1>/<language 2>	`lang: en/ja`
Multilingual glossary 1	<source language>/<target language 1>/<target language 2>. . .	`lang: src:en-US/tgt:ja/tgt:zh-CN`
Multilingual glossary 2	<language 1>/<language 2>/<language 3>. . .	`lang: en-US/ja/zh-CN`

6.2 Language tags for `term (src/tgt)` fields (multilingual)

Language tags for `term (src/tgt)` fields in a multilingual glossary are specified in a similar way as the language declaration.

Examples of translation direction notation	Description
`src:en | tgt:ja | tgt:fr`	One source language and two target languages.
`src:en | src:fr | tgt:ja`	Two source languages and one target language.
`term:en | term:fr | term:ja`	The source/target distinction is not required.

7. Appendix A: UTX content guidelines

This guideline provides some best practices and recommendations for creating an effective UTX glossary.

7.1 General guidelines

1. In general, a UTX glossary should contain only technical terms in a specific domain. A UTX glossary should not contain common words with common meanings. Common words may be included if they have special meanings in the domain.

   Correct example: XML declaration/XML宣言

   Correct example: window/ウィンドウ

   Incorrect example: window/窓

2. Use the basic form of a word (singular form for a noun, root form for a verb–as you would see in a commercial dictionary).

   Correct example: define, flow

   Incorrect example: defined, flows

3. For alphabetic characters and numbers, use single-byte characters, not multi-byte characters.
4. Do not include a variable (any words that can change depending on the context) indicated by an ellipsis (…) or any other characters within an entry.

   Incorrect example: "prefer . . . to . . ."

5. Do not include any comments in a field that is not dedicated to comments. Add a comment by either adding a comment field to the glossary or by adding a comment line beginning with "#."
6. Addition of a prefix "x-" to a field or a field item can clarify that they are user-defined. This prefix distinguishes a user-defined field or field item from a field or a field item that might be added to the UTX specification in the future.

   Example (user-defined field): x-serialNumber

   Example (user-defined field item for the part-of-speech field): x-preposition

7.2 Guidelines particular to machine translation use

1. Clearly define the specific domain of the glossary.

   Example 1: `domain: medical electronics`

   Example 2: `domain: legal`

2. For each entry, choose and add the single most appropriate translation (approved term) corresponding to a source term (approved term). Human translators can choose from alternative non-standard terms. However, many MT systems can only use one best target term. The

remaining non-standard terms might not be used at all, or they can be only referenced and used manually.

3. Do not add terms that are relevant only to a specific MT system. Sometimes, certain terms are added only to address specific issues or limitations within an MT system. Such information may not be required in other systems. If such terms are necessary, create a separate glossary to manage them.

7.3 Guidelines particular to the English language

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

7.4 Japanese writing style sample

The writing style of terms should be consistent in a glossary. The following is an example of writing style.

1. 日本語として登録される英数字はすべて半角にする。
2. 半角カタカナや機種依存文字は使用しない。
3. サ変動詞は「する」で終わる。例：強調する
4. 形容動詞はadjectiveとして示す。
5. 形容動詞は「な」で終わる。例：静かな
6. 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
7. カタカナ複合語などの語の区切りを示す中黒や半角スペースは省略しない。
8. すでに定着しているものを除き、意味があいまいなカタカナ語は避ける。

8. Appendix B: Recommended implementation for UTX-enabled applications

To make the best of UTX, we recommend implementing certain features in UTX-enabled applications and UTX converters such as MT software and terminological tools. These are suggestions, not requirements.

For MT systems (for consumers), we recommend implementing the following features.

1. Provide a fine-tuned system dictionary for non-technical, general terms.
2. Use, import, or export approved terms of composite words with a maximum priority.
3. Provide a mechanism to balance the priority of terms. Especially, a shorter term with a priority should not be prioritized over an idiomatic expression (that contains it) in the system dictionary. For example, even when an imported entry "figure/図形" has approved (therefore prioritized) term status, it should not override the idiomatic expression "figure of speech/比喩" already present in the system dictionary.
4. Show non-standard terms as alternative choices (below the priority of approved terms).
5. Predict the plural form of a noun from the singular form.
6. Predict the basic conjugation of a verb.
7. Supplement missing information whenever possible.
8. Enable the use of multiple user dictionaries at the same time.
9. Suppress the use of forbidden terms.

It is preferable for a UTX-enabled application to have a mechanism to suppress the use of forbidden terms systematically. If no such mechanism is provided, forbidden terms should be excluded from use. It is possible that a term is forbidden in one glossary, but not in another. A UTX-enabled application should, preferably, have a mechanism to detect such conflicts.

A UTX converter should be able to exclude entries that have provisional terms (for source and/or target terms) when converting a UTX glossary.

Computer-assisted translation applications for professional translation may have advanced terminological functions. For these systems, we recommend implementing the following features in addition to the above-mentioned items.

1. A mechanism to manage and switch between multiple sets of dictionaries.
2. A mechanism to detect (and optionally replace) forbidden terms.