日英京都関連文書対訳コーパス (Version 2.0)

2010 年 12 月 23 日

目次

1. 概要

本コーパスは、高性能な多言語翻訳、情報抽出システム等の構築を支援することを目的に作成された、大規模な日英対訳コーパスです。Wikipedia の日本語記事（京都関連）を人手で英語に翻訳することにより作成されました。また、本コーパスでは、翻訳の過程（一次翻訳→二次翻訳→最終チェックの 3 ステップ）が記録されています。訳文が精緻化されていく過程を観察できるため、翻訳支援ツールの開発、人手翻訳における誤り分析等にもご活用いただけます。翻訳対象となった Wikipedia 記事は、京都に関する内容を中心に、日本の伝統文化、宗教、歴史等の分野をカバーしています。各種観光情報の英訳や通訳ガイドのための用語集作成などにも活用いただけます。

2. コーパスの作成方法

2.1 日→英翻訳のステップ

日本語記事の英訳は、以下の 3 ステップで行われました。

(1) 一次翻訳
日本語を母語とする翻訳者が日本語原文を英訳する

(2) 二次翻訳
英語を母語とする翻訳者が一次翻訳文における情報の過不足および流暢さ（可読性）をチェックし、必要な場合は修正する。本作業は日本語原文を参照しつつ行う。

(3) 最終チェック
日本語を母語とするチェッカーが二次翻訳文における専門用語および言及する専門分野の知識のチェックを行い、必要な場合は修正する。本作業は、専門分野の者が日本語原文を参照しつつ行う。

作業にあたっては、翻訳者間の品質の違いを少しでも減らすために、規範的な翻訳例や FAQ などを翻訳者間で共有し、疑問点については随時議論を行いました。また、記事ごとに作業指示日時・作業終了日時・作業者 ID を記録し、問題が生じた場合の原因究明やその後の品質維持の対策を取りやすい体制を整えました。

2.2 ファイルのフォーマット

コーパスに収録されている日英対訳データは、一記事ごとにファイル化されています。以下のとおり、XML タグによりテキスト構造が明示化されています。

```
<?xml version="1.0" encoding="UTF-8"?>
<art orl="ja" trl="en">
<inf>Wikipedia dump の ID</inf>
<tit>
<j>｛日本語タイトル｝</j>
<e type="trans" ver="1">｛一次翻訳タイトル｝</e>
<cmt>｛一次翻訳者によるコメント｝</cmt>
<e type="trans" ver="2">｛二次翻訳タイトル｝</e>
<cmt>｛二次翻訳者によるコメント｝</cmt>
<e type="check" ver="1">｛最終チェック済みタイトル｝</e>
<cmt>｛最終チェッカーによるコメント｝</cmt>
</tit>
<sec id="1">
<tit>
<j>｛日本語セクション見出し｝</j>
<e type="trans" ver="1">｛一次翻訳見出し｝</e>
<cmt>｛一次翻訳者によるコメント｝</cmt>
<e type="trans" ver="2">｛二次翻訳見出し｝</e>
<cmt>｛二次翻訳者によるコメント｝</cmt>
<e type="check" ver="1">｛最終チェック済み見出し｝</e>
<cmt>｛最終チェッカーによるコメント｝</cmt>
</tit>
<par id="1">
<sen id="1">
<j>｛日本語原文｝</j>
<e type="trans" ver="1">｛一次翻訳文｝</e>
<cmt>｛一次翻訳者によるコメント｝</cmt>
<e type="trans" ver="2">｛二次翻訳文｝</e>
<cmt>｛二次翻訳者によるコメント｝</cmt>
<e type="check" ver="1">｛最終チェック済み文｝</e>
<cmt>｛最終チェッカーによるコメント｝</cmt>
</sen>
・・・
</par>
<copyright>copyright (c) 2010 ｛著作権者リスト｝</copyright>
</sec>
</art>
```

2.3 ディレクトリ構造

本コーパスを構成する各ファイルは、その内容によって 15 のカテゴリに分けられ、ディレクトリに分割して格納されています。

記事カテゴリコード一覧

コード	カテゴリ名	内容
SCL	school	学校
RLW	railway	鉄道（交通関連）
FML	family	旧家
BLD	building	建造物
SNT	Shinto	神道
PNM	person name	人名
GNM	geographical name	地名
CLT	culture	伝統文化 （現代文化も含む）
ROD	road	道路
BDS	Buddhism	仏教
LTT	literature	文学
TTL	title	役職・称号
HST	history	歴史
SAT	shrines and temples	神社仏閣
EPR	emperor	天皇

2.4 ファイル名

本コーパスのファイル名は、以下のような規則に則って付けられています。

```
[記事カテゴリコード][カテゴリ内での通し番号（5 桁）].xml
```

例） `BDS00056.xml` → 「Buddhism（仏教）」カテゴリの 56 番目のファイル

2.5 翻訳の基本方針

日本語記事の英語への翻訳は、以下のような基本方針のもと行われました。

- 日本語原文の一文に対して一つの英語訳文を付ける。
  曖昧な表現を避け、一文一意の訳文を心がける。ただし、原文が長すぎるなど、翻訳に際して明らかな不具合が生じる場合は、原文を複数文に分割し、それぞれを訳出する。
- 原文のすべての内容を忠実に翻訳する。
  原文の内容をよく理解した上、文の内容と構成を英語の文法の範囲以内で表現し、自然な訳文となるよう訳す。ただし、見出しおよび文章を主な翻訳対象とし、箇条書き等による短い項目の羅列などは翻訳対象外とする。
- 文体を統一する。ただし、口語体にならないこと。
- 用語、句読点、数字等の表記法を統一する。
- 翻訳の際、市販あるいは公開されている機械翻訳システムを利用しない。

2.6 翻訳の際の注意事項

翻訳時、2.5 の基本方針に加え、以下のような注意事項も翻訳者に提示されました。

- 表記全般について
  - ローマ字表記はヘボン式で統一する。
    例）じゃ じゅ じょ → ja ju jo、ちゃ ちゅ ちょ → cha chu cho、
    しゃ しゅ しょ → sha shu sho
  - 長音がある場合、母音を重ねたり、長音記号を使ったりしない。
    例）一休宗純 → ×Soujun IKKYUU, ×Sōjun IKKYŪ, ○Sojun IKKYU
  - 括弧（）は必ず半角とし、前後がピリオドやコンマ記号でない限り、前後に半角スペースを挿入する。
  - 原文において、文学作品等のタイトルに『』や「」が付いている場合と付いていない場合がある。それぞれ以下の通り対処する。
    - ✧ 『』や「」が付いている場合
      - 『』は “ ”に、「」は ‘ ’ に、それぞれ置き換えて訳出する。
    - ✧ 何も付いていない場合
      - 訳文でも何も付けない。
  - 内容的に、漢字（or かな等）の意味や音、表記そのものが話題となっており、意訳や音訳が無意味だと判断できる場合は、適宜漢字（or かな等）も織り交ぜて訳出する。

- 記事タイトル、記事内のセクション見出しの訳語は先頭の文字を大文字で表記する。
- その他、疑問点等は全てコメント欄（<cmt>…</cmt>）に具体的に記録する。

- 全角記号の半角への変換について

  Wikipedia 記事原文内にて全角で表記されている記号は、可能な限り半角に変換する。ただし、半角への変換が難しい以下のような記号は全角のまま表記する。

  ℃　≒　‰　α　ε　μ　π　ω

- 固有名詞、専門用語等の翻訳について

  固有名詞はすべての記事にわたり一定に訳す。しかし、省略形や指示詞を使った方が自然な場合はそれらを使用する。英訳が不明な場合は、日本語原文の表現をそのまま翻訳文にコピーせず、辞書、インターネットその他資料を利用し定訳を確認する。定訳が見つからない場合は、その旨をコメントとして記録した上で、次のように対処する。

  - わかりやすい英語訳が可能な場合
    - ✧ 英語訳を表記し、初出時のみ括弧内に原文の表現を表記する。
  - わかりやすい英語訳も難しいが、ローマ字表記で馴染みのある表現の場合
    - ✧ ローマ字表記とする。
  - わかりやすい英語訳も難しく、ローマ字表記では却って読みづらくなる場合
    - ✧ 原文の表現のままとする。

- 人名の翻訳について
  - 日本人や中国人の人名でも「名 姓」の順番とし、名字を大文字で表記する。
    例）野口英世 → Hideyo NOGUCHI
  - 「紀（の）貫之」「藤原（の）道長」など、姓と名の間に「の」が入るのが慣例である人名の場合に限り、名と姓を逆順とし、間に “no” を挿入する。
    例）紀貫之 → KI no Tsurayuki、藤原道長 → FUJIWARA no Michinaga
  - 「天璋院」など女性の院号は、 “Tenshoin” のように表記する。

- ドラマ、書籍、論文等のタイトルの翻訳について

  ドラマ、書籍、論文等のタイトルについては、日本語名が有名である、的確な意訳が難しい、音訳で雰囲気が伝わる、などの場合はローマ字表記とし、可能なら初出のみ括弧付きでタイトルの意訳を併記する。それ以外の場合は意訳のみとする。

  - ローマ字表記のみの例

- ✧ みなもと太郎『雲竜奔馬』→ Taro MINAMOTO, “Unryu Honma”
- ➢ ローマ字と意訳併記（初出のみ）の例
  - ✧ 『功名が辻』→ “Komyo ga Tsuji” (Crossroads of the Achievement)
- ➢ 意訳のみの例
  - ✧ 『とんち探偵一休さん 金閣寺に密室（ひそかむろ)』(鯨統一郎)
    → “Witty Detective Ikkyu-san: The Secret Room in Kinkakuji” (Toichiro KUJIRA)

- 地名、施設名の翻訳について
  - ➢ 寺院と神社の名称は、「東大寺」→ “Todai-ji Temple”、「八坂神社」→ “Yasaka-jinja Shrine” などの表記とする。ハイフンも忘れずに付ける。
    （最近の英訳では、 “Shinjuku-dori Street”、 “Kanda-gawa River” など、二重に訳す傾向がある。英語しか分からない人にも何の名称なのか見当がつき、また、日本語しか分からない人がその名前で尋ねられても分かる、実用的な翻訳といえる。）
  - ➢ 住所の「県」「府」「市」は省略せず訳出する。
    例）京都府京都市 → Kyoto City, Kyoto Prefecture

- 年号の翻訳について
  年号は、西暦のみを訳出し、和暦の表記は不要とする。旧暦についても、旧暦そのものが話題となる場合以外は、表記不要とする。
  例）寛保 4 年（1744 年） → in 1744

- 度量衡の翻訳について
  - ➢ 日本独特の面積単位「坪」は、1 坪=3.3 sq.m. として換算し、訳出する。
    例）10 坪 = 33 sq.m.
  - ➢ 「加賀 100 万石」など石高の単位は、 “koku” と訳出する。
  - ➢ km（キロメートル)、m（メートル)、l（リットル)、kg（キログラム）などは、ヤード・フィート・ガロン・オンスなどに換算せず、そのまま訳出する。

- その他
  - ➢ 原文に含まれる事実誤認もそのままの内容で翻訳する。
  - ➢ 原文に誤字・脱字が含まれる場合、前後の文脈から本来の語を推測し、翻訳する。また、原文に誤字・脱字があることをコメントとして記録する。
  - ➢ 俳句は、韻律等は気にせず普通の「詩」として訳出するか、或いは情景（内容）を通常の文で具体的に書き下す。ただし、内容の解釈、翻訳がどうして

も困難な場合は、翻訳文入力箇所を空白にし、コメント欄に「<cmt>意味解釈が困難なため訳出できず</cmt>」と追記する。

- 古文全般に関しても、内容の解釈、翻訳がどうしても困難な場合は、翻訳文入力箇所を空白にし、コメント欄に「<cmt>意味解釈が困難なため訳出できず</cmt>」と追記する。
- 条例等で公式な英語版が存在する場合は、それをそのまま訳文としてコピーしてもよい。
- 例えば、「となった。」だけで一文になっているなど、前後の文とつながらない、意味のない原文は翻訳対象外とする。コメント欄に「<cmt>不完全な文のため対象外</cmt>」などと記録する。

2.7 その他ファイルに関する特記事項

- 一原文一対訳の例外対処について

  Wikipedia 記事における原文が長すぎて翻訳しにくい場合、翻訳者の判断で文を分割しています。その場合、コメント欄に「<cmt>一原文一対訳の例外対処</cmt>」などという記載をしています。
- コメント欄の用語について

  コメント欄に以下のような用語が含まれている場合があります。いずれも作業者間で共有されていた、作業のための参考資料等を指すもので、作業者間の連絡に必要であったものです。コーパス利用者の皆さまには不要な情報かもしれませんが、作業者間の情報伝達の履歴としてそのまま公開しています。
  - 「term リスト」「用語リスト」「統一リスト」等
    - 翻訳者間で共有されていた専門用語リストを指します。
  - 「FB」「フィードバック」など
    - 翻訳過程において、(独) 情報通信研究機構から翻訳会社に伝達されたフィードバックを指します。
  - 「split sample」など
    - 上記「一原文一対訳の例外対処について」にも記載の通り、Wikipedia 記事における原文が長すぎて翻訳しにくい場合、翻訳者の判断で文を分割しています。その際の指針として翻訳者間で共有されていた分割サンプル集を指します。
  - 「EXX（XX には数字が入る)」「セル」など
    - 翻訳作業は表計算ソフト上のシートを用いて行われました。これらの用語はシート内のセル番号を指します。
- 余分なスペースの削除について

  コメント欄に「<cmt>スペースの重複を削除</cmt>」等、余分なスペースの編集に関

する記述が多く見られます。これは、翻訳過程において挿入されてしまった余分なスペースの修正履歴を表すものです。しかし、当該の余分なスペースに対しては、翻訳作業終了後に一括削除操作を行ったため、訳文内に残されていません。コメントと実際の訳文内でのスペース挿入状況が矛盾いたしますが、作業者間の情報伝達の履歴としてコメントはそのまま公開しています。

- 本コーパス内で使用されている漢字コードはすべて UTF-8 です。
- 本コーパスの翻訳に使用された Wikipedia dump データは、jawiki-20080607-pages-articles.xml および jawiki-20090527-pages-articles.xml です。各ファイルの使用 dump ID はヘッダに記載されています。
- 各ファイルの末尾に表示されている著作権者リストは、Wikipedia の編集履歴から抽出したものです（ただし、IP ユーザは含んでおりません）。万が一抽出漏れ等がある場合はお手数ですが、E-mail にて、kyoto-corpus@khn.nict.go.jp までご連絡ください。ご指摘内容を検討の後、必要な場合は修正を施します。
- 全ファイルの一覧は、Wiki_Corpus_List_2.0.csv をご覧ください。ファイル名、Wikipedia 上でのページ ID、ページタイトル、文対数、一次翻訳者 ID が示されています。

3. 利用に関する注意

- 本コーパスは、Wikipedia の日本語記事を英訳することにより作成され、Creative Commons Attribution-Share-Alike License 3.0 の条件の下、一般公開されています。本コーパスのご利用に際しては、Wikipedia の著作権（英語）を熟読の上、著作権法を考慮の上、十分に注意をしてください。
- 本コーパスでは、Wikipedia 記事に記載されたあらゆる情報をそのまま翻訳しています。本コーパスをご利用の際は第三者への誹謗中傷、差別用語、個人情報などに十分な注意をお願いいたします。
- （独）情報通信研究機構では、本コーパスにより獲得される情報の信頼性について責任を持ちません。また、本コーパスの使用に関連して生ずる損失、損害等について、いかなる場合においても一切責任を負いません。
- 本コーパスの内容における著作権侵害やその他問題を発見された場合は、お手数ですが、E-mail にて、kyoto-corpus@khn.nict.go.jp までご連絡ください。ご指摘内容を検討の後、必要な場合は修正を施します。