用語データを共有・再利用するための用語集形式の標準規格 UTX

山本ゆうじ／AAMT（アジア太平洋機械翻訳協会）http://cosmoshouse.com/mail.htm （個人連絡先）

概要 自然言語処理研究者とプロ翻訳者の関心が乖離していることは長年指摘されてきた。標準化された用語集形式は、その溝を埋める一つの方法である。AAMT が策定したシンプルな用語集形式 UTX は、企業や特許庁などの組織で活用されている。本論では、UTX 1.20 の特徴を説明し、ISO 規格 TBX と比較しつつ、UTX の用語データ収集、用語データ交換、用語データ抽出時の出力形式としての利点を論じる。UTX には用語管理の観点が盛り込まれているため、用語集の合意形成のツールとしても優れている。さらに、標準化されていながらシンプルで扱いやすいため、研究・教育用のデータ形式としての活用も期待される。

対訳キーワード

term:ja	用語管理	体系的翻訳	コンピューター支援翻訳	機械翻訳	用語集	用語ベース	オープンデータ
term:en	terminology management	systematic translation	computer-assisted translation	machine translation	glossary	termbase	open data

UTX, a Standardized Glossary Format for Sharing and Reusing Glossary Data

Yamamoto Yuji/AAMT (Asia-Pacific Association for Machine Translation)

Abstract There has been a huge gap between the interest of NLP researchers and professional translators. A standardized glossary format, if properly defined, may be able to bridge the gap. UTX is a simple glossary format that was established by AAMT. It is used by enterprises and organizations including the Japan Patent Office. This paper clarifies the characteristics of UTX 1.20 and analyzes its advantages as a format for collection, exchange, and extraction of term data. Since UTX incorporates a terminology management feature, it excels as a tool of agreement formation for a glossary. UTX's simplicity and ease of handling make it useful as a data format for both research and education.

1 はじめに

自然言語処理研究と実務分野での体系的翻訳の現場が乖離していることは長年指摘されてきた。標準化された用語集形式は、研究と実務翻訳現場を言語資源で結びつける方法である。UTX は、AAMT（アジア太平洋機械翻訳協会）が策定した用語集形式の標準規格である。AAMT は非営利団体であり、UTX 仕様を無償で提供している<http://www.aamt.info/japanese/utx/>。UTX は企業・組織での用語管理、機械翻訳に使われている。

本稿では、UTX 1.20 を含む UTX の特徴を説明し、ISO 規格 TBX と比較しつつ、同義語管理、標準化されたタブ区切り形式などによる UTX の用語データ収集、用語データ交換、用語データ抽出時の出力形式としての利点を論じる。UTX 用語データにより複数の機械翻訳システムで BLEU スコアが改善されることは、Bond ら(2009)が示した。さらに大倉ら(2011)は、わずか 51 語の UTX 用語データでも、文書に対して適切であれば、複数の機械翻訳システムで精度を向上できることを示した。その後、特許庁では 2014 年に 220 万語の中日辞書が作成された。調査報告書では、UTX による翻訳精度の向上について以下のように述べられている。

「『中日対訳辞書データ』を機械翻訳辞書に追加することにより、用語(名詞)の翻訳精度に関して一定の向上効果が得られることが確認できた。」(特許庁, 2012)

2011 年 5 月に UTX 1.11 正式版仕様書、そして 2015 年 12 月に UTX 1.20 ベータ版仕様書が公開された。UTX 1.20 ベータ版では、多数の改良が施された<http://www.aamt.info/japanese/utx/utx1.20.htm>。たとえば、多言語多方向の用語集が作成可能になり、用語ステータスをみなおすことで TBX との互換性が向上した。

2 問題意識—研究と体系的翻訳との乖離

UTX が策定され、アップデートされている背景には、研究と体系的翻訳の乖離に対処するという問題意識がある。UTX の特徴の一つは、用語データに体系的翻訳の観点を取り込むことで、それまでの「翻訳ソフトのためのユーザー辞書」という位置付けを、汎用的な用語集形

式として見直した点である。これは、UTX 1.10 の時点で用語ステータスが導入されたときにすでに現れていたが、UTX 1.20 では用語ステータスを改善することでさらに強化された。

これまで、日本の機械翻訳の開発と研究では、翻訳者を介さずに機械翻訳で処理を完結する方向性が主流であった。翻訳支援の観点、つまり体系的翻訳手法に基づき、プロ翻訳者チームを支援するという発想は限られる [i]。一方、翻訳学でも、翻訳支援の活用を含む翻訳実務技能は、翻訳学全体のごく一部としてしかみなされていない。[ii]

実務翻訳での体系的翻訳は、表記規則、用語集に基づく用語管理、翻訳メモリーの 3 つの柱からなる。翻訳支援では、用語や表記の一貫性、正確性が重要になり、概訳とは要件が異なる。企業や組織では大量の多言語文書をすばやく高品質に翻訳する必要性があるが、これは体系的翻訳によってはじめて可能となる。このような体系的翻訳では、多くの場合、翻訳者は個人ではなくチームの一部として作業する。つまり、翻訳に関わる全員で、用語や表記を統一し、一貫性を維持・管理する必要がある。

欧米では、翻訳メモリーに類似する概念は 30 年以上前に示唆されている（Kay, 1980）。だが、日本では、グローバルに展開している企業でも、表記規則、用語集、翻訳メモリーの体制を完備した体系的翻訳は普及していない。また、自然言語処理で求められる良質なコーパスには、表記や用語に一貫性があることが期待されるが、コーパスが作成される翻訳過程が体系的翻訳に基づいていなければ、高品質は望めない。体系的翻訳を行い、用語管理をすることはコーパスの品質を高めるためにも必須である。

研究と翻訳支援の立場の根本的な違いには、以下のような点もある。自然言語処理研究では「特定の翻訳プロジェクトにのみ役立つ言語資源」より「さまざまな翻訳に役立つ言語資源」のほうが有用であろう。だが、翻訳実務や翻訳支援の観点からは翻訳メモリーや用語集は、「特定の翻訳プロジェクトのみに役立つ」ことが当然であり、またそうでないとむしろ不都合がある。このような異なる立場を明確に意識して理解することにより、言語資源をより有効に活用できる。

3 UTX の特性と改善点

用語集標準規格としては、ISO の TBX（ISO 30042）が広く使われているが、TBX と UTX はいくつかの点で異なる。TBX は用語集マークアップ構造を定めた TMF（ISO 16642）に基づき、項目よりも「概念」を上位に置いている。これは論理的な考え方ではあるが、用語データの構造が複雑になる。翻訳支援では翻訳者や翻訳ツールが必要としているのは、確実に使用できる原語と訳語である。TBX の概念指向に対して、UTX はいわば「項目指向」であるといえる。項目が表形式で展開できるため、編集が容易であるとともに、タグが不要なため、後述のようにデータのサイズも縮小できる。

3.1. UTX での同義語の効率的な管理

UTX は、用語ステータスと概念 ID を使用することで、概念指向の用語集形式よりもシンプルな構造を実現できる。同義語を処理する際に、必要最小限の情報量で、訳語の位置付けを明確にできる。

用語集形式では、同義語をどう表記するかが課題であり、特に表形式用語集では慎重な扱いが必要である。UTX では、原則として、複数の同義語が存在する、「別の語だが同じ意味」の場合にのみ、概念 ID（concept ID）を付与する。表 1 は UTX 1.20 ベータ仕様に基づく表記である（私的に作成した ISO 用語集のデータより）。

表 1 UTX 1.20 ベータ版の例（UTX 1.11 互換用語ステータス）

#UTX 1.20beta			
#term:en	term:ja	term status	concept ID
expert	エキスパート	approved	1
expert	専門家	forbidden	1
directive	指令	approved	2
directive	ディレクティブ	forbidden	2
liaison	リエゾン		

概念 ID が同じ語群の間でも、用語ステータスによって、日本語に対する訳語の位置付けが明確になる。つまり、この特定の用語集では、「英日翻訳の場合、expert の訳語は『エキスパート』であり、『専門家』という訳語は使用しない」ということが示されている。

翻訳支援に使用する用語集では、同義語の割合はそれほど多くはない。同義語がどれだけ含まれるかは、分野や用途に依存する。またどの語を訳語とし、またどの語を同義語とするかは、特定の観点、プロジェクト、用語集作成者などに依存する。だが、同義語を積極的に記載している用語集でも、その割合は 1 割以下 [iii]であることが多く、対訳用語集の用語の大半は 1:1 対応する。また用語管理の観点からは、むしろ 1:1 対応するようで用語を選ぶことで利用者の混乱を防ぐべき

[i] 2016/01/02 時点で、学術文献検索サービス Google Scholar での"翻訳支援"の検索結果は 227 件、"computer-assisted translation" は約 3000 件。"用語管理"+"翻訳"の検索結果で、翻訳での用語管理に触れた文献は数点のみ。"terminology management"+"translation"では 2820 件ヒットする。

[ii] たとえば Munday, J.による翻訳学入門 *Introducing Translation Studies* 第三版でも、翻訳の文化側面と理論が中心である。

[iii] 現時点では用語集内の同義語の割合の統計的根拠は不十分である。今後、オープンな用語集が増加すればより正確な割合が判明すると期待される。

という必然性がある(これは特定目的の用語集が網羅的な辞書と異なる点でもある)。微妙な使い分けが重大な問題となる用語は、用語管理の観点から、採用すべきか慎重に判断する必要がある。UTX は表形式で一覧表示でき、複数用語を比較しながら用語ステータスを付与できるので、機械抽出した用語データで、用語を取捨選択する際の合意形成に適している。

3.2. UTX 1.20 での言語ごと用語ステータス

UTX 1.20 では、各言語の用語について用語ステータスを付ける方式(「言語ごと用語ステータス」)が追加された。この「言語ごと用語ステータス」により、TBX との互換性が高まったほか、(3 以上の言語を含む)多言語用語集が可能になり、さらに翻訳方向を逆にした場合でも情報が的確に保持されるようになった。UTX 1.11 では、二言語一方向(たとえば英語を日本語に訳す場合)のいわば「非対称」の翻訳が多いことを踏まえ、用語ステータスを1列で表記するようにしていた(表1)。これはシンプルであるという利点があったが、たとえば日本語を英語に訳す必要が出てきた場合は、注意が必要であった。UTX 1.20 で言語ごと用語ステータスでは、英語と日本語のそれぞれに用語ステータスを付与する。

表 2 UTX 1.20 ベータ版の例(言語ごと用語ステータス)

#UTX 1.20beta				
#term:en	term:ja	term status:en	term status:ja	concept ID
expert	エキスパート	approved	approved	1
expert	専門家		forbidden	1
directive	指令	approved	approved	2
directive	ディレクティブ		forbidden	2
liaison	リエゾン			

これにより、英語を日本語に訳す場合も、日本語を英語に訳す場合も、同じ形式、つまり反転させても同じ「対称的」翻訳ができる。ただ、用語ステータスは言語の数だけ必要になる。たとえば、対訳(二言語)用語集では、用語ステータスは 2 列必要になる。

3.3. 標準化されたタブ区切り形式による軽量化

UTX は、タブ区切り表形式で用語データを簡潔に表現できる。タブ区切り形式の利点の一つは、XML のようなタグが不要で、ファイルサイズを縮小できることである。ファイル構造が比較的単純なら、タグを使用しなくても表形式で的確に用語情報を記述できる。表 3 は、UTX 用語集を Glossary Converter <http://www.cerebus.de/glossaryconverter/>で SDL 用語ベース形式に変換した後、TBX 形式に変換した時のサイズ比較である。これらの用語集はすべて AAMT が Creative Commons ライセンスで公開しており、無償で利用できる <http://www.aamt.info/japanese/utx/download.htm#glossaries >。この変換では、各形式が持つ情報量は同じであるが、UTX は、TBX ファイルの 1/4～1/5 のサイズである。

ファイルサイズは、より語数が多い用語集では無視できない問題となる。たとえば、特許庁が作成した JPO 中日対訳辞書(220 万語)は、約 190MB である。TBX 形式で 4 倍のファイルサイズになると仮定すると、約 760MB となる。ファイルサイズを抑えることにより、用語データが扱いやすくなり、利用を促進できる。

ただし、UTX は TBX と競合するものではない。UTX は一覧性に優れるが、TBX は複雑な用語集、多言語用語集に適している。用途に応じて使い分け、変換ツールを用いて TBX 他の各種形式に変換することで、用語データを最大限に活用できる。

3.4. 断片化した用語データの統合とオープン データ言語資源としての活用

UTX は、オープン データとしての言語資源の形式としても最適である。現状では言語資源が断片化し、有効に蓄積できていない。単なるタブ区切り形式というだけでなく、データ構造を標準化することで、断片化した言語資源の集約が可能になる。2011 年 8 月の時点では、言語処理学会ウェブサイトでも日英対訳用語集が CSV 形式で公開されていた。たとえば、今後、学術論文のキーワードを UTX 形式の対

表 3 用語集形式によるファイル サイズの比較

	語数	SDL用語ベース形式(KB)	TBX形式(KB)	UTX形式(KB)	TBXとの比
法律用語集	5,451	22,648	1,585	430	27.13%
医学用語集	27,122	102,704	7,973	1,003	12.58%
サッカー用語集	9,883	39,388	2,924	789	26.98%
平均	14,152	54,913	4,161	741	22.23%

訳で提出することを義務づけることにより、高品質の対訳用語集を作成できる。本論文の冒頭では、UTX 形式に近いが、縦横を入れ替えた形で、試験的に対訳記載をしている。このように言語間の対応関係を明確にすれば、簡単に用語抽出できる。

また、言語資源の著作権を明確化し、標準化することで、利便性が高まる。データベースに対応した Creative Commons 4.0 などのライセンス形式も有用である。用語集を UTX 形式にして、用語を参照、編集、管理、共有するためのさまざまな障害を取り除き、有益な言語資源を構築することが、翻訳研究や翻訳業界の発展に資するものと思われる。

4 結論と今後の展望—翻訳教育での活用

本論では、自然言語処理研究と体系的翻訳のギャップを埋めるべく UTX がバージョンを重ねてきたことを示した。また、UTX が、用語ステータスと概念 ID を使用することで、シンプルな構造を実現できることを示した。さらに、UTX 1.20 での言語ごと用語ステータスの導入により、TBX との互換性向上、多言語用語集および双方向翻訳が実現したことを示した。そして、UTX が用語データの軽量化に貢献することを示した。さらに、標準化したタブ区切り形式がオープン データ言語資源として有用であることを示した。上記の点により、UTX 1.20 は、シンプルさと利便性を両立できると言えよう。

今後とも、UTX チームは、TBX グループとの連係を継続していく。AAMT が開発した UTX 変換ツール <http://www.aamt.info/japanese/utx/tools.htm>はオープンソースであり、開発者が機能拡張や他製品に組み込むことができる。ただ、このような開発が行われるには、今後、企業での体系的な翻訳支援技術の研究が必須となる。企業に体系的翻訳の重要性を啓蒙する活動が重要である。用語集の作成や活用が行われていないのは、用語集管理が複雑だからである。今後も、企業にシンプルな UTX 用語集の利点を周知し、利用を働きかける努力が必要となる。

UTX は、用語管理でのベストプラクティスであり、翻訳教育でも活用できる。体系的翻訳が日本に根づいていないのは、これまでは大学での翻訳学科が少なかったことも一因だろう。翻訳過程そのものを体系的、工学的に処理する研究、すなわち翻訳工学が今後進展するために、用語管理を含む体系的翻訳への取り組みが期待される。UTX 用語集は、翻訳の現場でどのように用語集が作られ、利用されるかを、商業翻訳に触れる機会がない研究者や翻訳初学者が理解する一助となる。2015 年現在では、立教大学、関西大学、津田塾大学などで翻訳学科が開設されているが、用語管理や翻訳メモリーを含む体系的翻訳の教育と教育がさらに活発化することが望まれる。

5 参考文献

AAMT. (2011).「UTX 仕様 Version 1.11」Retrieved January 8, 2016, from http://www.aamt.info/japanese/utx/utx1.11-specification-j.pdf

AAMT. (2015).「UTX 仕様バージョン 1.20 β」 Retrieved January 8, 2016, from http://www.aamt.info/japanese/utx/utx1.20beta-specification-j.pdf

Bond, F., Okura, S., Yamamoto, Y., Murata, T., Uchimoto, K., Kato, M., Shimazu, M., & Suzuki, T. (2009, February). Sharing user dictionaries across multiple systems with UTX-S. In *Proceedings of the 2009 international workshop on intercultural collaboration* (pp. 147-154). ACM.

ISO. (2003). Computer applications in terminology -- terminological markup framework

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

Kay, M. (1997). The proper place of men and machines in language translation. *Machine translation*, *12*(1-2), 3-23.

Lommel, A., Melby, A., Glenn, N., Hayes, J., & Snow, T. (2014). TBX-Min: a simplified TBX-based approach to representing bilingual glossaries. In *Terminology and knowledge engineering 2014*

Melby, A. (2015). TBX: A terminology exchange format for the translation and localization industry, *Handbook of terminology* (Vol. 1) (pp. 393-424). John Benjamins Publishing Company.

Okura, S., Yamamoto, Y., Ito, H., Kato, M., Shimazu, M., & Bond, F. (2011). UTX 1.11, a simple and open user dictionary/terminology standard, and its effectiveness with multiple MT systems. In *MT Summit 2011*

Wright, S. E., Rasmussen, N., Melby, A. K., & Warburton, L. (2010, October). TBX Glossary: a crosswalk between termbase and Lexbase formats. In *Proceedings of developing, updating and coordinating technologies, dictionaries and lexicons for terminological consistency workshop*.

特許庁. (2012).「平成 24 年度 中国特許文献の機械翻訳のための中日辞書整備及び機械翻訳性能向上に関する調査 調査報告書 概要版」

山本ゆうじ. (2015). ISO glossary. Retrieved December 28, 2015, from https://goo.gl/ZyBvm9