UTX-Simple 仕様 Version 1.00

0. 文書情報

●保有者： AAMT 共有化・標準化ワーキンググループ

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

●更新日付： 2010年4月6日

1. はじめに

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フォーマットを提供することである。同じUTX-S辞書を、さまざまなベンダ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、辞書フォーマット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、もしUTX-S形式が採用されれば、違うベンダ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。

1.1. 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辞書の作成にあたり、言語学、文法、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。

1.2. 対象分野

UTX-S辞書はどの分野でも作成し、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。UTX-Sは、専門性がない一般的・汎用的な内容の翻訳で作成、使用する場合は、予期できる効果は限定される。

2. 仕様

2.1. UTX-Sファイル

UTX-Sファイルの文字コードは、UTF-8（BOMなし）であり、改行コードは“\r\n”（CR+LF）、拡張子は“.utx”である。UTX-Sファイルは、ヘッダーと本文から構成される。

1. ファイル情報を記述するヘッダー（1、2行目）
2. 実際の項目（タブで区切られたテキスト）

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。

2.2. UTX-Sヘッダー

UTX-Sヘッダーは、“#”からはじまる2行によって表される。

2.2.1. UTX-Sヘッダー（1行目）

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 <バージョン>; <起点言語>/<目標言語>; <作成日付>; <その他の情報 （作成者、ライセンスなど）>

- 起点言語／目標言語の表記は、ISO 639、3166に準拠する。
- モノリンガル辞書の場合は、目標言語は記載しない。
- 作成日付の表記方法は、ISO 8601に準拠する。
- その他の情報（作成者、ライセンスなど）
-

例1:

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

例2:

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

2.2.2. UTX-Sヘッダー（2行目）

UTX-Sヘッダーの2行目は、3つの必須列に加え、省略可能なユーザー定義の列をタブで区切って記述する。

2.2.2.1. 必須列

- 1列目: **#src**: 起点言語の単語
- 2列目: **tgt**: 目標言語の単語
- 3列目: **src:pos**: 起点言語の単語の品詞

UTX-Sでは次の品詞が定義されている。

noun / properNoun / verb / adjective / adverb / sentence

詳細は0節「品詞」を参照のこと。

2.2.2.2. 省略可能な列

4列目以降は省略可能で、必要に応じてユーザーが定義したい情報を記述する。

- 起点言語の単語の活用形を定義する場合: src:<活用形>
- 目標言語の単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

pastp: 過去分詞形

comparative: 比較級

superlative: 最上級

例:

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

モノリンガル辞書の場合、目標言語の情報は省略できる。

半角ハイフン“-”は、単語に対するその情報がないことを表す。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。

2.3. UTX-S本文

UTX-S本文は、1つ以上の項目から成る。1項目はタブで区切られた形式で記述される。1列目、2列目、3列目は必須フィールドである。UTX-S本文は1つ以上のコメントを含むことができる。コメントは行頭の“#”により表される。

2.3.1. 1列目（必須）

起点言語の単語。

2.3.2. 2列目（必須）

目標言語の単語。

2.3.3. 3列目（必須）

起点言語の単語の品詞。

2.3.4. 4列目以降（省略可能）

ユーザー定義の属性。

2.3.5. コメント

コメント行は行頭に“#”を記述する。

2.4. 品詞

UTX-S辞書で使用できるのは、以下の品詞のみである。

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

もし品詞が不明な場合、空白のままとする。

`sentence`は必要な場合のみに使う。対訳文のペアの項目は、原則として、単語単位での処理を前提とする翻訳ソフトではなく、文単位での情報管理を前提とする翻訳メモリー用のデータに記述すべきである。

3. UTX-Simple作成ガイドライン

3.1. 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。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。

多くの場合、UTXの項目は名詞、特に複合名詞である。たとえば、"XML declaration"という複合語は"XML 宣言"としてユーザー辞書に登録する意義がある。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。

なお1文単位の翻訳は、一種の単語と見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。

- 辞書の分野を明確に定義する。複数分野を単一の辞書内に混在させない。
- 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。
- 基本辞書に含まれるような基本的な単語は除外する。
- 単語の基本形を記述する（例：名詞なら単数形、動詞なら基本形）。

- 必須フィールドの中で直接コメントを追加してはならない。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、“#”ではじまるコメント行に記述する。
- 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。
- 特定の翻訳ソフトの処理方法に依存するような単語は含めない。
- 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。

3.2. 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。

3.3. 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「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「・」は省略しない。もしくは半角スペースで代用する。例：テキスト ファイル
- 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。