UTX-Simple 仕様 Version 1.10

目次

文書情報

作成者/所有者：AAMT 共有化・標準化ワーキンググループ

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

状態：最終版

更新日付：2011年3月10日

文書の言語：日本語

0. はじめに

0.1. 背景

CAT（翻訳支援）ワークフローでルールベースの翻訳ソフトウェアを使用するとき、原文文書の専門用語、人名、地名は、基本的なシステム辞書には一般的に含まれておらず、期待するほど適切には翻訳されない。しかし、用語が吟味され、特定の分野で適切であれば、追加の詳細情報がなくても、中核的な用語情報（原語および訳語）のみで、機械翻訳の適切性と正確さを向上するには十分であることは広く知られている。だが、ユーザーによって作成された辞書（ユーザー辞書）は、異なるMTシステム間でしばしば互換性がなく、辞書の労力が報われないことが多い。この問題に対処するため、AAMT（アジア太平洋機械翻訳協会）<www.aamt.info>は、異なるMTシステム間で使用できる共有可能な辞書に関する、一連の仕様を策定した。1995年に、AAMTは、IPAの支援を受けてUPF (Universal PlatForm)と呼ばれる最初の仕様を策定した。その後、機械翻訳の技術や利用方法のさまざまな変化を反映するために、2006年から新しい仕様の策定を開始した。新しい仕様は、2007年8月に「UTX (Universal Terminology eXchange)」と呼ぶことが正式に決定された。UTXはオープン標準であり、仕様を利用するにあたってAAMTが課金することはない。2009年に、AAMTは、単純なタブ区切り形式のUTXであるUTX-Simpleを策定した。より複雑な機能は、既存の用語集の仕様であるTBXとTBX-Basicで実現できることが分かったため、AAMTはUTX-Simpleの開発に重点を置くようになった。

また、AAMTは、専門分野のオープンなユーザー辞書データの作成と収集も行う。また、持続可能な方法で、ユーザー辞書を生成、共有、蓄積するユーザー コミュニティーを作成することを計画している。

0.2. 目的

UTX-Simple（以下、UTX-S）の目的は、ユーザーの立場から、簡単に作成でき、簡単に使える機械翻訳用辞書の形式を提供することである。UTX-Sは、高度な管理性や可逆変換よりも、使いやすさと簡潔さを重視している。

同じUTX-S辞書を、さまざまなメーカーの翻訳ソフトで使用できる。さらに、UTX-S辞書は機械的処理に適した形式でありながら、人間にとっても読みやすい形式であるため、翻訳ソフト用途以外の一般的な用語集として使うこともできる。

翻訳ソフトのユーザーがユーザー辞書を作成するとき、個人として各自、分散して辞書を作成するのでは効率的でない。また、辞書形式が標準化されていないと、単純なプレーン テキスト ファイルでさえも、共有したり再利用したりすることが難しい。しかし、UTX-S形式が採用されれば、違うメーカーから提供される翻訳ソフトなどさまざまなツールで共有辞書が広く使われるようになり、再利用性が非常に高くなる。

UTX-Sは、たとえば、複数の情報源から用語集をすばやく作成する場合や、さまざまなアプリケーション間で使用される用語集を配布し再利用する場合に適切である。より詳細かつ長期の用語管理では、TBXが適切である。

1. 対象ユーザー

UTX-Sは、特に、翻訳ソフトのエンド ユーザーあるいは翻訳者向けに設計されている。UTX-S辞書の作成、編集および使用にあたり、言語学、文法、XML、翻訳ソフトなどの特別な知識は必要ない。UTX-S辞書は、起点言語（原文言語）と目標言語（訳文言語）についての最小限の知識があれば作成することができる。

2. 対象分野

UTX-S辞書はどの翻訳分野でも作成し、使用できるが、たとえばICT、医学、法律、エンジニアリングなど、専門性の高い分野で特に効果が高い。理想的には、「ICT分野の中のRuby（スクリプト言語）辞書」、「医学分野の中の心臓外科辞書」などのように、それぞれの専門分野の中でも特定の領域ごとに個別の辞書を作成することで効果が高まる。1つの辞書に明らかに異なる複数の分野が含まれる場合、それぞれの分野について個別の辞書を作成する必要がある。こうすることで、辞書を管理することが簡単になり、また、その他の辞書と組み合わせることによって、容易に再利用することができる。

UTX-Sは、専門性がない一般的・汎用的な内容の翻訳には適していない。このような一般的な内容で使用した場合、UTX-Sの期待される効果は限られる。UTX-Sのフレームワークは、使用するMTシステムが充実したシステム辞書を持っていることを想定している。UTX-Sは、既存のシステム辞書のみでは不可能なほど、翻訳の適切性を向上することができる。非専門用語は、それらが分野内で特定の意味と訳語を持つときのみ、UTX辞書に含めることができる。

3. 定義

3.1. 辞書管理者

1つの辞書に対して複数の用語提出者がいる可能性はあるが、最終的には辞書管理者が辞書の責任者であり、辞書の枠組みを定義する。辞書管理者は、「6.4.2 承認」で説明されるように、ある項目がその辞書の用語として承認されるか否かを決定する。辞書が個人によって作成される場合、辞書管理者と用語提出者は同一となる。

3.2. UTX変換ツール

UTX変換ツールとは、UTX辞書をその他のファイル形式に変換、あるいはその他のファイル形式をUTX辞書に変換するツールの総称である。

3.3. 翻訳方向

UTX辞書は、基本的には、1つの起点言語と1つの目標言語を持つ（すなわち翻訳方向は片方向となる）。完全な双方向辞書とは、すべての用語が逆の翻訳方向でも使用できることが確認されている辞書である。この場合、UTX変換ツールは、その辞書全体を、翻訳方向を逆にした片方向辞書としてエクスポートすることができる。以下で説明する「承認」ステータスが、すべての項目にではなく、一部の項目にのみある場

合、それらの項目だけを逆の翻訳方向の辞書としてエクスポートすることができる。

例

日英のUTX辞書を例とする。いくつかの項目が「承認」ステータスを持つとき、適切なUTX変換ツールを使うことで、これらの項目を英日辞書としてエクスポートすることができる。

3.4. 用語

用語とは、起点あるいは目標言語の見出し語である。用語は、辞書の見出し語のような単語の基本形（レンマ）である必要がある。「8 UTX-Simple作成ガイドライン」を参照。UTXの観点からは、用語とは、ある特定分野の専門用語である。それに対して、単語とは、UTX辞書に必ずしも含まれない、概念を表す文法的な単位を指す。

3.5. 項目

UTX辞書では、項目は1行の論理行で表される。UTX辞書は、「一語一義」の原則に従う必要がある。すなわち、辞書は「1つの用語は、その辞書の特定分野においてはただ1つの意味しか持たないということである。したがって、1つの原語は、理想的には1つの訳語のみを持ち、これが1つの項目（1行）を構成する。訳語が複数必要である場合（つまり複数の意味がある場合）、なぜそれらが複数必要なのか、その根拠を明確に示す必要がある。これにより、体系的な方法で用語集を作成することが促進される。複数の訳語が必要な場合は、同じ概念を示すために概念IDを指定することができる（6.5 概念ID（省略可能）を参照）。

UTXは、組織的なオーサリングおよび翻訳のプロジェクトで使用される場合に、最も効果的である。テクニカル ライターのような、起点言語の文書の作成者は、スタイル ガイドに沿って一貫した用語を使うよう務めていることが前提となる。UTXが組織的でないプロジェクトで使われる場合、その効果は限定される。

4. UTX-Sファイル

UTX-Sファイルの文字コードは、UTF-8（BOMなし）である。改行コードは、″¥r¥n″（CR+LF）である。拡張子は、″.utx″である。

UTX-Sファイルは、以下の要素から構成される。

1. 辞書情報。「5.1 辞書情報」を参照。
2. 列定義。「5.2 列定義」を参照。
3. 項目から構成される本文（タブ区切りテキスト）。「6 本文」を参照。

1と2から、ヘッダーが構成される。

項目は、行頭に″#″を記述することによりコメントアウトすることができる。

5. ヘッダー

UTX-Sヘッダーは、″#″から始まる必須の2行によって構成される。辞書の説明のために、必須の2行の間に、省略可能な行として、1行以上のコメント行を追加することができる。

5.1. 辞書情報

ヘッダーの1行目はUTX-Sファイルについての必要な情報を、セミコロンを区切り記号として次のように記述する。

#UTX-S <version>; < source language >/< target language>; <date created>; <creator>; <license>; <bidirectional (optional)>; <dictionary ID (optional)>; <other optional fields>

- 起点言語/目標言語：ISO 639、3166に準拠
- 作成日付：ISO 8601に準拠

例:

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

5.1.1. ライセンス（省略可能）

辞書のライセンスは、クリエイティブ コモンズ、パブリック ドメイン、あるいは他のライセンスの形式で宣言することができる。ユーザーが辞書をどのように共有し、使用することができるかを明確にすることが強く奨励される。

5.1.2. 双方向（省略可能）

辞書中のすべての用語を逆方向でも使うことができる場合は、辞書のヘッダーに「双方向」フラグを含めることができる。この場合、すべての用語が承認されたとみなされるため、個々の用語ステータスは不要である。

例:

#UTX-S 1.00; en-US/fr-FR; 2008-03-15T10:00:00Z+09:00; copyright: AAMT (2010); license: CC-by 3.0; bidirectional

5.1.3. 辞書ID（省略可能）

辞書IDは、辞書の一意な識別子である。4文字の英数字（大文字小文字の区別なし）からなり、辞書管理者が決定する。複数の辞書を統合するとき、個々の辞書を区別するために必要となる。2つの辞書が同じ概

念ID（6.5 概念ID（省略可能）を参照）の項目を持つ場合、関係のない項目が1グループにまとめられてしまうことが起こりうる。一意な辞書IDにより、このような状況を回避することができる。辞書IDは、必須ではない。複数の辞書を統合する必要が発生したときに、辞書IDを追加することができる。

例：

統合されたとき、辞書を区別するIDがないと、関係のない項目がグループ化されてしまうことがある。

	項目	概念ID
辞書A	outlet	76531
辞書B	instantiate	76531

辞書IDを使用すると、辞書が統合されても、同じ概念IDを持つ項目を区別することができる。

	項目	概念ID	辞書ID
辞書A	outlet	76531	AD64
辞書B	instantiate	76531	5d32

5.2. 列定義

UTX-Sヘッダーの2行目、あるいは、追加の説明行がある場合はヘッダーの最後の行は、列の定義を含む（いずれの行も"#"で始まる）。列定義は、タブで区切られ、3つの必須の列に加えて、ユーザー定義の列を必要に応じて追加できる。列定義と本文は密接に関連するため、詳細は「6 本文」で説明する。

6. 本文

UTXの本文は、1項目ごとの1論理行の集合から構成される。

6.1. 起点用語（必須）

1列目のsrcは、原語（起点言語の用語）を含む。

6.2. 目標用語（必須）

2列目のtgtは、訳語（目標言語の用語）を含む。単言語辞書の場合、目標用語は空白にする。

6.3. 品詞（必須）

3列目の、src:posは原語の品詞を含む。

UTX-Sでは次の品詞が定義されている。

noun / properNoun / verb / adjective / adverb / sentence

品詞が不明な場合は、空白にする。

sentenceは必要な場合のみに使う。文のペアの項目は、辞書ではなく翻訳メモリーに入れる必要がある。

6.4. 用語ステータス（省略可能）

項目は、用語の状態を示すために、4つの用語ステータス（暫定、承認、非標準、禁止）のうちの1つを持つことができる（省略可能）。用語ステータス（term status）は、辞書の主な翻訳方向についてのみ管理される。逆方向のために用語ステータスを管理する必要がある場合は、別の辞書を作成する。

6.4.1. 暫定

用語ステータス「暫定（provisional）」は、用語提出者によりその項目が追加されたが、辞書管理者によってまだ確認されていないことを意味する。暫定の用語ステータスは一時的なものであり、辞書管理者は、用語が以下に示す「承認」「非標準」「禁止」のどれに該当するかすみやかに決定する必要がある。辞書管理者は、必要に応じて、その用語を辞書から除外（削除）することもできる。

6.4.2. 承認

用語ステータス「承認（approved）」は、項目が辞書管理者により承認されていることを意味する。承認ステータスはその用語が必ず使われなければならないことを示す。その根拠はさまざまだが、その用語が、特定の分野における専門用語であったり、ある組織の用語集であったりする場合である。その用語の単語表記に異形がある場合（たとえば″plug-in″と″plugin″）、その中で承認済みの表記のみが使われるべきである。

明らかな理由がある場合、原語に対して複数の訳語を割り当てることは可能であるが、用語ステータスが承認となるのはそのうちの1項目だけである。

用語ステータスが「承認」であるとき、用語を「逆」にし、辞書に定義されている翻訳方向と反対方向にして使うことができる（5.1.2 双方向（省略可能）を参照）。

承認語は常に双方向である。同一の起点言語に対して複数の訳語があり、それらを逆方向にして使う場合、承認語は、それらの項目のうち、唯一有効な用語になる。

用語提出者が１名である場合、（同時にその辞書の辞書管理者でもある）その用語提出者は、辞書に項目を追加した直後に承認ステータスを割り当ててもよい。あるいは、その用語提出者は、該当する翻訳プロジェクトにおいてその新規の項目が適切に翻訳されることを確認できるまで、用語ステータスを空白にしておくか、暫定ステータスを割り当てることができる。

6.4.3. 非標準

用語ステータス「非標準（non-standard）」は、非標準の原語であることを意味する。非標準語は、単に原語の異形に対応するためにのみ許可される。非標準語は、訳語としては使われてはならない。UTX辞書が翻訳のためではなく、文書のオーサリングのための用語集として使用される場合、非標準語は用語として使ってはならない。非標準語が辞書に登録されているのは、起点言語の文書の作成者が承認されていない不適切な語を使用した場合でも、機械翻訳システムが翻訳できるようにするためにすぎないからである。

6.4.4. 禁止

用語ステータス「禁止（forbidden）」は、項目に**使ってはならない**訳語が含まれていることを意味する。そのような語は、言語的、社会的、用語的、その他の観点や、企業のブランド イメージの観点から明示的に禁止される。翻訳ツールが複数の辞書間の優先順位を適切に扱わない場合は、異なる分野別辞書との競合を避けるため、訳語を抑制する必要が生じることもある。

例

ICTの文脈では、英語の"window"は、たとえば日本語の「窓」と翻訳されることは極めて少ない。この「窓」という訳語を、機械翻訳システムが適切に扱えないのであれば、この訳語は明示的に抑制する必要があることがある。

禁止語は、逆方向の項目では承認ステータスとなることがある。また、非標準語は、逆方向の項目では禁止ステータスになることがある（たとえば概念IDが同一である場合）。

翻訳ツールは、禁止語の使用を抑制できることが望ましい。用語が、ある辞書で禁止されていて、別の辞書では禁止されていないということもある。言い換えると、翻訳ツールにおいて複数の辞書を使う場合、禁止のステータスは辞書間で異なることがある。翻訳ツールやUTX変換ツールは、そのような競合を検出する仕組みを持つことが望ましい。

禁止語は、翻訳ツールの外で用語チェックに使用するために抽出することができる。

6.5. 概念ID（省略可能）

異なる用語ステータスを持つ複数の項目がある場合、それらが同じ概念を共有していることを示すために「概念ID」を使用する。1つの原語に対し、1つの訳語しかない場合、概念IDは不要である。

概念IDは、同一辞書内で一意の通し番号からなる。概念IDは、10桁までの整数値である。複数の辞書を統合するとき、同一の概念IDを持つ項目はその辞書IDで区別できる。

例（英語から日本語）：

（用語番号）	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

（ここでの用語番号は単に説明の目的で使われている。これらはUTX辞書には存在しない。またUTX辞書では品詞が必須であるが、ここには挙げていない。）

上記の例では、用語番号1、2、3は同じ概念を指している。用語番号6と7に関しても同じである。4と5は、「一語一義」の原則に従い、概念IDの欄が空白になっている。したがって、区別が必要な項目は他に存在しない。なお、"outlet"の訳語としては「アウトレット」は禁止されているが、「アウトレット ストア」という語の一部としては承認されていることに注意されたい。また、複数の項目が同一の概念IDを共有している場合、承認用語は1つしかないことにも注意されたい。その他の用語はすべて禁止か非標準のいずれかとなる。

例（日本語から英語）

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	

6.6. その他の省略可能な列

5つ目以降の列もすべて省略可能である。ユーザーは必要に応じて情報を定義することができる。単語の活用について、追加の列は以下のように記述する。

- 原語の活用形を定義する場合：src:〈活用形〉
- 訳語の活用形を定義する場合：tgt:〈活用形〉

英語用には、〈活用形〉の種類として次の表記があらかじめ定義されている。

- plural：複数形
- 3sp：3人称単数形
- past：過去形
- presp：現在分詞形
- pastp：過去分詞形
- comparative：比較級
- superlative：最上級

半角ハイフン"-"は、単語に対するその情報がないことを明示的に表す。たとえば、英語の単語"information"には複数形がないため、これを明示的に示す場合は"-"を使う。

あらかじめ定義されている記号

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

uperlative

実際の例

#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	

7. 望ましいMTの機能

機械翻訳システムが以下の機能を持つ場合、UTX-Sは効果的に機能する。

- 非専門用語用の高品質のシステム辞書
- 複数のユーザー辞書が使用可能である
- 短い語よりも長い複合語を優先する
- 禁止項目の使用が抑制可能である

8. UTX-Simple作成ガイドライン

8.1. 一般的なガイドライン

一般的に、UTX辞書は専門分野の専門用語のみを含むべきである。多くの場合、UTXの項目は名詞、特に複合名詞である。翻訳ソフトの基本辞書に含まれない、よく調整された辞書を集め、共有し、再利用することにより、翻訳精度が向上できる。なお1文単位の翻訳は、一種の単語とみなすのが妥当だと思われる場合を除き、UTX辞書に含めない。原則として、UTXは対訳文のデータベースである翻訳メモリーと区別されるべきである。

たとえば「XML declaration」のような語は、辞書に登録するだけで「XML宣言」などと正しく訳せる。一方で、windowのような基本語彙は、すでに翻訳ソフトの基本辞書に含まれているため、含めるべきではない。

- それぞれの項目には、1つの訳語だけを含める。
- 基本辞書に含まれるような基本的な単語は除外する。
- 辞書の分野を明確に定義する。
- 単語の基本形を記述する（例：市販辞書の見出しの形式のように、名詞なら単数形、動詞なら基本形）。
- 原語、訳語そのもの以外のコメント情報は、コメント欄に記す。
- 1つの原語に対して、1つの最適な訳語を記述する（一語一義の原則）。合理的な使い分けが必要な複数の訳語がある場合、それらは別の項目として作成する。
- 特定の機械翻訳システムの処理方法に依存するような単語は含めない。
- アルファベットおよび数字は、半角の文字で記述し、全角で記述してはならない。
- 項目の中で、訳語「…研究所」の「…」のように、単語の一部の入れ替えを前提とする表記は避ける。
- コメントを追加するには、ユーザー定義でコメント用列を定義するか、″#″で始まるコメント行に記述する。

8.2. 英語特有のガイドライン

- 固有名詞を除き、1文字目は常に小文字とする。
- 冠詞（a、an、the）は、それが固有名詞の一部分である以外は記載しない。

8.3. 日本語特有のガイドライン

- 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。

9. 改版履歴

UTX 1.00 (2009年11月10日)

- 初版。

UTX 1.10 (2010年11月22日)

- 全体的な改版。
- 用語ステータス、辞書管理者、双方向、概念ID、辞書ID、および望ましいMT機能の項目を追加した。