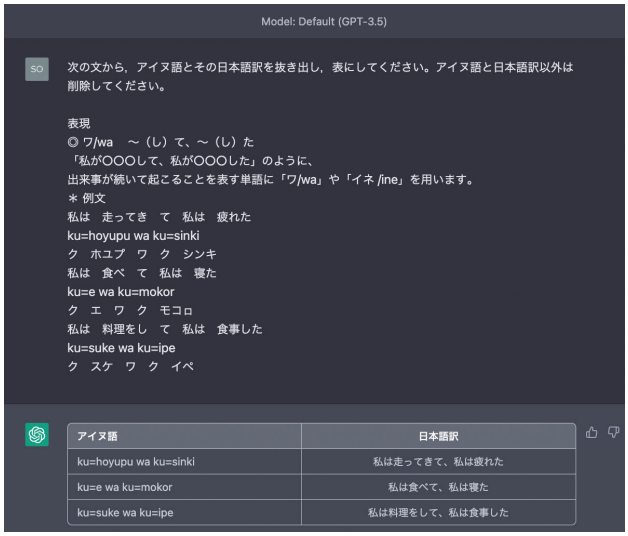
ChatGPTを活用してコーパスを構築する

ことばをコンピュータで扱うといえば、ChatGPTが最近話題です。ChatGPTは、アメリカのOpenAI社が開発した対話型生成AIで、ブラウザやアプリで質問や指示を文章で入力すると、回答が返ってきます。専門知識がなくても手軽に使え、さまざまな分野や目的、場面に対応できる能力を持っています。コーパス構築にもChatGPTが使われ始めています。どう使う？ 利点は？ 宮川 創助教に聞きました。

テキストデータ出典：『令和４年度「アイヌ語ラジオ講座」テキスト Vol.3』（公益財団法人アイヌ民族文化財団）



宮川 創
国立国語研究所
研究系 助教

Q 宮川さんは、ChatGPTをよく使っているのですか？

常識的・倫理的に問題なさそうな範囲で、文章の要約や翻訳の参考に使ったり、また自分が書いた文章を校正したり文体を変えたり、さらにはプログラミングやデータの整理などに使ったりしてみて、その結果を参考にすることがあります。GPT-4という非常に優秀な次世代大規模言語モデルを利用できる有料版のChatGPT Plusを使っています。

Q ChatGPTをコーパス構築に活用しているそうですね。どのようにChatGPTを使うのですか？

コーパスとは、ラテン語のcorpus「体」が語源ですが、言語学では言語のテキストデータセットを指します。私は最近、自身の機械翻訳の研究用にアイヌ語の対訳テキストのデータを収集しました。使用したのは、国立アイヌ民族博物館のアイヌ語アーカイブや、ほか複数箇所のコーパス、アイヌ語の教科書です。

アイヌ語アーカイブなどの既存のコーパスは、テキストがデータ化され、アイヌ語とその日本語訳が構造化されているので、そのままコンピュータで扱うことができます。

アイヌ語の教科書は印刷物なので、まずOCR（光学的文字認識）にかけてPDF形式にして、そこからテキストをコピーして抽出します。しかし、そのデータをプレーンテキストとして別のソフトウェアに貼り付けると、アイヌ語と日本語訳の対応がバラバラになってしまいます。そのため、アイヌ語とそれに対する日本語訳というようにテキストデータを整理して、表などコンピュータで扱いやすい形式にする必要があります。その構造化にChatGPTを使いました。

ChatGPTで「次の文から、アイヌ語とその日本語訳を抜き出し、表にしてください。アイヌ語と日本語訳以外は削除してください」と指示をして、その下にアイヌ語の教科書をOCRにかけて抽出したテキストデータを入力します。すると、表になって返ってきます（右上写真）。これらのデータを使い、機械翻訳の学習用データセットを構築しました。

Q ChatGPTをコーパス構築に使うことには、どういう利点があるのでしょうか？

この例では教科書をOCRにかけて抽出したテキストデータを使いましたが、私が研究しているコプト語では、対象が中世の写本など手書きの場合もあります。手書きテキスト認識（Handwritten Text Recognition：HTR）のソフトウェアには機械学習可能なもの（TranskribusやeScriptoriumなど）があって文字認識の精度を手軽に向上させることができるので、手書きの資料をテキストデータにすることは近年どんどん簡単になってきています。しかし、OCRやHTRで得られたテキストデータから原文と現代語訳のペアを抽出したり、分析しやすいように表などにしたりして構造化するには、これまでは手作業で膨大な時間がかかっていました。ChatGPTは、そうしたバラバラのデータも柔軟に構造化でき、コーパス構築の効率が大幅に向上します。

アイヌ語の例を紹介しましたが、日本には消滅の危機にある言語・方言がたくさんあります。ChatGPTなどの最新技術を活用することで、コーパス構築を効率化していきたいと考えています。