翻訳支援のためのシンプルでオープンな辞書仕様 UTX-Simple 1.10

大倉清司[†]、山本ゆうじ[††]、伊藤肇[‡]、加藤マイケル孝仁[‡‡]、島津美和子[+]

AAMT（アジア太平洋機械翻訳協会）　機械翻訳課題調査委員会

共有化・標準化ワーキンググループ

[†]富士通研究所　[††]秋桜舎　[‡]株式会社インターグループ
[‡‡]ラーニングコンサルタント　[+]東芝ソリューション株式会社

1. はじめに

機械翻訳を活用した翻訳支援の研究において、人間による実務的な翻訳の知見が十分に活かされているとはいえない。特に用語管理については、翻訳品質が問題となるプロ翻訳者による翻訳では必須といえるが、その研究は十分でない。

機械翻訳のルールベースと統計ベースの方式では、翻訳支援の方法論も異なる。ルールベース機械翻訳においては、ユーザー辞書に登録した用語を翻訳結果にほぼ反映することができるため、用語管理が必要な翻訳には向いている。一方、統計的機械翻訳においては、用語が確実に反映されないため、厳密な用語管理を必要とする翻訳には不向きである。この場合、後編集作業として用語の統一作業が発生する。いずれの機械翻訳の方式においても、用語管理が重要である。「機械翻訳では十分な訳質は得られない」という考えもあるが、用語管理を行うことで、機械翻訳精度の向上、そして翻訳支援による翻訳効率の向上が可能であると考える。

本稿では、用語管理の観点からAAMT（アジア太平洋機械翻訳協会）が策定した用語集形式UTX-Simple 1.10について説明する。UTX-Simpleは翻訳現場での相互運用性が高く、オープンで、対象を専門用語に限定した仕様が特徴である。2010年末に策定したUTX-Simpleの新バージョン1.10では「暫定」「禁止」「承認」「非標準」の4つの用語ステータスを導入し、用語管理における実用性を高めた。過剰な情報をそぎ落とした良質の用語集は、機械翻訳の精度を効果的に上げるだけでなく、翻訳支援にも使える。今後、UTX-Simpleは、ローカライゼーション、オープンソース、教育、行政、医療、法律などのさまざまな分野で活用が期待される。

2. UTX-Simpleについて

機械翻訳システムを実用的に使用するには、ユーザー辞書が必須だが、辞書の仕様が異なると相互利用できない。そのため、AAMT（アジア太平洋機械翻訳協会）[1]はどの機械翻訳システムでも共通に利用できる、共有辞書の仕様を策定している[2]。より具体的には、1995年にIPAの支援を受けて策定されたUPFをベースに、その後の技術や利用方法の変化を反映し、2006年から新しい仕様策定を開始した。2007年には「UTX (Universal Terminology eXchange)」という新名称に改め、2008年にそのシンプルな形式であるUTX-Simple 1.00の仕様を策定し、公開した[2, 3, 4, 5]。これまでに、医学辞書、法律辞書、言語学辞書をUTX-Simple 1.00形式に変換し、公開している[2]。

翻訳辞書の標準化された仕様としては、LISA(Localization Industry Standards Association）[6]のTBX(Term Base eXchange）[7]が挙げられるが、仕様が複雑であるため作成と管理の手間がかかり、広く普及していないのが現状である。より軽量で分かりやすい形式が求められている。AAMTはLISAとも連携して、標準化をすすめている。

UTX-Simple の主な特徴は以下の 4 つである。

1. **シンプル：**複雑な仕様はむだにユーザーの負担を増やし、結局は使われることがない。ユーザーの立場に立ち、使いやすく分かりやすい、シンプルで実際的な仕様とする。
2. **オープン：** UTX-Simple の仕様は公開されている。仕様に基づいた辞書を誰でも自由に作成して使用できる。
3. **特定分野内で高密度**：「専門用語」という観点から、辞書の分野を明確化し、「一語一義」とする。その分野内で、網羅的に密

度の高い辞書を目指す。無意味に訳語を増やさず、使い分けを厳密に規定する。該当分野で訳語が一義的に定まる語を登録対象とする。

4. **汎用的**：既存のテキスト エディターや表計算ソフトで迅速に共有・再利用ができる。翻訳精度向上に欠かせない辞書の構築がより効率的にできるようになる。

この他、今回策定した UTX-Simple 1.10 から、一方の翻訳方向だけでなく、双方向にも対応している。また単言語辞書として、用語の統一など校正支援ツールなどの入力としても使えるように仕様を策定している。特にオープンソースの翻訳は個人で個別に行っているために訳語の統一が難しい。UTX-Simple の導入により、共通の辞書を共有・再利用することで、翻訳作業の効率化が期待できる。

3. UTX-Simple 1.00 における課題

UTX-Simple 1.00 は、原則的に一方向の辞書フォーマットの仕様であり、例えば英日の辞書を日英として使うことは想定されていなかった。

また用語管理の観点からも課題があった。UTX-Simple 1.00 においても、翻訳支援の要素として一語一義の原則があったが、登録せざるをえない訳語と正式な訳語が、同一概念に含まれることを示す手段がなかった。また、単語を管理する手段もなかった。例えば、複数の翻訳者から用語が提示された場合、どれが暫定的で、どれが最終的な用語かが判断できなかった。また、業務としての翻訳では、政治的要因、社会的要因、企業のブランドイメージ、用語統一などのさまざまな理由から「使ってはならない語」がある。このような語を適切に管理できないと、用語集としての価値は低くなる。

これらの用語管理に関する課題を解決するため、仕様の拡張が行われた。

4. UTX-Simple 1.10 の仕様策定

UTX-Simple 1.10 での最大の向上点は、「用語ステータス(term status)」を導入し、用語管理を通じた翻訳支援での実用性を高めたことである。これは必須項目ではない。

UTX-Simple の機能向上にあたり、当初は、英日辞書を日英辞書としても使用するために、各項目にフラグを付けることが提案された。後に、この目的は、用語管理の機能に含められることが分かった。

用語管理のために、UTX-Simple 1.10 では、「暫定」「禁止」「承認」「非標準」の 4 つの用語ステータスを導入した。それぞれの用語ステータスを持つ語を、暫定語、禁止語、承認語、非標準語と呼ぶ。また、辞書管理の観点から、辞書管理者と用語提出者の概念を導入した。辞書管理者は、辞書の責任者であり、辞書の枠組みを定義する。用語提出者は、新しい用語を辞書に追加する。辞書管理者は、用語提出者が追加する用語が適切かどうかを判断し、用語ステータスを決定する。用語提出者が 1 人の場合はその人が辞書管理者となる。

- **暫定(provisional)**：その項目がまだ辞書管理者（後述）によって承認されていないことを意味する。辞書管理者は、可能な限り、暫定語を以下で説明する「禁止」「承認」「非標準」のいずれかに決定するか、削除することが望ましい。
- **禁止(forbidden)**：用語管理の観点から、項目の中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。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、目標言語の用語を抑制する必要が生じることもある。例えば ICT の文脈では、英語の用語である”window”は日本語の単語である「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だろう。禁止用語は、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。
- **承認(approved)**：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。その用語が必ず使われなければならないことを示す。何か明らかな理由がある場合、起点言語の 1 つの用語に対して複数の目標言語の用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。承認語は常に双方向性をもつ。つまり、辞書に定義されている翻訳方向と反対方向にして使える。同一概念（後述）に属する複数の項目が存在する場合、

翻訳方向を逆にしたときに、唯一有効な項目である。

- **非標準(non-standard)**：1 つの項目に対し、1 つ以上の非標準の起点言語の用語があることを意味する。非標準用語は、単に起点言語の用語の異形に対応するためにのみ許可される。非標準用語は目標言語の用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、非標準用語は用語として使ってはならない。非標準用語が辞書に登録されているのは、起点言語の文書の筆者が承認されていない不適切な語を使用した場合でも、自動翻訳が翻訳できるようにするためにすぎないからである。

複数の訳語を管理する観点から、概念 ID、辞書 ID も定義した。

- **概念 ID (concept ID)**：同じ概念を共有する複数の項目を、10 桁までの数値で表す。つまり、単語の訳が複数ある場合、概念 ID が必要となる。概念 ID は辞書内で他と重ならない通し番号からなる。表 1 は、用語ステータスと概念 ID の例である。

表 1　用語ステータスと概念 ID

src	tgt	term status	concept ID
outlet	コンセント	approved	73
outlet	アウトレット	forbidden	73
power point	コンセント	non-standard	73
PowerPoint	PowerPoint	approved	
outlet store	アウトレットストア	approved	
plugin	プラグイン	approved	245
plug-in	プラグイン	non-standard	245

- **辞書 ID (dictionary ID)**：辞書のための一意的な識別子で、4 文字の半角英数字（大文字・小文字を区別しない）で辞書管理者が定義する。複数の辞書を統合するときに必要である。2 つの辞書に同じ概念 ID を持つ項目がある場合に区別できる。

例えば、辞書 ID がないと、違う概念を表すが概念 ID が同じ単語が統合されてしまう（表 2）。この場合、辞書 ID があれば、概念 ID が同じ項目があっても区別できる（表 3）。

表 2　同じ概念 ID が違う辞書に存在する例

	Entries	Concept ID
Dictionary A	outlet	76531
Dictionary B	instantiate	76531

表 3　辞書 ID の例

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

5. 今後の展望と課題

UTX-Simple の最大の利点のひとつは、編集が簡単にできることである。高機能な用語集形式は、複雑で扱いにくく、翻訳現場での翻訳知識を効率よく収集することができない。今後、UTX-Simple と既存の用語データベースを自動で同期できる仕組みが開発されれば、翻訳の内容に即した多数の用語を、翻訳の現場から収集して、自動翻訳の精度向上に活かすことができる。UTX-Simple と各種用語集形式との変換ツールの開発も期待される。

アジア太平洋機械翻訳協会の共有化・標準化ワーキング グループでは、オープンな辞書コンテンツの収集と公開も行っている。すでに公開中の医学、法律、計算言語学に加え、その他の分野の辞書も追加したい。複数の用語提出者からの用語を、辞書管理者が管理する仕組みができたことから、CGM（消費者生成メディア）的手法による辞書作成が現実的になった。辞書をオンライン コミュニティーで共同作成する場合は、動機付けまたはインセンティブの問題がある。有効な動機付けまたはインセンティブの仕組みの研究は、今後の辞書コンテンツの充実に必須であるだろう。

用語集作成での最善慣行の周知も今後の課題

である。現状では、英語と比較して、日本語の標準的な表記にはばらつきが多い。テクニカルライティングの訓練も不足しており、難解な書き方がむしろ奨励されていることもある。論理的で明快な書き方は、日本語として読みやすいのはもちろんだが、機械翻訳の精度も大きく向上させる。原文としての日本語作文の最善慣行には、日本語表記の整理と統一も含まれる。例えば、日本翻訳連盟では、翻訳における表記の整理が検討されている[8]。

UTX の次バージョンでは、用語ステータスが適切に機能しているかを検証し、翻訳分野の管理、多言語対応などについて検討する。実際の翻訳プロジェクトでの使用実績を積んで、フィードバックを反映させていきたい。

参考文献

[1] http://aamt.info/
[2] http://www.aamt.info/japanese/utx/
[3] 大倉清司他(2008)「共有ユーザー辞書仕様 UTX の現状と今後の展開」言語処理学会第 13 回年次大会（東京）
[4] Francis Bond et al. (2009) “Sharing User Dictionaries Across Multiple Systems with UTX-S” in Second International Workshop on Intercultural Collaboration (IWIC2009), Stanford
[5]http://www.aamt.info/japanese/utx/utx-simple-1.00-specification-j.pdf
[6] http://www.lisa.org/
[7] http://www.lisa.org/standards/tbx/
[8] 日本翻訳連盟の表記検討フォーラム SINAPS http://jtf-forum.jp/

問い合わせ先

AAMT（アジア太平洋機械翻訳協会）では、UTX の仕様策定や辞書作成、評価にご協力いただける方を募集しております。現時点では、日本語、英語、中国語を優先しています。興味のある方は下記ページからお気軽にご連絡ください。

UTX についての説明 URL:
http://www.aamt.info/japanese/utx.htm

UTX メーリングリストについて:
http://groups.yahoo.co.jp/group/UTX/
（どなたでも参加できます）

UTX 1.10 のサンプル

	A	B	C	D	E	F
1	#UTX-S 1.10; en-US/ja-JP; 2010-12-20T17:00:00Z+09:00; copyright: AAMT (2010); license: CC-BY 3.0					
2	#description: This is a sample dictionary for AAMT-related terminology. It is not an official dictionary. / この辞書はサンプル用のAAMT関連の用語辞書です。AAMTの公式の辞書ではありません。					
3	#src	tgt	src:pos	term status	comment	concept ID
4	Asia-Pacific Association for Machine Translation	アジア太平洋機械翻訳協会	properNoun	approved		
5	source language	起点言語	noun	approved	必要に応じて「原文言語」と注記する。	
6	target language	目標言語	noun	approved	必要に応じて「訳文言語」と注記する。	
7	dictionary administrator	辞書管理者	noun	approved		
8	provisional	暫定	adjective	approved	用語ステータスの一つ。	
9	approved	承認	adjective	approved	用語ステータスの一つ。	
10	non-standard	非標準	adjective	approved	用語ステータスの一つ。	
11	forbidden	禁止	adjective	approved	用語ステータスの一つ。	
12	concept ID	概念ID	noun	approved		
13	dictionary ID	辞書ID	noun	approved		
14	merge	統合する	verb		辞書について。	3
15	merge	マージする	verb			3
16	dictionary	辞書	noun	approved		2
17	dictionary	ディクショナリー	noun	forbidden		2
18	optional	省略可能	adjective	approved		4
19	optional	オプショナル	adjective	forbidden		4
20	contributor	用語提出者	adjective	provisional	要確認	