

Common Format for User Dictionary UTX

(Universal Terminology eXchange)

"Translation software is useless"

Are you sure?

Advantages of UTX

- Significantly improve the translation accuracy
- Accumulate, share, and reuse translation assets
- Cut the time and cost to check up terminology

Introduction

UTX (universal terminology eXchange) is a common format for user dictionary, which is established by AAMT (Asia-Pacific Association for Machine Translation). In 2009, AAMT has established UTX-Simple (later renamed to "UTX"), which is an open format in a tab-delimited text. AAMT is comprised of three entities: researchers, manufacturers, and users of machine translation systems. Machine translation is the core technology for translation software.

Characteristics

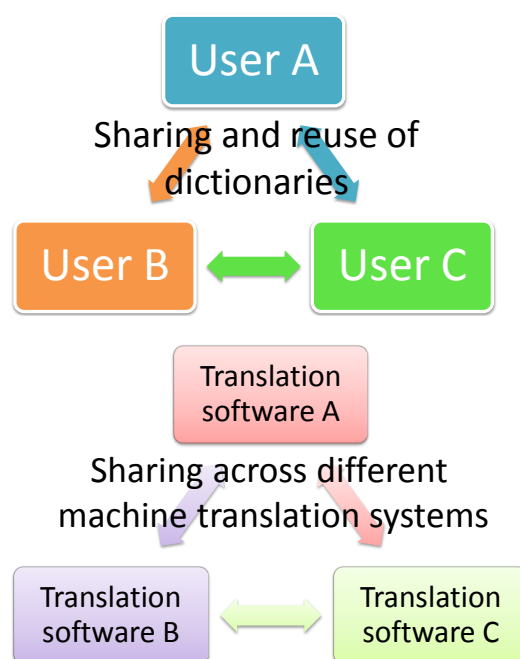
UTX greatly improves the accuracy of translation software by sharing the knowledge of terminology through dictionaries in a bilingual format. The goal of UTX is **to create a simple, easy to make, easy to use dictionary** from a user's viewpoint, not a developer's. A user can easily convert a UTX dictionary into various formats. With or without such conversion, the content of the same UTX dictionary can be used with various translation software or CAT (computer aided translation) tools. In addition, a UTX dictionary can also be used as a glossary without involving translation software.

Why UTX?

With UTX, a user can easily share and reuse user dictionaries of translation software. Have you ever thought that translation software can produce only strange translations? When translation software fails to translate correctly, the problem is often that it **doesn't have sufficient translation knowledge of certain words and phrases** that should be translated. UTX can greatly improve the accuracy of translation software by supplementing translation knowledge as user dictionaries.

Until now, even if a user of translation software makes the effort to prepare user dictionaries, the knowledge is fragmented and dispersed, and thus not effective. Also, even a simple plain text file is difficult to share or to reuse, if its format is not standardized as a dictionary. Although many glossaries are

available on the Internet, their formats are not readily usable out-of-the-box. A time-consuming correction is required to use them in various tools. However, if a single standard like UTX is adopted widely, dictionaries can be widely shared among various tools, such as translation software or terminology management/search tools from different manufactures, and they can also be reused easily.



Who uses UTX?

UTX is specifically designed to be used by end users of translation software, or translators. It does not require any advanced technical knowledge of linguistics, grammars, or machine translation software etc. to create or use it. It can be made only from minimum data such as basic parts of speech (noun, verb, etc.), and the plural form, if the entry is a noun.

In which domains?

UTX can be used in any domain, if it is specialized to a certain extent, such as ICT, medicine, law, engineering translation etc. It may not be suitable for translation of non-specialized, general contents.

What kind of words should we include?

A UTX dictionary contains only **technical terms of specific**

domains, such as names of products, parts, diseases, medicines, and laws. It also contains **proper nouns**, such as names of persons, places, and facilities. In many cases, entries would be nouns, especially compound nouns. 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 words are already contained in the system dictionaries of translation software. 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 be included, 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, but not words.

Is it multilingual?

Yes. The character code of UTX is Unicode (UTF-8 without BOM) and all the languages that can be handled by UTF-8 are accepted. Normally, a UTX dictionary includes only entries of the single source language A, and their translations in the single target language B. One of four statuses, provisional, forbidden, approved, and non-standard, can be specified to each entry. The term with "authorized" status can also be used for translation of the reversed direction (from language B to A).

How do we make UTX?

UTX dictionaries can be **easily created and edited with any text editor or spread sheet application**. A tool is also under development to convert various formats to and from UTX format.

How do we use UTX?

A UTX dictionary is converted in a snap, and can be imported into various tools. In tools such as OmegaT (a translation memory tool) and ApSIC Xbench (terminology reference tool), it can be used with a very little change.

To open source developers and translators - Why not release your glossary in UTX format, and share it with others?

By making a bilingual glossary into UTX format, and publishing and sharing it, you can multilingualize your software quickly and accurately. Thus, many potential users around the world would be able to use your software with a minimum effort.

UTX mailing list

Anyone can participate in the discussions on UTX through a mailing list.

Please search the Web with the keywords: "UTX mailing list."

Reference materials

- Okura Seiji et al. (2008) "Introduction to UTX, a specification for a shared user dictionary" 13th Annual Meeting of Association for Natural Language Processing
- Francis Bond et al. (2009) "Sharing User Dictionaries across Multiple Systems with UTX-S" in Second International Workshop on Intercultural Collaboration (IWIC2009), Stanford

Example

(An excerpt from a UTX 1.11 dictionary for ICT)

#UTX 1.11; en-US/ja-JP; 2011-05-15T10:00:00Z+09:00; copyright: AAMT (2011); 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	オプション的な	adjective	forbidden	
save	保存する	verb	approved	

denotes a comment line.

- 1st line (comment line): Basic information on the dictionary. Each item is divided by a semicolon and a space.

UTX <version number>; <source language>/<target language>; <last update timestamp>; copyright: <Copyright holder>; license: <License>; <Additional data> (optional)

- 2nd line (comment line): The names of properties. Each item is tab-delimited.

In the above example:

<Source word> <Target word> <Part of speech of the source word> <Term status> <Plural form of the source word>

- The 3rd and subsequent lines contain actual entries. Each entry is tab-delimited.

Tips for making a UTX dictionary (glossary)

- Define the domain of the dictionary clearly.
- Only use upper case for proper nouns.
- 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.

Disclaimer

By using the specifications of UTX, UTX-Simple, and UTX-XML (hereinafter collectively called "UTX Specifications") or the dictionaries based on UTX Specifications (hereinafter called "UTX Dictionaries"), you agree to be bound by the following terms. The invalidity or unenforceability of this disclaimer shall in no way affect the validity or enforceability of any other provision herein.

1. To the authors of UTX Dictionaries and related tools from the AAMT and its members:

(1) UTX Specifications are made public, and anyone can use them. The AAMT, however, does not waive any rights thereof and no one may alter UTX Specifications nor make them public.

(2) THE AAMT AND ITS MEMBERS PROVIDE UTX SPECIFICATIONS "AS IS," WITH NO GUARANTEES WHATSOEVER. YOU SHOULD USE UTX SPECIFICATIONS AND UTX DICTIONARIES AT YOUR OWN RISK.

(3) THE AAMT AND ITS MEMBERS SHALL NOT ASSUME ANY RESPONSIBILITY FOR UTX SPECIFICATIONS AND THE RESULT OF THEIR USE INCLUDING, BUT NOT LIMITED TO, THE EXISTENCE OF INFRINGEMENT OF THIRD PARTIES' RIGHTS AND THE ACCURACY, ADEQUACY AND QUALITY OF THE TRANSLATION.

(4) THE AAMT AND ITS MEMBERS SHALL NOT ASSUME ANY RESPONSIBILITY FOR VERIFYING NOR DO THEY GUARANTEE THE LEGITIMACY OF THE COPYRIGHT FOR EACH UTX DICTIONARY. YOU AND THE ORIGINAL AUTHOR OF EACH UTX DICTIONARY ARE RESPONSIBLE FOR THE LEGAL PROBLEM IF IN ANY CASE THAT THE ORIGINAL AUTHOR OF THE UTX DICTIONARY IS NOT THE LEGITIMATE HOLDER OF THE APPROPRIATE COPYRIGHT.

(5) The AAMT and its members grant you the permission to stipulate the terms and conditions for the use of UTX Dictionaries by their users for commercial or non-commercial purposes as long as you have the appropriate copyright; provided, however, that the author of UTX Dictionary is solely responsible for verifying the legitimacy of the copyright for data used in the UTX Dictionary.

(6) THE AAMT AND ITS MEMBERS SHALL NOT ASSUME ANY RESPONSIBILITY FOR THE RESULT OF USE OF THE TOOLS RELATED TO UTX DICTIONARIES.

2. To the users of UTX Dictionaries from their authors:

The users of UTX Dictionaries may make use of UTX Dictionaries, in accordance with their license terms and conditions. Since the license terms and conditions of UTX Dictionaries are varied, please confirm the license indicated in the UTX file header.

3. To the users of UTX Dictionaries and related tools from the AAMT and its members:

THE AAMT AND ITS MEMBERS SHALL NOT ASSUME ANY RESPONSIBILITY IN CONNECTION WITH THE UTX

SPECIFICATIONS AND THE RESULT OF THEIR USE INCLUDING, BUT NOT LIMITED TO, THE EXISTENCE OF INFRINGEMENT OF THIRD PARTIES' RIGHTS AND THE ACCURACY, ADEQUACY AND QUALITY OF THE TRANSLATION. You should resolve such problems between you and the author of the UTX Dictionaries.

MT Research Committee

Members of Sharing/Standardization Working Group (not in a particular order)

YAMAMOTO Yuji (leader)	CosmosHouse
ITOU Hajime	Inter Group Corp.
MURATA Toshiki	Ok Electric Industry Co., Ltd.
SHIMAZU Miwako	Toshiba Solutions Corporation
OKURA Seiji	Fujitsu Laboratories Limited
Michael Konin Kato	Learning Consultant

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

Contact: aamt-info@aamt.info

May 2011 edition