

# 翻訳品質と JTF 翻訳品質評価ガイドライン

## — 生産ベース評価の品質の考え方 —

西野 竜太郎

(グローバルイノベーションデザイン研究所)

山田 優

(関西大学)

### 1 はじめに

機械翻訳の翻訳(訳文)の品質評価では BLEU などの自動評価が頻繁に使われている。これらの評価方法では、あらかじめ「正」とされる参照訳を作成しておいて、その参照訳にどれだけ機械翻訳の訳文が近づけるのかを測定するものである。

他方で、実務翻訳を含む翻訳業界では、当然のことながら、毎回、新しい原文を翻訳しなければならず、あらかじめ参照訳が用意されていることはないため、自動評価の方法を用いることは稀である。

しかしながら、これまでの翻訳業界においても、翻訳の「品質」という言葉の内容やその定義ははっきりせず、翻訳のユーザーや評価者によって何を指しているのか異なるケースが多々あった。そこで業界団体である日本翻訳連盟(JTF)は、翻訳業界の関係者どうして品質に関する共通認識を持てるようにすることを目的として、2018年11月に「JTF 翻訳品質評価ガイドライン」[5]を策定し公開した(以下、JTF ガイドラインと呼ぶ)。JTF ガイドラインでは、品質に関する言葉の定義に加え、実際に人手評価作業を実施するための「JTF 翻訳品質評価モデル」も提供している。

後述するように、近年、翻訳品質の評価は、翻訳の発注者によって決定される翻訳の目的(あるいは、それが明示的に決定されない場合であっても、そのテキストが属するジャンルやテキストタイプに規範的に内在する目的)を考慮して行う考え方が広がってきた。具体的には、

翻訳の「仕様」を作成し、それに従って翻訳を行い、出来上がった翻訳物は仕様通りであるかを確認する方法である。JTF ガイドラインもこの方法を採用している。

そこで本稿では、まず翻訳の「品質」の考え方について概観する。続いて特に近年重要さを増してきた生産ベース評価の詳細と、そこで用いられる「仕様」について触れる。最後に JTF ガイドラインの品質評価手法を説明し、今後の機械翻訳開発と翻訳産業における品質評価に関する提言を行う。

なお筆者の一人は JTF 翻訳品質評価ガイドライン検討会の委員長として制作に携わった。

### 2 翻訳の「品質」とは

#### 2-1. Garvin に基づく 5 分類

これまで翻訳業界で用いられてきた「品質」という言葉の内容はまちまちであったと述べたが、それを整理しようとする動きが出てきた。2014年に Fieldsら[2]は経営学者 Garvin の論文[3]に基づき、品質の評価方法を次の5つに分類した。

- A. 超越的: 専門家による主観評価
- B. ユーザーベース: 最終読者による主観評価
- C. プロダクトベース: 客観指標による評価
- D. 価値ベース: 費用を考慮した相対的価値
- E. 生産ベース: 合意した仕様を満たす程度

これら 5 分類の関係を明示したものを図1に示す[4]。

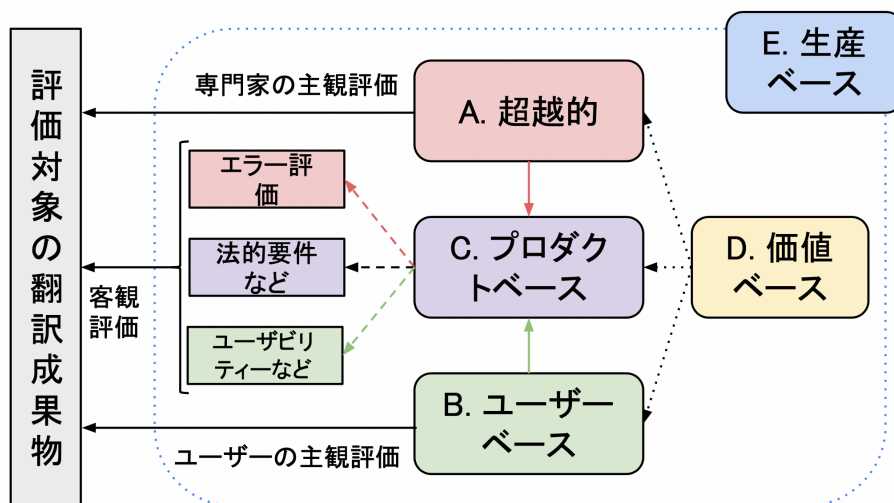


図 1. 品質 5 分類の関係図

5つのうち、直接的に翻訳成果物を評価できるのはA～Cである。ただしAの超越的、Bのユーザーベースのいずれも主観的評価であり、評価項目や基準が定められていることは少ない。これに対して、Cのプロダクトベースは、翻訳成果物をできるだけ客観的に評価するために、エラー評価のための分類表や、法的要件に準拠しているかを確認できる項目などが、あらかじめ決められていることが多い。

しかし、主観評価が、主観的であるからという理由だけで客観評価に劣るということは意味しない。専門家による主観評価は、その翻訳物の重要で本質的な品質を見抜いていることもあるだろうし、ユーザーによる主観評価も、翻訳を使う者の立場からは大切であることもある。決められた項目についてしか評価をしないプロダクトベース評価は、確認項目があらかじめ十分に用意されていない場合、たとえ客観的であっても、正しく品質を評価できないこともありうるからだ。別の言い方をすれば、超越的評価やユーザーベース評価でなされている内容は、既存のプロダクトベースの評価項目にはない側面を評価している可能性がある。たとえば、翻訳物が「かたい」というのは、AやBに該当する主観評価である。た

だしそのような主観評価であっても、場合によってはCの客観評価に変換することもできる。訳文が「かたい」と主観評価される場合、その実情を突き詰めて「文章中の漢字の含有率が一定よりも多い」ことがわかったとすれば、それを計量可能なルールに落とし込むことにより、以降の評価はその数字で客観評価できる。

Dの価値ベース評価は、直接的に翻訳成果物を評価するのではなく、その使用や作業コストなどの指標に置き換えて評価を相対化する。例えば、翻訳物そのものが品質の高いものであったとしても、翻訳時間が多くかかれば、その評価が下がるといった具合である。

Eの生産ベース評価は、A～Dすべてを包含した「仕様」で評価するため、上位レベルの項目と見做すことができる。以下で詳述するが、例えば、翻訳プロジェクトで「エラー評価」(プロダクトベースの一種)を採用するのか、採用するならどのようなエラー・カテゴリーを使うのか、合否しきい値は何点にするのか、といった点を仕様としてまとめておき、最終的な評価にも用いる。つまり、あらかじめ準備した「仕様」に適合しているのかを評価できるように、評価対象項目を仕様に折り込んでおく考え方である。

このように同じ「品質」という言葉を使っているにもかかわらず、直接的に成果物を評価しなかったり、別のレベルを指したりするため、これまでの品質の議論は混乱しがちであったが、その全体を大まかに整理できた。

## 2-2. 仕様と品質

上述したように、E の生産ベース評価では、仕様 (specification) を用いて翻訳の品質を評価する。仕様とは、翻訳成果物が持つべき要件について関係者 (典型的には発注者と受注者) 間で合意形成をしてまとめた内容である。この仕様に基づいて評価するという考え方は、比較的近年翻訳業界に広まってきた。たとえば国際標準の ASTM F2575 (2006 年 / 2014 年) や MQM (2015 年) では仕様が強く意識されているし、最近の ISO 17100 (2015 年) でも用いられている。JTF ガイドラインでも仕様を重視し、「品質」という言葉も「**翻訳成果物が、関係者間で事前に合意した仕様を満たす程度のこと**」と定義している。つまり、事前合意した仕様がなければ品質は評価できない。

仕様にはさまざまな要素が盛り込まれる。JTF ガイドラインでは例として言語方向 (例: 英日)、翻訳の目的 (例: 製品を購入してもらう)、価格、納期、さらには評価方法 (例: エラー評価) などを挙げている。評価部分に注目すると、翻訳開始前 (入口) に仕様で評価方法について合意し、翻訳後 (出口) にそれに基づいて評価を実施する流れとなる。

注意したいのは、仕様によっては一般的な直感と反する評価がなされる点である。たとえばいわゆる直訳は低品質とされることが多い。しかし直訳調の「原文が透けて見える訳」を欲する発注者もいる。その場合、直訳調であった方が仕様を満たす程度は高くなるため、最終的には高品質とみなされる。仕様は関係者間の合意で決まるので、何が高品質の翻訳かは案件ごとに変わりうる。

なお仕様の考え方は、翻訳学における機能主義に近い関係にある。機能主義では「翻訳の目的」(翻訳が何に使われるのか) など、テキスト外部の要素を重視する。翻訳において学术界と産業界との接点はあまり多くないが、機能主義は実務での応用可能性の高い接点の 1 つとされる [1]。

## 3 JTF ガイドラインの品質評価

### 3-1. 生産ベースとプロダクトベースの組み合わせ

JTF ガイドラインでは仕様を品質評価の前提にしている。また前述のように JTF 翻訳品質評価モデル (以下、JTF モデル) も掲載している。JTF モデルはエラーベース手法 (エラー評価) である。エラーベース手法は一般的に、訳文内でエラーをカテゴリ別に見つけて重大度に応じて点数 (ペナルティ) を付け、合計が所定の点数を上回ったら不合格といった判断を下す。図 1 で言うと C のプロダクトベースの一種に該当する。

つまり、JTF ガイドラインは E の生産ベースを基本とし、具体的な評価手法として C のプロダクトベース (のエラー評価) を備えていることになる。

### 3-2. エラー・カテゴリと重大度

この JTF モデルでは、上位レベルのエラー・カテゴリに「正確さ」、「流暢さ」、「用語」、「スタイル」、「地域慣習」、「デザイン」、「事実性」を設けている。このうち「正確さ」や「流暢さ」は機械翻訳の人手評価でもよく用いられる。最後の「事実性」は「訳文の情報に、事実や現実世界との齟齬があるかどうかに関する問題」を扱うカテゴリとなる。たとえば「発売中」と英日翻訳されていたとする。原文が書かれた米国で販売されていても、もし訳文が読まれる日本で販売されていなければ、事実性のエラーとなる。

JTF モデルでエラーの重大度は、翻訳の目的 (例: マニュアルなら操作を完了できるか) を達成できるかとい

う視点から判断する。言い換えれば、テキスト上(字面上)の整合性よりも、あるエラーが現実に影響するかどうかを見る。たとえば数字を1桁間違えて翻訳したとする。これが医療関係の文書で発生したなら、人命に関わりうるので重いエラーとなる。しかし誰かの日記ブログに書かれたテレビゲームの点数であれば、現実への影響は小さいので軽いエラーと判断してもよい。字面上では同じ1桁の間違いであるが、評価には現実世界を参照する必要がある。

#### 4 課題と提言

JTF ガイドラインでは、品質とは「仕様を満たす程度」とであると定義し、具体的な評価手法としてJTFモデルを提供している。JTFモデルには正確さや流暢さといったエラー・カテゴリーに加え、「事実性」という項目がある。さらにエラーの重大度は、現実への影響によって判断する。

こういった品質の評価方法からから見ると「翻訳」が備えるべき要件が見えてくる。まず、翻訳とは単なる無条件のテキスト変換ではない。変換(翻訳)される前段階で定められた「仕様」に基づいて、それに記述された要件を具体化すべく、必要に応じてテキストの外部にある項目をも参照してテキスト変換が行われ、最終的な目的言語のテキスト上に「仕様」の要件が反映されていなければならない。留意すべきは、その仕様は関係者間の合意で決まるため、何が良い翻訳かは案件ごとにより変わるという点である。

このように考えるならば、今の機械翻訳は、仕組み上、当然のことながら、仕様を考慮して訳出が行われているわけではなく、また機械翻訳システムの評価においても、こういった側面は全く考慮されていない。

また機械翻訳を産業翻訳で活用するポストエディットのような手法は、その案件の仕様と機械翻訳の出力とのギャップを埋める作業であるとも考えられるわけであるが、

必ずしも、事前に仕様が用意されているとも限らず、ポストエディットを行う作業への負担も増大していると考えられる。

今後、機械翻訳開発や、産業翻訳での機械翻訳の活用を検討する場合であっても、本稿でのべた品質の考え方を考慮した研究や取り組みが行われていくことを期待したい。

#### 5 謝辞

本研究の一部は、日本学術振興会科研費補助金基盤(S)「翻訳規範とコンピテンスの可操作化を通じた翻訳プロセス・モデルと統合環境の構築」(研究課題番号:19H05660)の支援を受けて行われました。

#### 6 参考文献

- [1] J. Drugan, *Quality in Professional Translation: Assessment and Improvement*. Bloombury Publishing, 2013.
- [2] P. Fields, D. Hague, G. S. Koby, and A. Melby, “What Is Quality? A Management Discipline and the Translation Industry Get Acquainted,” *Rev. Tradumática*, no. 12, pp. 404-412, 2014.
- [3] D. A. Garvin, “What Does ‘Product Quality’ Really Mean?,” *Sloan Management Review*, vol. 26, no. 1, pp. 25-43, 1984.
- [4] 西野竜太郎, “翻訳品質のランチボックス#2: 翻訳の「品質」とは(2),” *日本翻訳ジャーナル*, no. 284, 日本翻訳連盟, pp. 22-23, 2016.
- [5] 日本翻訳連盟, “JTF 翻訳品質評価ガイドライン,” 2018. [Online]. Available: [https://www.jtf.jp/tq/pdf/jtf\\_translation\\_quality\\_guidelines\\_v1.pdf](https://www.jtf.jp/tq/pdf/jtf_translation_quality_guidelines_v1.pdf). [Accessed: 2019-12-15]