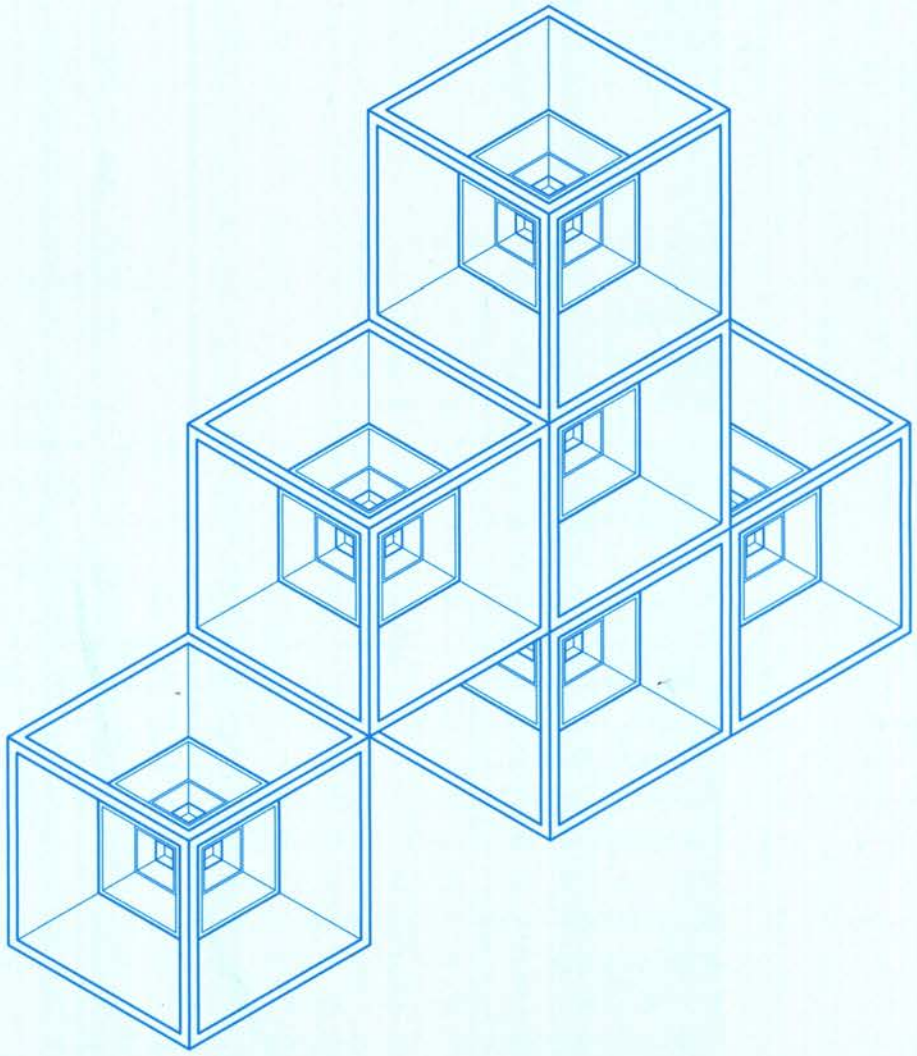
ISSN 1883-1818

AAMT Journal

Asia-Pacific Association for Machine Translation



November 2010 No.48

アジア太平洋機械翻訳協会

目　次

CONTENT

巻頭言

機械翻訳への期待

慶應義塾大学環境情報学部
教授　石崎　俊

翻訳支援技術は、機械翻訳システムそのものだけでなく翻訳メモリなども含めて、多言語で効率的に翻訳情報を得て利用するための手段として期待されていて、英語圏に限らず多数の国を対象とした有用な手段と考えられています。

近年の我が国の国際競争力の停滞を挽回するための手段として、例えば、鉄道や電力設備などに関係する大規模なインフラの輸出の可能性が考えられ、大いに期待されています。そのような機会には、大規模システムの説明書類やマニュアルばかりでなく、双方の国の法令や国内外の標準などを関係者が正確に理解する必要があり、背景となる情報も含めると、膨大な量の情報の翻訳を効率よく実施する必要があります。

これは、機械翻訳の研究開発の段階というよりも、実際に直面する問題に使える機械翻訳システムが欲しいという段階です。もとより完璧な機械翻訳システムである必要はありませんが、よく言われているように、すべてを人手で翻訳する場合に比べてメリットがあればよいわけです。費用や時間が削減できれば喜ばれます。

ところで、我が国で機械翻訳システムが製品として初めて発売された頃から見守っていまして、AAMTの事務局から送っていただいている機械翻訳関係の最新の記事などをいつも拝見していますが、近年の進歩・発展には顕著なものがあると感じています。特に、コーパスが整備されて統計的な手法が適用されて以来の翻訳システムの進歩・改良には目覚ましいものがあるように思います。そのような機械翻訳の様々な最新の技術や翻訳用辞書、システム評価などについては、昨年12月に出版された言語処理学事典の第3部統合技術・応用システムにまとめられています[1]。この事典は自然言語処理の先端の研究者たちのご協力で、数年かけて幅広い分野の研究成果がまとめられています。

一方では、統計的な処理によって統計上のメリットが出るのは、例外的な処理が必要なデータを除くと、全体の80%程度かもしれないと言われていて、自分自身でもそのように感じています。特に言語データは表現の多様性が大きいので、それ以上の高い精度を得るのは容易ではないかもしれません。

そのような意味で、性能が格段に進歩した機械翻訳システムであっても、実際問題に使用する場合は、単体で使うシステムというよりも、Web上にある翻訳環境を活用する翻訳支援機能のメリットが大きいと考えられます。Web上に多言語の翻訳用辞書が整備され、使用者にとって使い勝手が良く、効率的に翻訳作業ができるようなトータルなシステムが重要です。

さらに、機械翻訳で意味をしっかりと翻訳し精度を上げるために意味の曖昧性なども扱う場合は、必然的に文脈が関係し常識や背景的な知識が必要になると思われます。これは、人間にとっては基本的な機能ですがコンピュータにとっては高度で難しいため、一歩一歩進む地道な研究努力が今後期待されていると思います。

参考文献

[1]言語処理学会編、言語処理学事典、共立出版、2009年12月(同、デジタル言語処理学事典、同、2010年6月)

シンポジウム報告

第１回産業日本語研究会・シンポジウムの開催概要と今後の取り組み

一般財団法人　日本特許情報機構
調査研究部　前部長　渡邊　豊秀

１．はじめに

平成 22 年 2 月、わが国の産業・技術情報基盤としての日本語のあり方に着目した新たな取り組みが始まった。第１回産業日本語研究会・シンポジウムである。本稿では、このシンポジウムの概要について、開催の背景と今後の取り組み等を交えつつ報告する。

２．シンポジウム開催の背景

本シンポジウムの開催の背景には、英語ベースで進展しつつあるグローバル社会とわが国の情報基盤としての日本語のあり方に対する危機感がある。ご存じの通り、我が国産業界にとっては、情報発信や情報取得に際する翻訳コストの低減、機械翻訳の有効活用が急務である。また、爆発的に増える情報に対し、必要な情報へより効率的にアクセスし得る環境を作ることが求められている。しかしながら、表記上の自由度の大きさや曖昧な表記が慣行されてきた日本語テクニカルライティングの現状は、機械翻訳や情報検索等のコンピュータを活用した日本語処理の効率化を阻んでおり、知識管理の高度化・効率化の妨げとなっているという危機感である。こうした危機感を払拭するためにも、新たな日本語の枠組みについて、国や、言語・言語処理に関する研究・開発を進める関係学会、関係団体、関係企業といった All Japan 体制で幅広く総合的な議論を行う場が必要である。それが産業日本語研究会である。産業日本語とは、産業活動の諸側面を媒介する産業・技術文書を人に理解しやすくかつコンピュータにも処理しやすく表現するための日本語であり、一般財団法人日本特許情報機構（以下、Japio と略記する。）が、特許情報を記述する日本語のあり方の検討を通してそのコンセプトを作りあげてきた造語である。この産業日本語をベースに新たな日本語の枠組みを産業界へ普及させ、わが国の情報基盤として根付かせために、Japio、高度言語情報融合フォーラム（ALAGIN）及び言語処理学会が協力して、本シンポジウムを開催することとなった。シンポジウムの開催に先立ち、わが国における言語研究・言語処理研究を先導してきた先生方からなる世話人会が設置された。AAMT そして言語処理学会の初代会長であられ、現在は国立国会図書館館長であられる長尾眞先生に世話人会の顧問をしていただき、AAMT 会長を務められ、現在は高度言語情報融合フォーラム（ALAGIN）会長等を務められている東京大学教授の辻井潤一先生や、AAMT の現会長であられる豊橋技術科学大学教授の井佐原均先生には世話人として議論を先導していただいた。こうした体制で、シンポジウムのプログラムや研究会設置の構想についての議論、関係諸団体への協力の呼びかけ等、約１年間もの期間をかけて、準備が進められた。

３．シンポジウムの概要

「第１回産業日本語研究会・シンポジウム」は、平成 22 年 2 月 24 日に、東京大学福武ラーニングシアターで開催された。シンポジウムは、会場を埋め尽くす 184 人もの参加者を得て、大盛況であった。

シンポジウムの開催にあたり、経済産業省、特許庁、総務省、大学共同利用機関法人人間文化研究機構国立国語研究所、独立行政法人工業所有権情報・研修館、独立行政法人情報通信研究機構といった関係省庁の他、AAMT をはじめ、社団法人情報処理学会、社団法人人工知能学会等の多数の関係団体にご後援をいただいた。シンポジウムでは、長尾眞先生による「産業日本語を育てよう」と題された基調講演に続き、言語分野、司

法分野、医療分野といった様々な分野での日本語を明晰に使用するための取り組みや産業日本語に関する取り組み等が多数紹介された。その後に行われた総合討論の場では、一般参加者も加わり、活発な議論が行われた。そして、産業日本語の研究・開発・普及活動を先導する場としての産業日本語研究会への参加呼びかけが行われた。以下にシンポジウムのプログラムを紹介する。

第1回産業日本語研究会・シンポジウムプログラム

(1) 開会挨拶

橋田浩一　言語処理学会副会長／産業技術総合研究所社会知能技術研究ラボ長

(2) 基調講演「産業日本語を育てよう」

長尾　眞　国立国会図書館長

(3) 招待講演「言語学と機械翻訳をつなぐ語彙意味論」

影山太郎　国立国語研究所長

(4) 招待講演「法廷用語を市民の手に～日弁連における日常語化プロジェクトの経験から～」

酒井　幸　弁護士

(5) 取組・研究事例の紹介・総合討論

コーディネータ　井佐原　均　豊橋技術科学大学教授

① 日本語処理技術の研究開発における特許情報の可能性

藤井敦　東京工業大学准教授

② 病院の言葉を分かりやすくする提案

田中牧郎　国立国語研究所准教授

③ 特許版・産業日本語と機械翻訳向け特許ライティングマニュアル

渡邊豊英　Japio　特許情報研究所調査研究部長

④ 産業情報としての特許情報と産業日本語

横井俊夫　東京工科大学名誉教授／Japio 特許情報研究所顧問

⑤ 総合討論　参加者全員（会場含む）による総合討論

なお、産業日本語研究会やシンポジウムの詳細は以下の Web Site で紹介されている。

＝　産業日本語研究会 Web Site　＝

http://www.tech-jpn.jp/

4．今後の取り組み

産業日本語研究会・シンポジウムは、産業日本語研究会のメインイベントとして、産業日本語に関する諸方面での一年間の取り組みを総括し、産業日本語の普及・啓発を進め広く世論を喚起するために、今後とも、定期的に開催する予定である。

また、産業日本語研究会の具体的取り組みとして、産業・技術情報が生まれる様々な現場にて、産業日本語による文書作成や情報発信を、IT を用いて支援する仕組みの検討がはじまった。その検討の場として、産業日本語研究会の中に設置されたのが、産業日本語プラットフォーム委員会である。委員長には、AAMT の現会長で、産業日本語研究会の代表幹事であられる井佐原均先生が就任され、わが国における自然言語処理の研究開発現場の最先端でご活躍されている研究者の方々や企業の方々にご参加いただいている。

産業日本語プラットフォーム委員会では、わが国の情報基盤としての産業日本語プラットフォームのあり方について、詳細な議論をしていただき、プラットフォーム構築の実現に向けた開発計画の詳細化等の具体的検討が進められる予定である。産業日本語プラットフォームの開発には、網羅すべき分野の広さや深さから、未だ多数の研究課題を含んでいると考えられるため、膨大な人的資源と予算が必要となることが予想される。まさに、国をあげて取り組むべき一大プロジェクトである。産業日本語プラットフォーム委員会は、その成果を国等へのプロジェクト提案書の形にまとめ上げることを視野に入れ活動する予定である。

本年度末に開催予定の「第2回産業日本語研究会・シンポジウム」では、関係各者からの産業日本語に関する取り組みの紹介に加え、産業日本語プラットフォーム委員会での検討状況や Japio で進められている特許分野への産業日本語の導入に関する検討の状況等も報告され、より充実したシンポジウムとなることが期待される。

シンポジウム報告

第4回「機械翻訳技術のイノベーション」シンポジウムレポート

独立行政法人 情報通信研究機構
知識創成コミュニケーション研究センター
言語基盤グループ　風間　淳一

2010年3月1（月）に、東京大学本郷キャンパスの福武ホールにて、第4回「機械翻訳技術のイノベーション」シンポジウムが行われました。

本シンポジウムは、平成18年度から5年間の予定で、独立行政法人情報通信研究機構、独立行政法人科学技術振興機構、東京大学、静岡大学、京都大学が実施している、科学技術振興調整費による重点課題解決型研究「日中・中日言語処理技術の開発研究」が主催するシンポジウムです。今回で4回目を迎え、参画機関からのプロジェクトの進捗報告や、機械翻訳に関わる海外研究者による招待講演などが行われました。筆者も本プロジェクトに参加しているため、進捗報告を行いました。本稿では、シンポジウムの簡単な参加報告をさせていただきます。

まず、開会のあいさつとして、プロジェクト代表者である情報通信機構の井佐原 均より、プロジェクトの概要が紹介されました。本プロジェクトは、アジア諸国の科学技術情報の日本での活用を容易にし、また、我が国の最先端の科学技術情報の各国への流通を促進することにより、アジア諸国と日本の科学技術の発展に資することを目的とし、その実現のために、言語の構造をより深く利用した汎用的な用例翻訳手法、科学技術文献の翻訳・情報検索用辞書を半自動作成する手法を確立し、特に、中国語に焦点をあてて、大規模な言語資源を構築し、その言語資源を用いた機械翻訳プロトタイプシステムの開発をすることを目指しています。

開会のあいさつにつづいて、翻訳に関わる海外の著名な研究者や企業関係者による招待講演が行われました。今回は、Language Weaver および ISI/USC の Daniel Marcu 博士、Asia Online の Dion Wiggins 博士、Edinburgh 大学の Miles Osbone 教授の三氏を迎えての講演となりました。

Marcu 博士は、「Lessons Learned in Ten Years of Statistical Machine Translation」というタイトルで、米国の DARPA TIDES、GALE プロジェクトから得られた経験、教訓などをお話されました。特に印象的だったのは、機械翻訳に当初課された目標値が、翻訳者の一回目の訳に相当する値よりも高かったといったエピソードです。また、技術的な点では、より精緻なモデルの利用を可能とする賢く効率的な探索アルゴリズムが精度向上に重要であることを説かれていました。

Wiggins 博士は、Asia Online 社の CEO で、「Beyond Machine Translation: Collaboration, Integration, Quality, Change and Jobs Synopsis」というタイトルで講演されました。Asia Online は、機械翻訳と翻訳者による翻訳を活用して、多くの言語対に対して翻訳サービスを提供している企業です。講演は、「50年間望むものを提供するのに失敗してきた機械翻訳業界がなぜいまも存在するのか」という刺激的な問いかけから始まりました。その答えは、翻訳に対する無限に近い要求です。全世界での人間の翻訳者による翻訳のマーケットは毎年順調に伸び続けており、2012年には240億ドル市場へと成長すると予想されているそうです（一方、純粋な機械翻訳のマーケットは、最近の精度向上を考慮しても15億ドル程度です）。しかし、現状では人手翻訳のコストが高いため、翻訳サービスは、高品質を求めるような企業文書向けなどにとどまっており、E-mail やブログなどの潜在的な大量の翻訳ニー

ズの 0.5%しか満たしていません。そこで、氏は、機械翻訳をうまく活用することをうったえます。例えば、機械翻訳を前処理に用いることにより、人手翻訳の効率を大幅に向上させることができることを数字をまじえて説明されました。Asia Online では、実際に、非常にシステマチックな機械翻訳と人手翻訳の融合がすでに行われており、非常に感銘を受けました。

Osborne 教授は、「Stream-based Randomized Language Modelling for Machine Translation」というタイトルで講演されました。統計的機械翻訳では、対象言語としての出力の自然さを測るため、言語モデルを用います。有名なものは n-gram モデルと呼ばれるもので、これは、対象言語で各単語列（例えば 3 単語列）がどのくらいの頻度で出現するかを記録したものです。頻度を計測する元の文書が多ければ多いほど、n-gram モデルも正確となり、機械翻訳の質も上がります。近年では、Google 1T n-gram データなど、Web 等から抽出した非常に大規模な n-gram データも入手できるようになりました。問題は、大きすぎることで、これまでのデータ構造ではうまく扱えません。そのため、最近使われるようになってきた手法が、randomized language model というものです。n-gram モデルの本質的な機能は、ある単語列の頻度を問われたときにその頻度を答えることですが、この手法では、Bloom filter と呼ばれるデータ構造などを使い、ごく小さな確率で間違った頻度を答えてしまうことを許します。そのように妥協することで、これまでに比べて大幅に小さな容量で n-gram モデルを格納することができます。大規模になる利点と間違いの可能性のトレードオフとなりますが、実験では翻訳性能が向上することが示されています。本講演では、さらに、このデータ構造をストリームデータへ拡張する方法も紹介されました。ストリームデータとは、データは届いたときに 1 回だけ見ることができ、後からは見返せないというタイプのデータです。Web などは日々変化しているため、ストリームデータと考えることもできます。今後はこのようなデータの扱いも非常に重要になってくると考えられます。

招待講演につづいて、プロジェクトの参画機関からの進捗報告が行われました。タイトルと報告者は以下のとおりです。

- 「構文情報を利用した日中用例ベース機械翻訳」黒橋　禎夫（京都大学）
- 「機械翻訳のための高精度中国語解析技術と評価」風間　淳一（情報通信研究機構）
- 「The Development of Chinese HPSG Grammar from the Penn Chinese Treebank」Kun YU（東京大学）
- 「Automatic Extraction of Domain Term Translations From A Chinese-Japanese Parallel Corpus」中川　裕志（東京大学）
- 「コンパラブルコーパスを用いた訳語選択」梶　博行（静岡大学）
- 「日中機械翻訳用言語資源の構築と品質管理」菊池　俊一（科学技術振興機構）

今年が、プロジェクトの最終年度ということもあり、各者、これまでの成果や最後の 1 年間の研究の方向性について発表をしました。

京大の黒橋教授は、本プロジェクトで開発されている構文情報を利用した用例ベースの機械翻訳エンジンについて発表しました。統計的機械翻訳の研究の初期には、単語レベルでの対応を考えるモデルが多かったのですが、日英、日中のように構造が大幅に変わる言語対に対しては非力であることが認識され、近年では、国際的にも単語レベルではなく、句構造レベルなどで対応付けをするモデルが盛んに研究されています。このプロジェクトでは、計画段階からいち早くその必要性を認識し、両側言語を構文解析（係り受け解析）して、その部分構造の対応を用例として抽出し翻訳を行う用例ベースのシステムが志向されました。特に、日本語の処理に適した係り受け構造を利用することが特徴の一つです。提案手法では、単語ベースの対応付け

よりも高精度で翻訳の対応付けが行えること、また、翻訳の主観評価による質が向上することなどが報告されました。

筆者（情報通信研究機構 風間）からは、上記翻訳エンジンで使用されている中国語側の形態素解析、係り受け解析技術について発表を行いました。日本語や英語に比べて、中国語の研究は歴史が浅く、その精度もまだ十分とは言えません。日本語や英語では、形態素解析からはじめて係り受け解析まで通して解析を行った場合、90%程度の正解率となりますが、中国語の場合は80%に届きません。そこで、我々は最新の機械学習アルゴリズムを採用したり、近年有効性が示されている半教師あり学習を係り受け解析に適用するなどして、精度向上を目指してきました。後者の技術は次のようなものです。まず、手元にある正解データから係り受け解析器を学習します。学習された解析器を用いて、Web などから集めた大量の文書を自動解析して、そこから係り受けの部分木を抽出します。自動解析なので間違った木も含まれますが、大まかには、頻度が大きい木はもっともらしい木、逆に頻度が小さい木はありそうもない木であるということが分かります。この情報と最初の正解データを一緒に用いて、もう一度学習すると解析精度が向上することが示されています。現在は、中国語処理に関する研究が非常に盛んに行われていますが、我々の解析器は、現時点で、世界最高性能を保持しています。プロジェクトが対象とする科学技術論文に対する通しでの精度も、73.%から 77%程度までに向上しました。もちろん、まだ、日本語や英語の精度とは比較にならないため，継続した開発が必要です。

次に、東大の Yu 博士からは、中国語の HPGS 文法を獲得する方法について発表が行われました。HPSG はいわゆる深い構文解析を行う文法枠組みで、文法構造と意味構造を同時に組み上げることが考慮されている理論的に有望な枠組みです。しかし、そのぶん必要な情報も多いため、文法を人手で一から作成することは非常にコストがかかります。この研究では、より単純な句構造文法のツリーバンク（文に対し正解の木構造が付与されたデータベース）を HPSG の学習に適した形に低コストで変換して、そこから HPSG の文法を自動学習します。今後は獲得した文法を使用して実際に中国語の解析器を作成する予定となっています。この成果により、翻訳技術、解析技術、辞書獲得技術をさらに進化させることができると期待されます。

東大の中川教授からは、日中の対訳データから分野固有の専門用語の対訳辞書を自動で獲得する手法が発表されました。専門用語は日々生成され、またその分野の専門家でないと対訳辞書の作成が難しいといった問題があります。翻訳においては、専門用語の訳し間違いは、意味の伝達を阻害する大きな要因となります。したがって、専門用語の対訳辞書の自動構築が強く望まれています。提案手法では、まずそれぞれの言語で、中川教授が開発している「言選 Web」で専門用語を自動で認識して抽出します。次に、統計的機械翻訳で用いられる単語対応付け手法を用いて、両言語の専門用語の対応を得て、最終的に辞書として抽出します。対応付けをよりよくするため、単語の分割を修正したり、専門用語の対応のスコアを求める式などに工夫をすることにより、90%以上の精度を持つ高精度な対訳辞書を抽出することに成功しています。

静岡大の梶教授からは、コンパラブルコーパスを用いた訳語選択の方法について発表がありました。言葉はあいまい性を持っているので、翻訳においても適切に取り扱う必要があります。例えば、tank という語は、「戦車」「タンク（容器）」と訳せますが、どちらが適切かはその文章によります。この研究で利用しているアイデアは、ある言語で共起する語の訳語はやはり（翻訳先言語で）共起する、というものです。例えば、tank の近くに soldier などの単語があれば、tank はおそらく「戦車」と訳すべきだということが分かります。ただし、これだけでは問題もあるので、様々な工夫を加えて、関連語 - 訳語関連行列を反復計算するアルゴリズムを提案し、適切に訳語が選択できるモデルを構築しています。

科学技術振興機構の菊池氏からは、日中機械翻訳用言語資源の構築と品質管理についての発表がありました。これまでに、「環境工学」「情報工学」「生物科学」「医学」「エネルギー」「農業」「材料」「化学」の８分野を対象に、2,700 万文字を超える日中対訳コーパス、約２万８千語の専門用語辞書が作成されたことが報告されました。また、対訳コーパスの分析から、翻訳者が日本語を母語としない場合、日本語の助詞の「より」などのあいまい性の高い助詞の誤訳が多いことなどが報告されました。

報告につづいて、最後に、東京大学の辻井教授から閉会のあいさつが行われ、シンポジウムは閉会しました。海外の著名な研究者の招待講演あり，これまで４年間行われてきた大プロジェクトの成果の報告ありと、(手前味噌ですが) 非常に内容の濃い有意義なシンポジウムであったと思います。

なお、シンポジウムの発表スライドなどは、http://www.congre.co.jp/imttsympo/program/index.html で見ることができます。私の理解や認識に間違いもあるかと思いますので、詳細にご興味のある方は、是非、上記の URL をご覧ください。

AAMT長尾賞

「みんなの翻訳」は何を目指すか

影浦　峡、阿辺川　武、内山　将夫

1. はじめに

「みんなの翻訳」（http://trans-aid.jp/）が2010年第5回AAMT長尾賞受賞の栄誉に浴したことを機に、「みんなの翻訳」についての文章を寄稿することになりました。概要はすでにAAMT Journal第45号「みんなの翻訳 – ボランティア翻訳者の支援と翻訳の共有のためのWebサイト」で紹介しましたし、その後、様々な機能強化と拡張を行いはしましたが、それらは使っていただいた方が説明するより理解しやすいと思いますので、ここではむしろ「みんなの翻訳」の考え方や理念をいささか思弁的にまとめてみたいと思います。

2. 道具としての言語処理

モハメッドが、山を3度呼び寄せても来なかったため、「自分が行くより外に仕方があるまい」と言って自ら山へ向かって歩いて行ったエピソードは広く知られています。漱石も『行人』中「塵労」の章で登場人物のHさんをしてこのエピソードに対するその先輩の「ああ結構な話だ。宗教の本義は其処にある」という言葉を引用せしめています。抽象的ですが、「みんなの翻訳」は、「宗教」を「研究」そして「応用」に置き換えてもまさにこの言葉はあてはまるという認識をプロジェクトの理念的基盤としています。自己意識を肥大させず、坦々と対象に触れ、対象と関わること。とはいえ、ボロを出す前に、少し具体的なことに話を移しましょう。

人間の翻訳者を支援するために言語処理技術は実際に使えるはずだ。それにもかかわらず、本来使われてもよさそうなところでも使われていない。2005年、「みんなの翻訳」プロジェクトの前身である「椎茸プロジェクト」を始めるにあたって私たちが抱いていた問題意識はこのようなものでした。もちろん、欧州諸言語を扱う産業翻訳業界では当時でも機械翻訳はプロセスの一部として使われていましたが、私たちが対象とするオンラインの翻訳ではあまり使われておらず、日本語を扱う場合は辞書引きがオンライン化・電子辞書化された程度、というのが実状だったのです。

そうした状況に対して、私たちが目指したのは、使えるようになるまで精度をあげることではなく（そもそも現状の言語処理技術は使えるはずだと考えていたわけですから、それにもかかわらず使われないとするとそれは精度の問題ではないはずでした）、オンライン翻訳者に言語処理技術が使われるようなかたちでアプリケーションの配置を考えることでした。

そのためにオンライン翻訳者の作業プロセスを観察し、翻訳で用いている道具の性質を整理して、言語処理技術を適用することにより改善できる部分を改善し、柔軟な熟語の検索や辞書の構成、用語対訳抽出など（多）言語処理の基本技術を組み込んで（これらについてはAAMT Journal第45号の紹介をご覧下さい）、統合的な環境として公開したのが「みんなの翻訳」です。

つい最近、名古屋大学の佐藤理史先生から教わった便利な言葉を使うと、開発の際のポイントは、「道具としての言語処理」にありました。もちろん言語処理技術を応用に使う場合、技術は道具として位置づけられます。その際、大切なことは、よくできた道具は、例えば手で使うものならば、いわば手の延長として自然に位置づけられるという点です。道具が道具として当たり前に使われるためには、対象プロセスの中に自然に組み込まれ、いわばプロセスにおける一つの基盤となることが求められるのです（例えばかな漢字変換はそのようなものになっています）。これに対して、言語処理技術の多くは、オンライン翻訳者の作業プロセスの外にあるものでした。

外にある限り，使われるためには圧倒的な精度が求められます。ところが、翻訳プロセスの中で使われるような位置づけを確立できれば、ノコギリが使うにつれ手に馴染むように技術の改善も進めることができるのではないか。機械翻訳も含め、「みんなの翻訳」では言語処理技術をこう考えて取り込んできましたし、これからもこのように取り込んで行くことになります。このような道具はまた、人間の能力をいっそう引き出すものになりうると私たちは考えています。

3. 言語表現と言語

前節で、「対象プロセス」、「作業プロセス」という言葉を使いました。技術を対象プロセスの中に位置づけるためには、そのプロセスを把握する必要があります。翻訳者と話していて明らかになった最大の点は（翻訳者が自分から明言したわけではありませんが）、翻訳とは徹頭徹尾具体的な言語の表現からなるテキストを対象とするプロセスであって、言語に関するプロセスではない、という点です。

例えば、「山」という単語は言語において一つの単位を構成します（言語学の言語に対する見方も基本的にこのようなものです）。これに対して、文学の翻訳で最も顕著ですが、より一般に翻訳においてはシェークスピアの「山」とトーマス・マンの「山」は別のものです。言語学ではそのような違いを一般に「文脈」の観点から扱いますが、翻訳の場合、本来（究極的には）、シェークスピアがこの作品のここで言った「山」として、唯一単独的な「山」が対象となるのです。翻訳者が感じている言語表現のあり方はこのようなものです。

そうは言っても、そんな風に捉えられた「山」はもうそれ以上操作しようがありませんから、実際には翻訳のプロセスの中で「言語」的視点を少し入れて、こちらの「山」とこちらの「山」は似ているとか、あちらの「山」は違うとか、そういったかたちで言語表現を扱っていくことになります。コーパスを用いた言語学や言語処理も具体的な言語表現を使うのだから、その限りでは双方が近づくわけですが、言語学や言語処理では言語が本質的な対象であって言語表現はそこに到達するための手段（データ）であると捉えるのに対し、翻訳においては、両者の関係はまったく逆で、あくまで言語表現が本質であってその操作のために言語という概念を必要に応じて導入することになります。技術的な論文としてはなかなか発表できませんし言葉にしにくいのですが、「みんなの翻訳」では、言語の技術を言語表現のプロセスに内在させるためのマクロなメカニズムに実はかなり工夫を凝らしています。

4. 現状と具体的な展開

幸い、2009年4月に公開して以来、「みんなの翻訳」はそれなりに認められ、2010年10月12日現在、登録者は1407名に、翻訳文書数は5210、うち公開文書数は2282にのぼっています。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、GlobalVoicesOnline日本チームといった著名なNGOやネットニュースグループが利用しており、グリンピースジャパンや日本母乳育児支援ネットワークも利用の準備を進めています。いくつかの大学のゼミでも英語記事購読紹介に使われています。2010年には「いつでもどこでも『みんなの翻訳』」を実現するために、自分が読んでいるページを翻訳したいと思ったときにすぐ「みんなの翻訳」の翻訳支援エディタを立ち上げることができる「QReditブックマークレット」を公開しました。

2010年10月には中国語に対応し、現在、対象となる言語は英語、日本語、中国語となりました。さらに、カタロニア語と英語の間の翻訳にもまもなく対応する予定です。

また、翻訳教育のプラットフォームとして「みんなの翻訳」を展開するサブプロジェクトも始まり、現在、リーズ大学翻訳研究所と共同で開発を進めています。さらに、長尾賞受賞をきっかけに、World Digital Library（http://www.wdl.org/）が提供している世界の文化遺産の多言語展開に「みんなの翻訳」を活用する可能性も現在検討されています。これに限らず、社会的に興味深いプロジェクトには積極的に関わっていきたいと思っています。

5. 言語スケープとメディアの生態系

私たちが日常的に接し、目にし、読み、聞き、書き、話し、触れ、感じる言語表現の光景は、近年大きく変わっています。多様な言語が当たり前に存在していると同時に、Google Translateをはじめとする機械翻訳が一般に広まったことからフランス語のようでフランス語でないもの、日本語のようで日本語でないもの（あるいはフランス語でなさそうだけどフランス語のもの、日本語でないようでありながら日本語のもの）等々も目につきます。

このような新しい言語スケープの中では「みんなの翻訳」が想定している翻訳対象文書は、どちらかというと保守的なかちっとしたものの側に属しています。言語スケープがどのように変容するにせよ、変容の起点となり変容を認識する基盤となるものは、近代のメディア生態系に支えられた言語編成であり、現在でもほとんどの言語表現活動はそれに対応するかたちで捉えることができるからです。新しい言語スケープとメディア生態系に対して「みんなの翻訳」をどのように展開していくかは今後の興味深い課題です。

6. おわりに

「みんなの翻訳」プロジェクトは、前身から数えて現在 6 年目になります。その間、研究成果も公表してきましたが、何よりもよかったのは、とにかく楽しくてたまらないという点でした。それが今回認められ、AAMT長尾賞をいただけたことは本当に光栄に思っております。ご推薦下さった方々、ご審査に関わった方々にあつく御礼申し上げます。これからも楽しく本プロジェクトを展開し、機械翻訳をはじめとする言語処理技術と言語表現を扱う翻訳とをいっそう緊密に結びつけることに貢献していきたいと考えております。

レポート

「N の学生」の訳し方

信州大学　加藤　鉱三
早稲田大学総合研究機構　黒田　航

1. はじめに

名詞二つが助詞「の」で結ばれている表現は日本語では非常に生産性が高い言い方である。だが，それを翻訳するのはとても難しい。それは，日本語の「N_1 の N_2」と英語の "N_1's N_2" もしくは "N_2 of N_1" はともに属格表現である点では同じであるのだが，しかし①英語と日本語では，属格表現の取り得る意味範囲が違う，②意味解釈は N_1 と N_2 の関係で決まるので，属格であるということ自体が意味を決定するのではない，という２つの理由による。本稿では「N_1 の N_2」を機械翻訳ソフトがどう訳すのかを概観し，訳し方の改善策を提案することを目的とする。ただし「N_1 の N_2」では対象が広すぎるため，単純化のために N_2 を「学生」に固定して議論する。

2. 翻訳ソフトの英語訳

今回の考察で用いた翻訳ソフトは市販されている次の３点である。

[LV]　『LogoVista 2009 PRO』，LogoVista.
[Atlas]　『ATLAS V14 翻訳スタンダード』，富士通.
[PC]　『PC-Transer 翻訳 Studio 2009』，クロスランゲージ.

以下に興味深い訳出文を示す。英語訳の前にある記号は，○が「訳として使える」，×が「訳として使えない」ことを表わす[1]。ただし，本稿では冠詞や数の不自然さは指摘せず，主に前置詞の選択を考察する。

(1) 加藤先生の学生
[LV]: ○Mr. Kato's student
[Atlas]: ○Mr./Ms. Kato's student
[PC]: ×A student of Mr. Kato

(2) 経済学の学生
[LV]: ○The student of economics
[Atlas]: ○Student of economics
[PC]: △An economic student

(3) 東京の学生
[LV]: ×The student of Tokyo
[Atlas]: ○Student in Tokyo
[PC]: ×A student of Tokyo

(4) 教室の学生
[LV]: ×The student of a classroom
[Atlas]: ○Student in classroom
[PC]: ×The student of the classroom

(5) 底辺の学生
[LV]: ×The student of a base
[Atlas]: ×Base student
[PC]: ×A student of the base

(6) 男の学生
[LV]: ×A man's student
[Atlas]: ×Man's student
[PC]: ×The student of the man

(7) フルタイムの学生
[LV]: ○A full-time student
[Atlas]: ×Student of full-time
[PC]: ×A student of the full-time

(8) SNS ユーザーの学生
[LV]: ×A SNS user's student
[Atlas]: ×SNS user's student
[PC]: ×The student of the SNS user

(9) アルバイトの学生
[LV]: ×Part-time job's student

[1] ○,×,△は Google 検索の該当数を参考にして決めている。

[Atlas]: ×Student of part-time job

[PC]: ×The student of the part-time job

(10) 欠席の学生

[LV]: ○An absent student

[Atlas]: ×Student of absence

[PC]: ×A student of the absence

(11) 不本意入学の学生

[LV]: ×The student of involuntary attendance

[Atlas]: ×Student of unwilling entrance

[PC]: ×A student of the unwilling entrance to school

(12) ピアスの学生

[LV]: ×The student of a pierced earring

[Atlas]: ×Pierce's student

[PC]: ×The student of pierced earrings

(13) 和服の学生

[LV]: ○The student in Japanese clothes

[Atlas]: ×Student of Japanese clothes

[PC]: ×The student of the kimono

3. 問題の所在

機械翻訳ソフトは一般に「N_1 の N_2」という入力に対し "N_2 of N_1" という出力を出す強い傾向がある。出力から推測する限り翻訳ソフトのアルゴリズムは次であろう。

(14) a.「N_1 の」が辞書で形容詞に対応するという情報が得られる場合，形容詞が選択される ((7), (8)の LV を参照)

b.N_1 が<人>である場合に "N_1's N_2" が選択される ((1)の LV，Atlas を参照)

c.そうでない場合に "N_2 of N_1" が選択される

(14a, b)は基本的に適切な戦略であると考えられるが，(3)から(12)の事例を見れば(14c)は十分と言うには程遠いことは明らかである。本稿では(14)の戦略を改善する方向性を模索することを目的とする。

4. 問題の分析

「N_1 の学生」という表現がどのように言い換えられるかという観点から，(3)から(13)は次のように分類することができる。

・(3)-(5)は「N_1 にいる学生」と言い換えることができる

東京にいる学生

教室にいる学生

底辺にいる学生

・(6)-(9)は「N_1 である学生」と言い換えることができる

男である学生

フルタイムである学生

SNS ユーザーである学生

アルバイトである学生

・(9)-(12)は「N_1（を）している（または「した」）学生」と言い換えることができる

アルバイト（を）している学生

欠席（を）している学生

不本意入学（を）した学生

ピアス（を）している学生

・(12)(13)は「N_1 を身につけている学生」という内容で，「つけた」「着ている」等の動詞で言い換えることができる

ピアスをつけた学生

和服を着ている学生

ここで３種類の言い換えを考えたが，これらの言い換えの共通点をまとめると次のようになる。

(15) N_1 - (助詞) - V - 学生

つまり，属格の典型的な用法である<所有>や格関係[2]はない(3)以降の「N_1 の N_2」においては，①どのような意味になるかは N_1 と N_2 の意味関係で決定され，②その意味関係は(15)の V で表示される。(3)-(5)は N_1 が<場所>であり N_2 が<人>であるため，助詞と V を「に

[2] (2)では勉強の対象としての格関係，つまり目的語としての関係が認められる。

いる」とした言い換えになる。(6)-(9)[3]は N_1 が<属性>であり N_2 が<人>であるため，助詞とVを「である」とした言い換えになる。(9)–(12)においては N_1 と N_2 の意味関係は一様ではないのだが，それにも関わらず，「N（を）している」という表現の多義性に依存した言い換えに依存した用法のまとまりを形成している[4]。

5. 改善の方向性

前節で私たちは「N_1 の N_2」を言い換え可能性の観点で同値類に分類し，この同値類を基盤にした翻訳の改善法を提案した。ここでは問題の一般的条件を考察する。

「N_1 の N_2」という表現の英語への翻訳が難しいのは，対応する英語表現が一様でないことによる。特に“N_2 P N_1”という前置詞を使った翻訳を考える場合，of以外が正解訳である時に，それをどうやって見つけるかは非常に難しい。そのための方策として，次を提案したい。

(16) Step 1：入力「N_1の学生」を
「N_1 - (助詞) - V - 学生」という連鎖に
言い換える
Step 2：「N_1 - (助詞) - V - 学生」
という連鎖を，
「学生は N_1 - (助詞) - Vする」
という文の形の連鎖に言い換える
Step 3：その文を入力にして訳出する
Step 4：その訳出文を，主語の
studentsの直後にwhoを挿入し関係
節構造にする

(16)は，「N_1 の学生」をそのままの形で英語にするのは非常に難しいが，対応する文の形ならばまだ現実的な要求として機械翻訳に期待できる，ということを言っている。それは過度に多義な「の」をそのまま訳すことを回避するための措置である。

ここで提案している訳すべき文の形は，前ページで言い換えとして示したものを，例えば「ピアスをしている学生」を「学生はピアスをしている」といったん言い換えたものになる。実はそのような文の形にしても，現状のままでは本稿で示した全ての事例で正しい訳文が出力されるとは限らないが，そのような文の形の訳文は本稿で問題にしている「N_1の学生」の訳とは独立に，どの道正しい訳に導く努力がなされなければならないものである。本稿の主張は，そのような努力を前提とし，「N_1の学生」という形の入力に対しても，そのような文の形における努力の果実をそのまま応用しよう，という提案であると言える。

6. 言い換えをどう選別するか？

以上の本稿の提案には次の重大な但し書きが必要である。

(17) 「N_1のN_2」の言い換えとしての「学生は N_1 - (助詞) - V」において，翻訳プログラムがVをどう選別するか，という課題が解決済みである

より具体的に言えば，私たちは「N_1 の N_2」を i)「N_1 にいる N_2」，ii)「N_1 である N_2」，iii)「N_1 を身につけている N_2」への言い換えという観点で分類したが，言い換えの可能性がこれらに限られないのは明らかである(例えば他にも「角の本屋」のように「N_1 にある N_2」と言い換えられる場合もある)。従って，問題を一般的に解決するためには，言い換え可能性を自動的に認識する手段が必要である。課題を解決するには次のような手法が考えられる。

(18) 「学生はN1 - (助詞) - V」という連鎖のV（と助詞の組み合わせ）のうち，参照する作業コーパスの中で最も生起数が高いものを選択する

---

[3] なお，(9)の<属性>の解釈は，例えば「そんなことは正社員にやらせることであって，アルバイトの学生に期待してはいけない」というような状況を想定している。

[4] 本稿は，ここで，「N1 の N2」の取り得る意味範囲について，「『N1（を）している』で言い換えられることが条件である用法がある」というかなり強い言語学上の主張をしていることになる。

(19) その V（と助詞の組み合わせ）は有限であり，可能な選択肢の範囲内で選択される

例えば「東京」という名詞はいろんな助詞と動詞の組み合わせを許すが，「東京の学生」という表現の言い換えを考える際には，それら全ての可能性を考える必要はないはずである。おそらく4節で考察した「にいる」「である」「をしている」の3種類（もしくは少数の他の選択肢）から選べばいいものと思われる。もしそうならば，コーパスで「学生は東京である」とか「東京をしている」等は検出されないであろうから，望まれる「にいる」が正しく選択されるだろう。この方式は，対象となるN_1（この例では「東京」）の辞書記述を詳細なものにする，という作業を回避している点に注目されたい。

7. まとめ

以上，「N_1 の N_2」という名詞表現を，N_1 を便宜上「学生」に固定して考察してきた。改善の方向性として，「N_1 の学生」をそのまま人力として機械に翻訳させるのは無謀であること，そしていったん文の形にしてからそれを英語にし，関係節化を経由して名詞表現に戻す，という方策を提案した。

Report

Enabling Multilingual Applications of 'Controlled Language': the DITA framework

Anthony Hartley
Centre for Translation Studies, University of Leeds, UK
a.hartley@leeds.ac.uk

1. Overview

A controlled language (CL) is a version of a human language that embodies explicit restrictions on vocabulary, grammar and style for the purpose of authoring technical documentation (Kittredge, 2003; Nyberg et al., 2003). With roots in the Simplified English of the 1930s (Ogden, 1930), the initial objective was to minimize ambiguity and maximize clarity for human readers, including non-native speakers of English, and so to avoid the need for translation altogether. Today it is widely acknowledged that the costs of post-editing the target documents can be cut significantly if the source documents for MT comply with CL rules (e.g., Bernth & Gdaniec, 2001; Roturier, 2009). This is apart from improved comprehension by readers of the source documents themselves (O'Brien, 2010). Nevertheless, adoption of the CL concept remains largely confined to English, notwithstanding developments in European languages, such as French (Barthe et al., 1999) and Swedish (Sågvall Hein et al., 1997). This paper suggests that the absence of wider uptake of the CL principle in other languages is due above all to an under-specification of the contexts in which rules should apply and the fact that existing CL rule sets are modelled on English. It further argues that the Darwin Information Typing Architecture (DITA) XML standard (dita.xml.org) provides an excellent framework for creating fine-grained CL specifications both for English and for other languages. A Japanese CL, for example, could help redress the current shortfall in MT quality out of Japanese relative to the quality achieved by MT between European languages.

2. Controlled Language Benefits, Principles and Limitations

Even before the advent of translation memory (TM), the adoption of CL authoring with subsequent MT was capable of reducing overall translation costs by 50%-70%, halving production time and yielding a return on investment (MT purchase and dictionary creation, CL training of authors) within two years (Pym, 1990). Greater consistency at the authoring stage can only increase the re-use of text segments within today's TMs.

CLs comprise a lexical component and a syntactic component. Within the lexical component, CLs distinguish the terminology specific to a particular domain of activity and even a particular company, and a common technical vocabulary in wide use across one or more sectors. The latter set typically numbers 1,000-1,200 items. Both sets are governed by the same principles of 'one term – one meaning' and 'one meaning – one term', to eliminate ambiguity and synonymy. Thus, for example, the AECMA standard (AECMA, 1995) allows the use of 'free' only as an adjective meaning 'moves easily' (e.g., *Ensure the fasteners are free.*); its use as a verb is proscribed in favour of 'release' (e.g., *Release the fasteners.*). In addition to providing definitions of prescribed terms, the lexical component of a CL usually offers a reverse dictionary which maps commonly used but deprecated expressions into prescribed terms. So 'replace' will point to both 'renew' and 're-install' (with their distinct definitions) in order to forestall a potentially dangerous confusion (e.g., *Replace the master cylinder seal.*). Of the two lexical sub-components, it is the common technical vocabulary that presents the greater challenge for attempts to define a CL; specialized terms are *de facto* less likely to be ambiguous.

The syntactic component may comprise between 30 and 60 rules, stated, depending on the particular CL, to varying degrees of specificity; thus it is not always possible to identify shared rules (O'Brien, 2003). It is possible, however, to distinguish a common underlying ambition, which

is to eliminate syntactic complexity and ambiguity by restricting sentence length and complexity. So, sentences are limited to 20 words and the maximum number of clauses is normally two, with constraints on the connectives that are used. Local dependencies are flagged by such devices as obligatory use of the complementiser 'that' and of pronouns (e.g., 'which'), while English noun/verb homonymy is resolved by the insertion of a determiner (e.g., *[V] Access the [N] data* vs. *[V/N] Access [N] data*). However, some widely enforced rules, such as the proscription of the supposedly ambiguous '-ing' form, have been shown to have very little effect on current MT performance, including into Japanese (Aranberri-Monasterio & O'Brien, 2009).

The fundamental problem with the current formulation of CL rule sets is that they are almost exclusively specified at the level of the sentence rather than at the level of the document (Hartley & Paris, 2001). This is because the notion of document element is itself very coarse-grained: the AECMA standard recognises only three document elements: procedures (instructions for use or action), descriptions and warnings/cautions. There is a very sparse linkage between syntactic rules and document elements: for example, the constraint of 20 words per sentence is relaxed to 25 in descriptions, while warnings are required to begin with a simple, clear command.

This narrowly syntactic focus biased to the structural peculiarities of English has hampered the development of CLs in other languages (Barthe et al., 1999). Generalising the principles of CL across languages requires a language-independent model of document structure that has been lacking hitherto. Such a model is necessarily functionally-oriented; it needs also to be sufficiently fine-grained to allow rules to be tied to functional elements that can provide context-sensitive guidance both to authors and to an MT system in its translation decision making. DITA, an XML standard for defining documents in terms of their functional constituents, affords just such a framework.

3. DITA as an Enabling Framework

DITA is an authoring tool which is increasingly widely used by technical communicators in Europe, North America and Japan. It defines an inheritance hierarchy of topic-types consisting of required and optional content elements, represented as DTDs and schemas and instantiated with chunks of information meeting a specific communicative goal. Figure 1 illustrates their functional nature.

```
<task id="zz">
<title> Creating a new ZZ file </title>
<taskbody>
  <context> Once ZZ is set up, you must create a new ZZ file. </context>
  <prerequ> You must be logged into the ZZ server. </prerequ>
    <steps>
      <step><cmd> In the editor, create a new file. </cmd></step>
      <step><cmd> Enter a query. </cmd>
        <choices>
          <choice> You can use ZZ syntax. </choice>
          <choice> Or you can use the query builder. </choice>
        </choices></step>
    </steps>
  <result> The server displays the 'file created' dialog. </result>
</taskbody> </task>
```

Figure 1: DITA functional elements

Preferred or required syntactic features can now be mapped into these element types, such as a gerundive '-ing' form for the initial word in the <title>. Similarly, <prerequ> (prerequisite) could

be made to require 'must', while <choice> could require 'can', with the second and subsequent choices beginning with 'or'.

The mappings will be language(-family) dependent, so that the corresponding requirement for <title> in many European languages would be for a nominalised form. CLs for different languages can be developed independently, while referencing a common functionally-defined framework.

Implemented in XML, the DITA framework is by design extensible. Inspiration for such extensions can be found in Natural Language Generation projects, many of which have focused on automating the production of technical documentation (procedural and descriptive) in multiple languages (e.g., Hartley & Paris, 1997; Danlos et al., 2000; Kruijff et al., 2000). It is open to users to define sub-sentential elements and require, for example, that the <goal> of a user action or a <condition> on the action must be stated before the <action> itself (e.g., <goal>*To increase the pressure.* </goal><action>*turn the screw to the right.*</action>).

The generic and language-specific constraints can in principle inform the decision-making of an MT system, whatever its architecture. The XML markup could be used to parameterise rules in RBMT and EBMT systems and exploited as features in the SMT paradigm (assuming sufficient training data). Upstream, at the point of authoring, the DITA DTDs could be visualised to authors in an interface similar to WYSIWYM (What You See Is What You Meant – Power et al., 2003) or, again assuming sufficient data, supported by an auto-complete or authoring memory function similar to that pioneered by TransType (Macklovitch, 2006).

4. Conclusions

Given that Controlled Language has been shown to provide a highly cost-effective solution in reducing MT post-editing effort, this short article has attempted to explain why the CL concept has failed to establish itself beyond English. It is not because other languages are inherently more complex and do not provide writers with the means for simple yet clear communication of technical procedures. It is rather that, hitherto, we have lacked a suitable language-independent, functionally-defined framework for articulating context-sensitive constraints on a particular CL and for defining the translation relationship between CLs in different languages. The article contends that DITA affords an excellent basis for this task and that useful constructs for extending DITA have already been piloted in the field of Natural Language Generation. Combining MT and post-editing is reported to yield up to a 400% increase in words-per-hour between European languages, yet only a 50% gain from Japanese to English (Marcu, 2010). Could a Japanese CL help to bridge this gap?

References

AECMA (1995) *A Guide for the Production of Aircraft Maintenance Documents in the Aerospace Maintenance Language AECMA Simplified English. PSC-85-16598.* AECMA: Paris.

Aranberri-Monasterio, N. & S. O'Brien (2009) "Evaluating MT output for -ing forms: a study of four target languages" *Linguistica Antverpiensia* 8: 105-122.

Barthe, K. et al. (1999) "GIFAS Rationalized French: a controlled language for aerospace documentation in French" *Technical Communication* 46, 2: 200-224.

Bernth, A. & C. Gdaniec (2001) "Mtranslatability" *Machine Translation* 16, 3: 175-218.

Danlos, L., G. Lapalme & V. Lux (2000) "Generating a controlled language" *Procs. INLG2000.* Israel.

Hartley, A. & C. Paris. (2001) "Translation, controlled languages, generation" In E. Steiner and C. Yallop (eds) *Translation and Multilingual Text Production.* Berlin: Mouton. 307-325.

Kittredge, R. (2003) "Sublanguages and controlled languages" In R. Mitkov (ed.) *Oxford Handbook of Computational Linguistics.* Oxford: Oxford University Press. 430-437.

Kruijff, G-L. et al. (2000) "A multilingual system for text generation in three Slavic languages" *Procs. COLING 2000.* Saarbrücken.

Macklovitch, E. (2006) "TransType2: the last word" *Procs. LREC2006.* Genoa. 167-172.

Nyberg, E., T. Mitamura & W. Huijsen (2003) "Controlled language for authoring and translation" In H. Somers (ed.) *Computers and the Translator*. Amsterdam: Benjamins.

Marcu, D. (2010) "Language technologies for Asian languages: Language Weaver" *TAUS Executive Forum*. Tokyo, April 2010.

O'Brien, S. (2003) "Controlling controlled English: an analysis of several controlled language rule sets" *Procs. EAMT2003 / 4th Workshop on Controlled Language Applications*. Dublin.

O'Brien, S. (2010) "Controlled language and readability" *Translation and Cognition* 15: 143-165.

Ogden, C. (1930) *Basic English* London: Treber.

Power, R., D. Scott & A. Hartley. (2003) "Multilingual generation of controlled languages" *Procs. EAMT2003 / 4th Workshop on Controlled Language Applications*. Dublin.

Pym, P. (1990) "Pre-editing and the use of simplified writing for MT" In P. Mayorcas (ed.). *Translating and the Computer 10*. London: Aslib. 80-95.

Roturier, J. (2009) "Controlled language for MT in action" *Procs. Translingual Europe*. Prague.

Sågvall Hein, A., Almqvist, I. & Starbäck, P. (1997) "Scania Swedish - a basis for multilingual machine translation" *Procs. Translating and the Computer 19*. London: Aslib.

レポート

Microsoft Translator の紹介

マイクロソフトリサーチ
相川　孝子

Microsoft Translator は、当社研究所で開発された統計学を利用した機械翻訳機で、開発当初は、社内のローカリゼーションプロセスの効率化を目的としていたが、現在では、ウェブ上で一般の人が無料で利用できる自動翻訳サービスや技術開発者たちが自由に使える API なども提供している。

http://www.microsofttranslator.com

本稿では、この Microsoft Translator の機能について紹介しながら、今後の機械翻訳のあり方などについて簡単に考察を述べてみたい。

Microsoft Translator ホームページでは、文章翻訳以外に次のような機能が体験できる。

- バイリンガルビューアーによるウェブページの対応翻訳
- 音声合成による翻訳文の読み上げ
- バイリンガル辞書による単語の辞書引き

最初のバイリンガルビューアーだが、これは自分が翻訳したいウェブサイトの URL をホームページのテキストボックス中にタイプすると、そのウェブサイトの対訳が自動的になされるというサービスである。このバイリンガルビューアーは、文レベルで対応翻訳を示してくれ、他言語で書かれたウェブページの内容をさらっと把握したいような場合には、大変便利である（図1参照）。

図1：バイリンガルビューアーによるウェブページの対応翻訳

http://russianfooddirect.com/store/4/food/grocery_page1.html

次に、音声合成による読み上げ、バイリンガル辞書による単語の辞書引き機能だが、最近こうした言語支援ツールを組み合わせる自動翻訳サービスが増えている。これは、機械翻訳が単に「翻訳」という領域にとどまらず,「外国語学習」のための支援ツールとしても利用できることを示唆する。また、音声合成、音声認識技術の開発などにより、機械翻訳が今後さまざまなアプリケーションで登場してくることは、間違いない。

Microsoft Translator は、当社の製品内にもいろいろな所に入っている。例えば、Office の様々なプログラム（Word, Outlook, PowerPoint など）内での翻訳機能とか、また、Internet Explorer 8 のアクセラレータでの翻訳アドオン（図 2 参照）の一つとしても使われるようになっている。当社製品内での使われ方については、次のサイトに記載されている。

http://windows.microsoft.com/ja-JP/windows-vista/Try-Microsoft-Translator

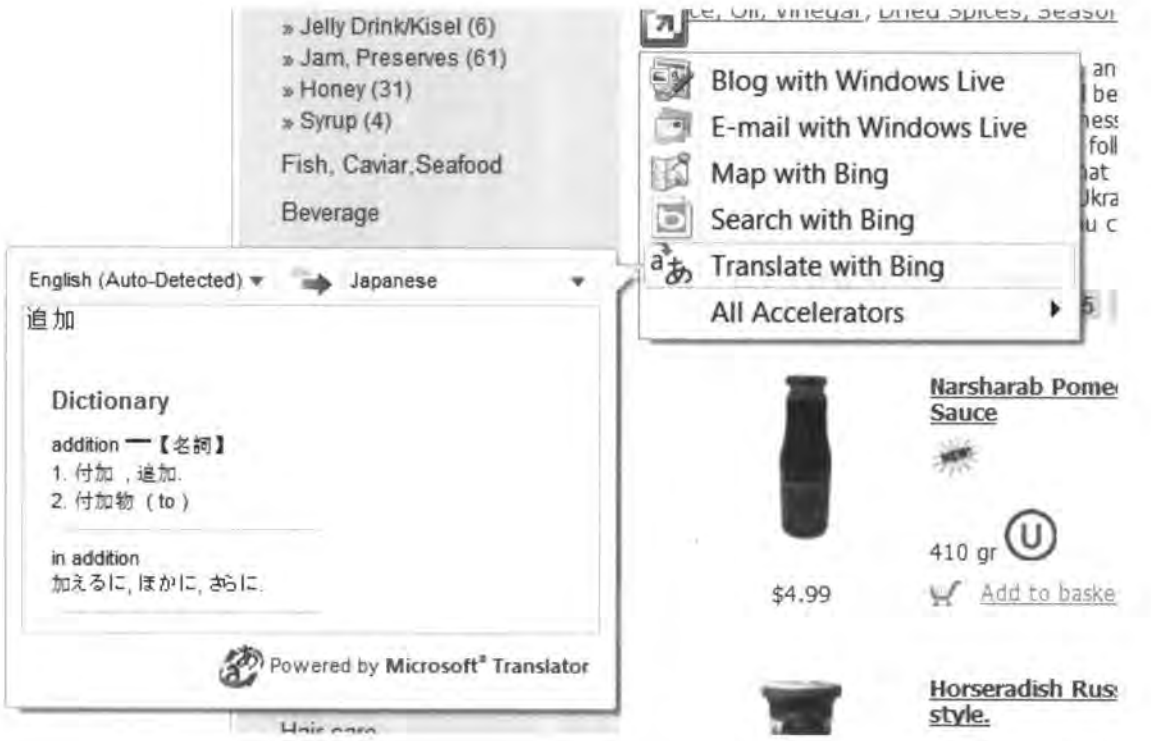


図 2 : Internet Explorer 8 のアクセラレータでの翻訳機能

開発を担当している機械翻訳チームでは、こうした既存する機能、また翻訳精度の向上、そして、対応言語の拡張などに取り組みながら、同時に新しい機能の開発にも努力している。その一つとして、最近公表された共同翻訳フレームワーク（Collaborative Translations Framework: CTF)がある。CTFは、「機械と人間の共同作業による多言語情報共有」という大規模なゴールを想定している。機械翻訳の精度が最近いくらかよくなったとは言え、誤訳は避けられず、機械翻訳は、人間の翻訳に勝てるものではない。しかし、我々が機械翻訳による誤訳をただ単に批判しているだけでは、前進できない。むしろ、「機械（コンピューター）と手に手をとって、翻訳精度を高め、多言語による情報共有を推進させていこう」という態度をもたなければ、この多言語による情報化社会には、対処できなくなってしまう。CTFは、こうした壮大なゴールを達成させてくれるユーザーインターフェイスを提供してくれるもので、次のようなプロセスで動く。

- Widgetのウェブページへの埋め込み
- ユーザーによる機械翻訳の編集
- 修正翻訳の確認

まず、ウェブマスターが自分のウェブサイトにWidgetを埋め込み、そのサイトを訪れたユーザーが自分の言語に機械翻訳。そして、誤訳があった場合は、CTFを使い、修正または、編集（図3参照)。こうしてユーザーから集められた翻訳は、ウェブマスターの承諾がなされれば、次回からそのサイトの翻訳文として使われるわけである。ユーザーが編集してくれた翻訳文を自分のサイトに使うことにより自然に翻訳の精度が高まり、ウェブマスターにとっても、またユーザーにとってもメリットが得られるというわけである。

図3：CTFの編集機能とWidget

CTF は、個人レベルのウェブサイトに限らない。例えば、当社は国立大学法人豊橋技術科学大学と協力し、CTF を使った多言語情報処理の発展寄与に関する共同研究などにも踏み込んでいる。また、アメリカの方でも地元の教育委員会などと協力し、CTF のプロジェクトを進めてきている（図４参照）。CTF がフルに浸透し利用されるのには、まだ時間がかかると思うが、人間が「機械翻訳はダメだ」と頭ごなしに拒否するのではなく、機械と協力しながら多言語で書かれているウェブ上の情報共有が可能になる時期がいち早くくることを期待する。

図 4 : Lake Washington School District での CTF 利用

http://www.lwsd.org/News/Pages/Connections-Newsletter.aspx

http://snow-white.cocolog-nifty.com/first/2010/05/microsoft-trans.html

プロジェクト・レポート

M^3（エムキューブ）とTackPad（タックパッド）
― 用例対訳を用いた多言語医療対話支援システム ―

和歌山大学
吉野　孝、宮部　真衣、福島　拓

1. 情報通信技術を用いた多言語対話支援

日本語で流暢なコミュニケーションのできない外国人が病気になった際に、病院内で必要なコミュニケーションをするためのシステムとして、多言語医療受付対話支援システムM^3（エムキューブ）と多言語用例対訳共有システムTackPad（タックパッド）を開発しています。

日本に在住する外国人は約220万人（人口の約2%弱）となり、2007年には外国人入国者数は900万人を超えました。労働人口の急速な減少にともない、日本は急速に多文化共生社会へ進んでいます。このような急激な社会変化にともない、就労、教育、医療などにおいて、様々な問題が顕在化してきています。

医療の現場において特に問題となるのは、言葉の壁です。外国人患者が医療機関を利用する場合，受付、問診、診察、会計など、あらゆる場所で言葉の壁に直面します。一部の医療機関において、医療通訳や多言語による医療情報の提供が実施されています。地域によって必要な言語は異なりますが、ポルトガル語（ブラジル）、スペイン語（チリ）、中国語の通訳のニーズが高くなっています。医療通訳者は、医師とのやり取りの通訳だけでなく、外来案内、医療費の説明や相談、手術の立ち会いなど、一人の患者に対して、医療機関の全ての場所における通訳に対応する必要があります。

医療通訳者が雇用されている病院は極めて少ないですが、NPOや国際交流団体などが地方自治体と連携し、ボランティアを派遣している例があります。しかし、多くの場合には、外国人患者の友人や知人など、身近な者が通訳として同行しています。通訳を友人や知人に依頼した場合、知られたくない個人的な情報を知られるなどの問題が発生していますが、十分な対応は行われていません。

近年、外国語での対応が困難という理由により、病院の外国人受け入れ拒否が発生し、それによるたらい回しの問題なども表面化してきています。

これらの問題を少しでも軽減するために、情報通信技術を使うことができないかと考えました。私たちは、NPO多文化共生センターきょうとと協働で、2004年から多言語医療対話支援システムの研究開発を始めました。

このレポートでは、実際に医療機関において利用されている多言語医療受付対話支援システムM^3と多言語用例対訳共有システムTackPad、Web上で公開している多言語問診システムM^3（Web版M^3）、現在開発中のスマートフォン上で動作するM^3（モバイル版M^3）を紹介します。

2. 多言語医療受付対話支援システムM^3（エムキューブ）

多言語医療受付対話支援システムM^3は、医療受付における対面対話支援や受診支援を行うためのシステムです。タッチパネルを用いて操作するシステムであり、次の6つの機能の提供により、外国人患者の受診を支援します。図1は、M^3の画面例です。

(1)　対話機能

医療従事者と外国人患者は、予め用意された質問文と回答文の組み合わせを用いて、対話を行うことができます。

(2)　問診機能

人体図を用いて症状のある部位と症状を選択することにより、患者は自分の症状を伝えることができます。

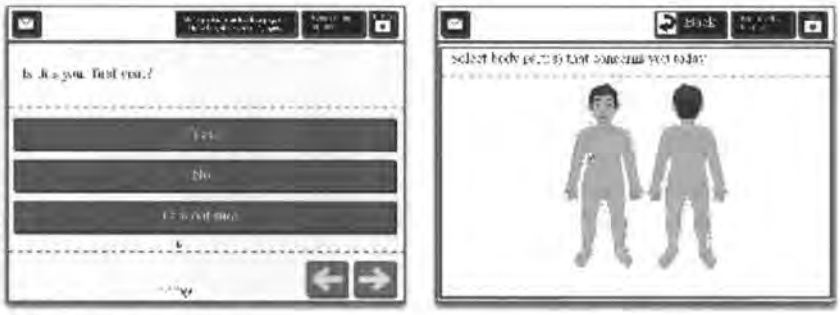
(3)　受診科選択機能

診察を受けたい受診科を選択し、医療従事者に伝えることができます。

(4)　道案内機能

病院内の目的地までの経路を、多言語で確認することができます。

(5)　Q＆A機能



（1）対話機能　　（2）問診機能

図1　多言語医療受付対話支援システムM³の画面例

病院に関するQ＆Aを確認することができます。

(6)　受診手続き支援機能

医療受付における定型化された質問の流れに回答していくことにより、必要な手続きを調べることができます。

このシステムでは、翻訳精度による問題の発生を防ぐために、医療通訳者により正確に翻訳された用例対訳（実際に使用される文例を多言語で集めたもの）を、言語グリッド（情報通信研究機構（NICT）を中心に開発されているインターネット上の多言語サービス基盤）を介して利用しています。しかし、事前に用意された用例対訳を利用しているため、自分の伝えたい内容の用例対訳が存在しない場合に、コミュニケーションができないという問題があります。

そこで、多言語医療受付対話支援システムは問題点のフィードバックの仕組みを備えています。図２は、システムで問題が発生した場合の連絡画面です。

用例対訳の不足や、用例対訳に間違いがあった場合、ユーザからシステム管理者へと連絡することができます。また、用例対訳不足のフィードバックについては、多言語での登録を可能とするために、音声データを用いて登録する機能も備えています。この機能は、多言語医療受付対話支援システムで用例対訳の不足が発生した際に、不足した用例を発言してもらい、その音声を録音することにより、不足用例データを作成し、サーバへと送信します。不足用例データは、用例対訳の収集・共有を行う多言語用例対訳共有システムへとフィードバックされます。不足用例データのフィードバックにより、多言語用例対訳共有システムにおいて、医療現場で実際に必要とされる用例対訳が作成される可能性が高まり、用例対訳を用いた対話の実現可能性を高めることができると考えています。

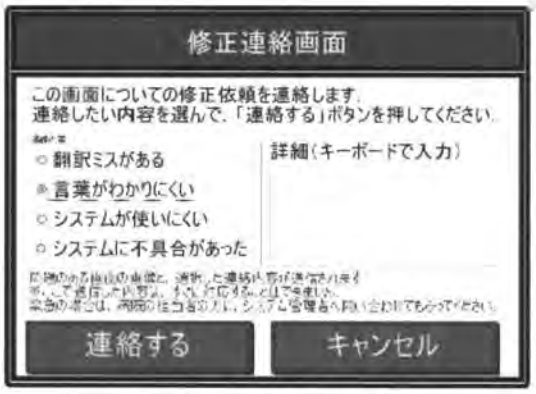


図2　M³の修正連絡画面例

図3　京都大学医学部附属病院に設置されているM³の様子

多言語医療受付対話支援システムM³は、現在、京都市立病院、京都大学医学部附属病院、洛和会音羽病院に設置されています。図３は、京都大学医学部附属病院に設置されているM³の様子です。

3. 多言語用例対訳共有システムTackPad（タックパッド）

多言語医療受付対話支援システムM³は用例対訳を使用しています。用例対訳の数が少ないと、コミュニケーションが制限されます。この問題を解決するために、

多様な用例対訳の収集と共有を行う多言語用例対訳共有システムTackPadを開発しています。

TackPadは、様々な言語を使う人々を利用者として想定しているため、画面の表示言語を他の言語へ切り換えられます。図４は、TackPadの画面例です。現在の収集言語は、日本語、英語、中国語、韓国・朝鮮語、ポルトガル語、スペイン語、ベトナム語、タイ語、インドネシア語の９言語です。

TackPadにおける収集と共有に関する基本機能は次の３つです。

(1)　用例の提案

医療従事者や患者などが他の言語に翻訳してほしい用例を提案する機能です。実際の用例の利用者がそれぞれの立場から提案するため、必要な用例を集めることができます。また、用例の提案には翻訳作業が不要なため、理解できる言語が１言語の利用者も用例対訳の収集に貢献することが可能です。

(2)　対訳の作成

「用例の提案」で提案された用例を翻訳する機能です。医療分野では正確な翻訳が必要なため、本機能は多言語話者の利用を想定しています。

(3)　用例対訳の検索

TackPad内の用例対訳を検索する機能です。本機能は、医療従事者や患者、翻訳者などすべての利用者が利用可能です。

TackPadを実際に運用したところ、初めのうちは用例を入力してもらえるものの、しばらくすると用例の登録が少なくなることがわかりました。これは、入力に対する十分なインセンティブがないことが理由として考えられます。そこで、用例登録数のランキング機能等を提供しましたが、入力状況はあまり改善しませんでした。そこで、用例登録を促すための工夫として、プロジェクト型用例収集支援機能と機械翻訳を利用した用例登録促進機能を開発しました。

(1)　プロジェクト型用例収集支援機能

図５は、プロジェクト型用例収集支援機能の画面例です。「用例の提案」では、用例を利用する場面を考えてから用例を提案します。しかし、場面を考えることは用例作成者にとって大きな負担になっていました。そこで、この機能では、用例登録者に収集する用例のテーマを提供することで、用例の登録を促しています。現在では、プロジェクトをきっかけとして新規用例が収集されるようになっています。

(2)　機械翻訳を利用した用例登録促進機能

図６は、機械翻訳を利用した用例登録促進機能の画面例です。図６の上部には、用例対訳が表示されていますが、図６の下部には、対訳が登録されていないときに、機械翻訳による翻訳結果が表示されます。もともとTackPadでは、医療現場を対象とした用例対訳の収集・共有を行っており、機械翻訳の利用は想定していませんでした。しかし、TackPadの利用者から、「最初から文を翻訳する作業は大変だけど、間違いのある

図4　多言語用例対訳共有システムTackPadの画面例

図5　プロジェクト型用例収集支援機能の画面例

文をみると修正したくなる」というコメントを受けて、この機能を開発しました。

近年、機械翻訳の精度は向上しており、機械翻訳による翻訳結果がそのまま利用可能な場合や、若干の改善で十分な精度になる場合もあります。但し、用例の精度を担保する必要があるため、人手による確認が必要です。そのため、TackPadの参加者にその確認作業に協力してもらうことを想定しています。

4. Web上で動作する多言語問診システムM³(Web版M³)

多言語医療受付対話支援システムは、医療受付に設置したタッチパネルを用いて操作を行うシステムです。そのため、システムの設置されていない病院において、外国人患者は支援を受けることができません。そこで、Web上で動作する多言語問診システムM³（Web版M³）を開発しました。図７は、Web版M³の画面例です。医療従事者に自分の症状を伝えるためのシステムをWeb上で提供することにより、システムの設置されていない病院における外国人患者の支援を行えると考えています。外国人患者は、Web版M³を用いることにより、自宅で自分の症状についての情報を多言語で作成することができます。作成した症状データは、印刷することができます。

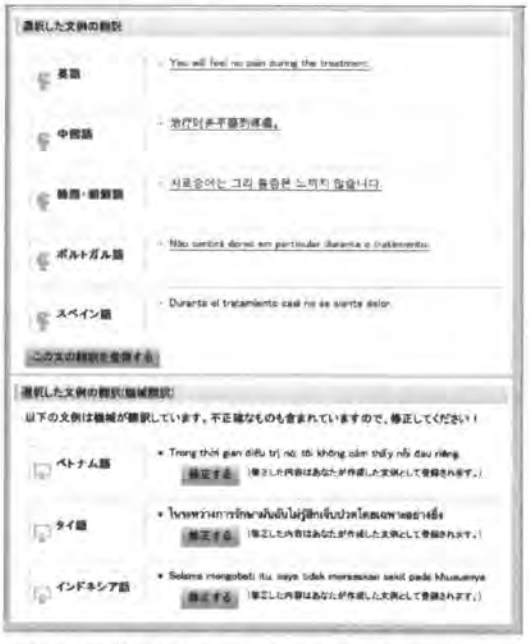


図6　機械翻訳を利用した用例登録促進機能の画面例

図８は、作成した症状データの印刷画面例です。症状データは、ユーザ自身の言語およびユーザが選択した翻訳先言語の2言語で併記されており、症状データを印刷して持っていくことにより、医療機関において自分の症状を正確に伝えることができます。また、利用時にユーザはコメントを登録することができ、登録したコメントは開発者へフィードバックされ、今後の開発に利用されます。

図7　Web上で動作する多言語問診システムM³の画面例

図8　症状データの印刷画面例

図9 モバイル版M³の画面例

5. スマートフォン上で動作するM³(モバイル版M³)

スマートフォン上で動作するM³は、Web版M³をスマートフォン上で実装したものです。Web版M³を、利用するためには、PCが必要となります。また、病院等へ行く前に印刷が必要になっています。そこで、利用者自身のスマートフォン上で利用可能になるとさらに便利になると考え、モバイル版M³を開発しました。

図9は、モバイル版M³の動作中の画面例です。モバイル版M³は、Android上とiPhone上で動作します。現在、Web版M³と同等の機能を持っていますが、公開に向けてスマートフォンの特徴を活かした新しい機能を開発中です。

公開サイト

- 多言語医療受付対話支援システムM³は、希望病院への導入が可能です。詳細は、次のサイトに情報を示しています。
  http://sites.google.com/site/tabunkakyouto/m3/
- Web上で動作する多言語問診システムM³（Web版M³）は、次のサイトで無料公開しています。
  http://sites.google.com/site/tabunkam3/
- 多言語用例対訳共有システムTackPadは、次のサイトで無料公開しています。また、用例登録のボランティアを求めています。
  http://med.tackpad.net/

謝辞

M³（エムキューブ）とTackPad（タックパッド）は、多くの皆様に支えられて研究開発を進めております。ここに記して謝意を表します。

多文化共生センターきょうとの重野亜久里氏、前田華奈氏には、多言語医療受付対話支援システムM³と多言語用例対訳共有システムTackPadの開発において、開発当初から現在に至るまで、システムの設計や改良、多言語化において、多大な協力を頂きました。

京都大学の石田亨先生には、研究開発を進めていくにあたり、多くの助言を頂きました。情報通信研究機構（NICT）の言語グリッドプロジェクト関係者には、開発において多くの技術協力を頂きました。

M³とTackPadの開発の一部は、総務省戦略的情報通信研究開発推進制度（SCOPE）の助成により行われています。

M³の導入病院の関係者の方には、貴重なコメントを多数頂きました。TackPadのボランティアの方には、多数の用例対訳を登録して頂きました。

TM Business

New and innovative products from SDL and a pioneering open platform

Yasuo Arai - SDL Japan K.K.

SDL is the leader in Global Information Management (GIM) solutions that empower organizations to accelerate the delivery of high-quality multilingual content to global markets. Its enterprise software and services integrate with existing business systems to manage the delivery of global information from authoring to publication and throughout the distributed translation supply chain.

SDL has developed a pioneering open and unified environment for the visionary GIM Platform™ based on the scalable and open architecture of SDL Common Enterprise Application Framework™ (CEAF). The GIM Platform will hold all future SDL technology products.

In addition to this, SDL is releasing three revolutionary new desktop technology products:

New in Translation Memory
- SDL Trados™ Studio 2009. Innovation Delivered

New in Terminology Management
- SDL MultiTerm® 2009. Because Brand Matters

New in Software Localization
- SDL Passolo™ 2009. Designed with Software in Mind

What is Translation Memory?

A translation memory is a linguistic database that continually captures your translations as your work for future use.

All previous translations are accumulated within the translation memory (in source and target language pairs called translation units) and reused so that you never have to translate the same sentence twice. The more you build up your translation memory, the faster you can translate subsequent translations, saving time and money.

SDL Trados Studio 2009. Innovation Delivered.

Since the acquisition of Trados in 2005 by SDL, there has been anticipation in the market about a unified translation memory software product. SDL Trados Studio is the culmination of 4 years of research, development and significant financial investment and not only combines the best of both SDLX and SDL Trados, it is the next generation translation memory software.

SDL Trados Studio combines decades of translation technology experience with new and innovative features, meaning it is the most revolutionary software on the market today. With one integrated environment for all translation, review and project management needs, it offers radical new features on an open, standards-based platform. SDL Trados Studio significantly enhance productivity and maximize performance throughout the translation supply chain.

Key new features include:

- **AutoSuggest™** - Benefit from sub-segment matching by using your translation memories further. Maximize your assets and watch productivity skyrocket!

- **Automated Translation** - SDL Automated Translation Solution, SDL Language Weaver and Google Translate are fully integrated into your translation environment.
- **RevleX™ translation memory (TM) engine** - A ground-breaking TM engine which delivers a host of new time-saving features such as Context Match, improved concordance searching, plus many more.
- **Context Match** - Provides *"beyond 100%"* matches by recognizing location and context to deliver the best translation. No complicated set-up or configuration required!
- **QuickPlace™** - All formatting, tags, placeables and variables at the translator's fingertips. Smart suggestions based on source content make translating any file type a breeze.
- **Real-time Preview** - See the final document as you translate! Currently available for DOC/DOCX, HTML and XML. On-demand preview available for PPT/PPTX. Traditional preview available for all other formats.
- **Open standards for open communication** - Even more recognized industry standards such as XLIFF (bilingual files), TMX (translation memories) and TBX (terminology databases).
- **Support for PDF files** - You asked, we listened! A new filter so that you can accept PDF documents when the original source files are not available.
- **One integrated application for all your tasks** - Translate, review and manage your projects all in the same place. Fully customizable to suit your needs enabling you to work more productively.
- **Multiple translation memory sequencing** - Access multiple TMs at the same time and control how you work with them for ultimate flexibility.

For more information on SDL Trados Studio 2009 please go to
http://www.sdl.com/en/language-technology/

What is Terminology?

Terminology is the study of terms and their use. Terms are words and phrases which describe products, services or industry jargon. They frequently drive competitive differentiation. Most companies use an increasing number of industry- or organization-specific words which need to be accurately stored, shared and translated. Terms could be anything from a product name to a marketing tagline.

Terminology management is growing in importance as organizations are growing globally and looking to convey a unified brand message across the globe, but in local languages. The incorrect usage of terminology can lead to inconsistent company branding and ultimately leads to poor customer satisfaction. It is vital that both content creators and translators manage and share terminology to achieve this consistency and accuracy in customer communications.

SDL MultiTerm 2009. Because Brand Matters.

SDL MultiTerm 2009 is the new terminology management software from SDL. Built on SDL CEAF, it provides one central location to store and manage terminology and integrates with both the authoring environment and SDL Trados Studio. By providing access to all those involved with applying terminology, including engineers and marketing, translators and terminologists, it ensures consistent and quality content and branding from source through to translation.

What is Software Localization?

Software localization is the process of adapting a software product to the linguistic, cultural and technical requirements of a target market. This process is labour-intensive and often requires a significant amount of time from the development teams. Traditional translation is typically an activity performed after the source document has been finalized. Software localization projects, on

the other hand, often run in parallel with the development of the source product to enable simultaneous shipment of all language versions.

SDL Passolo 2009. Designed with Software in Mind.

SDL Passolo 2009 is specifically designed with the software localizer in mind. Providing one visual environment for software localization, it enhances the speed, quality and efficiency of the localization process. This latest version is easy-to-use, requires no programming experience, and is the fastest version of SDL Passolo ever thanks to QuickIndex™ technology. It enables software companies to accelerate the delivery of their products to global markets and helps them achieve a simultaneous global release.

The real power of SDL Passolo is its tight integration with the SDL Trados translation environment. This ensures maximum leverage of previously translated content through translation memory, centralized terminology use for brand consistency and the ability to plug-in to enterprise-wide solutions such as SDL Translation Management System®.

What is Automated Translation?

Automated Translation is the translation of text by a computer, with no human involvement. Pioneered in the 1950s, automated translation can also be referred to as machine translation, automatic or instant translation.

How does machine translation work?

There are two types of machine translation system: rules-based and statistical:

- Rules-based systems use a combination of language and grammar rules plus dictionaries for common words. Specialist dictionaries are created to focus on certain industries or disciplines. Rules-based systems typically deliver consistent translations with accurate terminology when trained with specialist dictionaries.
- Statistical systems have no knowledge of language rules. Instead they "learn" to translate by analysing large amounts of data for each language pair. They can be trained for specific industries or disciplines using additional data relevant to the sector needed. Typically statistical systems deliver more fluent-sounding but less consistent translations.

When would I use machine translation?

When translating with SDL Trados Studio, any segments not leveraged from translation memory can automatically be machine translated for the translator to review, then accept and amend if necessary, or decide to manually translate instead.
As a translator, you can configure how much machine translation is used and which machine translation engines to use.

Respecting client confidentiality

If the projects you work on are commercially sensitive, your customer may require that information is not disclosed to any third parties. Carefully consider how and when to use machine translation as you will be sharing segments of the source with a third party.
Audit files are automatically generated by SDL Trados Studio 2009 which record the use of machine translation.

What machine translation can I use?

SDL Trados Studio 2009 supports 3 machine translation engines that are available over an internet connection – SDL Enterprise Translation Server, Language Weaver, and Google

Translate. By default, these provide a free of charge, baseline (untrained) translation for generic rather than specific domains. Language pair support varies by machine translation engine.
With SDL Enterprise Translation Server™ and Language Weaver, it is also possible to use your customer's trained machine translation systems to deliver even faster.

What are the benefits of using machine translation?

Increased productivity – deliver translations faster

- Use freely available machine translation to pre-translate new segments that are not leveraged from translation memory.
- As a translator, you can connect to and use a customer's or supplier's trained SDL Enterprise Translation Server or Language Weaver machine translation system to deliver even faster.

Flexibility and choice – to suit all types of project

- Select from 3 different machine translation engines
- Choose from over 50 languages and more than 2,500 language pairs to suit your project
- Compare rules-based and statistical machine translation engines

What are the differences between SDL Automated Translation System, Language Weaver and Google Translate?

SDL Trados Studio 2009 supports 3 machine translation engines that are available over an internet connection – SDL Enterprise Translation Server, Language Weaver, and Google Translate. By default, these provide a free of charge, baseline (untrained) translation for generic rather than specific domains.

- The SDL Enterprise Translation Server ensures that the text sent to it for translation is not shared, pooled, or used for any other purpose.
- Language Weaver provides a secure (HTTPS) connection to their service and anonymously uses the data it receives for Language Training.
- Language-pair support varies by machine translation engine.
- SDL Enterprise Translation Server is also available for hosted or on-site deployment that can be trained for a particular customer or domain.
- Language Weaver is also available on a Software-as-a-Service basis that can be trained for a particular customer or domain.
- SDL Trados Studio can be configured to access a trained SDL or Language Weaver system instead of the default system.

	SDL	Language Weaver	Google Translate
Type	Rules-based	Statistical	Statistical
Content	Secure (on location + SaaS)	Secure SaaS	Non-secure (public)
Languages	16 languages 26 language pairs	75 language pairs	51 languages 2,550 language pairs
Audit Trail	Yes (SDL XLIFF)	Yes (SDL XLIFF)	Yes (SDL XLIFF)
Access	Via Studio	Via Studio	Via Studio
Cost	None	None	None

SDL Acquires Language Weaver, Affirming its Leadership in Machine Translation and Global Information Management

The acquisition of Language Weaver not only delivers best-of-breed automated translation technology into SDL's Global Information Management Platform, it does much more. Integration of secure machine translation technology into the translation supply chain at all levels will allow enterprises and governments to translate significantly larger volumes of content faster and more efficiently to meet the needs of the vast content in today's increasingly online world.

The amount of content available through the web today has increased dramatically and includes marketing content, manuals and support content as well as user generated content such as blogs, tweets. The need for people to read this vast amount of content in their own language is greater than the availability of human translators. Machine translation, as part of an overall strategy for creating and managing multilingual content, is the solution to that problem. The acquisition of Language Weaver's machine translation technology puts SDL firmly in place for ensuring the effective provision of secure multilingual content into the future's digital age through:

- Language Weaver's success in government, combined with SDL's blue chip client portfolio and their collaboration with the translation community
- SDL's leadership in content and language technology solutions, combined with Language Weaver's authority in machine translation technology

"Only a small percentage of content is translated today," said Mark Lancaster, Chairman and CEO of SDL. "The digital universe is set to rise 10 fold in the next five years and this expansion will be across the globe. Research has proved that internet users are significantly more likely to read and react to content in their own language. However there are simply not enough translators in the world to translate the text we need in local language at the speed and quantities required. We believe machine translation will become an integral part of companies' content creation and management strategy. Within the next 5 years we expect over 30% of all translated content to utilize machine translation in the process of translating one Language to another. We regard Language Weaver as the best-in-class machine translation technology available in the world today. Integrating secure and customized machine translation technology into SDL's Global Information Management technology stack positions SDL well to support our customers in creating global content in the future."

"While Google Translate has set the standard for ad-hoc translations by consumers, we have found that most enterprises want to own their automated translation technology," said Mark Tapling, President and CEO of Language Weaver. "When using Language Weaver, your content remains secure and confidential; it follows a workflow for translation and easily integrates into your other systems. It can also provide quality ranking and trained systems so that you have a trusted level of quality that respects things like corporate branding rules and consistency of

translations. The Language Weaver R&D team has continuously pushed the boundaries in statistical machine translation research while developing human communication solutions for enterprises and governments. The SDL acquisition significantly expands Language Weaver's ability to address the problem sets that the team tackles, to bring to market unique high-value machine translation products and solutions."

Mark Tapling went on to say, "The combination of our technology and SDL's translation and content management technologies gives enterprises a unique opportunity to intensify their engagement with prospects and customers. This is very much in line with SDL's vision of Global Information Management."

Today, automated translation only represents around 1% of the total translation market (estimated at $10 to $15 billion – source IDC), but market analysts expect that both the total market, as well as the market share of automated translation will continue to grow at a substantial rate. SDL has seen that automated translation allows its customers to lower translation costs by 30% to 50%, while at the same time cutting time-to-market for translated content by more than 50%.

For more information on the acquisition please visit www.sdl.com/languageweaver

MT 活用事例

ローカル主導での機械翻訳の導入と今後の展望

CA Technologies
菊池　邦明

1. 経緯

筆者の所属する部署では、社内の翻訳やローカライズ関連業務を他の部署から請け負っている。対象とする分野は、営業、マーケティング、教育、製品ローカライズなど多岐にわたるが、分量としては製品のローカライズが最も多い。

これらの翻訳対象コンテンツが増加する一方で、コスト削減も同時に達成する必要性に迫られ、弊社でも英語からターゲット言語への機械翻訳の導入を検討するようになった。日本市場向けの翻訳ニーズは他国市場向けよりも高いという事情もあり、弊社では日本語は重要な言語の一つとして考えられている。

2. 初回の取り組み

弊社が最初に機械翻訳の評価を真剣に検討し始めたのは 2005 年にまで遡る。このときの対象言語はドイツ語、フランス語、日本語の 3 つで、米国本社が選定した機械翻訳エンジンに対して各国語の翻訳チームがトライアルを行うことになった。この機械翻訳エンジンは複数の言語をサポートし、元来欧州言語での対応から発達したものではあったが、日本語もサポートされていた。米国本社の指示に従い、各言語とも足並みをそろえながら機械翻訳辞書の作成などを行った。

訳出結果の品質が直接的な理由ではなかったが、この最初の取り組みは 2006 年の年頭に中止されることになった。この初回の取り組みでは、以下のような問題点が見られた。

- **訳出される日本語の品質が低い**
  日本語として正しい構造になっていない、必要な助詞がない、など。
- **日本語への対応が十分でない**
  辞書における訳出先言語（日本語）側の設定で、不要な前置詞に関する属性は存在するが、助詞に関する属性が存在しない、など。
- **辞書の登録方法がやや複雑**
  動詞の場合には、あるコードを記述して辞書登録を行うが、期待どおりの結果が得られない場合が多い、など。
- **日本語に関する議論がしにくい**
  翻訳エンジンのサポート窓口に対しては、英語で問題を説明する必要があったため、日本語の詳細な文法事項などを説明しにくい。
- **サポートによる解決が遅れる**
  翻訳エンジンのサポート窓口は、日本から離れたタイムゾーンにあった。質疑応答を繰り返して最終回答を得るまでに時間がかかる。

これらの事柄は、次に機械翻訳エンジンの選定を行う機会が訪れたときに考慮されることになる。

3. 新しい取り組み

弊社が再度機械翻訳の評価を検討し始めるのは 2007 年になってからであった。企業で使用するための機械翻訳エンジンの選定では、純粋な翻訳品質だけではなく、エンジニアリングおよび保守の側面（既存のコンテント マネジメント システムや翻訳メモリ システムとの統合のしやすさ、管理の手間の削減など）も重要になってくる。しかしながら、弊社では前回の結果をふまえ、今回はローカル（日本）の翻訳チームが翻訳エンジンの選定や評価を主導することになった。つまり、実際に翻訳品質の責任を担う翻訳チームからの視点に重点を置いた翻訳エンジンの選定を行うことになった。

翻訳エンジンの第一要件として、弊社内で翻訳サーバとして使用できるものという事項があった。候補の選定にあたっては、主に Web サイトから情報を収集

し、評判のよい翻訳エンジンを中心に電子メールで問い合わせを行った。実際に製品を借用して翻訳品質や辞書管理についてより詳細な評価を行った翻訳エンジンは 4 つである。総合的に判断した結果、弊社では東芝ソリューション株式会社の The 翻訳サーバ™ Enterprise Edition を導入した。

また、今回の取り組みでは、米国本社ではなく日本に日本語用翻訳サーバを配置し、常にオンサイトでのサポートが可能なローカルの社内エンジニアを機械翻訳担当として割り当ててもらった。これらの結果、以下のような効果があった。

- **翻訳品質が向上した**
  翻訳元の英文が比較的長くても正しく解釈される場合が多い、助詞の抜けなどがほとんど発生せず自然な日本語が生成されやすい、など。
- **辞書の取り扱いが容易**
  複数の種類の辞書を活用することで多くのケースに対して翻訳品質を改善できる一方で、辞書の登録はそれほど複雑ではなく、すべて GUI で属性を指定できる。意図したとおりの訳出結果が得られやすい。
- **日本語でローカルのサポート**
  日本語で日本語に関する問題を説明できるため、誤解が少ない。翻訳エンジンのサポート窓口がローカルにあるため、1 日に何度も通信でき、解決に至るまでの時間が短い。
- **迅速なエンジニアリング**
  ローカルに翻訳サーバがあり、エンジニアのサポートも得られるため、直接話し合いながら新しいアイディアを実装して試すことができる。また、翻訳エンジンではまだ解決できない問題に対しても、詳細な対応ができる。弊社が開発した翻訳支援ツール上で機械翻訳を直接適用できるようになった。

他方で、予見されていたことではあるが、不便な事柄も存在する。弊社では日本語だけではなく他の言語または言語グループでも同様なアプローチで翻訳エンジンの選定を行ったため、各言語または言語グループの翻訳サーバは、基本的にそれぞれ対応するローカルオフィスに配置されている。これらのサーバの管理は、各言語または言語グループごとに行う必要がある。また、すべての言語に対応可能な機械翻訳エンジンであるなら、他の関連システムと統合するためのエンジニアリングは 1 回で済むであろうが、こちらに関しても存在するエンジンごとの統合エンジニアリングが必要になる。

4. 機械翻訳適用の流れ

製品のローカライズでは、翻訳対象として性質の異なる 2 つのコンポーネントがある。1 つはソフトウェア リソースと呼ばれるもので、ダイアログ ボックスなどの GUI やメッセージなどでユーザが見かけるものである。もう 1 つはドキュメントで、これには PDF のマニュアルやオンライン ヘルプが含まれる。

ソフトウェア リソースの翻訳では、弊社は社内で開発された翻訳支援ツールを使用しており、すでに大きな成果を得ている。現在は翻訳過程でこのツールから直接翻訳サーバに接続して機械翻訳の訳出結果を翻訳環境に反映できるようになっている。ドキュメントの翻訳では、現状ではまだソフトウェア リソースのようには翻訳サーバの訳出結果を翻訳支援ツール上に直接反映できるようにはなっていない。

いずれのコンポーネントの場合にもプリプロセスとポストプロセスが重要な役割を果たしており、これらの処理は実装方法によって 2 つの種類がある。

1 つは弊社のツールにソフトウェア的に組み込まれたプリ/ポストプロセスで、これらはツールのオプションによってそれぞれオン/オフを切り替えることができる。たとえば、弊社のソフトウェア リソースに出現する変数に対処する処理（図 1 参照）は、このプリ/ポストプロセスの一例である。この処理では、変数として使用される文字列が機械翻訳によって翻訳されないようにしている。

もう 1 つは翻訳者が任意にカスタマイズすることが可能な文字列置換テーブルによる文字列処理である。たとえば弊社のスタイルでは、「～しなければならない」という表現の使用は避けているが、文字列置換テーブルを使用して「～する必要がある」という表現に置換されるようにしている。文字列置換テーブルでは正規表現が使用できるため、多くのケースに柔軟に対応できる。

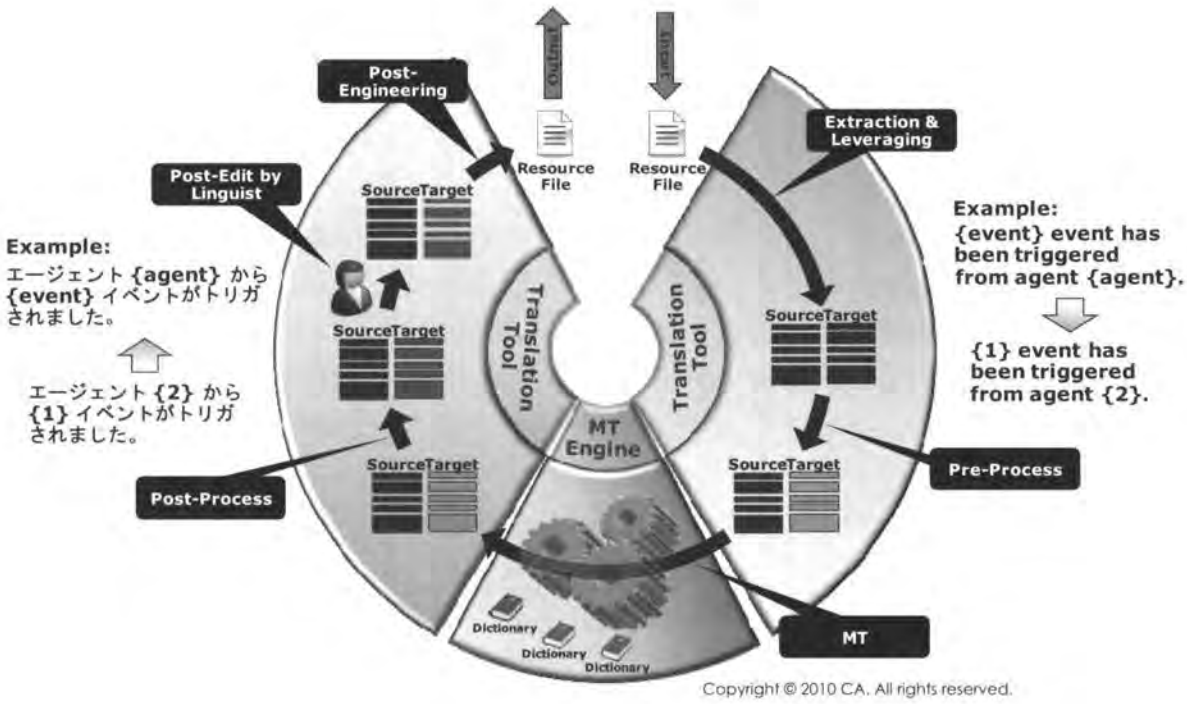


図 1： ソフトウェア リソースにおける機械翻訳適用の流れ

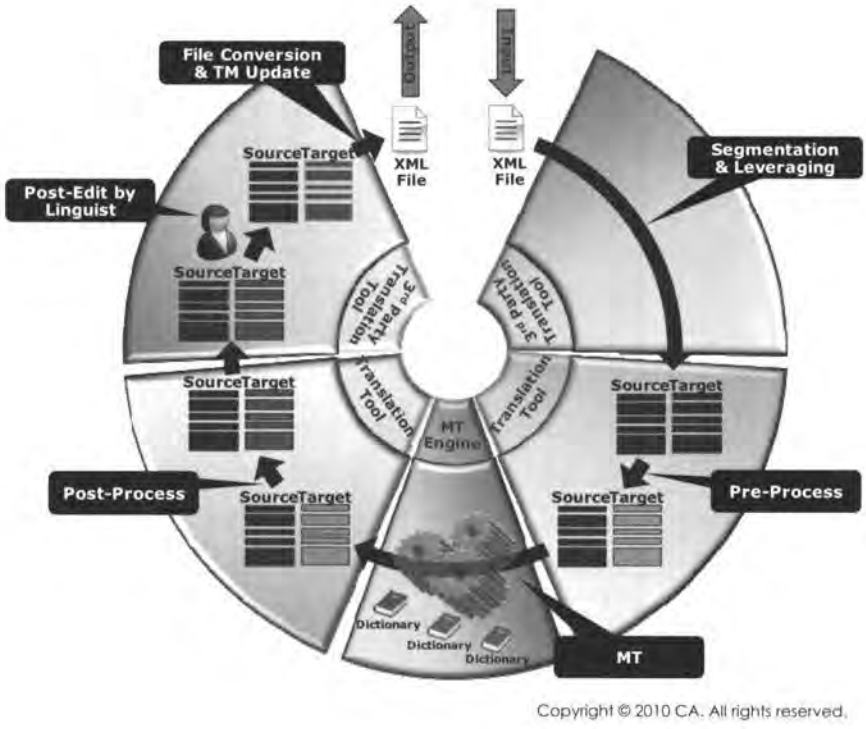


図 2： ドキュメントにおける機械翻訳適用の流れ

5. まとめと今後の展望

弊社における機械翻訳に対する新しい取り組みで有効であったと思われる点を以下にまとめる。

- 各言語ごとに最も有効な機械翻訳エンジンを選択した。
- ローカルの翻訳チームが機械翻訳エンジンの選定を行った。
- ローカルでのサポートが可能なエンジニアを配置した。

今後は、以下のような点で機械翻訳の改善に取り組む予定である。

- 原文となる英語を機械翻訳向けに最適化する。
- 機械翻訳辞書をさらに改善する。
- ポストエディットを請け負えるベンダーの数を増やす。
- 弊社およびTDA（http://www.tausdata.org/）の翻訳メモリの再利用性を高めるため、統計的な手法も採り入れた Advanced Leveraging を調査する。
- ポストエディットの成果を自動的に学習し、機械翻訳出力に適用できるようにする方法を調査する。まずは日本語から日本語への統計機械翻訳を利用した自動ポストエディットがどのくらい有効か検討する。

※The 翻訳サーバは、東芝ソリューション株式会社の商標です。

MT活用事例

機械翻訳＋ポストエディットの実証研究：先行研究レビュー

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

実務翻訳の世界では、近年、機械翻訳(MT)を翻訳支援ツール(CAT=computer assisted translation)として利用できるか否かの議論は非常に活発化してきている。翻訳メモリ(TM)とMTを融合したツールが実用化されたりと(SDL Tradosの日本語MT対応、OmegaTのGoogle Translateとの連動など)、翻訳支援ツールとしてのMT活用が非常に注目されている。

MTをCATとして利用する方法としては、大きく分けて前編集(pre-edit)と後編集(post-edit)があるが、実用的には後編集、いわゆるポストエディット（以下PE）が主流だ。PEは説明するまでもなく、機械翻訳の訳出結果を後から人の手によって修正することである。厳密にはPEにも種類があり、原文の意味が分かればよい程度に目標言語に仕上げるためだけの簡易的なRapid post-editingから、出版や実務のレベルまでに品質を上げるFull post-editingがある(Allen, 2003を参照)。

さて、ここで実際問題になるのは、機械翻訳のPEは、TMやツールを使わない翻訳よりも、質・効率ともに優れているのかどうか、という事だろう。PEを前提として、機械翻訳の訳は使えるのか使えないのか、実務翻訳者の間でよくある議論である。しかし実務経験に基づく様々な意見や議論が交わされている一方で、客観的かつ実証的な研究は、ほとんど行われていない。少なくとも、日本においては、それが研究環境の現状であろう。そもそも「翻訳」というものがアカデミックの場で、研究の対象として認知されていないことに起因する。本稿の趣旨とずれるので詳述は避けるが、欧州では、翻訳学(translation studies)という分野が学問的に認知されている歴史があり、翻訳や通訳研究が大学等の機関で行われている。日本でもようやく学術的に翻訳研究を行う大学院や学会が増えてきているものの、まだまだ発展途上である。筆者は今、立教大学大学院の博士課程で翻訳研究を行っているが、このような研究ができる学術機関の数はまだまだ少ない。今後の発展と普及を願いたい。

以上、前置きが長くなったが、本レポートは2回の連載を予定しており、1回目の今回は、機械翻訳のポストエディット(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する[注1]。海外での近年の研究結果を鑑みると、諸国語の組合せでは、MT+PEは実用レベルに達しているようである。次回のレポートでは、先行研究に基づき英語・日本語の組合せで、筆者が行った実験結果を掲載する。では、先行研究の詳細を見ることにしよう。

ALPAC報告書(1966)

MTの歴史において色々な意味で有名なALAPC報告書ではあるが、この報告書中にもポストエディット(PE)に関する記述があるので(Appendix 19)、その検証結果から見てみよう。ALPACは、23名の被験者（プロとアマチュア翻訳者が混在）に対し、英語・ロシア語の（技術文書）のPEについて実証検証を行っている。最初のアンケート調査では、翻訳者がPEを行った主観的な感想を、ツールを使わない普通の翻訳の場合との比較で調べている。感想の結果は、8名がPEの方が楽であったと回答した一方で、別の8名はPEよりも普通に翻訳をした方が楽だと回答した。また6名はPEも普通の翻訳も同じであると答えた。この結果自体が特に示唆することは無いのだが、興味深いのは、普通に翻訳をした方が楽だと回答した8名の翻訳者のうち6名が経験の長い翻訳者であり、逆にPEの方が楽だと答えた8名中6名は初心者の翻訳者であった点である。

実際の翻訳速度の結果と照らし合わせても、面白い結果が見られた。熟練翻訳者は、普通に翻訳を行う速度とPEの速度はあまり変わらず、初心者の翻訳者はPEを行うことにより普通に翻訳する速度の1.5倍から数倍程度の向上がみられた。つまり、熟練の翻訳者にとって機械翻訳のPEはあまりメリットがないが、初心者にとっては翻訳効率の向上に非常に役に立つということになる。

しかし、PEの平均翻訳速度は被験者全体で均一化する傾向にあり、初心者がPEによって翻訳速度が上昇したとしても、熟練者が普通に翻訳する速度と大差がない。この結果のみから結論づけるならば、MT+PEは、プロの実務レベルではあまり役に立たなということになる。当時の機械翻訳の精度を考えれば、ある程度納得のいく結果ということだろうか。以下では、もう少し最近の研究を見てみることにしよう。

Krings (2001)

Krings (2001)は、機械翻訳のPE作業プロセスを検証した近年では最も意欲的な研究である。Think Aloud Protocolを用いて、翻訳者のPEプロセスと機械翻訳を使わない翻訳プロセスとの比較検証を詳細に行った。使用した機械翻訳はルールベース型(SystranおよびMetal)で、言語の組合せは英語から仏語／独語であった。

相対的な作業効率(時間)では、普通の翻訳よりもMT+PEの方が20%程度上昇した。興味深いのは、機械翻訳の訳出と(raw MT output)とPE後のテキストとの類似度(similarity level)を比較すると、4割弱程であったという報告である。つまりMT訳の6割近くが、PEにおいて修正されたことになる。Kringsの実験が行われた10年前の機械翻訳の精度はALPAC報告書当時よりも向上していると思われるが、それでも6割という修正量は、実用化のレベルとして考えるには、少し多いようにも見受けられる。そうであるとしても、MT+PEによって翻訳に要した時間が2割減少したのは、むしろ驚くべき結果であろう。英語・日本語で2割の効率アップが達成できるとすれば、実務では歓迎されるものかもしれない。

Bowker & Ehgoetz (2007)

翻訳の品質の評価は難しい問題だ。品質の定義の仕方しだいでは、作業効率も変わってくるからだ。Bowker & Ehgoetz(2007)の研究では、この品質をユニークかつ実践的に扱い、機械翻訳の検証を行った。

翻訳の品質は、Skopos（翻訳の目的）によっても左右されると考えられるが(Vermeer, 1989/2000等参照)、Chesterman & Wagner (2002:80)は、翻訳を「サービス（業）」と捉え、その品質を測るには顧客の満足度を調べるのも一つの方法になりうる、と提案している。これは、受容者評価(recipient evaluation)と言われ(Trujillo, 1999)、Bowker & Ehgoetzは、実務で重きの置かれる3要素＝CQD（コスト、品質、納期）と関連づけて翻訳の品質を評価した。

大学の事務関連業務で発生する文書の翻訳を検証対象とした。大学や企業のように予算と時間の限られた状況では、翻訳の需要があっても、その全てを外注できない。そこで安価でスピーディーなMT+PEを利用できないか、調査するのがこの研究の目的であった。

同じ原文に対して3種類の訳出物を用意し、翻訳のユーザーとなる大学教授に対してアンケートを実施した。3種類の翻訳とは、(1)翻訳者がゼロから翻訳したもの、(2)MT+PEしたもの、(3)機械翻訳のみを行ったもの、である。品質の順位は、当然、(1)>(2)>(3)の順になる。しかし、これにコストと納期の条件を加える。(1)が一番高価で納期も長い。これに対して(2)は、(1)の5ー10分の1程度。(3)は更に少なく100分の1程に設定した。数字の割合は実務の予想工数に基づいている。この条件において、ユーザーはどれを選択するのかが焦点だ。

結果は、(1)を選んだのが全体の32.3%、(2)が67.7%、(3)を選んだ人は誰もいなかった。つまり、使用目的が限定された翻訳であれば、7割弱の人はMT+PEの品質レベルで満足できるという。逆に、機械翻訳そのままでは、いくら安価かつ短納期であっても、実用レベルに達しないということである。また、(1)を選んだ3割の人は、人間の翻訳者による翻訳を必要としていたわけだが、この中には文学部や外国部学部など言語に関わる学部の教授らが多く含まれていたこともあり、言葉・言語に対する意識の違いが結果に反映されていた

と、Bowkerらは分析する。

この調査で実施されたPEはRapid post-editingなので、通常のFull post-editingよりも品質は落ちていたにもかかわらず、条件次第では実用化レベルになるというのは、非常に興味深い結果である。

Fiederer & O' Brien (2009)

では、コストや時間的の制限を設けずにポストエディットを行った場合の品質は、普通の翻訳に比べてどうなのだろうか。Fiederer & O'Brien (2009)は、翻訳品質のみに焦点を当てた評価を行った。Hutchins & Somers (1992, p.163)によると、翻訳品質には3側面ある。Accuracy（正確性）、clarity（理解しやすさ）、style(スタイル)である。これら3点から、O'Brienらは MT+PE後の翻訳の品質と普通の翻訳とを11人の評価者の評価結果に基づき調査した。

まず、clarity、すなわち、訳文が読みやすいかどうか等の基準では、PEも普通の翻訳もほぼ同等の評価であった。翻訳者が訳文の読みやすさに注意を払うのは当然であるが、PEでも普通の翻訳同様にclarityは修正されるようである。

次に、accuracyであるが、これはPEの方に軍配が上がった。Accuracyは訳抜けや原文への忠実性ということになるので、機械翻訳を使うと普通の翻訳よりもaccuracyについては良い結果になることが分かった。普通の翻訳では、意外と訳抜けや誤訳が多いということの証明とも言える。

最後のstyleは、目標言語の言語使用域などに適合した訳文となっているかという基準となる。Clarityとの違いが微妙ではあるが、styleの方は、短文評価でなく文章（テキスト）としての一貫性や、特定分野での言い回しに準じているかという要素が評価対象になる。結果は、普通に翻訳をした方がPEよりも優れていた。

以上、品質3要素の比較をまとめると、clarityはPEも普通の翻訳も同等、accuracyはPEが有利、styleでは普通に翻訳をしたほうが有利ということであった。品質の詳細分析では、PEも普通の翻訳もどちらも甲乙つけがたい結果であった。

しかし興味深いのは、総合的に判断してPEと普通の翻訳のどちらの品質が良かったか、と質問したところ、ほぼ全員の評価者が「普通の翻訳」が良かったと答えている。これは、詳細分析結果とは裏腹に、MT+PEの品質に対する悲観な結果とも解釈できる。しかしO'Brienらは、この理由として、styleの点数差が大きかった点を指摘し、評価者はstyleの要素を過大評価している可能性があることを示している。つまり、翻訳の品質評価を行う場合には、style的要素が重視されてしまい、それによりaccuracyやclarityが軽視されうるということである。逆に言えば、翻訳者自身も翻訳中にstyleに多大な注意を払っているということでもあり、PEの作業とは少し異なるのかもしれない。この点は、筆者が行った実験結果とも絡んでくるので、翻訳品質の評価方法とポストエディットのやり方を考える上で非常に重要となる要素であろう。

O' Brien (2006a)

機械翻訳にかける前に、原文に含まれる文法的曖昧性などを取り除いておけば、機械翻訳後の訳出精度が上がり、結果としてPEに要する労力が低減し効率アップにつながると、予想できる。O'Brien(2006a)は、この原文の前編集(pre-edit)に制限言語(controlled language)を使用することによって、その後に生成される機械翻訳の結果をPEすることにより、どの程度の効率化が図れるかを調べた。

O'Brienは、この「PE効率」を、時間的(temporal)、技術的(technical)、認知的(cognitive)側面から検証した。IBMのWebsphere(ルールベース型)を使用して、制限言語で書き直した原文（前編集有り）と書き直さない原文（前編集無し）とを機械翻訳にかけ、それぞれのPEの作業効率を調査した。

時間的な処理速度(総ワード数÷所要時間)の比較では、予想通り、前編集した機械翻訳結果をPEしたほうが速かった。ただ、分節(segment)を個別に見た場合、前編集をした方が全ての分節で速かったかと言えば、そうでない箇所も観察された。O'Brienはこの理由を次のように説明する。PEを行う場合は、単語の位置などを変えるだけで良いことがある。この操作を行うために「カット&ペースト」機能を使えば効率が上がるが、翻訳者

（後編集者）の多くは、新たにキーボードから文字入力をしていた。入力作業は、認知的に負荷がからないからなのかもしれないが、このような冗長な技術的作業は、時間的な効率性からは無駄である。全ての分節で時間が短縮できなかった理由を、このような技術的操作が関与していたとした。しかし、原文に対応する訳語をキーボード入力するというプロセスは、ひょっとすると翻訳という基本行為となんらかの関係があるのかもしれないと、筆者は考えている。

さて、認知的な負荷の問題であるが、通常、翻訳者が翻訳の問題に直面すると、入力の手を止めて考えたり、調べ物をしたりと、訳出作業が一時中断する。つまり、一時中断(ポーズ)の割合が多ければその分だけ、翻訳者が難問に直面する割合が高くなり、認知的負荷も高くなると言われている。O'Brienは両方のPEのケースについて、ポーズの割合を調査したが、違いは全く見られなかった。実験参加者の実験後のコメントの中に、「PEは、普通に翻訳をするより疲れる」という感想が散見された。Kringsの調査でも指摘されていたことだが、PEは、原文と訳文を行き来する回数が増えるために、直線的な作業になりづらいらしい。つまり、前編集により機械翻訳の下訳の精度が上がったとしても、「PE」という作業の性質上、原文と訳文と照らし合わせるための認知負荷は、さほど変わらないのかもしれない。

いずれにしても、目に見える結果として、前編集とPEを組み合わせれば、時間的な作業効率が向上することは、この実験で実証されたといえる。

O' Brien (2006b)

MT+PEの作業が、実際に翻訳者の認知負荷にどのくらい影響しているのかを、技術的なキーボード入力のポーズの割合だけからでは観察不可能であることが先の実験から判明した。そこで、O'Brien(2006b)では、人間の瞳孔の動きと開き具合を測定できるアイトラッキング装置を用いて、PE作業に要する作業者の認知負荷を測定した。実験は、もともと翻訳メモリにおけるファジーマッチ(Fuzzy Match)のマッチ率と瞳孔の開き具合との相関を調査する目的で行われたのだが、メモリ内にMTの訳文も混ぜて行ったのが、この研究のユニークな点であった。

結果は、大方の予想通り、翻訳メモリの70%ー100%マッチ前後までは、マッチ率に従って瞳孔拡張は減少し続けた。またノーマッチ(No Match)で瞳孔拡張は最大になった。つまり、ゼロからの翻訳（ノーマッチ状態）では翻訳者の認知負荷が最も大きくなり、近似箇所を修正するだけの作業（ファジーマッチ状態）では、認知負荷も小さくなることが証明された。

この結果は想定内なのだが、特筆すべきは、機械翻訳のPEの作業における認知負荷が、予想以上に低かったという結果である。驚くことに、機械翻訳の修正作業（PE）で、瞳孔拡張は、85ー90%ファジーマッチとほぼ同等だったのだ。翻訳メモリを使ったことのある方なら想像できるだろうが、85%マッチの場合は、大抵、1つか2つの単語を入替える程度の修正作業でしかない。非常に単純な作業なので、認知的負荷が低いのは頷ける。これが機械翻訳のPEでも同じだということだ。つまり機械翻訳の訳出精度が高いということの裏付にもなる。この実験で使用された言語ペアは、英語→仏語／独語であった。英語と日本語の組合せならば、まだこのレベルにはならないだろう。

Guerberof (2009)

O'Brien(2006b)の結果を受けてGuerberof(2009)は、統計的機械翻訳(SMT)を使った英語→西語での、PEの作業時間と品質に関する追試を行っている。彼女の実験も、翻訳メモリのファジーマッチとSMTの訳文をメモリ内に混在させて比較検証を行った。結果は、機械翻訳を修正する場合の方が、翻訳メモリのファジーマッチを修正するよりも、時間と品質ともに優位であった。

この理由として、翻訳メモリの修正の場合は、（人間の）翻訳者の訳文が近似文（下訳）として表示されるため、文章がこなれていて自然であるために、差分箇所を見つけ出すのに時間が掛かってしまうというものであった。また品質（この場合は、誤訳や訳漏れがないという基準を用いた）についても、翻訳メモリの文章がこなれているために、訳抜けがあったとしても、見逃してしまうことがあると指摘された。これに対し

て機械翻訳の訳文は、ぎこちない直訳が多いので、原文と訳文の一対一対応が比較的容易になり、品質的にも有利になるというものであった。

MT+PEと、ゼロからの翻訳（ノーマッチ）との品質の比較では、僅かながらゼロからの翻訳が優勢であったものの、所要時間とのバランスを考慮した総合的評価では、MT+PEに軍配が上がる。つまり、英語→西語での翻訳は（分野が制限されるという条件はつくものの）もはやゼロから翻訳するよりも、そして翻訳メモリを使うよりも、MT+PEが一番良いということが言えるのだ。

実は、Guerberofの研究の動機は、翻訳者へのワード単価をいくらに設定すべきか悩んでいたことに端を発する。というのも、すでに彼女が働く翻訳会社では機械翻訳を導入しており、この実験のようにPEの作業だけを翻訳者に発注していたからだ。もしも、このGuerberof研究結果とO'Brien(2006b)の結果が採用されることになれば、MT+PEの作業は、翻訳メモリ85ー90%マッチと同じ単価、すなわち、通常のノーマッチの4分の1程の値段になってしまう。実際の効率はここまで向上はしていないので、この数字が額面通り使われることはないとしても、翻訳業界の横行する単価の値崩れは、この方面からも押し寄せていることを、改めて実感させられた研究結果である。

Garcia (2010)

これまでの実証研究は、英語とヨーロッパ言語の組合せであったが、Garcia(2010)は英語→中国語での検証を行った。時間と品質について、MT+PEとゼロからの翻訳とを比較した。品質基準にはNAATIの試験基準を使用した。

結果は、MT+PEもゼロからの翻訳も、どちらも時間、品質ともにほとんど変わりがなかった。これはまだアジア言語での機械翻訳の精度が、ヨーロッパ言語との組合せよりは、劣っているということを暗示しているのかもしれないが、仮にそうだとしても、機械翻訳を使うことが決してマイナスに働くことのないレベルまでは近づいているとも解釈することができる。

この研究の特徴は、実験にGoogle翻訳者ツールキットという環境を使った点にあった。Garciaが指摘するように、翻訳支援ツールの歴史は翻訳メモリの単体使用から機械翻訳との融合というように変容してきた。Googleが用意する翻訳者ツールキットでは、機械翻訳に主眼をおき、翻訳メモリは副次的にしか機能しない。また、このようなツールを使うことが、翻訳者の翻訳への考え方にも影響を与えている。

実験参加者に対して行った調査では、「Google翻訳者ツールキットを使った翻訳のほうが、使わないよりも翻訳しやすい」という意見が、実験後には増えていた。興味深いのは、そういった意見と実際のデータとの関係である。ツールを好むと述べた翻訳者の訳出物の品質は、ツールを使わなかった時の品質よりも優れている場合が多かった。逆に品質が悪くなったケースもあることはある。また「ツールを使わないほうが良い」と回答した翻訳者は、己を良く理解してか、その翻訳品質は、ツールを使うと確かに悪くなっていた。

テクノロジーに対する向き不向きはあるにせよ、全体としては、機械翻訳の活用を前向きに受け入れる翻訳者の数が上回っており、機械翻訳に敵対心を抱いていないというのは、翻訳の未来を考えるうえで何かヒントを与えてくれそうな結果だと思う。

まとめ

以上、まばらではあるが、MT+PEに関する文献を見てきた。このように欧州言語ペア間では、分野や使用目的をすれば、ほぼ実用レベルに達していることが実証データからも分かった。特にO'Brien(2006b)で示されたようにPEの認知負荷がTMの85%マッチ相当というデータは非常に衝撃的である。また、Garcia(2010)が言うように、若い世代ではPEという作業を積極的に受け入れる傾向があるものも面白い。ALPACの悲劇以前のようにMTに対し盲目的に期待を寄せるのではなく、PEという作業を介して、MTを見ていくことが、より現実的かもしれない、ということだろうか。

ということで、次回は、英語→日本語でのPEの実験結果を書きます。

【註】

（注1）本稿は，2010年9月1日発行の『翻訳通信』（発行人：山岡洋一氏）に掲載した記事に加筆および修正をしたものである。

【参考文献】

Allen, J. (2003) Post-editing. In H. Somers (Ed.), *Computers and translation: A translator's guide* (pp. 297-317). Amsterdam/Philadelphia: John Benjamins.

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=R1

Bowker, L. (2005). Productivity vs quality: A pilot study on the impact of translation memory systems. *Localisation Focus 4(1)*, 13-20.

Bowker, L. and Ehgoetz, M. (2007). Exploring user acceptance of machine translation output: A recipient evaluation. In D. Kenny and K. Ryou (Eds.), *Across Bound- aries: International Perspectives on Translation* (pp. 209–224). Newcastle-upon-Tyne: Cambridge Scholars Publishing.

Chesterman, A. and Wagner, E. (2002). Can Theory Help Translators? A Dialogue Between the Ivory Tower and the Wordface. Manchester: St. Jerome Publishing.

Fiederer, R. and O'Brien, S. (2009). Quality and machine translation: A realistic objective? *The journal of specialised translation, 11*, 52-74.

Garcia, I. (2010). Is machine translation ready yet? *Target, 22(1)*, 7-21.

Guerberof, A. (2009.) Productivity and quality in the post-editing of outputs from translation memories and machine translation. *Localisation Focus, 7(1)*, 11–21.

Hutchins, J. and Somers, H. (1992). *An introduction to machine translation*. London: Academic Press Limited.

Krings, H. (2001). *Repairing texts: Empirical investigations of machine translation post-editing processes* (G. S. Koby, Ed). Ohio: Kent State University Press.

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.

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.

Trujillo, A. (1999) Translation Engines: Techniques for Machine Translation, London: Springer.

Vermeer, H. J. (1989/2000). Skopos and commission in translational action. In L. Venuti (Ed.), Translation studies reader (pp. 227-38). London: Routledge.

AAMT会員のひろば

AAMT 会員の新たな交流の場を AAMT Journal 誌面上で提供するべくスタートいたしました「AAMT 会員のひろば」、会員の皆さまのご助力をいただきまして、第一回の No.41 のスタートから第七回を迎えることができました。今号では、法人会員一社、個人会員二名の皆さまからのご寄稿をいただいております。

独自のお取組みのご紹介、機械翻訳研究への提言、AAMT の活動へのご要望など、今回も貴重なご意見をお寄せいただきました。

AAMT Journal では今後も引き続き、会員の皆さまからのご寄稿を心よりお待ちしております。

ご寄稿・お問い合わせは AAMT 事務局(E-mail: AAMT-info@AAMT.info)まで宜しくお願いいたします。

法人会員（敬称略・50 音順）

会員名

成田　一（大阪大学大学院　教授）／NARITA Hajime　(Osaka University)

自己紹介

言語構造の解明が私の研究の出発点であった。65年頃から75年頃までは現代言語学の基盤となる言語理論「生成文法」が「標準理論」のほか「格文法」、「生成意味論」といった３グループに分かれ、おもちゃ箱をひっくり返したように実に多様な言語現象が構造操作面を中心に解明されたが、その後70年代後半頃からは理論の形式的な先鋭化に走り、（「照応」など）ごく限られた言語現象しか扱わなくなった。このため、84年頃から私は言語構造を明確な操作で扱う機械翻訳に関心を寄せ、日英語構造の対照研究に基づいて「機械翻訳における言語処理」をいろいろ提案するとともに「翻訳システムの解析・生成力の評価」を専門的に行い、日本電子工業振興協会の機械翻訳専門委員会ほかの学術委員を務めた。だが、94年頃からバブル崩壊の余波で電気通信系企業の多くが機械翻訳開発部門を閉鎖して行ったことから、機械翻訳に関する講演や新聞・雑誌・専門誌などでの啓蒙・執筆は続けつつも、徐々に英語教育に研究と活動をシフトさせ、近年は英語学や言語学（言語習得論など）、音声学といった基礎科学と脳研究など隣接科学の成果を英語教育に統合することを目指す英語教育総合研究会（HP参照）を主催し、「旬」なテーマのシンポジウムを年二回開催して問題提起・提案を行っているが、来年度には英語教育総合学会を設立する。

単著に『パソコン翻訳の世界』(講談社)、共著に『名詞』(研究社)、『日本語の名詞修飾表現』(くろしお出版)、『ことばは生きている』(人文書院)、『私のおすすめパソコンソフト』(岩波書店)、編著に『こうすれば使える機械翻訳』(バベルプレス)、『英語リフレッシュ講座』(大阪大学出版会) などがある。

社内英語化の誤謬と英語教育の蹉跌
Fallacy of "English as the Official Language in Companies" & Failure of "the Current English Education in Japan"

最近、大いに憂慮し、阻止しようと新聞（朝日新聞「私の視点」『英語の社内公用語　思考及ばず、情報格差も』2010/9/18）や雑誌（「社内英語と英語教育」『英語教育』大修館 12 月号 2010、『新英語教育』3 月号、4 月号巻頭エッセイ 2011）、シンポジウム（『英語運用力の底上げ』英語教育総合研究会 2010.12）などでキャンペーンを張っていることがある。何かと言うと、日本の経済界に「英語を社内公用語化」しようという動きがあることだ。私はそれが如何に多くの問題を孕んでいるかを、外国語で話す際の①「脳内処理の負担」と、それに伴う②「思考レベルの低下」及び③「発言の減少と劣化」、さらに④「情報共有の危うさ」など、本質的な欠陥に焦点を当てて指摘するとともに、「ゆとり教育」と「オーラル・コミュニケーション偏重」に起因する日本の英語教育の改悪と学力低下についても警鐘を鳴らしている。

グローバリゼーションを履き違えた社内英語

楽天、ユニクロが 2012 年を目処に「英語を社内公用語化する」と発表した。グローバル化に伴う海外展開を睨んで、「社内の会議、文書に英語を使う」というのだ。残念ながら、TV の報道番組のキャスターや新聞の社説なども、これが日本の企業が世界を相手に事業展開するにあたって目指すべき方向ではないかとする見解が多い。「英語の社内公用語化により、社員が話せるようになる」という幻想を抱いているのかもしれないが、社内英語化を押し付けたところで英語力の乏しい社員が「実務で英語が使える」ようにはならない。「グローバル化と言語の関係」についても勘違いしているのではないだろうか。「英語を使うこと＝グローバル化」ではない。（英米の旧植民地など）英語が生活に根付いた地域ならともかく、国内の本社は別だ。楽天、ユニクロは役員に外国人はいないし、一般社員も外国人は欧米人より（日本語に習熟した）中国人や韓国人が多い。そうした企業が会議を英語でするのは、活発な意見のやり取りや内容の掘り下げが難しいなど、弊害が大きい。欧州諸国の英語が堪能な社員も同国人同士では英語を使わない。外国人がいるときだけ、通じる言語に切り替えるのだ。

「ネイティブが交じる会議は英語にする」というのも、合理的なようで、実は問題がある。脳の思考活動はワーキング・メモリー（作業記憶）の機能によって遂行されるが、複雑な内容になると、日本人は（母語とかけ離れた特徴を持つ外国語の聴取・理解や発話における文構成など）言語処理に脳の作業記憶の多くが占有され、（論点を分析し対案を提示するなど）論理的な思考への割り当てがあまりできなくなる。日常的なやりとりはともかく、会議での討議では、英語を聴き取ったり発話を構成する作業に気を取られ、討議の中身をじっくり考えることができないし、言いたいこともなかなか英語にならない。

一方、母語であれば、聴取や発話構成などの言語処理は意識下でほぼ自動的に行われ、ワーキング・メモリーにはあまり負荷がかからないので、余裕を持って思考に振り向けられる。このため、英語での討議は母語話者が主導する。「国際会議で頻繁に発言し、精神的に有利な立場で交渉を進める」という報告もある。英国の宰相チャーチルは首脳会談では英語を通した。フランス語は堪能だったが、「フランス語を使うとフランス人もどきの思考回路になり相手のペースに巻き込まれてしまう」らしい。欧州議会の議員も「英語で話すと、思うように言葉が使いこなせないせいで、本来自分が言いたかったことではなく、自分の英語で表現できることしか言えない」と洩らしている。日本人が英語を使う場合には、文成分の配列が逆転していることから、リアルタイムの脳内処理

において、欧州人より遥かに不利な状況に置かれる。込み入った内容の討議において「英語で考え意見を述べる」域に達している日本人が果たしてどれだけいるだろうか。英語を使うことでしっかり思考できず、実務に支障をきたすのならば、通訳を使えば良いのだ。

「情報を正確に共有する」という問題も重要だ。母語ならば可能な深い内容の討議が、外国語を使ったのでは困難であり誤解を生み易い。日本の会社で英語を公用語にすると、多くの社員の間で「情報が正確に共有できない」恐れがある。英語習熟レベルにより、情報格差（ないし情報歪曲）が起こり得るのだ。日本企業の社員の英語力からすると、オーラルな実務英語だけでなく、文書の英語も正確に理解されるとは限らない。放映された社内の英会話の授業のやりとりを見る限り、国際ビジネス英語力テストTOEICの成績が（日本の大卒新入社員平均の）450点（ごく初歩的な最低限の会話力）もかなり疑わしい社員が多い中で、社長の掲げる英語公用語化の目標年度2012年までに英語で実務ができるようになる見込みはない。楽天では社長が英語で「朝会」をしているが、話の内容が分からない社員も相当多いらしい。日本人だけの役員会は英語に堪能な社長の独壇場で、反対意見も出ないのだろう。「会議を英語で行う」という方針は、社員の英語力を考慮すると不見識も甚だしい。裸の王様の決断で「自由闊達な議論がなくなる」ことが危惧される。かつて米国に出張し排ガス関連の仕事で環境保護局や技術者と会議した某自動車メーカの方が、その時の経験から英語運用力の必要性を痛感し、英語が堪能な部下10人ほどに「全員、日本語厳禁。すべて英語で仕事しよう」と指示したが、それ以降、部下は口を閉ざし「沈黙の職場」と化したという。三木谷社長は、自身が米国の小学校で（言語が自動的に習得できる低学年に）学んだだけでなく、母親もニューヨークの小学校に通い英語で教育を受けている。英語習得に格別恵まれた環境に育ったということだが、おそらく帰国後も家では英語を使い英語力を維持・伸長できたと思われる。このため英語で思考しそのまま話せるようだが、そこまで恵まれた社員はいないだろう。

ほとんどの社員は、「日本語との言語差が極めて大きく思春期以降には習得が困難な」英語を、苦労を重ねて学んだのだ。努力してもなかなか運用力が伸びない。思ったことの半分も英語で発言できない。英語を苦もなく習得し駆使できる社長には、そうした社員の悩みが推し量れないのだろう。少なくてもTOEIC 800点程度ないとネイティブを相手にまともな討議はできないとされるが、実務的な英語力を昇進の条件にした場合、「仕事はできるが英語ができない」人よりも「仕事はできないが英語はできる」人が高い地位に就く。『週刊東洋経済』のインタビューに「英語が出来ない役員は2年後には首にします」と社長は答えたが、創業時から社長を支え会社を育ててきた役員はその多くが「英語力不足で失職」することになる。楽天は99%が国内収益で海外での事業実績はない。いくつか海外企業の買収を始めたとは言え、実質的な経験もない中「世界を相手にする企業は社員に英語が不可欠」といった短絡的な信条に振り回されるのではなく、「適材適所」を念頭に、「社員それぞれの能力を最大限に生かす職場環境を整える」ことが企業の発展につながる。業務で英語を使うこともない社員にまで一律英語運用力を課し、英語習得の負担がストレスを引き起こしたり、本務に専念する時間と気持ちの余裕を失くすることにつながるとしたら、本末転倒ではないだろうか。

オーラル・コミュニケーション偏重の英語教育の誤謬

英語運用力を過度に重視する企業の勘違いは、文科省の教育行政の誤謬と軌を一にする。「英語の使える日本人の育成」構想に沿って、文法や読解が大幅に削減され、オーラル・コミュニケーション偏重への流れが進み、昨年には、高校の学習指導要領で「授業は英語で行う」という基本方針が示された。ところが、「授業は英語で」という意見は中央教育審議会の外国語専門部会の議事録には記載がない。委員間では全く討議もされていないのに、文科省の役人が勝手に仕立てた方針なのだが、学校現場ではこの方針を巡り混迷を深めている。

どの教師にも英語で授業ができるのかと言うと、指示や質問、そしてコミュニケーション科目の教科書に出ている「場面ごとの定型的な表現」を英語で言うことができたとしても、英文の内容とその背景にある文化や社会の解説、表現や構文の文法的な説明を「生徒に分かるようなやさしい英語」で明快に行えるとは考えにくい。外国語でも欧州語ならば言語基盤が共通な欧米でさえ、文法など複雑な内容の説明は母語で行うのが原則だ。仮に教師が英語で説明したとしても、その英語を聴き取り理解する生徒は何人いるのだろうか。日本の平均的な公立高校の生徒の英語力を考慮すると、落伍する生徒が満ち溢れる事態が予想される。オーラル・コミュニケーション偏重の学習指導要領の下で3年間学習した最初の高校生がセンター試験を受験した97年には、成績が偏差値換算で一気に10点急落（大学入試センター職員の研究報告）し、中学生も高校入学時の成績が95年から11年間で7点低下（茨城県下全公立高校で模擬試験実施）した、との研究報告もある。ゆとり教育を推進した文科省の役人は学力低下の責任を取らないままに私大の教授に天下りしたが、「英語で授業」を方針とした役人も責任を取るとは思えない。

日本語とかけ離れた言語的な特徴を持つ英語習得には、「文を構成し理解する仕組み」としての文法は不可欠であり、これが英語運用力の基盤となる。文法が脆弱なままでは「読み書く」どころか「聴き話す」能力も育つはずがない。韓国や中国では文法もしっかり教えている。文法を疎かにしてきた近年の日本の英語教育は英語力の目に余る劣化を招いてしまった。また、オーラル・コミュニケーションを偏重しながら、どういうわけか肝心の発音教育を怠ってきた。英語は、他の欧州諸語にも見られないほど、発音がダイナミックに変化する。このため、そのメカニズムを理解し、自動的に発音できるまで練習しなければならない。ところが、教職課程において「音声学」の取得が義務付けられていないのだ。教師自身が発音のメカニズムをしっかり学び、そのエッセンスを生徒に教え訓練を通して体得させることで、生徒の「聴き話す」能力は飛躍的に向上する。むやみにオーラルな運用力にこだわる経済界や世間の声に惑わされず、言語教育の原点に返って、文法・語彙・発音などの基盤能力を育てる英語教育を推進しなければならない。それによって、ネット情報の活用、メール交換を含む実務をこなすのに必要な英語力が育成できるのだ。

機械翻訳の有効活用

「社員は母語で情報を正確に把握する」という原則で、日本語の文書やメールは英語や韓国語、中国語などに翻訳して外国人社員に周知するのが良いが、その時に機械翻訳ソフトを利用することが翻訳の時間と費用負担を抑えるのに有効だ。ただし、日本の上場企業で働く一般業務の韓国人、中国人に関しては、日本語能力検定試験で1級ないし2級の保持者が多いので、文書の翻訳の必要はない。翻訳ソフトを使うとすれば、一般に日本語力の低い欧米人や（ITなど）技術系のインド人などが対象になる。海外支社からの文書やメールでの連絡を日本の本社で日本語に改めて情報共有を図る際にも、機械翻訳の利用が時間と経費の削減に不可欠となる。日本にいる外国人社員の場合は、日本語と外国語の文書を対訳形式で提示することが、意味の確認だけでなく日本語習得にも好影響を及ぼすだろう。

機械翻訳だと「専門用語などの訳語の信頼度が高くなるほか、訳語に揺れを生じない」という統一性の確保ができ、処理速度が速い点でもメリットがある。英日・日英翻訳の場合、翻訳ソフトで粗訳を行い、翻訳部門の翻訳者が仕上げをするという工程になる。なお日韓翻訳の場合は仕上げ工程がいらない。日韓翻訳は95%以上の精度になるので、へたな翻訳者の訳より精度が高いのだ。ただし、韓日翻訳は翻訳者が手を入れなければならない。韓日翻訳に際して、韓文に含まれる同音異義語の機械的な選択によって誤訳が起こるためである。（同音異義語は本来漢字で表記されていたが、日本からの独立後は表音文字ハングルで表記されるようになり、多義性を解消し

なければならなくなったのだ。）日韓翻訳ならば、「漢字仮名混じり」で入力した段階で、人間の判断により同音異義が解消しているのだ。このため、韓日翻訳は日韓翻訳より 5～10%ほど精度が下がる。英語と欧州主要語（仏語、西語、独語、露語）との間の翻訳も、言語的な近さから、実用的な精度になっている。文書の分野にもよるが、英仏翻訳、英西翻訳が 93%～98%、英独翻訳が 90%～95%、英露翻訳が 85%～90%程度の精度だ。近似言語は文法や語彙が似ているだけでなく、修飾関係が曖昧な構造をそのまま訳しても翻訳が成り立つためだ。欧州語は一旦翻訳ソフトで英訳したものをさらに英日翻訳する（「ブリッジ／リレー翻訳」）ことで大体の意味は把握できる。

翻訳ソフトの精度は、一義的には、①「文法・語彙に用例データを加えたシステム」による言語処理能力で決まる。だが、言語差が大きい場合には、修飾関係の曖昧さを解消する②「文脈処理機能」がなければ、適切な翻訳ができない。日本語では「文脈上同一指示の要素が現れる場合、これを削除する」という談話上の任意規則があるため、原文に主語や目的語などの主要成分が欠落したものが多く含まれ、「文脈情報から欠落成分を復元する」仕組みを持たない機械翻訳では、適切な英文に翻訳できない。これに対し、英文では「文脈上同一指示の要素が現れる場合、これを（削除するのではなく）代名詞に換える」という談話上の規則がある。このため、翻訳対象の文に主要成分の欠落が見られない。欠落成分を復元する必要がない点で、ソフトの精度や文書の種類にもよるが、（冠詞や時制など些細な間違いを別にすれば）英日翻訳は良いもので 84～88%程度の翻訳率で日英翻訳はこれより 5～10%落ちる。ただ、近似言語間と違い、修飾関係が曖昧な構造を文脈に沿った適切な関係に翻訳しなければならないが、日英語間の翻訳ソフトには、文脈情報を活用する仕組みがないことから、「翻訳の壁」は残る。

AAMT への要望

関西でも研究会・報告会を開いて欲しい。

個人会員（敬称略・50 音順）

会員名

塚田　元(日本電信電話（株）コミュニケーション科学基礎研究所／TSUKADA　Hajime)

自己紹介

私は現在、統計的機械翻訳（以下、統計翻訳）の技術開発に従事しております。企業研究者としてのキャリアは二十年ちょっとになりますが、機械翻訳に関わるようになったのは、ここ十年弱に過ぎません。AAMT の会員の中では、ほんの「ひよっこ」かと思います。

入社して最初のまとまった仕事は音声合成でした。今でも、証券会社の音声応答サービスなどで、我々が開発した音声合成器が使われているそうで、企業研究者冥利につきます。その後、ＡＴＲ音声翻訳研究所に出向し、音声認識に従事しました。その時期（１９９９年）に滞在した、米国ＡＴ＆Ｔで出会ったのが、統計翻訳です。

統計翻訳は、互いに翻訳となってる文の対（対訳文）を大量に集めたものから統計モデルを学習し、そのモデルを使って自動的に機械翻訳システムを構築する技術です。正直、こんなものまともに動くはずがないと思いました。しかし、評価してびっくり。その当時のＡＴＲの翻訳システム（用例に基づき人間が翻訳ルールを構築）に負けない翻訳文が出てきたのです。これは、音声認識の研究などしている場合ではないと思いました。とはい

え、その後なかなか研究の機会が得られませんでした。２００３年頃にようやく希望がかない、現職場で統計翻訳研究を立ち上げて、現在に至っております。

機械翻訳および翻訳業界に期待すること

私自身は、統計翻訳のアルゴリズム開発を生業としており、AAMT の多くの会員の皆様と比べると現場からは遠い人間です。そんな人間の戯言として、私の思い描く近未来の姿をお読みいただければと思います。

近年、Google や Microsoft など、統計翻訳を用いた実サービスが開始されました。これらのサービスは、多言語間の自動翻訳を提供するものですが、従来のルールベースのアプローチでは開発コストが膨大なものとなってしまいます。それを削減するために統計翻訳が導入されました。学習用の対訳データさえあれば、開発コストはほとんどかからないからです。

これらのサービスは、一般ユーザ向けのものですが、翻訳者向けのサービスに統計翻訳はどうでしょうか？　統計翻訳の最大の特徴は、学習型であるということです。実は、この特徴が、今後、機械翻訳システムと翻訳者のコラボレーションをより密なものにしていくのではないかと考えています。現在、多くの翻訳者の方が、自分らの作成した翻訳文を再利用するために、翻訳メモリを活用しています。また、一部の翻訳者の方は、辞書登録をするなどして、機械翻訳システムをカスタマイズして使っておられるかと思います。統計翻訳を活用すれば、自分が作成した翻訳文や辞書を使って、自分用にカスタマイズした機械翻訳システムをより容易につくることができます。さらにユーザインタフェースを高度化することで、ちょうど携帯電話の仮名漢字変換のように、翻訳文を部分的にタイプするとそれに続く表現をサジェストしてくれるような翻訳支援システムも実現できるでしょう。日本語文章を書く人にとっての仮名漢字変換のように、学習型の機械翻訳は翻訳者にとって空気のような存在になるに違いありません。

翻訳者支援の統計翻訳システムの本質的な部分（統計モデル）はそのまま一般ユーザ向けの自動翻訳システムとしても利用可能です。両者の統計モデルを共通化すると、翻訳者が新しい表現を翻訳するだけで、一般ユーザ向けのシステムでも翻訳できるようになります。特許翻訳やマニュアル翻訳サービスを考えてみましょう。メインサービスは翻訳者が訳した文章を提供することになるでしょう。そしてサブサービスとして、まだ訳せていない新しい文書を、自動翻訳サービスで提供することになります。従来、この二つのサービスは、独立したものでした。翻訳者が自分の翻訳業務のために行うカスタマイズが一般ユーザ向け自動翻訳サービスに反映されるようになれば、メンテナンスフリーでユーザ向けサービスを実現できるようになります。

技術的にはまだ多くの課題が残っておりますが、統計翻訳をベースに機械翻訳システムと翻訳者の win-win の関係を築くことで、波及的に翻訳サービス全体を大幅に効率化できるのではないかと考えております。そして、国内外の情報格差が少しでも小さくなることを願っております。

AAMT へのご要望

以上のような思いを現実のものとするため、引き続き機械翻訳技術者と翻訳者の架け橋を担っていただきたいと思います。また、特許翻訳に関する取り組みはその一つかと存じますが、新たな翻訳マーケットの創出にも力を注いでいただければと思います。たとえば、国内外の情報格差の大きい、医療分野は社会的に重要な翻訳マーケットかと思います。

協会活動報告

(2010 年 6 月～2010 年 10 月)

第 20 回通常総会

2010 年 6 月 14 日

第 1 号議案　2009 年度事業報告（案）　　第 2 号議案　2009 年度決算報告（案）
第 3 号議案　2010 年度事業計画（案）　　第 4 号議案　2010 年度収支予算（案）
第 5 号議案　役員改選について
その他・会員提案事項

● 報告会

2010 年 6 月 14 日

①機械翻訳課題調査委員会　　②AAMT/Japio 特許翻訳研究会
③インターネット WG　　④編集委員会

講演会

2010 年 6 月 14 日

講演 I :『Can Darwin ensure the revival of contolled Languages?』
Prof.Tony Hartley 氏(Professor of Translation Studies,Centre
for Translation Studies,University of Leeds)

講演 II :『マイクロソフト機械翻訳（支援）ツール：その使用例と今後のあり方』
相川孝子氏（マイクロソフトリサーチ 機械翻訳チーム）

第 5 回 AAMT 長尾賞受賞式・記念講演会

受賞者：（1）『みんなの翻訳』構築・運用グループ
影浦　峡　　東京大学
阿辺川　武　国立情報研究所
内山　将夫　（独）情報通信研究機構

受賞理由：（みんなの翻訳）の利用者コミュニティを立ち上げ、翻訳家の無報酬の翻訳活動を支援し、対訳コーパスの蓄積に大きく寄与すると共に、利用者の機会翻訳の理解と普及に貢献した功績が顕著なため。

懇親会

2010 年 6 月 14 日　ホテルアジュール竹芝　12 階　　白鳳

決算理事会

2010年6月14日

第1号議案　2009年度事業報告（案）　　第2号議案　2009年度決算報告（案）

第3号議案　2010年度事業計画（案）　　第4号議案　2010年度収支予算（案）

第5号議案　役員改選について

その他・会員提案事項

機械翻訳課題調査委員会

2010年4月26日（2010年度　第1回）

①今年度事業計画及び来年度活動計画について　②前回委員会の議事録の確認

③各WGの活動について（各WGに分かれて議論）

④活動内容の報告（各WGから）　　⑤活動内容についての議論

⑥まとめと次回委員会について

2010年6月4日（2010年度　第2回）

①今年度事業計画及び来年度活動計画について　②前回委員会の議事録の確認

③各WGの活動について（各WGに分かれて議論）

④活動内容の報告（各WGから）　　⑤活動内容についての議論

⑥まとめと次回委員会について

2010年7月5日（2009年度　第3回）

①前回委員会の議事録の確認　②各WGの活動について（各WGに分かれて議論）

③活動内容の報告（各WGから）

④活動内容についての議論　　⑤まとめと次回委員会について

2010年8月6日（2009年度　第4回）

①前回委員会の議事録の確認　②各WGの活動について（各WGに分かれて議論）

③活動内容の報告（各WGから）

④活動内容についての議論　　⑤まとめと次回委員会について

2010年9月8日（2009年度　第5回）

①前回委員会の議事録の確認　②各WGの活動について（各WGに分かれて議論）

③活動内容の報告（各WGから）

④活動内容についての議論　　⑤まとめと次回委員会について

2010年10月13日（2009年度　第6回）

①前回委員会の議事録の確認　②各WGの活動について（各WGに分かれて議論）

③活動内容の報告（各WGから）

④活動内容についての議論　　⑤まとめと次回委員会について

AAMT/Japio 特許翻訳研究会

2010 年 5 月 14 日（2010 年度　第 1 回）
①前回議事録の確認　　②今年度の研究内容について
③次回の開催について
2010 年 6 月 11 日（2010 年度　第 2 回）
①前回議事録の確認　　②今年度の研究内容について
③次回の開催について
2010 年 7 月 23 日（2010 年度　第 3 回）
①前回議事録の確認　　②今年度の研究内容について
③次回の開催について
2010 年 9 月 17 日（2010 年度　第 4 回）
①前回議事録の確認　　②今年度の研究内容について
③次回の開催について
2010 年 10 月 22 日（2010 年度　第 5 回）
①前回議事録の確認　　②今年度の研究内容について
③次回の開催について

編集委員会

2010 年 7 月 26 日　①AAMT Journal No.47 進捗確認　　②その他

インターネット WG

① AAMT ホームページの更新（毎月）
② AAMT Forum メーリングリリストの管理
③ AAMT Forum への情報発信（月数回）
④ 委員会活動におけるメールングリストの管理
⑤ 会員専用ホームページ開設にむけた準備
⑥ その他事務局ネットワークのインフラ管理全般

第 20 回通常総会および関連行事の報告

AAMT 事務局

当協会の第 20 回通常総会が 2010 年 6 月 14 日（月）13 時 20 分より・ホテルアジュール竹芝にて開催されました。総会後、各委員会からの報告会、講演会、そして第 5 回 AAMT 長尾賞授与式と受賞者による記念講演会が盛況のうちに行われました。

第 20 回通常総会

1. 開会の辞
2. 会長挨拶　豊橋技術科学大学・情報基盤センター教授・井佐原 均
3. ご来賓挨拶
4. 出席会員の確認
5. 議案

   第 1 号議案　2009 年度事業報告（案）　第 2 号議案　2009 年度決算報告（案）
   第 3 号議案　2010 年度事業計画（案）　第 4 号議案　2010 年度収支予算（案）
   第 5 号議案　役員改選について（案）
   その他・会員提案事項
   理事・監事交代、機械翻訳課題調査委員会委員長交代
   事務局移転のお知らせ
6. 総会後の組織、人事について（案）
7. 閉会の辞

報告会

開会挨拶　　会長・豊橋技術科学大学・情報基盤センター教授　　井佐原 均

1. 機械翻訳課題調査委員会	委員長	長瀬 友樹（(株)富士通研究所）
2. AAMT/Japio 特許翻訳研究会	副委員長	横山 晶一（山形大学）
3. 編集委員会	委員長	宇津呂 武仁（筑波大学大学院）
4. インターネットワーキング	リーダー	富士 秀（(株)富士通研究所）

講演会

○講演 I：「Can Darwin ensure the revival of controlled languages?」
Prof. Tony Hartley (Professor of Translation Studies,
Centre for Translation Studies, University of Leeds)
○講演 II：「マイクロソフト機械翻訳（支援）ツール：その使用例と今後のあり方」
相川孝子氏（マイクロソフトリサーチ機械翻訳チーム）

第 5 回 AAMT 長尾賞授与式・記念講演会

受賞者：「みんなの翻訳」構築・運用グループ

影浦 峡　　東京大学

阿辺川 武　国立情報学研究所

内山 将夫　（独）情報通信研究機構

受賞理由：「みんなの翻訳」の利用者コミュニティを立ち上げ、翻訳家の無報酬の翻訳活動を支援し、対訳コーパスの蓄積に大きく寄与すると共に、利用者の機械翻訳の理解と普及に貢献した功績が顕著なため。

選考委員長 ： 飯田 仁（東京工科大学）
選考委員　 ： 横山 晶一（山形大学）
　　ウィラット（NECTEC）
推薦者　　 ： 安達 久博（(株)サン・フレア）
同意人　　 ： 木道 嘉之（(株)ATR-Trek）

授賞式

懇親会

本会後の懇親会は、多数の参加者にお集りいただき、学識経験者、研究者、翻訳家等、幅広い分野の方々が活発な意見交換を行い有意義な交流の場となりました。ご参加誠に有難うございました。

編集後記

AAMT ジャーナル編集委員会委員長
筑波大学大学院システム情報工学研究科
知能機能システム専攻
宇津呂　武仁

AAMT ジャーナル 48 号をお送りします。

今号では、第 5 回の AAMT 長尾賞の選考結果を受けまして、受賞者である東京大学 影浦峡先生、NII 阿辺川武先生、NICT 内山将夫様よりご寄稿して頂きました。

第 5 回の AAMT 長尾賞は、「みんなの翻訳」の利用者コミュニティを立ち上げ、翻訳家の無報酬の翻訳活動を支援し、対訳コーパスの蓄積に大きく寄与すると共に、利用者の機械翻訳の理解と普及に貢献した功績に対して授与されました。

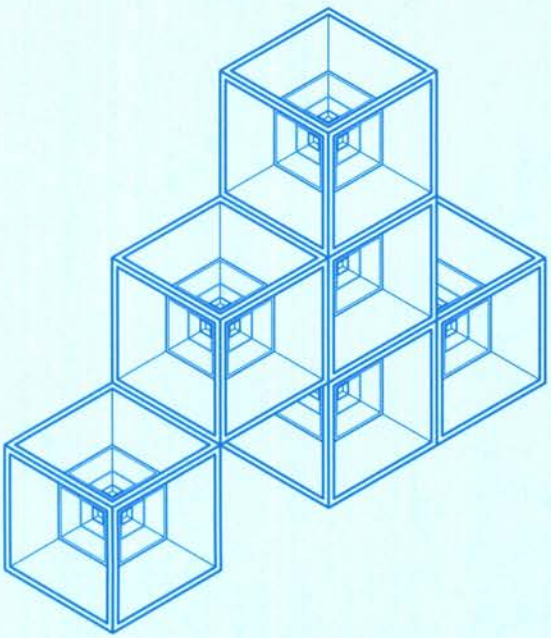
今号では、産業界における機械翻訳、および、翻訳メモリの動向を中心にご寄稿を頂きました。マイクロソフトの相川孝子様からは、Microsoft Translator の紹介記事をご寄稿頂きました。日本 CA の菊池邦明様からは、社内における機械翻訳導入事例についてご紹介頂きました。翻訳メモリ業界においてよく知られた SDL-TRADOS の新井康郎様からは、同社の製品・サービス動向についてご寄稿頂きました。

また、JAPIO の渡邊豊秀様より、「第 1 回産業日本語研究会・シンポジウムの開催概要と今後の取り組み」についてご寄稿頂きました。

その他、「AAMT 会員のひろば」の企画におきましては、個人会員の紹介文 2 件を掲載しました。

Memo

AAMT



AAMTジャーナル No.48

発行：アジア太平洋機械翻訳協会（AAMT）
ホームページ：http://www.aamt.info
住所：〒441-8580　愛知県豊橋市天伯町雲雀ヶ丘1−1
豊橋技術科学大学　情報メディア基盤センター内
phone：0532-44-6620　fax：0532-44-6620
編集委員会：宇津呂 武仁　小谷 克則　大倉 清司
鈴木 博和　三浦 貢　村上 嘉陽
事　務　局：神崎 享子　服部 絹依
印　刷　所：株式会社ナビックス

Asia-Pacific Association for Machine Translation
c/o TOYOHASHI University of Technology Information and Media Center
1-1 Hibarigaoka, Tempaku-cho, Toyohashi-City, Aichi, 441-8580 Japan
Phone:+81-532-44-6620　FAX:+81-532-44-6620
URL:http://www.aamt.info