統合翻訳ホスティング・サイトを用いた協調作業による下訳・修正訳データの収集

影浦峡 †, 阿辺川武 ‡, 内山将夫 §, 隅田英一郎 §
† 東京大学大学院教育学研究科
‡ 国立情報学研究所連想情報学研究開発センター
§ 情報通信研究機構 MASTAR プロジェクト

1　はじめに

近年、単一言語のパラレル・コーパスの重要性が一般に認識されている（Barzilay and McKeown, 2001; Tono, 2009）。単一言語のパラレル・コーパスの一種として、人間の翻訳者による下訳と修正訳のコーパスがあるが、これは、(i) 経験のない翻訳者の自習や翻訳教育などに活用できる（NGO や NPO など、ボランティア翻訳者に依存するところでは経験の少ない翻訳者の自習は死活問題である）ほか、(ii) 機械翻訳や翻訳支援の高度化にも有用な情報を提供する、重要なコーパスである。例えばリーズ大学の MeLLANGE は翻訳者訓練のために下訳・修正訳データを提供しており、注目を集めている（MeLLANGE, 2009）。発表者らも、出版社と交渉して下訳・修正訳データを入手しコーパス構築を行ってきた（Abekawa & Kageura, 2008a）。

しかしながら、こうしたデータは、これまで、翻訳学部や政府の翻訳機関といったところでしか体系的に構築・入手できないものであった。一般に、翻訳者は下訳データを提供したがらず、またデータを入手できた場合でも、言語処理の応用に向けてタグ付け等必要な処理を施す作業も手間がかかる（Abekawa & Kageura, 2008b）。一方、経験のない人間の翻訳者は、下訳と修正訳のデータから直接学ぶことができるにもかかわらず、人間が利用する経路やメカニズムはこれまでのところ体系化されていなかった。これら問題を解決するための一歩として、筆者らは、翻訳ホスティング・サイト「みんなの翻訳」（Utiyama, et. al., 2009; 内山他, 2010）を利用して、原文、下訳、修正訳（完成訳）からなるコーパスを構築する手法を開発してきた。本発表では、これを紹介するとともに、人間の翻訳者によるデータの活用方法、現在の統計数値を紹介する。

2　「みんなの翻訳」概要

「みんなの翻訳」は翻訳支援を備えたオンラインの翻訳ホスティング・サイトで、2009 年 4 月以来、一般公開されている。サイトは無料で利用することができる。登録したユーザは統合的翻訳支援エディタ QRedit（Abekawa & Kageura, 2007）を利用することで、三省堂『グランドコンサイス英和辞典』（三省堂, 2001）のような高品質辞書の参照をはじめ、Wikipedia や Google 検索を活用しながら、オンライン文書および任意の電子テキストを翻訳し、「みんなの翻訳」上に保存し、（著作権が許せば）公開することができる。

ユーザの翻訳活動を通して、「みんなの翻訳」には、自然に翻訳文書が蓄積される。2009 年 12 月 29 日の時点で、登録ユーザは 1021 人（うち翻訳を「みんなの翻訳」で公開しているユーザは 40 人）、登録文書数は 3087（うち公開文書は 1478）である。アムネスティ・インターナショナル日本やデモクラシー・ナウ！ジャパン、アジア太平洋資料センターといった著名な NGO が「みんなの翻訳」を使っているほか、複数の大学のゼミでも「みんなの翻訳」が利用されている。

3　下訳・修正訳コーパスの構築

3.1　基本メカニズム

下訳と修正訳を蓄積する基本メカニズムは非常に簡単である。翻訳者は「みんなの翻訳」を利用して作成した文書を「みんなの翻訳」サーバに保存するが、その際、「みんなの翻訳」は 1 文書につき最大 10 バージョンまで翻訳を保存する。

翻訳者によっては、データが失われることを恐れて例えば一段落翻訳するごとに、翻訳文書を保存する場合がある。これを行うと、保存バージョンが 10 では重要なバージョンが失われる可能性がある。この問題を回避するために、2009 年 11 月から「みんなの翻訳」では、通常保存モードとスナップショット保存モードを導入している。通常保存モードで保存されたバージョンは、次のバージョンが保存された際に上書きされる。一方、スナップショット保存モードで保存された文書は、次のバージョンが保存されても、最大 10 バージョ

ンまではログが保存される。両モードの使い分けには利用者の意識的な操作が必要になるが、このモード識別は、サーバの容量負担を極端に大きくせずに必要なログを保持するだけでなく、翻訳者による下訳・修正訳データの利用を促すためにも重要である。モードの選択は簡単で、翻訳支援エディタ QRedit で作成した文書を保存する際、「スナップショット」（カメラのマーク）にチェックを入れればスナップショット保存、チェックをはずせば通常保存となる。

3.2 協調的構築

「みんなの翻訳」では協調的翻訳の機能を提供することで、同一翻訳者の下訳と修正訳を蓄積するだけでなく、異なる翻訳者が作成した下訳と修正訳を蓄積することができる。このメカニズムは、「みんなの翻訳」が提供する協調翻訳機能に支えられている。協調翻訳機能を実現するために「みんなの翻訳」は以下の機能を提供している。

1. 翻訳者は、別の翻訳者に、自分が翻訳した文書の編集許諾を与えることができる。この編集許諾は、(a) 「みんなの翻訳」に登録している全ユーザに対して与えることも、(b) 指定したユーザのみに与えることもできる。
2. 「みんなの翻訳」では、翻訳者のグループを定義することができる。編集許諾も、個別翻訳者を指定して許諾を与えるだけでなく、グループを定義しておけばグループに対して許諾を与えることができる。

これらの、ある意味で極めてオーソドックスな機能は、NGO や NPO、大学のゼミなどで、翻訳支援環境を使わずに進められている翻訳実践や英文購読実践を翻訳支援環境でトレースするために大切な役割を担う。NGO などでは、一般に、翻訳経験の少ないボランティア翻訳者が下訳を作成し、それを熟練した翻訳者が修正して最終的な翻訳を完成させるという手続きがとられている。大学のゼミなどで、学生の訳文を教員が修正する場合も同様である。「みんなの翻訳」では、経験の少ない翻訳者が熟練した翻訳者に編集許諾を出し、それを受けて熟練翻訳者がその翻訳を修正することで、異なる翻訳者による下訳と修正訳のデータが蓄積されることになる。大学のような学習環境では、経験の少ない翻訳者相互の翻訳修正も行うことができる。従って、グループでの利用が増えれば、「みんなの翻訳」に自然に蓄積する下訳・修正訳コーパスの一部として、異なる翻訳者による下訳・修正訳データが構築されることになる。ログには翻訳者の名前が付与されているので、どのような下訳・修正訳データかを容易に識別することができる。

4 翻訳者による下訳・修正訳の利用

本稿冒頭で述べたように、原文と下訳・修正訳からなるコーパスは、翻訳者の自習のために（MeLLANGE, 2009）、あるいは機械翻訳や関連技術を高度化するために（Abekawa & Kageura, 2008b）、利用することができる。機械翻訳に人間の専門知識を取り入れることの重要性はますます高まっているが（Casacuberta, et. al, 2009）、ここでは、人間の翻訳者による利用に焦点をあてる。

多くの NGO や NPO では、ますます増大する翻訳ニーズに対応するため、多数のボランティアに頼る必要がある。しかしながら、ほとんどの NGO や NPO は、経験を積んだボランティア翻訳者の不足という問題を抱えている。ボランティアで参加した翻訳者の継続性が低いこともあって、中核的に活動する経験を積んだ翻訳者の負荷がますます高まり、そのために経験の少ない翻訳者にアドバイスする時間が取れず、それは経験の少ない翻訳者の離脱の要因となる … という悪循環が、ここには存在する。

経験の少ない翻訳者の自習を促進するために、「みんなの翻訳」では異なる翻訳バージョン対を対比的に表示する「翻訳履歴」表示機能を提供している。図 1 に、二バージョンの対比表示例を示す。ここでは、追加された文言は赤で、削除された文言は緑で示されることで、どこがどのように修正されたかを簡単にチェックすることができる。また、逆の視点から相違をチェックすることもできる。二つのバージョン間の相違は、Google が提供している javascript のライブラリ、google-diff-match-patch を利用して表示されている（Google, 2009）。

必ずしも翻訳学習ではないが、言語学習のための SNS サイトとして、例えば、Lang-8（Lang-8, 2009）のようなサイトがある。「みんなの翻訳」が提供する自習向け機能は、「みんなの翻訳」自体が翻訳のホスティング・サイトとして本来的に翻訳の学習を目的としたものではなく、完成された翻訳文書を作成し公開するサイトであるという点で、こうしたサイトと異なっている。すなわち、まず、現実の翻訳の場とは離れたところで翻訳の学習と訓練を経て、それから実際の翻訳活動に参加するというのではなく、実際の翻訳活動の中で一定の役割を担いながら、自習を進めていくというかたちになっている。

とりわけ重要なのは、ボランティア翻訳、オンライン翻訳の現状がまさに現実の翻訳活動中心であり、そこに参加する翻訳者は実践を通した訓練のメカニズムが提供されない限り、ほとんどの場合、改めて翻訳の訓練を受ける場がないという点である。「みんなの翻訳」は、こうしたボランティア翻訳、オンライン翻訳

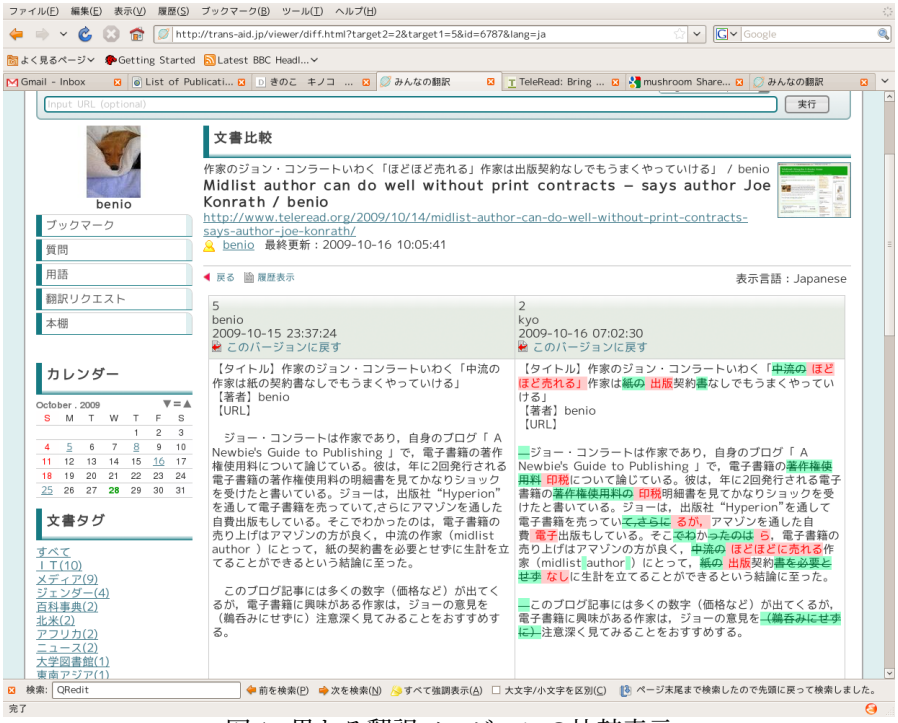

図 1. 異なる翻訳バージョンの比較表示

バージョン数	文書数（比率）	
1	1286	(43.30)
2	566	(19.06)
3	325	(10.94)
4	139	(4.68)
5	90	(3.03)
6	69	(2.32)
7	56	(1.89)
8	30	(1.01)
9	36	(1.21)
10	373	(12.56)

表 1 バージョン数毎の文書数

バージョン数	文書数（比率）	
2	4	(4.9)
3	5	(6.1)
4	9	(11.0)
5	3	(3.7)
6	12	(14.6)
7	2	(2.4)
8	3	(3.7)
9	4	(4.9)
10	40	(48.8)

表 2 バージョン数毎の文書数（複数翻訳者）

の作業の流れに、技術的な環境以外には変更を強いることなく、個人およびグループの翻訳活動を効率化し、その流れの中で経験の少ない翻訳者の自習を可能にするメカニズムを組み込んでいる点で、翻訳実務の効率も翻訳学習の効率も、翻訳サイクルの中で自然に改善するメカニズムを提供している。

5 下訳・修正訳データの現状

2009 年 12 月末の時点で、「みんなの翻訳」には 2970 文書の翻訳（英日方向のみの数値）が蓄積されている。一翻訳文書あたりの平均文字数は 1626 文字である。表 1 は、翻訳バージョン数（最大 10 バージョン）毎の文書数を示したものである。全文書のうち約 60 パーセントが少なくとも 2 バージョンの翻訳を有している。10 バージョンの文書数が多いのは、翻訳者によっては不要なバージョンまで保存対象としているためである。実は、前述したように通常保存モードとスナップショット保存モードとを区別したのは、公開後半年を経てからで、これは、不要なバージョンの保存により必要なバージョンが 10 バージョンから押し出されてしまう現象を認識し、それに対処するためであった。

「みんなの翻訳」を使って翻訳された 2970 文書のうち、元翻訳者が他の翻訳者に編集許諾を出した文書数は 290 である。表 2 に、複数の翻訳者が関与した文書の数値を示す。そのうち、実際に二人以上の翻訳者が翻訳に加わった文書の数は 82 であり（ただし、スナップショット保存モードを導入する前に失われた情報があるかもしれない）、今のところ、ほとんどすべてが、二人の翻訳者（つまり下訳者と修正訳者）によ

るものである。全文書と比べると、翻訳バージョンが8以上ある文書が、複数翻訳者が関与した文書では多い。詳細な分析は、今後、より多くのデータが蓄積されてから行っていきたいと考えている。

6 おわりに

本稿では、、現在一般公開され、実利用に供されているオンライン翻訳ホスティング／翻訳支援サイト「みんなの翻訳」を使った下訳・修正訳コーパスの構築メカニズムおよび人間の翻訳者によるその利用法を紹介した。構築および利用の技術的側面は比較的単純である。重要なポイントは、第4節でも述べた通り、「実世界における翻訳実践とは別の学習ステップと、それにより一定の技術を身につけたあとでの実践」という枠組みで人間の翻訳者による学習を定義するのではなく、また、システムの機能を優先しそれに対して人間の作業ステップを適用させる枠組みでもなく、「実世界における翻訳実践の実状に組み込まれた翻訳者の自習」という、実際にNGOやNPO等で求められている枠組みに基づき、実世界におけるオンライン翻訳の実践を支援する統合的な翻訳ホスティング・サイトの作業サイクルの中に、経験の少ない翻訳者の自習を可能にするメカニズムを組み込んだ点にある。すなわち、システムが介在しようがしまいが人間の間に存在していたあるいは存在すべきであると考えられていた協調関係を促進するためにシステムを用いたのであって、人間の側は、基本的に個々人がシステムの環境に慣れさえすれば、作業動線を変更する必要なく、経験の少ない翻訳者をフォローできる。

この機能が評価されて、2009年11月からは、英リーズ大学翻訳研究センターと神戸市外国語大学の翻訳者訓練共同プログラムで「みんなの翻訳」が利用されている。現在のところ、いくつかのNGOやNPO、そして大学関係と、本稿で紹介した機能を積極的に活用している事例は限られているが、枠組みとしてはもう一歩手前の学習段階、すなわち高校での英文和訳にも活用できる。もちろん、翻訳という行為と中学・高校での英文和訳という行為には大きな質的差異があるため、人間の側でそこから来る問題を十分見極めて利用する必要があるが、今後は、そうした方向への活用についても検討していく予定である。

下訳・修正訳コーパスの構築という観点からは、経験の少ない翻訳者の自習機能がより広く知られ活用されれば、自然にデータは蓄積されていくことになる。データが蓄積されれば、人間の自習用に自ら翻訳したテキストの修正を直接チェックするというだけでなく、コーパス全体を活用して機械翻訳技術の改善を進めたり、経験の少ない人間の翻訳者が陥りやすい翻訳上の問題とそうした問題が発生する条件を明らかにしたりなど、さらなる活用を展開することができるだろう。

謝辞

本研究は，日本学術振興会科学研究費補助金基盤(A)「包括的な翻訳情報資源を実現する統合翻訳支援サイトの構築」(課題番号00211152)の支援を得ている。

参考文献

Abekawa, T. and Kageura, K. (2007) “A translation aid system with a stratified lookup interface,” *ACL 2007 Demo and Poster Sessions*, p. 5–8.

Abekawa, T. and Kageura, K. (2008a) “Constructing a corpus that indicates patterns of modification between draft and final translations by human translators,” *LREC 2008*.

Abekawa, T. and Kageura, K. (2008b) “What prompts translators to modify draft translations? An analysis of basic modification patterns for use in the automatic notification of awkwardly translated text,” *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, p. 241–248.

Barzilay, R. and McKeown, K. (2001) “Extracting paraphrases from a parallel corpus,” *Proceedings of ACL2001*, p. 50–57.

Casacuberta, F., Civera, J., Cubel, E., Lagarda, A. L., Lapalme, G., Macklovitch, E. and Vidal, E. (2009) “Human interaction for high-quality machine translation,” *Communications of the ACM*, 52(10), p. 135–138.

Google (2009) google-diff-match-patch. http://code.google.com/p/google-diff-match-patch/

Lang-8 (2009) http://lang-8.com/

MeLLANGE (2009) http://corpus.leeds.ac.uk/mellange/ltc.html

三省堂編集所編 (2001)『グランドコンサイス英和辞典』 東京：三省堂.

Tono, Y. (2009) The JEFLL Corpus Project. http://jefll.corpuscobo.net/

Utiyama, M. et. al. (2009) “Hosting volunteer translators,” *MT Summit XII*.

内山将夫, 阿辺川武, 隅田英一郎, 影浦峡 (2009)「みんなの翻訳第二報」言語処理学会第16回年次大会.