

Qwen3.5 によるカタカナアイヌ語 OCR の開発

安岡孝一 (京都大学)・安岡素子 (京都外国語大学)

はじめに

2026 年 2 月 16 日に発表^{a)}された Qwen3.5 は、最小 0.8B (8 億) パラメータの生成 AI でありながら、マルチモーダルかつ多言語対応を謳っており、実際、日本語 OCR としても、そこそこの性能を発揮する。しかし、アイヌ語は読めない、多言語対応にアイヌ語が含まれておらず、アイヌ語の小書きカタカナ [1, 2] が読めないのだ (図 1 に Google Colaboratory でのプログラムと結果を示す)。では、Qwen3.5 にカタカナアイヌ語を追加学習するには、どのような手法が必要なのか。本稿では、カタカナアイヌ語 OCR の開発という視点から、その手法を探る。

```
!pip install "transformers>=5.2.0" accelerate jinja2
img="https://qiita-image-store.s3.ap-northeast-1.amazonaws.com/0/428317/55d10bae-10b0-4a18-8e27-0907bc16361f.jpeg"
from transformers import pipeline
nlp=pipeline("image-text-to-text", "Qwen/Qwen3.5-0.8B", max_new_tokens=1024, device_map="auto")
d=nlp([{"role": "user", "content": [{"type": "image", "image": img}, {"type": "text", "text": "OCR Ainu sentences."}]}])
print(d[0]["generated_text"][1]["content"])
```

ウエランカラ

令和7年度アイヌ工芸品展「アイヌの美・彩りと輝き」アニ アン レヘ チコレ ワ、ウサムベカ イコロ チアマ ワ エチヌカレン。

アイヌ アナクネ テエタ ワノ アン クネ ワ 20 イタ ネ バ エムコ バツノヤウンモシリ、ヤンケモシリ (カラフト)、ルトム オッタ オカイ、タネ ヤウンモシタ カイ サモロモシタ カイ、モシリ エビッタ オカイ。テエタ アナクネ チクニ、ニカブ、キナカブ、チコイキブ ボネ、キラウ、シキテ、ルシ、チュエカブ、セイ アニ ウサ オカイ ベ カラ。ネアンベ オッタ ケシト アネイワンケ ヲ カイ アン。イノミ オッタ アネイワンケ ヲ カイ アン。ピシカン モシリ エオカイ ウタツ トウラ、アイヌ ウイマム ワ、センカキ、ヌイト、カネ、タマ コロ ワ ネアンベ アニ ビリカ イカラカラ ビリカ イヌイアイヌ キ。

タバン イキ オッタ イロホ ビリカ フ、ヌベキ ビリカ ヲ チヌムケ ワ チサンケ。イコロ エチヌカレン カトウ エネアニ。センカキ トウラ アンニヌ ワ アンカラ イミカイ アン。センカキ トウラ アンテセ ワ アンカラ チタラベ カイ アン。ルトム ウンクル イカラカラ ワ カラ クッ カイ アン。ヤンケモシリ ウン クル コロ タマサイ カイ アン ワ、ボン コンコ コロ タマサイ カイ アン。ウサ カネ ウサ ユキウ アニ アントムテ イコロ、イカヨビコロ、タンバクオ、マキリ カイ アン。イノミ オッタ アネイワンケ ヲ トウキ、タカイサラ、パツチ カイ アン。

ネアンベ トウラ クスロツ タ シュス チカブ美恵子 (1948~2010) カラ ベ カイ エチヌカレン。エエバキタ、貝澤徹 (イヌイエル)、下倉洋之 (カネイエル)、藤戸康平 (イコロカラ) タネ ウサ オカイ ビリカ フ カラ ワ コアスル アシタン ネ ワ クスネアンベ カイ チサンケ ワ エチヌカレン。

タプ アナクネ、アイヌプリ ビラサ クニ、公益財団法人アイヌ民族文化財団 (アスコロトミ アニ アイヌプリ ビラサ ウタラ) ケシ バ キ ヲ ネ ワ、タネ 30 スイ チキ シッタンナ。タンベ エチヌカラ ワ、テエタ アイヌプリ、タネ アン アイヌプリ、アムキリ クル インネ チキ ビリカ。イヨッタ イオシ、タバン イキ チコロ テッサ、ビリカイコロ ウネルサ ア ウタラ ウサ ウサ ウンカサイ ウタラ、バセタラ チコヤイライケ シッタンナ。

北海道立釧路芸術館
京都府
京都府京都文化博物館
公益財団法人アイヌ民族文化財団

Qwen3.5-0.8B へのトークン追加

Qwen3.5-0.8B の Qwen2Tokenizer は、UTF-8 のバイト列からトークンを組み上げる手法を採っており、トークナイザ辞書にない文字に対しては、バイト列に対応するトークンを組み合わせる形で入出力をおこなう。たとえば、小書きカタカナ「プ」(UTF-8 バイト列は e3 87 b7 e3 82 9a) は、トークン 159 (e3), トークン 229 (87), トークン 115 (b7), トークン 3293 (e3 82), トークン 248 (9a) を組み合わせて、[159, 229, 115, 3293, 248] という 5 つのトークン列で表現される。

しかし、複数のトークンで表現される文字は、入力

令和7年度アイヌ工芸品展「アイヌの美・彩りと輝き」アニアンレチコレワ、サウムベカイコロチアマワエチヌカレン。

アイヌアナケテデワウウンクネワ20イケネバエムコバクノヤウンモリ、ヤンケモリ (カラフト)、ルトム オッタ オカイ。タネヤウンモシタカイサモロモシタカイ、モシリエビタオカイ。テデワアナケチニ、ニカブ、キナカブ、チコイキブボネ、キラウ、シキテ、ルシ、チュエナカ、セイアウウオカイベカラ。ネンベオットケストアインワンケワカイアン。イノオットアインワンケワカイアン。ピシカンモシリエオカイウツタラ、アイヌウィムワ、センカブ、ヌイト、カネ、タマコロワネンベアニビリカイカラカラビリカイスアイイキ。

タバンイコオットイロホビリカ、スベビリカフチムケワチンケ。イコロエチヌカレンカツエネニ。センカトトラアンヌワウンカライミカイアン。センカトトラアンテセワウンカラチタベカイアン。ルトムウンクイカラワカラクカイアン。ヤンケモリウンクワタマシカイアン、ボンコンコワタマシカイアン。サウカネサウチュウアイントメイコロ、イカヨビコロ、ダンバクオ、マキリカイアン。イノオットアインワンケトウキ、タカサイラ、パツチカイアン。

ネンベトラクスロツタシクスチップ美恵子 (1948~2010) カラベカイエチヌカレン。エバキタ、貝澤徹 (イヌイエル)、下倉洋之 (カネイエル)、藤戸康平 (イコロカラ) タネウサオカイビリカカラワコアスルアウンネワクスネンベカイチンケワエチヌカレン。

タバンアナケ、アイヌプリビラサクニ、公益財団法人アイヌ民族文化財団 (アヌスコロトミアニアイヌプリビラサウタラ) ケバキネワ、タネ30スイチシタンナ。タネエチヌカラワ、テデタアイヌプリ、タネアンアイヌプリ、アムキリクインネチキビリカ。イヨッタイオシ、タバンイコチコロテサ、ビリカイコロウネルサアウタラウサウウンカスイウタラ、バセタラチコヤイライケシタンナ。

北海道立釧路芸術館
京都府
京都府京都文化博物館
公益財団法人アイヌ民族文化財団

図 1: 『アイヌの美・彩りと輝き』ウエランカラ (左) を Qwen3.5-0.8B で OCR した結果 (右)

^{a)}<https://qwen.ai/blog?id=qwen3.5>

はできて、めったに出力されない。図1右で「プ」などのいわゆるアイヌ語小書きカタカナが出力されていないのは、これが原因である(表1)。

表1: Qwen3.5-0.8B のトークン (小書きカタカナ)

ア	149368	へ	[159, 229, 116]
イ	54198	ホ	[159, 229, 117]
ウ	158936	ム	[159, 229, 118]
エ	74607	ヤ	65203
オ	149185	ユ	68850
カ	247021	ヨ	74469
ク	[159, 229, 108]	ラ	[159, 229, 119]
ケ	158810	リ	[159, 229, 120]
シ	[159, 229, 109]	ル	[159, 229, 121]
ス	[159, 229, 110]	レ	[159, 229, 122]
ツ	24404	ロ	[159, 229, 123]
ト	[159, 229, 111]	ワ	[2687, 106]
ヌ	[159, 158830]		
ハ	[159, 229, 113]		
ヒ	[159, 229, 114]		
フ	[159, 229, 115]		
プ	[159, 229, 115, 3293, 248]		

Qwen3.5-0.8B は、248320 種類のトークンを、それぞれ 1024 次元の「単語ベクトル」で表している。トークン 0~248319 のうち、0~248043 が UTF-8 バイト列によるトークン、248044~248076 が制御用の特殊トークン、248077~248319 が未使用である。ここにアイヌ語小書きカタカナを追加するにあたって、われわれは、特殊トークンを 248063~248095 に移動し、該当する「単語ベクトル」も同時に移動した。その上で、半濁点 (e3 82 9a) をトークン 248044 に追加し、アイヌ語小書きカタカナ UTF-8 のアタマ 2 バイト (e3 87) をトークン 248045 に追加し、「ク」「シ」「ス」「ト」「ヌ」「ハ」「ヒ」「フ」「ヘ」「ホ」「ム」「ラ」「リ」「ル」「レ」「ロ」をトークン 248046~248061 に追加し、さらに「プ」をトークン 248062 に追加した。Qwen2Tokenizer のトークナイザ辞書に「プ」(e3 87 b7 e3 82 9a) を追加するには、「フ」(e3 87 b7) と半濁点 (e3 82 9a) を組み上げなければならず、しかも特殊トークンをまたいではいけないうえ、こういう手順でのトークン追加が必要だった。なお、各トークンの「単語ベクトル」の初期値は、たとえば「プ」であれば、「プ」の「単語ベクトル」から「ヤ」を引いて「ヤ」を加えた値とした。

トークン追加をおこなうと、モデルの動作が不安定となる(一度「プ」を出力すると、そのまま無限に「プ」を出力しつづけてしまう)ため、さらに追加学習をおこなった。追加学習に用いたのは、国立アイヌ民族博物館の「ごあいさつ」(図3)を段落で区切った画像 65 枚と、大小カタカナの比較画像 17 枚である。なお、各画像の長辺が 1024 ドット以下になるよう、一部の画像は縮小した上で、全ての画像に読み取った情報を手作業でテキスト付与した。trl 0.29.1 での追加学習 (3 epochs) に要した時間は、NVIDIA GeForce RTX 3080 Laptop (16GB) 1GPU で 19 分程度だった。

評価と考察

追加学習後のモデル Qwen3.5-0.8B-AinuOCR で、図1左の画像を OCR してみたところ、図2の結果が得られた。小書きの「プ」も出力されており、かなり良くなっている気がする。比較のため、原テキスト(「プ」を1文字とみなし、連続する空白と改行を1文字にまとめた結果、総文字数 1124 字)を手作業で準備し、レーベンシュタイン距離を見てみたところ、図1右が 412 (単純誤り率 36.7%) だったのに対し、図2は 144 (単純誤り率 12.8%) と、かなりの改善が見られた。

結果に気を良くしたわれわれは、トークン追加と追加学習を Qwen3.5-2B に適用し、さらに Qwen3.5-4B にも適用してみた。ただし、RTX 3080 LapTop の

ウエランカラ
令和7年度アイヌ工芸品展「アイヌ之美・彩りと輝き」アニアレヘチコレワ、ウサム
ベカイコロチアマワエヌカクレアン。
アイヌアナケネテエタワノアンクルネワ20イクネバエムコバノヤウンモシリ、
ヤンケモシリ(カララト)、ルムトオツカイ。タネヤウンモシリタ
カイ、モシリエビタオカイ。テエタアナケネチクニ、ニカブ、キナカブ、チコキブボネ、
キラウ、シケチ、ルシ、チュバカブ、セイアニウサオカイベカラ。ネアンベオタクシ
アネイワンケブカイアン。イノモオッタアネイワンケブカイアン。ビシカンモシリ
エオカイウタツトウラ、アイヌウイマムワ、センカヌ、ヌイト、カネ、タマコロワネアンベ
アニビリカイカラカラビリカイヌイエイヌキ。
タバンイキオッタイロホビリカワ、ヌベキビリカワチヌクワチチンケ。イコロ
エチヌカクレアンカトウエネアニ。センカヌトウラアンヌワアンカイイミカイアン。
センカヌトウラアンテセワアンカワチタラベカイアン。ルムトウンクルイカラカラ
カラクツカイアン。ヤンケモシリウンクルコロタマサイカイアンワ、ボンコンコロ
タマサイカイアン。ウサカネウサヌチラウアニアントメイコロ、イカヨビコロ、
ダンバクオプ、マカリカイアン。イノモオッタアネイワンケパートウキ、タカサイラ、パツ
カイアン。
ネアンベトウラノクスロツタシクスチカブ美恵子(1948~2010)カラベカイ
エチヌカクレアン。エバキタ、貝澤徹(イスイエクル)、下倉砂之(カネヌイエクル)、ウシラン平
(イコロカクル)タヌウサオカイビリカブカラワコアスルアシウタンネワクスネアンベ
カイチチンケワエヌカクレアン。
タバアナケネ、アイヌブリビラサクニ、公益財団法人アイヌ民族文化財団(アヌコロ
トミニアニヌブリビラサウタラ)ケシバキネフワ、タネ30スイチキシッタン
ナ。タバエチチカラワ、テエタアイヌブリ、タネアンアイヌブリ、アムキリクルインネ
チキビリカ。イヨッタイオン、タバンイキチコロテサザマ、ビリカコロウヌサウタラ
ウサウサウカスウタラ、パセタラチコヤイライケシッタンナ。
北海道立釧路芸術館
京都府
京都府京都文化博物館
公益財団法人アイヌ民族文化財団

図2: 図1左を Qwen3.5-0.8B-AinuOCR した結果

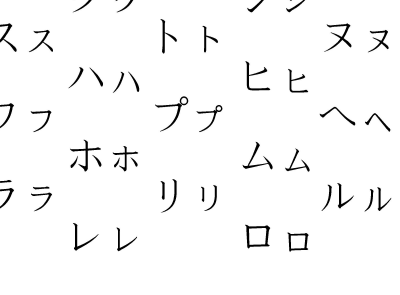
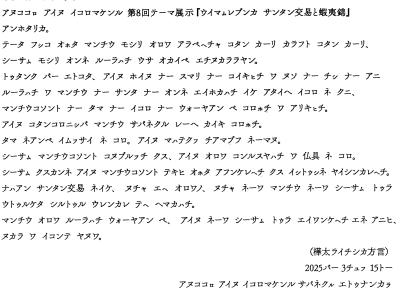
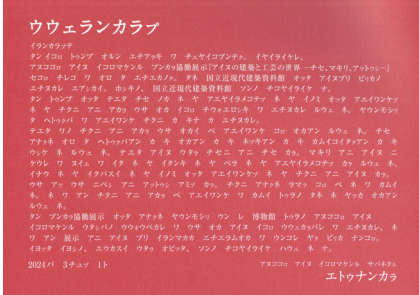
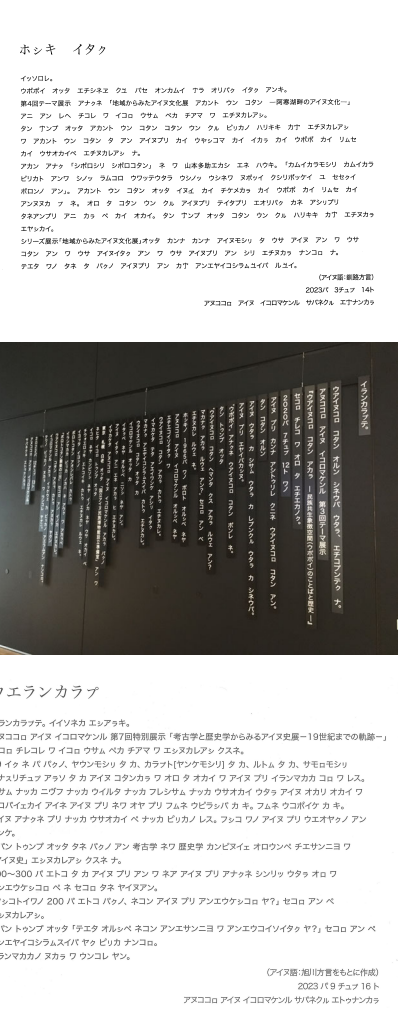
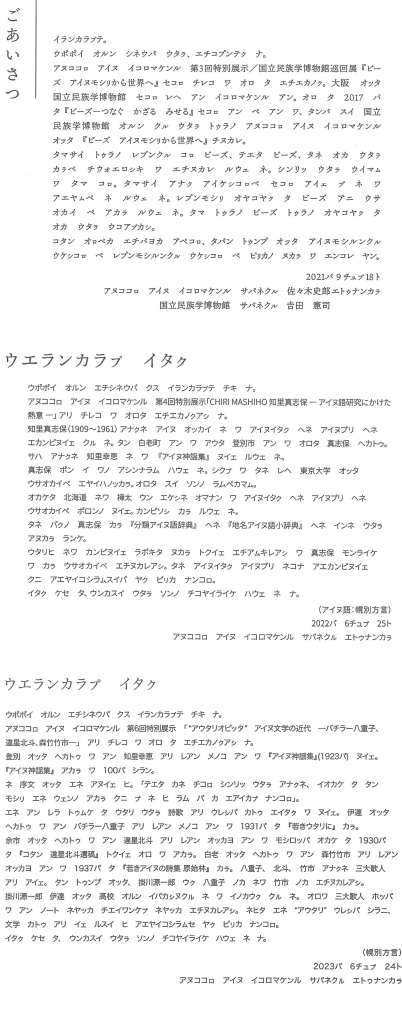
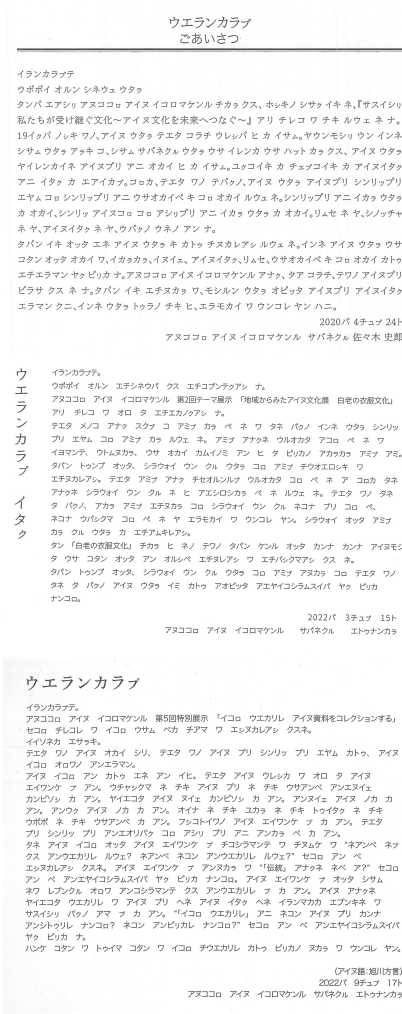


図 3: アイヌ語 OCR の追加学習に用いた画像

GPU メモリ (16GB) に乗り切らなかったことから、AMD Ryzen7 H255 (64GB)^{b)}で追加学習 (3 epochs) したところ、Qwen3.5-2B で 53 分, Qwen3.5-4B で 110 分を要した。

追加学習後の Qwen3.5-2B-AinuOCR・Qwen3.5-4B-AinuOCR で、図 1 左の画像を OCR した結果を、そ

れぞれ図 4 右・図 5 右に示す。いずれもレーベンシュタイン距離は 88 (単純誤り率 7.8%) となった。

こうなると、Qwen3.5-9B にも挑戦したくなるのだが、モデルの規模が大きくなると、メモリ不足で追加学習がうまくいかない。本稿執筆時点での各モデルの評価結果を、表 2 に示す。この結果を見る限り、解析精度とモデル規模のバランスがよいのは、Qwen3.5-2B-AinuOCR ということになる。

^{b)}H255 には Radeon 780M が搭載されているが、ROCm の gfx1103 がイマイチ安定しないことから、本稿では CPU で追加学習をおこなった。

ウエランカラブ 令和 7 年度アイヌ工芸品展「アイヌの美・彩りと輝き-」アニ アン レヘ チコレ ワ、ウサム ベカ イコロ チアマワ エチヌカレアン。 アイヌ アナクネ テエタ ワノ アンクル ネ ワ 20 イク ネ バエム コ バッノ ヤウンモシリ、 ヤンケモシリ (カラプト)、ルトム オッタ オカイ。タネ ヤウンモシッタ カイ サモロモシッタ カイ、モシリ エビッタ オカイ。テエタ アナクネ チクニ、ニカヂ、キナカヂ、チコイキバボネ、 キラウ、シキテ、ルシ、チェブカヂ、セイ アニ ウサ オカイ ベカラ。ネアンベ オッタ ケスト アネイワンケ ナ カイ アン。イノミ オッタ アネイワンケ ナ カイ アン。ビシカン モシリ エオカイ ウタツウラ、アイヌ ウイマム ワ、センカヂ、ヌイト、カネ、タマ コロ ワ ネアンベ アニ ビリカ イカラカラ ビリカ イヌイ エアイヌ キ。 タバン イク オッタ イロホ ビリカ ナ、ヌベキ ビリカ ナ チヌム ケ ワ チサンケ。イコロ エチヌカレアン カトウ エネ アニ。センカヂ トウラ アンニ ワ アンカラ イミカイ アン。 センカヂ トウラ アンデセ ワ アンカラ チタラベカイ アン。ルトム ウンクル イカラカラ ワ カラ クク カイ アン。ヤンケモシリ ウンクル コロ タマサイ カイ アン ワ、ボン コンコ コロ タマサイ カイ アン。ウサ カネ ウサ ユッキラウ アニ アントム テイコロ、イカヨビコロ、 タンバクオ、マキリ カイ アン。イノミ オッタ アネイワンケ トウキ、タカイサラ、パッチ カイ アン。 ネアンベ トウラ ノ クスロッタ シツヌ チカッ プ美恵子 (1948～2010) カラ ベカイ エチヌカレアン。エエバキタ、貝澤徹 (イヌイェクル)、下倉洋之 (カネヌイェクル)、藤戸康平 (イコロカラクル) タネ ウサ オカイ ビリカ プカラ ワ コアスル アシ ウタン ネ ワ クス ネアンベ カイ チサンケ ワ エチヌカレアン。 タナ アナクネ、アイヌ プリ ビラサ クニ、公益財団法人アイヌ民族文化財団 (アヌコロ トミアニ アイヌ プリ ビラサ ウタラ) ケシ バキ プネ ワ、タネ 30 スイ チキ シッタン ナ。タンベ エチヌカラ ワ、テエタ アイヌ プリ、タネ アン アイヌ プリ、アム キル クル インネ チキ ビリカ。イヨッタ イオシ、タバン イキ チコロ テッサマ、ビリカイ コロ ウネルサ ア ウタラ ウサ ウサ ウンカスイ ウタラ、バセタラ チコヤイライ ケ シッタン ナ。 北海道立 釧路芸術館 京都府 京都府 京都文化博物館 公益財団法人アイヌ民族文化財団	ウエランカラブ 令和7年度アイヌ工芸作品展「アイヌの美・彩りと輝き-」アニ アン レヘ チコレ ワ、ウサム ベカ イコロ チアマワ エチヌカレアン。 アイヌ アナクネ テエタ ワノ アンクル ネ ワ 20 イク ネ バエム コ バッノ ヤウンモシリ、 ヤンケモシリ (カラプト)、ルトム オッタ オカイ。タネ ヤウンモシッタ カイ サモロモシッタ カイ、モシリ エビッタ オカイ。テエタ アナクネ チクニ、ニカヂ、キナカヂ、チコイキナボネ、 キラウ、シキテ、ルシ、チェブカヂ、セイ アニ ウサ オカイ ベカラ。ネアンベ オッタ ケスト アネイワンケ プカイ アン。イノミ オッタ アネイワンケ プカイ アン。ビシカン モシリ エオカイ ウタツウラ、アイヌ ウイマム ワ、センカキ、ヌイト、カネ、タマ コロ ワ ネアンベ アニ ビリカ イカラカラ ビリカ イヌイ エアイヌ キ。 タバン イク オッタ イロホ ビリカ プ、ヌベキ ビリカ プ チヌム ケ ワ チサンケ。イコロ エチヌカレアン カトウ エネ アニ。センカキ トウラ アンニ ヌ ワ アンカラ イミカイ アン。 センカキ トウラ アンデセ ワ アンカラ チタラベカイ アン。ルトム ウンク イカラカラ ワ カラ クク カイ アン。ヤンケモシリ ウンクル コロ タマサイ カイ アン ワ、ボン コンコ コロ タマサイ カイ アン。ウサ カネ ウサ ユッキラウ アニ アントム テイコロ、イカヨビコロ、 タンブクオバ、マキリ カイ アン。イノミ オッタ アネイワンケ ボトウキ、タカイサラ、パッチ カイ アン。 ネアンベ トウラ ノ クスロッタ シクヌ チカッ プ美恵子 (1948～2010) カラ ベカイ エチヌカレアン。エエバキタ、貝澤徹 (イヌイェクル)、下倉洋之 (カネヌイェクル)、藤戸康平 (イコロカラクル) タネ ウサ オカイ ビリカ プカラ ワ コアスル アシ ウタン ネ ワ クス ネアンベ カイ チサンケ ワ エチヌカレアン。 タバ アナクネ、アイヌ プリ ビラサ クニ、公益財団法人アイヌ民族文化財団 (アヌコロ トミアニ アイヌ プリ ビラサ ウタラ) ケシ バキ プネ ワ、タネ 30 スイ チキ シッタン ナ。タンブ エチヌカラ ワ、テエタ アイヌ プリ、タネ アン アイヌ プリ、アム キル クル インネ チキ ビリカ。イヨッタ イオシ、タバン イキ チコヘ テッサマ、ビリカイ コロ ウネルサ ア ウタラ ウサ ウサ ウンカスイ ウタラ、バセタラ チコヤイライ ケ シッタン ナ。 北海道立 釧路芸術館 京都府 京都府 京都文化博物館 公益財団法人アイヌ民族文化財団
---	---

図 4: 図 1 左を Qwen3.5-2B で OCR した結果 (左) と Qwen3.5-2B-AinuOCR した結果 (右)

ウエランカラブ 令和 7 年度アイヌ工芸品展「アイヌの美・彩りと輝き-」アニ アン レヘ チコレ ワ、ウサム ベカ イコロ チアマワ エチヌカレアン。 アイヌ アナクネ テエタ ワノ アンクル ネ ワ 20 イク ネ バエム コ バッノ ヤウンモシリ、 ヤンケモシリ (カラプト)、ルトム オッタ オカイ。タネ ヤウンモシリ タ カイ サモロモシッタ カイ、モシリ エビッタ オカイ。テエタ アナクネ チクニ、ニカヂ、キナカヂ、チコイキバボネ、 キラウ、シキテ、ルシ、チェブカヂ、セイ アニ ウサ オカイ ベカラ。ネアンベ オッタ ケスト アネイワンケン カイ アン。イノミ オッタ アネイワンケン カイ アン。ビシカン モシリ エオカイ ウッタ トウラ、アイヌ ウイマム ワ、センカキ、ヌイト、カネ、タマ コロ ワ ネアンベ アニ ビリカ イカラカラ ビリカ イヌイ エアイヌ キ。 タバン イキ オッタ イロホ ビリカン、ヌベキ ビリカン チヌム ケ ワ チサンケ。イコロ エチヌカレアン カトウ エネ アニ。センカキ トウラ アンニ ス ワ アンカラ イミカイ アン。 センカキ トウラ アンデセ ワ アンカラ チトラベカイ アン。ルトム ウンクル イカラカラ ワ カラ クク カイ アン。ヤンケモシリ ウンクル コロ タマサイ カイ アン ワ、ボン コンコ コロ タマサイ カイ アン。ウサ カネ ウサ ユッキラウ アニ アントム テイコロ、イカヨビコロ、 タンバクオ、マキリ カイ アン。イノミ オッタ アネイワンケ ボートウキ、タカイサラ、パッチ カイ アン。 ネアンベ トウラ ノ クスロッタ シクス チカッ プ美恵子 (1948～2010) カラ ベカイ エチヌカレアン。エエバキタ、貝澤徹 (イヌイェクル)、下倉洋之 (カネヌイェクル)、藤戸康平 (イコロカラクル) タネ ウサ オカイ ビリカ ンカラ ワ コアスル アシ ウタン ネ ワ クス ネアンベ カイ チサンケ ワ エチヌカレアン。 タバ アナクネ、アイヌ プリ ビラサ クニ、公益財団法人アイヌ民族文化財団 (アヌコロ トミアニ アイヌ プリ ビラサ ウタラ) ケシ バキ プネ ワ、タネ 30 スイ チキ シッタン ナ。タンベ エチヌカラ ワ、テエタ アイヌ プリ、タネ アン アイヌ プリ、アム キル クル インネ チキ ビリカ。イヨッタ イオシ、タバン イキ チコロ テッサマ、ビリカイ コロ ウネルサ ア ウタラ ウサ ウサ ウンカスイ ウタラ、バセタラ チコヤイライ ケ シッタン ナ。 北海道立 釧路芸術館 京都府 京都府 京都文化博物館 公益財団法人アイヌ民族文化財団	ウエランカラブ 令和 7 年度アイヌ工芸品展「アイヌ之美・彩りと輝き-」アニ アン レヘ チコレ ワ、ウサム ベ カイ コロ チアマワ エチヌカレン。 アイヌ アナクネ テエタ ワノ アンクル ネ ワ 20 イク ネ バエム コ バッノ ヤウンモシリ、ヤン ケモシリ (カラプト)、ルトム オッタ オカイ。タネ ヤウンモシッタ カイ サモロモシッタ カ イ、モシリ エビッタ オカイ。テエタ アナクネ チクニ、ニカヂ、キナカヂ、チコイキバボネ、 キラウ、シキテ、ルシ、チェブカヂ、セイ アニ ウサ オカイ ベカラ。ネアンベ オッタ ケシ ト アネイワンケ プカイ アン。イノミ オッタ アネイワンケ プカイ アン。ビシカン モシリ エ オカイ ウタツウラ、アイヌ ウイマム ワ、センカキ、ヌイト、カネ、タマ コロ ワ ネアンベ アニ ビリカ イカラカラ ビリカ イヌイ エアイヌ キ。 タバン イキ オッタ イロホ ビリカ プ、ヌベキ ビリカ プ チヌム ケ ワ チサンケ。イコロ エチヌ カレン カトウ エネ アニ。センカキ トウラ アンニ ヌ ワ アンカラ イミカイ アン。センカキ ト ウラ アンデセ ワ アンカラ チタラベカイ アン。ルトム ウンクル イカラカラ ワ カラ ク プカ イ アン。ヤンケモシリ ウンクル コロ タマサイ カイ アン ワ、ボン コンコ コロ タマサイ カ イ アン。ウサ カネ ウサ ユエクラウ アニ アントム エイコロ、イカヨビコロ、タンバクオ、マキ リ カイ アン。イノミ オッタ アネイワンケ プートウキ、タカイサラ、パッチ カイ アン。 ネアンベ トウラ ノ クスロッ タ シクス チカ プ美恵子 (1948～2010) カラ ベカイ エチヌカ レン。エエバキタ、貝澤徹 (イヌイェクル)、下倉洋之 (カネヌイェクル)、藤戸康平 (イコロカ ラクル) タネ ウサ オカイ ビリカ プカラ ワ コアスル アシ ウタン ネ ワ クス ネアンベカイ チサンケ ワ エチヌカレン。 タバ アナクネ、アイヌ プリ ビラサ クニ、公益財団法人アイヌ民族文化財団 (アヌコロ トミ アニ アイヌ プリ ビラサ ウタラ) ケシ バキ プネ ワ、タネ 30 スイ チキ シッタン ナ。タネ プ エチ ヌカラ ワ、テエタ アイヌ プリ、タネ アン アイヌ プリ、アム キル クル インネ チキ ビリカ。 イ ヨクタイ オホ、タバン イキ チコロ テッサマ、ビリカイ コロ ウネルサ ア ウタラ ウサ ウサ ウ ンカスイ ウタラ、バセタラ チコヤイライ ケ シッタン ナ。 北海道立 釧路芸術館 京都府 京都府 京都文化博物館 公益財団法人アイヌ民族文化財団
--	---

図 5: 図 1 左を Qwen3.5-4B で OCR した結果 (左) と Qwen3.5-4B-AinuOCR した結果 (右)

表 2: 各モデルで図 1 左を OCR した際の評価

	Qwen3.5-xB	Qwen3.5-xB-AinuOCR
0.8B	412 (36.7%)	144 (12.8%)
2B	122 (10.9%)	88 (7.8%)
4B	126 (11.2%)	88 (7.8%)
9B	217 (19.3%)	89 (7.9%)

※レーベンシュタイン距離 (単純誤り率)

おわりに

小書きカタカナを含むアイヌ語 OCR モデルとして、Qwen3.5-0.8B-AinuOCR, Qwen3.5-2B-AinuOCR, Qwen3.5-4B-AinuOCR を開発した。これらのモデルは <https://huggingface.co/KoichiYasuoka> で公開している⁹⁾ので、ぜひ使ってみてほしい。なお本稿は、京都大学未踏科学研究ユニット「フィールドワークとデータサイエンスを融合した超学際研究ユニット」の成果である。

参考文献

- [1] 佐藤知己: アイヌ語を記述するのに必要な文字セットについて, JIS 符号化文字集調査研究委員会第 2 分科会 (WG2) 資料, JCS-2-8-02 (1996 年 11 月 25 日).
- [2] Addition of forty eight characters, JTC1/SC2/WG2 N2092 (September 13, 1999).
- [3] アカント ウン コタン —阿寒湖畔のアイヌ文化—, 白老: 国立アイヌ民族博物館 (2023 年 3 月 14 日).

ホシキ イタク

イッソロシ。
ウボボイ オッタ エチシネエ クム バセ オンカムイ テラ オリバク イタク アンキ。
第4回テーマ展示 アナタネ 「地域からみたアイヌ文化展 アカント ウン コタン —阿寒湖畔のアイヌ文化—」
アニ アン レヘ チコレ ワ イコロ ウサム ベカ チアマ ワ エチヌカレアシ。
タン テンブ オッタ アカント ウン コタン コタン ウン クル ビリカノ ハリキキ カツ エチヌカレアシ
ワ アカント ウン コタン タ アン アイヌプリ カイ ウヤシコマ カイ イカラ カイ ウボボ カイ リムセ
カイ ウサオカイベ エチヌカレアシ ナ。
アカン アナク 「シボロシリ シボロコタン」 ネ ワ 山本多助エカシ エネ ハウキ。 「カムイカラモシリ カムイカラ
ビリカノ アンワ シノッ ラムコロ ウワツチウタラ ウシノッ ウシネワ ノボッイ クシリボツケイ ユ セセケイ
ボロンノ アン」。 アカント ウン コタン オッタ イヌエ カイ チケメカラ カイ ウボボ カイ リムセ カイ
アンヌヌカ ヲ ネ。 オロ タ コタン ウン クル アイヌプリ テイタプリ エオリバク カネ アシップリ
タネアンプリ アニ カラ ベ カイ オカイ。 タン テンブ オッタ コタン ウン クル ハリキキ カツ エチヌカ
エヤシカイ。
シリーズ展示「地域からみたアイヌ文化展」オッタ カンナ カンナ アイヌモシリ タ ウサ アイヌ アン ワ ウサ
コタン アン ワ ウサ アイヌイタク アン ワ ウサ アイヌプリ アン シリ エチヌカラ ナンコロ ナ。
チエタ フノ タネ タ バッノ アイヌプリ アン カツ アンエヤシコシラムユイバ ルユイ。

(アイヌ語: 網路方言)
2023/3 3チュブ 14ト
アヌコロ アイヌ イコロマケル サバネクル エヌナンカラ

- [4] アイヌ・モシリ, オ 3 号 (1958 年 3 月), p.7.
- [5] 中川裕: アイヌ人によるアイヌ語表記への取り組み, 表記の習慣のない言語の表記 (東京外国語大学アジア・アフリカ言語研究所, 2006 年 3 月), pp.1-44.

付録 アイヌ語のカタカナに関して

本稿では、小書きカタカナを追加するやり方で、カタカナアイヌ語 OCR の開発をおこなったが、追加するのは小書きカタカナだけでいいのか。たとえば『アカント ウン コタン』[3] のホシキ イタク (図 3 右上) 末尾には、「エヌナンカラ」という文字列が見える。館長の佐々木史郎 (エトウナンカラ) が、この「ごあいさつ」では山本多助のアイヌ文字 [4, 5] を使用しており、「トウの合字」以外にも「シュの合字」「チャの合字」「イエの合字」が見える¹⁰⁾。

Qwen3.5 のトークナイザには、「トウ」(e3 83 88 e3 82 a5)を表すトークン 220263 が準備されており、これを使うという戦略もあるが、そうすると「トウ」という文字列と「ㇿ」の見分けがつかなくなる。その戦略は、「プ」と「ㇿ」を見分けるべく頑張る、という本稿の主旨に反する。かと言って「ㇿ」をトークナイザ辞書に追加しようにも、この字が UCS に含まれていない (つまり UTF-8 が決まっていない) ので、現状どうしようもない。

そんなわけなので、このような合字を含む段落は、とりあえず、追加学習から取り除いておいた。結果として、われわれのアイヌ語 OCR は、これらの合字が読めない (図 6)。今後の課題としておきたい。

ホシキ イタク

イッソロシ。
ウボボイ オッタ エチシネエ クム バセ オンカムイ ヤラ オリバク イタク アンキ。
第4回テーマ展示 アナタネ 「地域からみたアイヌ文化展 アカント ウン コタン —阿寒湖畔のアイヌ文化—」
アニ アン レヘ チコレ ワ イコロ ウサム ベカ チアマ ワ エチヌカレアシ。
タンヤンブ オッタ アカント ウン コタン コタン ウン クル ビリカノ ハリキキ カン エチヌカレアシ
ワン アカント ウン コタン タ アン アイヌプリ カイ ウヤシコマ カイ イカラ カイ ウボボ カイ リムセ
カイ ウサオカイベ エチヌカレアシ ナ。
アカン アナク 「シボロシリ シボロコタン」 ネ ワ 山本多助エカシ エネ ハウキ。 「カムイカラモシリ カムイカラ
ビリカノ アンワ シノッ ラムコロ ウワツチウタラ ウシノッ ウシネワ ノボッイ クシリボツケイ ユ セセケイ
ボロンノ アン」。 アカント ウン コタン オッタ アイヌエ カイ チケメカラ カイ ウボボ カイ リムセ カイ
アヌヌカ プ ネ。 オロ タ コタン ウン クル アイヌプリ テイタプリ エオリバク カネ アシップリ
タネアンプリ アニ カラ ベ カイ オカイ。 タンヤンブ オッタ コタン ウン クル ハリキキ カン エチヌカラ
エヤシカイ。
シリーズ展示「地域からみたアイヌ文化展」オッタ カンナ カンナ アイヌモシリ タ ウサ アイヌ アン ワ ウサ
コタン アン ワ ウサ アイヌイタク アン ワ ウサ アイヌプリ アン シリ エチヌカラ ナンコロ ナ。
チエタ フノ タネ タ バッノ アイヌプリ アン カン アンエヤシコシラムヰイバ ルヰイ。
(アイヌ語: 網路方言)
2023/3 3チュブ 14ト
アヌコロ アイヌ イコロマケル サバネクル エヌナンカラ

図 6: 『アカント ウン コタン』ホシキ イタク (左) を Qwen3.5-2B-AinuOCR した結果 (右)

⁹⁾各モデルのリポジトリに、開発に用いた `maker.py` も含めておいた。

¹⁰⁾<https://qiita.com/KoichiYasuoka/items/1be160bc15d73edf6593>