# JEITAにおける句対応付き多言語コーパスの作成

荻野紫穂　　佐田いち子　　佐々木美樹　　井佐原均

日本IBM(株)　　シャープ(株)　　沖電気工業(株)　　通信総合研究所

電子情報技術産業協会 言語処理技術専門委員会 言語資源グループ

shiho@jp.ibm.com　sata@isl.nara.sharp.co.jp　sasaki234@oki.com　isahara@crl.go.jp

## 1. はじめに

電子情報技術産業協会（以降 JEITA）言語処理技術専門委員会では、数年前から、対訳コーパスの作成に取り組んでいる。原言語と対象言語の対応を取り、句やさらに細かいレベルの名詞類などにタグ付けすることによって、言語処理の研究・開発に貢献する対訳コーパスの開発を目指している。タグ付けの仕様は、コーパスの種類や文体に依存する部分が大きいこともあり、年々変遷しているが、基本的には、タグの種類を増やし、より細かいレベルでの対応を取る方向を目指すことで、過去の仕様との整合を計っている。

本稿では、JEITAコーパスのこれまでの経緯と、2002年度に行った、2種のコーパスのタグ付け作業、及び、将来の展望について報告する。

## 2. JEITAコーパスの経緯

JEITAの前身である日本電子工業振興協会（以降JEIDA）では、1996年度から、日英対訳コーパスの試作を行ってきた。この章では、各年度におけるコーパス開発作業の概要を述べる。詳細な内容については、各年度の報告書[1,2]を参照されたい。

最初に、1996年度に、日英対訳コーパスの入手と、電子化形式の正規化が行われた。入手コーパスの対象分野には、日英対訳がある・商業出版物にある制限がない・広範囲の話題が含まれている、などの理由から、政府刊行物である白書が選ばれた。

1997年度から1998年度には、入手コーパスに対し、サポートツールを使った人手による文対応付け作業が行われた[3]。白書は、複文や重文が多く、一文がかなり長い。このため、長文をいくつかに分割して対応付けが行われるだけでも意義がある、という意見を基に、1998年度から1999年度にかけて、当店単位程度の大まかな句・節対応付けが行われた。また、用語抽出を視野に入れ、複合名詞・固有名詞には別途細分化タグが付与された。以降、対応付け作業はコーパスの翻訳方向を問わず、日本語文を見ながら行われる（英韓コーパスを除く）。

2000年度は、技術マニュアルを題材とした英日対訳コーパスの開発を行った。この文書は英語としての信頼性も高く、かなり正確に文対応が取れていた。このため、ある程度各構造を意識し、白書関連の対訳コーパスよりも細かい単位での対応付けを実施したが、この新方式に関しては委員の間でも意見が分かれた。

そこで、2001年度に入り、一旦細かい単位でタグ付けした対訳コーパスに対し、各要素間の繋がりを示すため、「述部」とそれに関する「目的語」「補語」や「名詞句」とそれに関する「修飾句」などの対応付けを行うタグを追加した。

2002年度は、この仕様を踏襲し、京大コーパス[4]とその英訳版、Penn Treebank[5]とその和訳版の各ペアを題材とした日英・英日の双方向コーパスの対応付けを行った。更に、これまで日本語・

英語間で行っていた対応付けを多言語に広げる試みとして、2000年度から2001年度の作業対象であった技術マニュアルの韓訳に対してタグ付けを行い、英日韓3ヶ国語対訳対応付けコーパスの開発作業を行った。

## 3. 日英・英日対訳コーパス

### 3.1. 開発経緯

1999年度から2000年度にかけて開発した白書類の日英対訳コーパスと、2001年度に開発した英日対訳コーパスとを比較すると、後者のほうが句の構成に言語間の差が少なく、句対応付け作業がやりやすいという差が見られた。これが、白書と技術マニュアルという分野の違いによるものなのか、原言語を基にした対応付けと対象言語を基にした対応付けの差なのか、日英・英日のそれぞれの方向の翻訳方式の差によるものなのかは判らない。要因がはっきりすれば、対応付け作業の際の手順を効率化できる可能性がある。このため、今年度は、日英方向は京大コーパスを元に、英日方向はPenn Treebankを元にして翻訳コーパスを作成し、双方新聞記事を対象として分野の条件を揃えたタグ付けを行い、その差を観察した。

### 3.2. 作業

日英・英日対訳コーパスの例を以下に示す。

```
====== 4 ======
彼女は髪が長い。
She has a long hair.
=====
P J4:0-2: 彼女
P E4:0-3: She
--------
PX J4:3-8 は髪が長い。
PX E4:5-20 has a long hair.
--------
```

コーパスは双方向とも、文ＩＤ行、日英対訳文に続き、対応タグと対訳文中の位置を添えた、句対応を示す形式になっている。作業者は、日本語を母国語としているが英語も理解し、英日・日英の句対応付けの経験者である。

今年度のコーパスに付与されたタグの一覧を以下に示す。対応タグは、原則として2002年度までのタグセットを踏襲している。詳しいコーパスの形式やタグの使用法などの情報については[2]を参照されたい。

| タグ | 対応関係 |
|---|---|
| P | 普通の句対応 |
| PX | 意味は対応、構文は非対応の句 |
| PL | 複数句対応 |
| PPRON | 人称代名詞の省略 |
| W | 固有名詞 |
| WN | 複合名詞 |
| NT | 対訳なし |
| NTS | 主語（句）省略 |
| NTO | 目的語（句）省略 |
| NTADV | 副詞（句）省略 |
| NTADJ | 形容詞（句）省略 |
| NTPREP | 前置詞（句）省略 |
| NTVB | 動詞（含分詞）省略 |
| TM1 | 綴り間違い |
| TMi | 必要な空白文字の欠落 |
| ER | その他エラー |
| C | コメント |

### 3.3. 知見と考察

日本語から英語に翻訳した京大コーパスでは、明確な対応が取りづらい・取れないといったケースが多く挙げられた。特にスポーツ記事や会話文では、新聞や雑誌などに特有の簡潔な表現や臨場感に溢れる表現が多いためか、こうした傾向が顕著に見られた。これは、自明な主語や文脈から判

断できる目的語の省略や体言止めなどが修辞技法として定着している、日本語の言語特性にも大いに依存していると思われる。

これに対し、Penn Treebank コーパスを英語から日本語に翻訳したものでは、京大コーパスの場合と比べて問題の件数が少なかった。

一つの仮説としては、同じ言語でも、原言語として使われる場合と対象言語となる場合では、文の自然さや構文が異なるかもしれない点を挙げられる。

翻訳専門家が文学作品など文体が重視される邦訳書を出す場合は別だろうが、通常新聞記事などを翻訳する場合は、原文に忠実に翻訳しようとする意識が働き、原文の構造を引きずった翻訳文になることが考えられる。また、翻訳における意味保存の制約をかけられ、いわばフィルタを通した状態で書かれた翻訳文は、原言語で考えてそのまま書かれる原文よりも、構造がシンプルになりやすい可能性も高い。この場合、今回の作業のように句範囲の選択を特定の言語（今回の場合日本語）に限定して句対応付けを行うと、句範囲選択を行う言語と原言語が一致している場合より、句範囲選択を行う言語が対象言語になっているほうが、問題が起きにくいことになる。

この仮説が正しい場合、双方向の翻訳の質が同程度に保証されるならば、

- 英語側を句範囲選択の言語に固定した場合、京大コーパスでも問題が減る。また、Penn Treebank でも問題が発生しやすくなる。
- Penn Treebank を、いかにも日本の新聞記事調の文体に翻訳した場合、問題が発生しやすくなる。

などの現象が観察される可能性が高い。これらについては、対訳コーパス開発作業を効率化する上で、今後の研究課題となろう。

## 4. 英日韓３ヶ国語コーパス

### 4.1. 経緯

JEITA では、2001 年度まで日本語と英語の対訳コーパスを開発してきたが、句対応付き対訳コーパス開発の過程として、今までに採用されたタグ付け方式や句構造の範囲が、日本語と英語以外のコーパスについても使用可能かどうか、検証したほうがよい。そこで今年度は、2000 年度から 2001 年度にかけて開発したコーパスの原文となった、英文技術マニュアルを韓国語に翻訳し、英韓対訳コーパスの対応付け作業を行った。前年度までに、同じ英文を原言語にした英日対訳コーパスの対応付けが行われているため、英韓対訳コーパスの開発作業が終了すると、英語を中心とした英日韓３ヶ国語の句対応付けコーパスとなる。

### 4.2. 作業

大規模なタグ付けプロジェクトでも、単言語コーパスへのタグ付けの言及はよく見られるが[6]、欧米言語とアジア言語の対応付けについては、作業手順の言及はあまりない。作業時間などの条件を考慮した上で、今年度は、全体の様子を見るために、作業者に英日句対応付きコーパスに韓国語の文翻訳を加えたデータを渡し、以下の４種類のタグを付与することにした。

| タグ | 対応関係 |
|---|---|
| PM | 英日韓が一致 |
| PN | 英韓一致、日不一致 |
| PO | 日韓一致、英不一致 |
| PQ | 英日一致、韓不一致 |
| PX | 英日韓不一致 |

作業者は、韓国語を母国語としているが、日本語・英語の双方を理解する。作業済みコーパスの例を以下に示す。

====== 43 ======

이미지 활성화

イメージの有効化

image enabling

======

PN K43:1-7 이미지 활성화

PN J43:0-8: イメージの有効化

PN E43:0-14: image enabling

--------

### 4.3. 知見と考察

現在、作業報告を分析中である。作業者から挙げられた英日韓の一致の傾向を以下に述べる。」

- 句対応付けの単位（句切り出し単位）については、英日対訳コーパスの作業結果を韓国語に適用しても、問題なく作業ができる。
- 英日が一致し韓だけが違うものは数が少ない。
- 英日韓３ヶ国語一致が見られるものは、名詞、その中でも専門用語が多い。
- 英韓が一致し日だけ不一致のものは、英韓が形容詞、日が名詞のように、品詞がずれているものが多い。また、日が助詞を伴って修飾句を形成しているために不一致になっているものも多い。
- 日韓が一致し英だけが不一致のものは、日韓の文法体系にはなく英にだけ見られる関係節を使った表現が多い。
- 日韓がどれも不一致のものは、日韓が一致し英が不一致の例のように、日韓と英とで文法構造が違う表現のなかで、日韓の表現も微妙にずれているものが多い。

これらを含む報告を解析した上で、句構造対応の多言語展開を更に考える予定である。

## 5. 今後の予定

本稿では、JEITA コーパス開発の経緯と今年度の作業について報告した。

日英・英日双方向の翻訳コーパスで、句対応が一致している文ほど、翻訳過程で特別な構造変換などが必要なく、逐語訳に近い方式での翻訳が適用できそうである。逆に、句構造が一致していない文は、一般的な方式では扱いきれない言い回しや構文を含み、翻訳が困難な文であろう。こうした句対応の一致・不一致の特徴は、難易度段階別の翻訳評価文セットの開発に利用できる可能性が高い。これまでの、文法チェックのための人工的な作文[7]とは違い、実際に使われている文で評価を行い、処理の難易度分けができる点で期待ができるため、こうした評価文セットの開発を視野に入れて活動する予定である。

## 参考文献


1. 自然言語処理システムの動向に関する調査報告書, 日本電子工業振興協会, 1997, 1998, 1999.
2. ヒューマンインタフェース技術に関する調査報告書, 電子情報技術産業協会, 2000, 2001, 2002, 2003(3月発行予定).
3. Isahara, H. and Haruno, M., Japanese-English aligned bilingual corpora, Parallel Text Processing, pp. 313 – 334, Kluwer Academic Publishers, 2000.
4. http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/corpus.html
5. http://www.cis.upenn.edu/~treebank/home.html
6. http://www.ilc.pi.cnr.it/EAGLES/home.htm
7. Isahara, H., JEIDA's Test-Sets for Quality Evaluation of MT Systems – Technical Evaluation from the Developer's Point of View, Proc. Of MT Summit V, 1995.