レポート

機械翻訳＋ポストエディットの実証研究（その２）：英日翻訳での実験結果

立教大学大学院 異文化コミュニケーション研究科 博士後期課程
山田　優

前回のAAMTジャーナル(No. 48)では、機械翻訳のポストエディット(以下MT+PE)を翻訳支援ツールとして活用した場合の、翻訳品質と作業効率に関する先行研究を概観した。これらの文献で検証対象とされていた欧州言語の組合せでは、翻訳の分野や目的を限定すれば、MT+PEは実務レベルの品質／効率にほぼ達していることが実証データから分かった。

Bowker & Ehgoetz (2007)が示したように、ユーザーが求める品質相応の納期とコストを評価条件に加えることにより、MT+PEによって生み出されたプロダクト（訳出物）が商品（実務）としての価値を持つようになる。すなわち、MT+PEは、例えば、出版翻訳には向かないとしても、日常業務に支障ないビジネス翻訳としては十分耐えうる品質とスピードを兼ね備えているということである。また、O'Brien (2006b)の実験により、MT+PEの作業が時間だけでなく、実際の作業者の認知負荷のレベルにおいて、翻訳メモリ(TM)85%マッチの分節を処理する負荷とほぼ同等相当だということも示された。

一方で、MT+PEのポジティブな検証結果はいずれも西欧言語の組合せでのみ実証されているように見受けられ、言語間の「距離」のある組合せ、すなわち英語と日本語のように統語や語彙の非似なるペアでも同様の結果が得られるかどうかは不明である。ということで、今回のレポートでは、英語→日本語でのMT+PEの検証を行う。筆者が行った一連の実験のうち、断片的にではあるが、MT+PEの作業負荷、効率性、品質に関する結果を本紙面で報告したい(注1)。

実験の目的と評価方法

本実験の主な目的は、英日翻訳のMT+PE作業におけるMT訳の受容レベルを検証することである。平たく言えば、MT+PE は実務に耐えうるのか、効率は上がるのか、品質はどうなのか、といった点を検証するのが目的だ。以下の3つの観点から検証を行った。

まず、翻訳者にとってMT+PEという作業は、普通に翻訳をするよりも楽なのか、という点である（ALPAC, 1966を参照)。被験者に作業後のアンケートを実施し、MT+PEの体感作業負荷を数値化してもらった。普通に翻訳する場合の体感工数を100(基準)として、MT+PEはどうだったか、例えば、MT+PEによって通常よりも２割程度の工数軽減を感じたならば80と記入(100マイナス20)する、逆に２割くらい増えたと感じた場合は120(100プラス20)という具合に回答してもらった。

主観的データなので信頼性に乏しく、またAPLAC報告書（ibid.)でも指摘されたように、翻訳者は概してMT+PEの作業を過大評価し過ぎるという危険性があるものの、その一方で、結局のところ、翻訳者自身がMT+PEをどう感じているのか、それが実務では一番重要な要素だとも言えるだろう。

２つ目の評価点は、修正量である。修正量（テキスト類似度）とは、PEを行う前のそのままのMT訳と、PE後の最終訳出物とが、どの程度似ているかということである。MT訳の品質が高ければ、PE量（すなわち修正量）が少なくなるだろう、という考えに基づいている。

修正量の測定には、MTの自動評価法のひとつであるGTM(General Text Matcher)を用いた。BLEUやWERなど広く使われているツールに比べて、比較的マイナーなGTMを選択した理由は、GTMではセンテンス単位での評価が可能だという点と、またTatsumi (2009)のPEの処理時間と自動評価点との相関性を調査した先行研究において、GTMの点数が最も高い相関を示していたことによる。

3つ目の評価ポイントは品質だ。効率UPや作業負荷の軽減が達成されても、最終品質が下がっては仕方ない。MT+PEの実践導入を検討する翻訳会社等にとっては、最も関心ある事柄であろう。ということで、品質

面も限定的ではあるが評価対象とした。

実験デザイン

実験では、2種類の原文（英語）を用意した。それぞれ約40ワード、内容は工業機械のマニュアル文書の抜粋を使用した。被験者には、両方のテキストをポストエディットしてもらった。用意した2つのテキストは、内容こそ似ているが、構文の複雑性などが異なっている。ひとつのテキストは、いわゆる制限言語(CL=controlled language)で書かれた原文で、ASD-STE 100 (Ishikawa, 2005)に準拠している（以降、こちらのテキストを「CLテキスト」と呼ぶ）。これに対して、もうひとつの原文は、通常のライティングによる英文テキストである（こちらを「Non-CLテキスト」と呼ぶ）。これら2つのテキストを、それぞれ機械翻訳にかけてから、PEを行った。

2種類のテキストを比較する実験デザインはO’Brien (2006a)と似ているが、ここでは厳密に制限言語を用いたMT+PEの作業効率を評価するのが目的ではなく、実務翻訳で言うところの「ローカリ分野」の対象となるマニュアル・テキストと、そうでない原文との違いをみるために用意したに過ぎない。つまり、CLテキストは「マニュアル文書」、Non-CLテキストは「マニュアルではない文書」といった状況を想定した。

それぞれのテキストが40ワードと少ない点については、実験自体がパイロット・レベルであり、本結果に基づいて追試等を別途行っている点は否めないが、近年の翻訳実務においてTMS（翻訳マネージメントシステム）等の導入により、プロジェクトが分割化され、翻訳者に一度に割り当てられるワードカウントが数十から数百というという状況も珍しくない現状があり、このような点を考慮すれば、用意したテキストのワードカウントは、翻訳現場の現状を反映しているともいえよう。

被験者は大学院で通翻訳を学ぶ学生（セミプロ）8名に依頼した。翻訳スキルや分野の背景知識では、プロのレベルに達していない点は、本実験の限界点として挙げておく。しかしながら先行研究でも、翻訳学科の修士課程の大学院生を「プロ」と看做すケースも多々ある（Lörscher, 1996等参照）。いずれにしても、本実験結果の解釈には、筆者がプロに対して行った別実験の結果とも照合しながら補足的に考察するように心がけた。

実験作業には、Google Translator Toolkit(以下GTT)を使って作業をした(Garcia, 2010参照)。上記2種類のテキストを同サイトにアップロードし、GTTのインターフェース上でPE作業を行った。これにより、実験で使用したMTエンジンは同社のSMT（統計的機械翻訳）ということになる。最近のMTの動向を考慮した。

GTTは割合と使い易いツールではあるが、実験参加者にツールに慣れてもらうために、実験前に練習用テキストを用意して、1時間程度のトレーニング時間を設けた。

作業は、教室環境で全員一斉に行った。最初にNon-CLテキスト、続いてCLテキストの順でPEを行い、作業後にアンケートに回答してもらった。

仮説（予測される結果）

検証ポイントとなる体感工数、修正量、品質の各項目について予測される結果を以下に述べておく。

仮説1：先行研究の実験では、MT+PEにより、実務に耐えうる効率や品質が得られることが観察されたものの、英語→日本語での使用に関しては、筆者自身はまだ懐疑的である。翻訳者が、普通の翻訳よりもMT+PEの作業の方が容易だと感じるかどうか、正直分からない。ということで、英日のPEでは、通常の翻訳と比較した場合の作業負荷の軽減はごく僅かだろうと予測する。具体的な数字としては、Krings (2001)が時間的に20%程度の短縮を達成できたという結果を受けて、英日でも10～20パーセント程度の負荷軽減であれば現実的な数字だろうと仮定する。

仮説2：Non-CLとCLテキストの比較では、おそらくCLテキストのPE作業のほうが楽に感じられるだろう。O’Brien (2006a)の結果でも、CLテキストの方が工数的には軽減されていた。ということで、CLテキストのMT+PEの主観評価点の方がNon-CLテキストよりも低くなる（すなわち、作業が楽になる）と予測される。

仮説3：上記の仮説2の結果に準じて、実際の修正量にも違いが見られると思われる。CLテキストのPE作業の方が楽に感じるのであれば、実際の修正量も減っていると予想される。ということで仮説3は、CLテキストのMT+PEの修正量は、Non-CLテキストよりも減少するとしておく。

仮説4：翻訳者は、出力されたMTの品質にかかわらず、各自の最終訳出物に対しては一定の責任を負うはずである。CLとNon-CLのMT訳の品質に差があったと

しても、PE後の両テキストの最終品質に差があるとは考えられない。むしろその差は、上記で予想したように、体感工数や修正量の差として現れるはずである。仮説4は、CLおよびNon-CLテキストの最終訳出物の品質の差はない、とする。

結果

それでは実験結果を、上記の仮説順に見ていくことにしよう。まず初めは、被験者の主観的工数負荷の評価点である。

主観的工数負荷の軽減

通常の翻訳(100)を基準としたNon-CLテキストのMT+PE作業の評価は、平均86.9であった（表1を参照）。単純比較では、普通の翻訳に比べ約13%程度の感覚的負荷の低減がもたらされたと言える。ただし標準偏差(17.5)を考慮すると、必ずしも優位な差だとは言い切れない。それでも、8名中6名の被験者がMT+PEは普通の翻訳よりも楽だと評価しており、多少なりとも工数軽減を感じているようだ。残りの2名(翻訳者AとC)は、普通の翻訳と変わらない(Cのポイントは100)、または普通よりも大変であった(Aは110)と回答している。Non-CLテキストの方の結果なので、筆者自身はもう少し悪い結果を覚悟していたが、MT+PEという作業そのものが邪魔になってしまうほどではないようである。

次にCLテキストの結果を記す。平均点は73.1と、Non-CLテキストよりも下がっている。Non-CLとの差は13.8ポイントで、統計的にも有意差がある(t=2.986)。被験者全員が、普通の翻訳よりも作業が楽になったと回答していることも含めて考察すると、CLテキストの方はNon-CLの場合よりも事実として作業が楽になっており、普通の翻訳との比較でも明らかに工数軽減を体感していると言える。

まとめると、MT+PEでの作業は、CLテキストのように、原文の構文構造をシンプルにしておけば、通常の翻訳作業よりも10～25%程度の作業負荷低減が実感できるようである。よって仮説1は証明されたといえよう。またNon-CLとCLテキストの比較でも、作業負荷の差を体感出来たので、仮説2も同様に支持されたことになる。

表１：体感工数低減結果

翻訳者	Non-CL テキスト	CL テキスト	Non-CL と CL の差
A	110	80	-30
B	80	80	0
C	100	70	-30
D	85	85	0
E	90	80	-10
F	50	30	-20
G	90	70	-20
H	90	90	0
平均	86.88	73.13	-13.75
標準偏差	17.51	18.70	
t			-2.986
有意差 = 5%（t=2.365）			

修正量（テキストの類似度）

では、実際の修正を見る。繰り返しになるが、修正量（テキスト類似度）とは、PE前のMT訳と、PE後の最終訳との差分をGTMを使って測量したものだ。1.0が最高点で、それに近ければ近いほど類似度が高い、すなわち修正量が少ないといえる。

Non-CLとCLテキストの平均点数は、0.217 vs. 0.483であった。CLテキストの点数が、Non-CLの約2倍であった。これらの点数と、先の工数負荷の結果とを合わせて解釈すると、Non-CLとCLテキストとの体感工数の差は、GTMで0.266点(0.483-0.217)の差がもたらしたとも換言できる。GTMスコアと工数負荷軽減との間に相関関係があるとまでは断言しないが、工数負荷の差と同様に、GTMスコアにも有意差を観察できた。

表２：GTM スコア（PE 修正量）

翻訳者	Non-CL テキスト	CL テキスト	Non-CL と CL の差
A	0.290	0.515	-0.225
B	0.197	0.493	-0.296
C	0.182	0.508	-0.326
D	0.269	0.492	-0.223
E	0.271	0.594	-0.323
F	0.179	0.500	-0.321
G	0.189	0.441	-0.252
H	0.156	0.317	-0.161
平均	0.217	0.483	-0.266
標準偏差	0.052	0.079	
t			-12.450
有意差 = 5%（t=2.365）			

では、ここでGTM点数自体の意味について考えてみたいと思う。補足データも交えながら、GTMスコアについて解釈を加えてみる。

まず、GTMの詳細アルゴリズムは無視して、便宜的に、CLテキストの0.483という点数を、そのままのMT

訳の完成度が訳48%であったと読み替えてみると、翻訳者はPE作業中に100%に足りない52%分のテキストを修正したことになる。同様にNon-CLのケースでは、1.000-0.217 = 0.783、つまり約78%分の書き直したと考えることができる。Kring (2001)の実験では、修正量の換算方法が異なるものの、大雑把に言えば、PEでの修正量が60%で、その作業時間は普通の翻訳に比べて20%減少したと報告されている。これを基にすると、本実験のCLテキストのPEでは、52%が修正され、体感工数が10～20%軽減したというのは妥当な数字と言えるかもしれない。他方、Non-CLの場合も、78%の修正量で普通の翻訳と比較した場合の工数低減に有意差がなかったというのも、それはそれで信憑性のある結果であった。

また、筆者はこの実験とは別に、翻訳メモリ(TM)に関する検証を行い、そこでTMのマッチ率とGTMスコア（TMで表示されるFuzzyマッチ訳と最終的な訳文とのテキスト差）、それに各マッチ率の処理時間等を測定した。一般的に、TMを使った翻訳の場合は70%マッチ以下になると、TM上に既存のFuzzyマッチ訳を表示させるメリットがないと言われているが、その理由として、Fuzzyマッチを見比べながら作業するよりも、最初から翻訳し直したほうが速いから、というものであろう。しかし、筆者の実験では、50～60%マッチくらいまでは、Noマッチよりも翻訳速度的に有意差があった。逆に言うと、TM50～60%マッチまでは、Fuzzyマッチ訳をベースに作業をしたほうが効率性は上がるということなのだ。

この50～60%マッチ付近を、効率性の分岐点と考えて、そのFuzzyマッチと最終訳出物との類似度をGTMで調べると0.460であった。この数字と本実験結果を比較してみると面白い。CLテキストはGTM0.483であり、若干ではあるがTMのFuzzyマッチの分岐点0.460を上回っていた。つまりCLテキストのMT+PE作業は、TMの50～60%Fuzzyマッチよりも少し高いマッチ率の作業と同等だったと考えることができ、速度的にも普通の翻訳よりも速かったと想像できるのである。これに対してNon-CLテキストの方は、GTM0.217であり、これは上記の分岐点よりも低いために、作業負荷的にも翻訳速度的にも普通に翻訳した時よりもメリットがほとんどなかったと解釈できるのだ。

品質

品質については、「正確性(accuracy)」のみを対象とし、筆者自身が評価を行った。結果は、作業工数および修正量の少なかったCLテキストの方には、特に目立ったエラーは見つからなかったが、Non-CLテキストには、1箇所致命的な誤訳が見つかった。驚くことに、8名中半分の4名が同じミスをしていた。原文のワード数が40と少ないにもかかわらず、5割の確率で同じミスが発生したのは、やや驚きの結果であった。以下に詳述する。

原文:
Gain access to blade.
After removing old blade, new blade may be fitted by proceeding in reverse order, using gloves to avoid injuries by teeth of blade.

MT訳:
ゲインのアクセスはブレード。
古い刃を削除した後、新しいブレードを装着逆の順序で進めて、ブレードの歯でけがを避けるために手袋を使用することがあります。

最終訳　(被験者Bの例):
刃を取り替えます。
古い刃を取り除いた後、刃先でけがをしないように手袋を使用しながら、新しい刃を装着時とは逆の手順で装着してください。

原文（英語）は、new blade may be fitted by proceeding in reverse order、つまり「取外時とは逆の順序で新しい刃を装着します」という意味である。しかし、MT訳が「装着逆の順序で進めて」となっているため、これに引きずられるようにPE後の訳が「装着時とは逆の手順で装着してください」となってしまっている。「装着時と逆の順序で装着する」のは理論的に不可能であり、取扱説明書の翻訳としては致命的な誤訳だ。なぜこのようなエラーが半数の被験者で発生したのだろうか。

MT+PE作業のように、提示された訳を修正するような作業の場合、提示訳（この場合はMT訳）にエラーが含まれていると、翻訳者（作業者）はそのエラーに気づかないことが多々ある。翻訳メモリの作業でも、TM内にミスがあれば、それが新規訳に残ってしまう、いわゆる「エラーの伝播(error propagation)」という現象が起きる。上記の誤訳も、伝播の一種と考えられるだろう。

Bowker (2005)やRibas (2007)では、TM作業でのエラー伝播を検証しており、100%マッチで残ってしまうのならともかく、マッチ率が低い場合でもエラーが伝播する点を指摘している。本実験をTMのマッチ率で考えた場合、50〜60%マッチ相当の作業であったと先述したが、結果的にエラーは修正されていないなかったことになる。

Guerberof (2008)は、MTとMTの内在エラー修正率を比較した実験を行い、TM作業の方がMT+PEよりもエラー伝播率が高いことを示した。その理由として、TMの場合は、人間の翻訳者による訳文がベースとなっているため、訳文が「自然な訳」であり、翻訳者は細かなミスに気付きづらい点を挙げている。また、Fiederer & O'Brien (2009)は、品質の3要素のうち、翻訳者は、styleの要素をaccuracyやclarityよりも重要視してしまう傾向を挙げている。つまり、本実験でPE作業をした翻訳者は、品質のstyleの部分、すなわち訳文がマニュアルとして正しい文体になっているかという側面に気を取られたために、accuracyの要素を見過ごしてしまったと考えることができるのだ。

更に、今回使用したMTエンジンがSMTであったということからも、従来のルールベース訳と比較して「自然な訳」であった、という可能性も見過ごせない。今後、SMTを用いたPEを行う場合には、考慮しなければならない品質的なデメリットがはっきりとした形で露呈された結果となった。

品質と作業負荷評価の関係

最後に、観察レベルでの分析でしかないのだが、品質に関する興味深い結果が見られたので記載しておく。エラーに気付いた翻訳者とそうでない翻訳者との違いを、彼らの作業負荷評価の結果から考察してみた。

ミスをした4名の翻訳者（A、B、C、D）のうち、AとBは、MT+PE作業の評価を、普通の翻訳と変わらないか、むしろ大変だったとしていた。彼らのNon-CLの点数は、それぞれ110と100であった。これに対して翻訳者BとDは、MT+PE作業を過大評価しすぎている感があった。彼らの点数は80と85だった。翻訳者Fが50という評価をしているのを例外とすれば、残りの翻訳者（ミスをしなかった翻訳者）は90という妥当な評価をしていた。

つまり、エラーをした翻訳者のMT+PEに対する評価は両極端であり、作業が楽ではなかったというように悲観的な見方をしている者か、または非常に楽であったと楽観的な評価をした者だということが分かる。MT+PEに対する苦手意識が強くても、得意気すぎた態度でも、品質的には悪くなる可能性があるので、より冷静な評価が必要であるということが示されたようである。今後、MT+PEの実践へ導入を考えている翻訳会社等にとって、翻訳者（ポストエディター）の選定の際に考慮したい結果であろう。

まとめ

以上、実験デザインとして不完全な部分は残るものの、限定的な状況における英日翻訳のMT+PEの受容性についての実験結果を述べてきた。被験者のサンプル数、テキストのワード数、原文テキストの分野等に関しては、今後の研究での精緻化を目指したい。

ということで、最後に現時点における英日のMT+PEの結果をもう一度まとめておく。CLテキストの結果で見たように、一般的なマニュアルのようなシンプルな文体で書かれた文章のMT+PEであれば、翻訳者（作業者）は、普通の翻訳と比較して10〜20%の作業負荷軽減を体感できるようである。その根拠として、PEでの実際のテキスト修正量が、翻訳メモリでの50〜60%Fuzzyマッチを若干上回るレベル（GTM0.460以上）に達していることが確認できた。しかし一方で、SMTエンジンを使うとMT訳が「自然」になるために、それに翻訳者がとらわれてしまい、accuracy等の基本的な誤訳に気づきづらくなる可能性が増すようだ。MT+PEの実践への応用には、以上のような点を考慮して検討していきたい。また研究面でも、より確実なデータを提供していけるように努めたいものである。

【註】

（注1）本稿は、2011年3月発行予定の『異文化コミュニケーション論集』（立教大学大学院 異文化コミュニケーション研究科）に掲載予定の筆者の論文(英語)に加筆および修正を行い日本語で書き改めたものである。

【参考文献】

ALPAC. (1966). Languages and machines: Computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington, D.C.: National Academy of Sciences, National Research Council. Retrieved from http://www.nap.edu/openbook.php?record_id=9547&page=R

Bowker, L. (2005). Productivity vs quality: A pilot study on the impact of translation memory systems. Localisation Focus, 4(1). 13-20.
Bowker, L., & Ehgoetz, M. (2007). Exploring user acceptance of machine translation output: A recipient evaluation. In D. Kenny and K. Ryou (Eds.), Across boundaries: International perspectives on translation (pp. 209–224). Newcastle-upon-Tyne: Cambridge Scholars Publishing.
Fiederer, R., & O'Brien, S. (2009). Quality and machine translation: A realistic objective? The Journal of Specialised Translation, 11, 52-74.
García, I. (2010). Is machine translation ready yet? Target, 22(1), 7-21.
Guerberof, A. (2008.) Productivity and quality in the post-editing of outputs from translation memories and machine translation. Localisation Focus, 7(1), 11–21.
Ishikawa, S. (2005). Potential application of XML [XML no motsu potential o kangaeru]. TC symposium, Japan Technical Communicators Association.
Krings, H. (2001). Repairing texts: Empirical investigations of machine translation post-editing processes (G. S. Koby, Ed). Ohio: Kent State University Press.
Lörscher, W. (1996). A psycholinguistic analysis of translation processes. Meta 41(1), 26-32.
O'Brien, S. (2006a). Controlled language and post-editing. MultiLingual, October/November, 17-19. Retrieved from https://216.18.156.115/multiligual/downloads/screenSupp83.pdf
O'Brien, S. (2006b). Eye-tracking and translation memory matches. Perspectives: Studies in Translatology, 14 (3), 185-205.
O'Brien, S. (2008). Processing fuzzy matches in translation memory tools: An eye-tracking analysis. In S. Gopferich, A. Jakobsen, & I. Mees (Eds.), Looking at eyes: Eye-tracking studies of reading and translation process. (pp. 79-102). Copenhagen: Copenhagen Business School.
Ribas, C. (2007). Translation memories as vehicles for error propagation: A pilot study. Minor Dissertation. Tarragona. Universitat Rovira i Virgili.
Tatsumi, M. (2009). Correlation between automatic evaluation metric scores, post-editing speed, and some other factors. MT Summit XII proceedings. Retrieved from http://www.mt-archive.info/MTS-2009-TOC.htm