日本語文章の課題と前編集手法
—用語集形式 UTX と実務日本語の観点から—

Issues of Japanese writing and pre-edit approaches to address these issues
- In view of UTX and Practical Japanese -

秋桜舎代表 **山本　ゆうじ**

PROFILE: 筑波大学を経てシカゴ大学修士号。実務翻訳業務、大規模翻訳・文書管理／作成の講習やコンサルを行う。近著に『IT 時代の実務日本語スタイルブック——書きやすく、読みやすい電子文書の作文技法』。

1 体系的翻訳手法の現状

本稿では、日本語文章（ビジネス文書や特許文書）のさまざまな課題の中から、CAT（コンピューター翻訳支援）での「前編集」に関わる問題を、用語集形式 UTX（後述）と実務日本語がどのように解決できるかを説明する。

大企業、大規模組織では、ICT を活用する体系的翻訳手法が必要になる。体系的翻訳手法とは、用語集、スタイル ガイド、翻訳メモリー、翻訳ソフトのすべてを活用して、品質と効率を最大化する手法を指す。これらすべてに ICT が関わるが、CAT という呼称では、特に翻訳メモリーや翻訳ソフトが中心となる。

実際には、体系的翻訳手法が正しく機能している事例は少ない。体系的翻訳手法に対する理解度も企業や個人で大きな差があり、共通理解が関係者間で共有されているとは到底いえない。企業で、用語集が不完全や矛盾だらけという例や、翻訳メモリーの基本機能を勘違いしている例はまだいい。翻訳メモリーやスタイル ガイドの存在すら知らない、用語集がまったく使われていないなど、重大な問題はいくらでもある。また、体系的翻訳手法は、企業では活用されているが、大学ではあまり使われていない。研究者は原文を直接読めるので翻訳が不要、というのは理由の一部でしかない。日本では、翻訳という作業が独立した分野と見なされてこなかった。研究者が行う翻訳は、個人レベルが多い。研究グループや複数の著者が関係する論文でも、個人の著作の集合である。用語および表記統一が厳密に求められるとは限らない。もしするとしてもそのプロジェクト限りであろう。Venice をヴェネツィア、ベネチア、ヴェニス、ベニスのどれで表記するかを、研究科が強制することはない。良くも悪くも研究者個人の判断しだいであるし、文書の性質によって変えることもある。このような大学での文書に対して、企業の文書は、個人名ではなく企業名で出されるため、多くの文書で用語や表記を統一することが重要になる。外部向けの文書ではもちろんだが、内部文書でも統一することが望ましい。このような大学と企業の翻訳に関する状況の違いがあるため、大学の研究者に CAT の話をしても、用語集、スタイル ガイド、翻訳メモリーの必要性が理解されないことも多い。

2 概訳と翻訳支援の違い

翻訳関連技術には、「概訳」と「翻訳支援」という 2 つの異なる方向性（または用途）がある。概訳とは、翻訳者でない人が外国語文章のだいたいの意味を知るときに必要とするレベルの翻訳である。一般的な流れは、Google Translate などの翻訳サイトでボタンをクリックして、出力された訳文を確認するだけ、というものである。機械翻訳の生の出力そのままであるため、処理速度は最も速い一方で、翻訳精度は低い。概訳が目的の場合、後編集（機械翻訳処理後の手作業の修正作業）はほとんど、またはまったく行わない。なお、概訳は、「抄訳」とは異なる点は注意が必要である。抄訳は、翻訳と要約を同時に行う、高度な作業である。

概訳に対して、翻訳支援とは、通常、プロ翻訳者が高

品質の訳文を完成させる過程を支援することである。翻訳支援では、翻訳メモリーが中心となるが、一部の翻訳ソフトも含まれる。翻訳関連技術の専門家同士が話をしていても、概訳と翻訳支援という前提の違いに気づかないと、話が食い違うことがよくある。概訳と翻訳支援は、同じ自然言語処理技術に基づくことはあるが、機械翻訳ユーザーの技能が大きく異なり、ユーザーが作成する訳文の品質もまったく異なる。

概訳で使われることも多い統計機械翻訳では、イタリア語とスペイン語、韓国語と日本語など、類似言語間では一定の精度が得られる。しかし、英語と日本語という言語構造と語順が大きく異なる言語の組み合わせでは、十分な精度は得られない。だが、最近の傾向では、実際の精度を無視して、異種言語での統計機械翻訳をむりやりに適用して翻訳コストを浮かそうとする試みがあり、問題を生んでいる。

3 悪文は体系的翻訳手法の足かせとなる

概訳と翻訳支援のどちらであっても、体系的翻訳手法がうまく機能しない大きな原因の一つは、品質の低い文書、つまり「悪文」である。これは、作文という分野が、企業・組織・教育機関で軽視されていることによるもので、文書作成者側の問題である。文章の問題は、数値化しにくく、発見しにくい。特に、「ある文章がどれだけ分かりやすいか」という点は放置されていることが多い。専門家同士の文書のやり取りで、互いに意味不明の専門用語を乱発し、互いによく理解していないのに、あえて放置していることがないだろうか。

文章の問題による現実的な悪影響は、発見しにくくはあっても確かに存在する。たとえば用語と表記の不統一により、読み手が混乱することや、翻訳工程で本来は不要な作業が発生し、コストが増大することなどである。

日本語能力は、基本的な能力であるため、根本的に改善するには日本の学校教育全体を変えるしかない。たとえば、理系学生への作文教育を改革し、論文の論理を改善できれば、国際的な論文発表の場でも大きな効果を上げうると思われる。文章の訓練は少なくとも中学生から開始すべきだろう。開始する時期が遅ければ負担も増える。作文習慣がすでに定着した社会人が、作文訓練により成果を上げるには、一定の困難を克服する必要がある。

4 文書作成者、リライター、翻訳者

作文改善では、文書作成者自身が自分の文章を自分で改善することが理想的である。機械翻訳の前処理としての作文改善をだれがするか、という役割は、便宜的に、文書作成者、リライター、翻訳者に３分できる。まず文書作成者自身が自分の文章を確認して改善できれば、不要な費用は抑えられ、効果も高い。そのためには文書作成者の訓練が必要となる。だが特許文書、法律文書、医療関連文書などの専門知識が必要になる文章では、それらの専門知識のほうが重要視され、文章技能は軽視されがちである。

文書作成者自身による作文改善や訓練が困難な場合は、作成者以外のリライト専門家、つまりリライターがその文書を書き直すことになる。リライターは、表記や文書の改善のポイントを熟知しているため、費用が問題でなければ、作文改善としては確実な方法とは言える。しかし、工程が増えることにより、追加の費用と作業時間が発生する。作文改善が必要な状況でも、実際にはリライターによる書き直しができるのは限定的かもしれない。

これらの役割を用途別に考えてみよう。この機械翻訳を概訳用途で使用する場合は、機械翻訳の結果が悪ければ、機械翻訳ユーザー自身がリライトをする、ということになる。機械翻訳を翻訳支援で使用する場合は、文書作成者に修正を依頼することは困難なことが多く、リライターが入る予算的余裕もないので、翻訳者がリライトをせざるをえないこともある。ただ、翻訳者にとっては機械翻訳の精度を上げるために役立たなければリライトをする意味がない。このように考えれば、「前処理」という段階に至る前に、文書作成者自身が自分の文書を自分で改善することが理想的である。

5 スタイル ガイドの必要性

作文改善方法の一つとして、表記を記載したスタイル ガイドは、機械翻訳など文章処理の精度を上げるのはもちろんだが、文章の外観を統一することで読み手にとっても利点がある。英語では表記を統一するのは常識だが、日本語では表記が多様なこともあり表記統一は後回しにされがちである。

ユーザー辞書に基づいて翻訳処理をするルールベース機械翻訳では、特に用語の表記統一がされれば翻訳精度が高まる。言語用語の表記にばらつきがあると、それぞれの表記に訳語を登録する必要がある（前述のVenice/ ヴェネツィア、Venice/ ベニス……など）。特許文書ではスタイル ガイドを参照して作成することはまだ一般的ではないが、今後、表記に関する関心が高まることが望ましい。スタイル ガイドの例として、日本翻訳連盟（JTF）が 2011 年に作成し、公開した翻訳用スタイル ガイドがある。

詳　細：<http://www.jtf.jp/jp/style_guide/styleguide_top.html>

表記ルール以外にも、作文ルールとして、格助詞や連用節の扱い方など、文法に基づいて文章を読みやすくするルールもある。このような文法的作文ルールは、論理的ではあるが、実践の場では十分に活用するには注意を要する。文書作成者が作文するときは、自分の文書の本題に集中しており、読み手にとっての読みやすさまでなかなか意識が及ばない。さまざまな文法用語を考えながら文書作成者に書くように求めるには、文書作成者を徹底的に訓練する必要があり、文書作成者にも動機付けが必要になるだろう。

別のアプローチとしては、文法的作文ルールよりも単純化した、より実際的な基準を使うことができる。著者が提唱する「実務日本語」では、文法的な基準をあえて使わず、「1 文が 100 字を超えたら分割する」（百半ルール）という、実際的なルールを採用している（山本ゆうじ『IT 時代の実務日本語スタイルブック――書きやすく、読みやすい電子文書の作文技法』、2012 年、ベレ出版、p. 90）。

6 スタイル チェック ツールの必要性

文章改善には、スタイル ガイドに加えて、スタイル チェック ツールとそれを組み込んだワークフローが必要である。前述したようにスタイル ガイドは重要ではあるが、それさえあれば適切な文章が書けるというわけではない。スタイル ガイドを隅々まで読み込んで遵守する書き手は少数派であろう。また、スタイル ガイドの恩恵を直接得られるのは、読み手であり、書き手はそれを直接感じにくい。また、ルールというものは、実行する以上は厳密でなければ、無視されがちである。

文章の問題を確実に確認して修正するには、スタイル ガイドに加えて、スタイル チェック ツールを必ず使用する必要がある（前掲書、p. 234「チェック リスト（置換リスト）とチェック ツール」）。スタイル ガイドの中で、文法的な面ではツールによるチェックが困難な項目もあるが、言い換え表現など、機械的にチェックし置換できる項目もまた多い。

文書作成では、電子文書作成の基本ルールが守られていないことも多い。たとえば、全角英数字の禁止、空白文字を使ってレイアウトしない、丸数字のような機種依存文字を使わない、箇条書きは記号を入力するのではなく書式設定で行う、などである（前掲書、p. 80「【内離ルール】内容とレイアウトを分離する）)。これらは電子文書に特有の事項ではあるが、スタイル ガイドに含まれることもある。これらもまた機械的にチェックできる。

このような、機械的にチェックできる項目をわざわざ人間にチェックさせるのは不確実であるうえに、効率が悪い。機械的にチェックできる項目はスタイル ガイドで明確に区別し、人間は、人間にしかできないチェック項目に集中すべきである。たとえば、前述の JTF スタイル ガイドについては、現在、3 つのスタイル チェック ツールが以下で公開されている。

＜http://www.jtf.jp/jp/style_guide/stylechecktool.html ＞

7 用語集の必要性

用語レベルで日本語文章を改善するには、スタイルガイドやスタイル チェック ツール以外に、用語集が必要となる。用語集では、どのような用語が分かりやすく、適切かという点が重要である。

一般に、漢語の用語は意味が推測できることが多い。一方、一部のカタカナ語、頭字語は説明不足で読み手の理解を妨げる要因になる。たとえば、「ベスト プラクティス」というカタカナ語の意味はよく伝わらないことも多いが、代わりに「最善慣行」という用語を使えばだいたいの意味は推測できる。このように翻訳での「不適切な訳語」がそのまま不適切な用語となることも多い。読み手が該当分野の専門家でない限り、TS、WD、NP、TMX、TBX などのアルファベットが説明抜きでぞろぞろ出てきたら、だれしも面食らう。しかも分野によって意味が異なる可能性もある。だが、これらの頭字語を日常的に使っている人にとっては、読み手が理解できるかなど意識すらしていないかもしれない。このような頭字語は、言い換える必要があるかもしれない。

また、難しい言い回しや専門用語を乱用することが、「特許文書らしい」「論文らしい」などと誤解されていることも多い。読み手にしっかり伝わらない用語を乱用するのは、自らの作文能力の貧しさをさらけだすことである。より確実に理解できる用語を常に探す姿勢が求められる。

8 UTX 用語集と機械翻訳の改善

文書作成や翻訳で用語集を使う場合や、機械翻訳を活用する際は、用語集形式UTXを活用できる（下図参照）。

UTX（Universal Terminology eXchange） は、AAMT（アジア太平洋機械翻訳協会）が策定した、シンプルかつ汎用的で、オープンな用語集形式である（筆者は、UTX を策定するチーム リーダーを務める）。

UTX用語集の図解（AAMT/UTX用語集）

#UTX 1.20; en-US/ja-JP; 2013-04-01T19:00:00Z+09:00; copyright: AAMT (2013); license: CC-BY 3.0

#src	tgt	src:pos	term status
Asia-Pacific Association for Machine Translation	アジア太平洋機械翻訳協会	properNoun	approved
dictionary administrator	辞書管理者	noun	approved
contributor	用語提出者	noun	provisional
domain	分野	noun	
glossary	用語集	noun	
bidirectional	双方向	adjective	approved
merge	統合する	verb	approved
原語	訳語	品詞	用語ステータス

データとしての形式を整え、さまざまな環境で共有・再利用できるようにする

UTX の詳細：

＜ http://www.aamt.info/japanese/utx/＞

翻訳の現場では、複雑な用語集ではなく、今すぐ使えるシンプルな用語集が必要とされている。UTX は、その要望に応え、ルールベース機械翻訳の基礎となる用語集データ（ユーザー辞書）となるとともに、機械翻訳とは別個の、人間翻訳でも扱いやすい用語集としても優れた形式である。UTX は、特許庁の機械翻訳調査での辞書作成に使われているほか、企業での翻訳工程改善に活用されている。

UTX には、４つの用語ステータス（暫定、承認、非標準、禁止）による用語管理の機能があり、複数の用語のうち、どれが正規の用語で、どれが許容の用語を区別できる。この情報に基づいて、用語のばらつきをなくすこともできる（下図参照）。

UTX は、本来は、翻訳用の対訳用語集形式だが、「日本語のみ」など単一言語用語集としても作成できる。

翻訳ソフトの辞書編集機能には用語管理の観点がなく、用語の整理がしづらいことがある。UTX は、分野を整理して必要な辞書のみを適切に組み合わせることで、最大限の効果を発揮する。各種の分野が混在した巨大な辞書では、UTX の本来の効果は得られない。

UTX などの適切な用語データに基づくルールベース機械翻訳では、対訳翻訳エディターを使えば、翻訳者の意図通りに翻訳を進めることができ、効率的な翻訳支援となる。用語データを管理する時点で、適切な用語適用がほぼ保証されるため、別途に用語をチェックする必要は少ない。

これに対して、統計機械翻訳では、どれだけ精度が上がっても「最後の一歩」が必ず不足する。統計である以上、文レベルで人間訳に「非常に近い」結果は得られることはあっても、なにかが違う。そしてそのなにかは、予測不能である。用語レベルでも用語適用の保証がされないため、用語チェックは、統計機械翻訳とは独立した過程として別途、行う必要がある。

統計機械翻訳を翻訳支援に使うことは、見方によれば、ソフトが統計処理可能な部分のみを処理してしまい、処理不可能だった点を人間翻訳者に押し付けるしくみということもできる。それで人間翻訳の単価が引き下げられては、翻訳者にとってはたまったものではない。統計機械翻訳は、前述の通り「類似言語間の概訳」には有効でも、日本語英語間のような異種言語のプロ翻訳者にとっては翻訳支援をするどころか、厄介ものでしかない。少なくとも日英・英日の場合、現状では、翻訳者が、主体的に活用でき、また活用すべきなのは、統計機械翻訳よりもルールベース機械翻訳である。将来的には、統計機械翻訳でも UTX による用語適用が必要になってくるだろう。

4つの用語ステータス

「使うべきか否か」の用語管理をする

日本語	English	意味
暫定	provisional	とりあえず使える訳語
承認	approved	必ず使う訳語
非標準	non-standard	本来は使うべきでないが 処理できるようにした訳語
禁止	forbidden	使ってはいけない訳語

9 必要なのは高品質のシンプルな対訳用語集

ルールベース機械翻訳のユーザー辞書に必要なのは、原語、訳語、品詞などの基本的な対訳情報のみであり、名詞の細かな概念属性などの付加情報は必要ない。名詞が固有名詞か否か程度は役立つが、それ以上の細かい特性の記述は、管理上の負担を増加させる割には、ルールベース機械翻訳の翻訳精度を向上させない。システム辞書としては有用な情報であっても、「翻訳者の立場」で作るユーザー辞書では必要ないのである。ここに、機械翻訳開発者の発想と翻訳者の発想の大きなギャップがある。開発者がこのことを理解するには、自分で翻訳ソフトを使って翻訳作業をしてみる必要がある。そうすればはじめて、本当に必要なのは高品質だがシンプルな用語集ということが実感できる。

シンプルな UTX 用語集は、特許文書などに必ず添付する用語データ形式として収集できれば、特許機械翻訳の精度を大きく向上できる。用語集をゼロから作るのは大変な作業である。だが、断片的な「ミニ用語集」であっても、UTX のような統一された形式で集めることにより、大きな力にすることができる。

UTX については以下の「よくある質問と回答」も参照されたい。

<http://www.aamt.info/japanese/utx/faq.htm>

10 UTX 変換ツールで開ける世界

AAMT の UTX チームでは、現在、初心者ユーザーにも使いやすい UTX 変換ツールを開発中である（下図参照。ソースコードを含め、無償で公開予定）。このツールを使えば、翻訳ソフト各社の独自形式や、用語ツール用の形式（MultiTerm 用語ベースなど）と UTX を相互に変換できる。

UTX 用語集から、用語ステータスが「禁止」となっている用語を抽出すれば、「使ってはいけない用語」のリストを作成することもできる。さらに、それらの禁止語と、用語ステータスが「承認」の語をペアにした置換リストを作ることもできる。この置換リストを前述のスタイル チェック ツールなどで使えば、手作業でスタイル ガイドを参照するよりも確実かつ効率的に用語を修正できる。たとえば、「褥瘡」などの難解な専門用語を、より分かりやすい「床ずれ」などの用語に置換できる。また、不必要に回りくどい表記を、簡潔な表記に置換することもできる。

UTX はシンプルではあるが、シンプルであればこその汎用性と可能性を秘めている。UTX 変換ツールが完成すれば、UTX は、前処理の手順を単純化するとともに、機械翻訳の用語データとして、また実用的な用語集形式として、活躍の場を大きく広げることだろう。

UTX 変換ツール完成予想図