オンライン協調翻訳環境におけるユーザ用語管理メカニズム

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

本発表では、筆者らが開発している翻訳ホスティング･サイト「みんなの翻訳」（Utiyama, et. al., 2009）が提供するオンライン協調翻訳環境が実装しているユーザの用語管理メカニズムを紹介する。

翻訳における専門用語管理の重要性は以前から広く認識されており（Hutchins, 1998）、用語管理のモジュールは職業翻訳者向け商用翻訳支援システムの重要な構成要素として高度化されてきている。「みんなの翻訳」は、以下の点で、これら商用翻訳支援システムとは、用語管理と活用の性格が異なっている。

1. オープンな利用環境。「みんなの翻訳」はオンラインで公開されており、誰もが登録できる環境であるため、相互に面識のない多数のユーザが参加している。
2. 活用のインタフェース。「みんなの翻訳」は経験の少ない翻訳者も支援の対象としているため、『グランドコンサイス英和辞典』（三省堂, 2001）、Edict（Breen, 2008）、Wikipedia 等複数の辞書・百科事典情報源の対訳情報を翻訳支援エディタ QRedit の中から表示する（Abekawa & Kageura, 2007）。QRedit は原文領域と翻訳文作成領域からなる 2 ペインのエディタで、原文領域の単語や熟語をマウスクリックすると辞書から訳語情報が表示されるほか、辞書情報全体の詳細表示、グーグル検索のシームレスな利用など、オンライン翻訳に最適化された機能を持つ。本稿で論ずる「専門用語リソースの利用」は、基本的に QRedit の原文領域から辞書・専門用語リソースが参照されることを意味する。図 1 に、翻訳支援エディタ QRedit の初期辞書引き場面を示す。ここで、「The Telegraph ／テレグラフ紙」の表示はユーザが登録した用語からの表示、Wi は Wikipedia、Gc は三省堂『グランドコンサイス英和辞典』からの表示、等々である。そのほかに、「みんなの翻訳」の用語検索ボックスからも、用語対訳を検索することができる（図 2）。

したがって、オープンな理念を保ちながら混乱しない用語活用メカニズムを導入し、用語を他のレファレンス情報資源と同一のインタフェースを通して、けれどもあくまでユーザ定義用語であることを示しながら提供する必要がある。以下では、現在「みんなの翻訳」に組み込まれている、こうした点を考慮した用語管理のメカニズムについて説明する。なお、その際、システムにユーザが登録したものとは別の専門用語レファレンス・リソースが追加される可能性も考慮に入れる。特に断らない限り、扱われている言語対は英日を想定するが、議論自体は一般性を持つものである。

2　専門用語リソースの性格と役割

一般語辞書の場合、商業的に出版されているものの数は多いが、それらはいくつかの類型に整理することができる。それに対して、専門用語リソースは別の意味で多様である。第一に、分野によっては標準的な専門用語辞書や専門用語集が存在し、通常それらには、対訳が付与されている。第二に、翻訳者グループや主題専門家、ターミノロジストが管理する「インハウス」の用語リストが存在する。例えば、アムネスティ・インターナショナル日本（アムネスティ, 2009）では人権関連用語の対訳リストを維持している。さらに、個々人で必要な用語を管理している場合もある。

専門用語は、利用の観点からも一般語辞書とは異なる性格を持つ。第一に、与えられた専門用語対訳を義務的に用いなくてはならない場合がある。これは、「インハウス」で管理されている用語リストを用いる場合に多い。一方、一般語の対訳と同様、リソースが示す専門用語対訳の利用が必ずしも義務的でない場合もある。分野の標準用語辞書やそれに相当する自動構築された辞書などを用いる場合、対訳の利用は義務的でない場合が多いだろう。いずれの場合でも、原則としてある専門用語の対訳が選ばれた場合には、翻訳テキスト中でその対訳を一貫して用いる必要がある。

これらの要因を考えると、専門用語リソースをその性質に応じて類別する必要があることがわかる。「みんなの翻訳」では、概念的な枠組みとして、専門用語

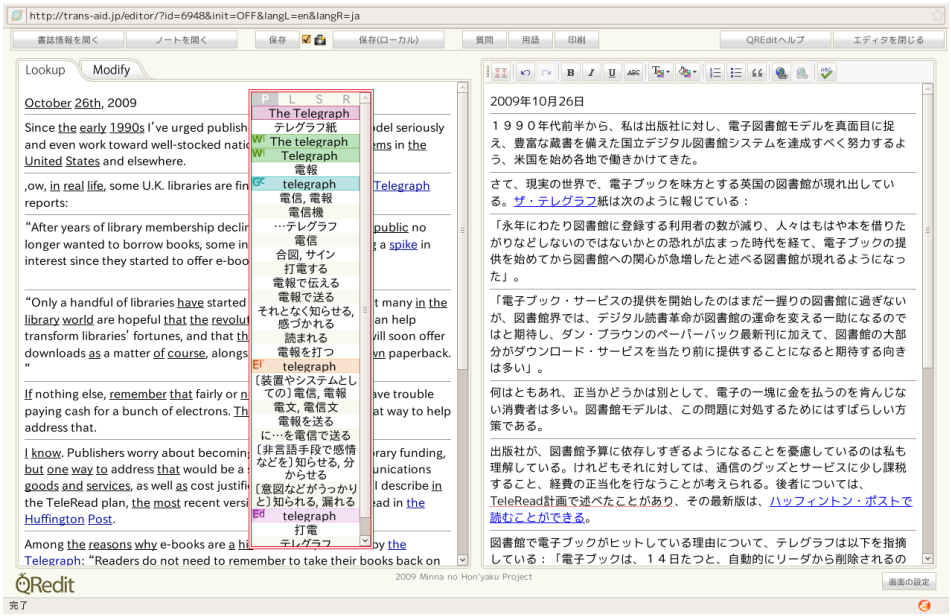

図 1. QRedit の初期辞書引き画面

図 2. 「みんなの翻訳」用語検索ボックスからの用語検索

リソースを４種類に分けて考える：

(i) 編集を経て作成された体系的な専門用語集。これらは、基本的に一般語辞書と同じ位置づけを有するものと考えることができる。

(ii) 「みんなの翻訳」利用者が登録したオープン・エンドな対訳用語リストの集積（下記 (iii) と (iv) はこの一部を構成する）。

(iii) 例えばアムネスティ・インターナショナルのようなグループが作成し管理している用語集で、「みんなの翻訳」にアップロードされたもの。

(iv) 「みんなの翻訳」ユーザが個人個人で登録した対訳用語。

グループ利用では、これらの用語集の組織化と調整が必要となる場合が多い。一方、個人ユーザは、理念的には何をどう使うか自分で判断し決められるのが望ましいが、実際には例えば大規模な用語集が別のユーザから登録された場合、あるいは多数のユーザが関連する用語を少数登録している場合、そこから何をどう使うかきめ細かく決めることは困難である。

3 用語登録メカニズム

上記専門用語リソースの (ii) から (iv) は利用者が登録するものである。ここでは、用語登録メカニズムを簡単に整理する。

「みんなの翻訳」では、大きくわけて以下の３種類の用語登録メカニズムがユーザに提供されている。

(a) 用語リストの一括登録。「用語, 対訳, 用語言語, 対訳言語, タグ, コメント, ペンネーム, 公開設定」の形式で一行一用語ずつ記述されたファイルを一括

して登録することができる。

(b) 用語抽出・登録。「みんなの翻訳」に登録された既訳文書から用語を抽出し、選択して登録することができる。ユーザは対訳対象データを既訳文書集合、タグ、翻訳者名などにより指定することができる。用語抽出・対訳抽出には、既往の手法を用いている（Nakagawa & Mori, 2003; Utsuro, et. al., 2006）。用語抽出・対訳抽出は、体系的な専門用語集に相当する専門用語リソースの構築には利用しにくいが、一定のテキスト集合に所属する用語リソースの構築には有用である。

(c) 個別の用語登録。気づいたときに、一語一語個別に用語と対訳を登録することができる。

以上のうち、(a) と (b) は、「みんなの翻訳」が提供する用語管理モジュールを通して登録する。(c) のような登録は、「みんなの翻訳」が提供する用語管理モジュールからもできるが、一般に個別の用語登録は、文書を翻訳している際に行われることが多いため、QRedit から、用語登録セッションを呼び出すことができるようになっている。ユーザは、用語を登録する際に、一括あるいは個別に用語の公開ステータスを指定することができる。これについては、次節で述べる。

4 「みんなの翻訳」の用語管理

4.1 システムが提供する用語リソース

「みんなの翻訳」が、編集を経て作成された専門用語リソース（第 2 節の (i)）を提供している場合、ユーザ側は一般語の辞書と同様に、まとまりごとに自分が利用するリソースを指定することで、専門用語リソースをコントロールすることができる。現在のところ、専門用語リソースに相当するものとしてシステム側からは Wikipedia が提供されているが（図 1 のポップアップにおける「Wi」で表示されたもの）、ユーザは「みんなの翻訳」メニューの「辞書引き対象設定」で、これを含め、リソースとして QRedit からの辞書引き対象とするもの・しないものを指定することができる。今後、システムとして提供する専門用語リソースを拡充していった際に、ユーザ側からの選択は利用しやすさの点で極めて重要になる。

4.2 ユーザが登録する用語リソース

第 3 節で述べたように、ユーザは、用語リストの一括登録、用語抽出・登録、個別の用語登録、という 3 種類の登録メカニズムを利用して用語を登録することができる。ユーザは、登録の段階あるいは登録後の任意の段階で、用語リストのまとまりごとにあるいは個別に、用語の公開ステータスを選択することができる。公開ステータスとして、「非公開」（自分のみの利用）、「限定公開」、「一般公開」がある。説明のために、ユーザ A が登録した非公開用語を AP、ユーザ A が登録した限定公開用語を AL、一般公開用語を AA としよう。ユーザ A 自身は、AP、AL、AA をすべて QRedit の辞書引きでも「みんなの翻訳」の用語検索ボックスからも参照することができる。A 以外のユーザ、例えばユーザ B からは、次のようになる。

AP ：QRedit からも「みんなの翻訳」の用語検索ボックスからも参照できない。

AL ：ユーザ A がユーザ B を参照許可対象ユーザに明示的に指定したときにのみ、QRedit からも「みんなの翻訳」の用語検索ボックスからも参照できる（従ってユーザ A の側からは、登録した用語を他の一部のユーザに提供する場合、まず用語を限定公開とし、その上で明示的に公開対象ユーザを指定することが必要になる）。

AA ：ユーザ B（および他のすべてのユーザ）は、「みんなの翻訳」の用語検索ボックスからは参照できるが、QRedit からの参照は、ユーザ A がユーザ B を公開対象ユーザに指定した場合にのみ参照できるようになる。

なお、図 1 に示したように、ユーザが登録した用語は QRedit における辞書引きの小ポップアップでは、一番最初に表示される。

現在のところ、受け手側ユーザからの用語参照制御メカニズムも開発されてはいるが、システム上では動かしていない。従って、ユーザ登録用語のステータス制御は登録したユーザ側ですべてが行われることになる。

これは、成員間で信頼関係が成り立っており、かつ情報流通経路が明確なグループでの利用に適している。例えば、アムネスティ・インターナショナル日本など、組織として用語対訳が管理されており、翻訳者は全員、その用語集が定めた対訳を用いなくてはならない場合には、中心となるユーザが組織の対訳用語集をアップロードし、アムネスティの翻訳に参加するユーザ全員を公開対象ユーザに指定することで、翻訳者全員が一律に同じ用語集を参照することになる。共訳で図書を翻訳する場合や、企業翻訳を複数人が請け負う場合なども、同様の状況が想定される。

5 問題

以上のような用語管理メカニズムは、実は、基本的に「性善説」を前提とし、また、類似の用語を必要とするユーザはそれなりにお互いを知っており、情報流通の透明性が一定程度確保されていることを前提としている。しかしながら、ユーザが増えるに従って想定

される以下のような問題に、現在のところ対処できていない。

- 悪意のユーザ。例えば、ユーザ B に悪意を持つ人が「みんなの翻訳」に登録し、多くの用語に「キノコ」と対訳を付けた用語リストを登録してユーザ B を公開対象ユーザに指定した場合、ユーザ B の辞書引きの度にポップアップの先頭に「キノコ」という訳語が現れることになる。受け手側からの制御機能を設けることが当然の対処策となる。
- 重複制御。複数のユーザが同一の用語を登録し、公開した場合、現在のところシステム側で重複制御がなされていない。例えば、図 2 から、二人のユーザが同一の用語対訳を登録していることがわかるが、一方のユーザから他方のユーザに公開指定が出されている場合、他方のユーザの QRedit 辞書引きでは同じ用語対訳が二重に表示されてしまう。
- 多数のユーザの用語供給・受け取りの相互関係。悪意のユーザに対する対策として、受け手側からの制御機能をあげたが、受け手が多数のユーザから用語を受ける場合、これは受け手側の負担を極端に高めることになる。また、ある翻訳者が異種のプロジェクトに複数参加する場合、同一ユーザ ID で「みんなの翻訳」を利用する限り、プロジェクト毎に利用する用語を区別するメカニズムが現在のところないが、そのメカニズムを入れると、受け手の制御がやはり複雑化する恐れがある。

起こりうる問題が以上で尽きているわけではないだろう。それも含め、現在対処できていない部分については、実際に問題が発生したときに、ユーザの声も取り入れながら対応していきたいと考えている。

6 おわりに

本稿では、現在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」における用語の管理メカニズムを紹介した。不特定多数のユーザが登録する「みんなの翻訳」では、既存の商用翻訳支援システムとは異なる要因を考慮した用語管理メカニズムが必要であり、その要因をどのようにとらえ、現在のところどのようなメカニズムを実装しているかを紹介してきた。「問題」でも述べたように、発表者らは、今後も利用実態を見ながら、メカニズムの改変・更新を続ける必要があると考えている。

最後に、ここまであまり強調しなかったが、「用語抽出・登録」で採用されている用語抽出・対訳抽出メカニズムは、独立した用語集の編成においてではなく、本システムのように特定の対訳文書集合に帰属しうる文献の翻訳のための既訳集合における用語対訳の参照という枠組みでは強力な実用ツールとなることが実利用の観点から確認されたと考える。ここから逆照射するならば、用語抽出や対訳抽出などの技術は、一見したところ「ニュートラル」なテスト的（擬似）評価でパフォーマンスを論ずるのではなく、どのような利用においてどのように有効かを、現実の応用と向き合いながら検討する段階に入っていると考えることができよう。

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)、および国立情報学研究所共同研究「異種情報源の特性を考慮した、実用的な専門用語対訳辞書の構築と活用」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.

アムネスティ・インターナショナル日本 (2009) http://www.amnesty.or.jp/

Breen, J. (2008) Edict. http://www.csse.monash.edu.au/~jwb/edict.html.

Hutchins, J. (1998) “Computer-based translation tools, terminology and documentation in the organizational workflow: Report from recent EAMT workshops,” *International Conference on Professional Communication and Knowledge Transfer*. p. 255–268.

Nakagawa, H. and Mori, T. (2003) “Automatic term recognition based on statistics of compound nouns and their components,” *Terminology*, 9(2), p. 201–219.

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.

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

Utsuro, T., Kida, M., Tonoike, M. and Sato, S. (2006) “Collecting novel technical terms from the Web by estimating domain specificity of a term,” Matsumoto, Y., Sproat, R., Wong, K-F., and Zhang, M. (eds.) *Computer Processing of Oriental Languages: Beyond the Orient: The Research Challenges Ahead*. Berlin: Springer, p. 173–180.