ローマ字・カタカナ・キリル文字併用アイヌ語RoBERTa・DeBERTaモデルの開発

AUTHOR(S):

安岡, 孝一

CITATION:

安岡, 孝一. ローマ字・カタカナ・キリル文字併用アイヌ語RoBERTa・DeBERTaモデルの開発. 情報処理学会研究報告 : 人文科学とコンピュータ(CH) 2023, 2023-CH-131(7): 1-7

ISSUE DATE:

2023-02-11

URL:

http://hdl.handle.net/2433/279486

情報処理学会研究報告
IPSJ SIG Technical Report
Vol.2023-CH-131 No.7
2023/2/18

ローマ字・カタカナ・キリル文字併用 アイヌ語RoBERTa・DeBERTaモデルの開発

安岡孝一 (京都大学人文科学研究所附属東アジア人文情報学研究センター)

書写言語としてのアイヌ語は，ローマ字 (ラテンアルファベット)・カタカナ・キリル文字など，多彩な文字と記法によって記述されてきた．その一方，抱合語としてのアイヌ語は，日本語や欧米諸語とは全く異なる言語構造を持つことから，これらの言語向けの RoBERTa・DeBERTa モデルは，そのままではアイヌ語に適用できない．本発表では，ローマ字・カタカナ・キリル文字で書かれたアイヌ語に対し，RoBERTa・DeBERTa モデルを開発する手法を示し，さらに形態素解析・係り受け解析への応用について考察する．

1 はじめに

アイヌ語 Universal Dependencies は，東京大学の瀬沼甫がローマ字 (ラテンアルファベット) 版を開発 [1, 2] し，筆者がカタカナ・キリル文字への拡張をおこなった [3, 4]．最初の公開は『アイヌ神謡集』[5] の「ホテナオ」のみだったが，これに『アイヌ語会話字典』[6] や『国立アイヌ民族博物館ガイドブック』[7] を加え，順調とは言えないものの徐々にコーパスを増やしている[a]．

筆者が研究代表者を務める学際大規模情報基盤共同利用・共同研究拠点公募型共同研究『単語間に区切りのない書写言語における係り受け解析エンジンの開発』(共同研究者: 山崎直樹・二階堂善弘・師茂樹・鈴木慎吾・Christian Wittern・池田巧・守岡知彦・白須裕之・藤田一乗) では，各言語ごとに RoBERTa [8]・DeBERTa [9] などの事前学習モデルを製作し，Universal Depedencies コーパスでファインチューニングする，というやり方で，係り受け解析エンジンを開発・公開[b]している．では，ローマ字・カタカナ・キリル文字を併用する形でのアイヌ語 RoBERTa・DeBERTa モデルは，どのようなやり方で製作すべきか，この点について本稿で議論する．

2 Universal Dependencies の概要

Universal Dependencies (UD) [10] は，書写言語における品詞・形態素属性・依存構造 (係り受け関係) を，言語に関わらず記述する手法である．句構造を考慮せずに係り受け関係を記述することで，言語横断性を高めており，全ての文法構造を単語間のリンクで記述するのが特徴である．

依存構造解析それ自体は，Tesnière の構造的統語論 [11] に源を発し，Мельчук の有向グラフ記述 [12] によって，一応の完成を見た手法である．その最大の特長は，いわゆる動詞中心主義によって言語横断的な記述が可能だという点にあり，Мельчук 依存文法をコンピュータ向けに洗練した UD においても，言語に関わらない記述，という特長が前面に押し出されている．UD における文法構造記述は，句構造を考慮せず，全てを単語間のリンクとして表現する．これは，Мельчук の有向グラフ記述が，単語間のリンクという形態を取っていたからであり，そういう割り切りの結果として，言語横断的な文法構造記述を可能としているのである．

UD 依存構造コーパスの交換用フォーマットとして，CoNLL-U と呼ばれるタブ区切りテキスト (文字コードは UTF-8) が規定されている．CoNLL-U の各行は各単語に対応しており，以下に示す 10 個のタブ区切りフィールドで構成される．

1. ID: 単語ごとに付与されたインデックスで，文ごとに 1 から始まる整数．縮約語に対しては，単語の範囲を示すのも可．
2. FORM: 語，または，句読記号．
3. LEMMA: 基底形，語幹．
4. UPOS: UD で規定された言語普遍な品詞タグ[c]．
5. XPOS: 言語固有の品詞タグ．
6. FEATS: UD で規定された言語普遍な形態素属性のリスト．言語固有の拡張も可．
7. HEAD: 当該の単語の係り受け元 ID．係り受け

[a] https://github.com/KoichiYasuoka/UD-Ainu
[b] https://huggingface.co/KoichiYasuoka

[c] ADJ・ADP・ADV・AUX・CCONJ・DET・INTJ・NOUN・NUM・PART・PRON・PROPN・PUNCT・SCONJ・SYM・VERB・X の 17 種類．

表 1: UD 係り受けタグ (DEPREL)

	Nominals	Clauses	Modifier Words	Function Words
Core arguments	nsubj 主語 obj 目的語 iobj 間接目的語	csubj 節主語 ccomp 節目的語 xcomp 節補語		
Non-core dependents	obl 斜格補語 vocative 呼称語 expl 形式語 dislocated 外置語	advcl 連用修飾節	advmod 連用修飾語 discourse 談話要素	aux 動詞補助成分 cop 繋辞 mark 標識
Nominal dependents	nmod 体言による連体修飾語 appos 同格 nummod 数量による修飾語	acl 連体修飾節	amod 用言による連体修飾語	det 決定詞 clf 類別詞 case 格表示
Coordination	**MWE**	**Loose**	**Special**	**Other**
conj 接続 cc 接続詞	fixed 固着 flat 並列 compound 複合	list 細目 parataxis 隣接表現	orphan 親なし goeswith 泣き別れ reparandum 言い損じ	punct 句読点 root 親 dep 未定義

```
# text = kamuy tura okay=an
1       kamuy       kamuy       NOUN    名詞        _   3   obl     _   _
2       tura        tura        ADP     後置副詞    _   1   case    _   _
3       okay        an          VERB    自動詞      _   0   root    _   SpaceAfter=No
4       =an         =an         PART    人称接辞    _   3   nsubj   _   _

# text = カムイ トゥラ オカヤン
1       カムイ      kamuy       NOUN    名詞        _   3   obl     _   _
2       トゥラ      tura        ADP     後置副詞    _   1   case    _   _
3-4     オカヤン    _           _       _           _   _   _       _   _
3       オカイ      an          VERB    自動詞      _   0   root    _   _
4       アン        =an         PART    人称接辞    _   3   nsubj   _   _

# text = камуй тура окаян
1       камуй       kamuy       NOUN    名詞        _   3   obl     _   _
2       тура        tura        ADP     後置副詞    _   1   case    _   _
3-4     окаян       _           _       _           _   _   _       _   _
3       окай        an          VERB    自動詞      _   0   root    _   _
4       ан          =an         PART    人称接辞    _   3   nsubj   _   _
```

図 1: アイヌ語 UD の CoNLL-U データ

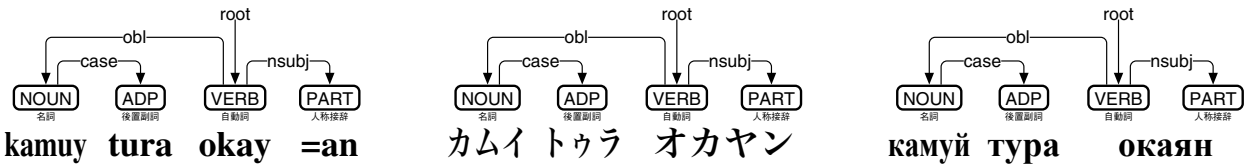

図 2: deplacy によるアイヌ語 UD の可視化

元が無い場合は 0 とする．

8. DEPREL: UD で規定された言語普遍な係り受けタグ (表 1)．HEAD が 0 の場合は root とする．言語固有の拡張も可．
9. DEPS: 複数の係り受け元を持つ場合，全ての HEAD:DEPREL ペア．
10. MISC: その他のアノテーション．

ID･FORM･LEMMA は，単語そのものに関するフィールドである．UPOS・XPOS・FEATS は，単語の品詞と形態素属性に関するフィールドである．HEAD・DEPREL・DEPS は，単語の係り受けに関するフィールドである．

UD における係り受け関係は，単語間の有向グラフを HEAD と DEPREL で記述する．HEAD は，その単語に入る有向枝のリンク元 ID を示しており，DEPREL は，その有向枝における係り受けタグである．ただし，HEAD が 0 の場合，その枝に入るリンク元は存在しない．リンクの本数は単語の個数に等しく，各リンクのリンク先は，全て互いに異なっている．すなわち，各単語から出るリンクは複数有り得るが，各単語に入るリンクは 1 つだけである．なお，リンクはループしない．

UD の係り受けリンクは，Мельчук 依存文法の後裔であり，いわゆる動詞中心主義である．動詞をリンク元として，主語や目的語へとリンクする．修飾関係においては，被修飾語から修飾語へとリンクする．ただし，側置詞 (前置詞や後置詞) を体言の修飾語だとみなす点 [13] が，Мельчук とは異なっている．ちなみに，コピュラ文においては，補語をリンク元として，主語へとリンクする．

アイヌ語 UD の例として，「kamuy tura okay=an」「カムイ トゥラ オカヤン」「камуй тура окаян」の CoNLL-U データを図 1 に示す．LEMMA と XPOS は，基本的に『アイヌ語沙流方言辞典』[14] に従っている．また，これらの CoNLL-U を比較すべく，deplacy [15] で可視化した (図 2)．UD 依存構造は全く同一だが，「オカヤン」や「окаян」は，文字の途中に単語境界がある点に注意されたい．

3 アイヌ語 RoBERTa・DeBERTa モデルの開発

アイヌ語 RoBERTa･DeBERTa モデルの開発に際し，筆者の頭を最も悩ませたのは，適切なトークナイザの設計だった．Transformers [16] は DeBERTa (V2) モデルに対し，SentencePiece [17] を基に Unigram トークナイザをサポートしている．単語間に空白やハイフンを含む言語に対しては，Unigram トークナイザの性能は非常に高く，ローマ字で書かれたアイヌ語だけを対象とするなら，Unigram トークナイザで十分だと考えられる．しかし，カタカナやキリル文字で書かれたアイヌ語は，単語境界が文字の途中に来てしまう場合がある．かなり手強い．

アイヌ語 (沙流方言) には閉音節があり，末子音として k, s, t, n, p, m, y, r, w が立ち得る．この直後に母音から始まる語が来ると，しばしばアンシェヌマンを起こす．カタカナで書かれている場合は，単語境界が文字の途中に来てしまう．たとえば「okay=an」は，カタカナでは「オカヤン」となる．キリル文字においては，末子音 y でのアンシェヌマンが，やはり問題となる．

そこで，アイヌ語 RoBERTa･DeBERTa モデルの内部ではローマ字だけを使い，入出力部でカタカナ・キリル文字との変換をおこなうことを考えた．変換には，Unigram トークナイザ同梱の Replace ノーマライザ (単文字変換[d]) を用いた．カタカナからローマ字への変換は，『ローマ字のつづり方』[18] 第 1 表の左半分に従うが，一部のカタカナは表 2(a) に変更している．ローマ字からカタカナへの逆変換は，通常は入力文字列をそのまま保持する形を取るが，運悪くカタカナの途中に単語境界が来てしまった場合は，末子音に表 2(a) を逆適用する．キリル文字からローマ字への変換は，ISO 9:1995 [19] に従うが，一部のキリル文字は表 2(b) に変更している．これらの変換

表 2: ローマ字変換表

(a) カタカナ

チ ci	ワ wa	ヰ wi	ヱ we	ヲ wo	ン n	ヂ ji
ヅ du	ー -	ア a	イ i	ウ u	エ e	オ o
ク k	シ s	ス s	ツ t	ト t	ヌ n	ハ h
ヒ h	フ h	ヘ h	ホ h	プ p	ム m	ラ r
リ r	ル r	レ r	ロ r			

(b) キリル文字

е ye	ё yo	ж j	й y	ъ ’	ы wi	ь ’
ю yu	я ya	і i				

(c) ローマ字縮退

b p	d t	f h	g k	j c	l r	q k
v u	w u	x h	y i	z s		

[d] 文字列からの変換も可能だが，現状では対応関係 (アラインメント) が狂ってしまう．小書きの「ㇷ゚」(U+31F7 U+309A) だけは，やむを得ず 2 文字から 1 文字への変換を用いているが，他は，1 文字から 1〜2 文字への変換とした．

に加えて，アクセント記号の除去，大文字から小文字への変換，さらに表 2(c) のローマ字縮退をおこなった上で，Unigram トークナイザ[e]へ渡される．この結果，たとえば「kamuy tura okay=an」という入力に対しては，モデル内部では「kamui」「tura」「okai」「=」「a」「n」とトークナイズされて，処理がおこなわれる．「カムイ トゥラ オカヤン」という入力に対しては，「kamui」「t」「o」「u」「r」「a」「okai」「a」「n」となる．「камуй тура окаян」に対しては，「kamui」「tura」「okai」「a」「n」となる．

このように構成した DeBERTa (V2) トークナイザを，国立アイヌ民族博物館アイヌ民話ライブラリ[f]・国立国語研究所アイヌ語口承文芸コーパス[g]・二風谷アイヌ文化博物館アイヌ口承文芸[h]・アジアアフリカ研究所アイヌ語資料[i]のアイヌ語テキストに適用し，deberta-base-ainu を製作・公開[b]した．また，RemBERT トークナイザを借りる[j]形で，roberta-base-ainu も製作・公開した．

4 アイヌ語形態素解析・係り受け解析向けファインチューニング

roberta-base-ainu と deberta-base-ainu に対し，アイヌ語形態素解析・係り受け解析向けのファインチューニングをおこなった．係り受け解析手法としては，esupar [20] の Biaffine [21] 実装と，ud-goeswith モデル [22] の Chu-Liu-Edmonds [23, 24] 実装を，比較してみることにした．

4 モデルの比較評価には，国立アイヌ民族博物館第 3 回テーマ展示『ウアイヌコロ コタン アカㇻ』(2022 年 12 月 13 日〜2023 年 2 月 12 日) の挨拶文から前半を抜粋[k]し，手作業で CoNLL-U 化したデータ (図 3) を用いた．評価指標には，CoNLL 2018 [25] の LAS (Labeled Attachment Score) / MLAS (Morphology-aware Labeled Attachment Score) / BLEX (Bi-LEXical dependency score) に加え，Tokens・UPOS の F1 値も見ることで，単語切り・品詞付与の精度も比較した．

評価結果を表 3 に示す．残念ながら Tokens の値が，いずれも不十分である．単語切りの段階で十分な精度が出ていないため，品詞付与・係り受け解析にかけて，結果がどんどん悪くなっている．処理結果を眺めてみたところ，6ヶ所も出てくる「ウアイヌコロ」が全く読めていない．「ウアイヌコロ」は，内部的には「u」「ainu」「kor」とトークナイズされるのだが，高頻度語である「ainu」と「kor」に引きずられて，「ainu」が NOUN，「kor」が VERB と品詞付与されてしまう．すると「u」が残ってしまうため，これが直前の語にくっ付いたりなど悪さをして，どんどん解析精度が下がっていってしまうのだ．悲しい．

表 3: 評価結果 (Tokens / UPOS / LAS / MLAS / BLEX)

RoBERTa esupar	76.23 / 68.86 / 32.23 / 25.81 / 12.90
ud-goeswith	65.84 / 61.04 / 31.33 / 24.24 / 18.18
DeBERTa esupar	78.16 / 69.89 / 40.15 / 29.67 / 16.48
ud-goeswith	59.17 / 56.45 / 31.45 / 28.74 / 17.96

ただ，『アイヌ語沙流方言辞典』での基本表記は「uwaynukor」であり，「uaynukor」が別表記として示されている．カタカナなら「ウワイヌコロ」と「ウアイヌコロ」ということになる．これらの別表記を両立させながら，うまく同一語として解析するような手法を編み出す必要がある，ということである．

5 おわりに

『ウアイヌコロ コタン アカㇻ』挨拶文の 3〜5 行目は，実際には「アヌココロ アイヌ イコロマケンル 第 3 回テーマ展示『ウアイヌコロ コタン アカㇻ —民族共生象徴空間 (ウポポイ) のことばと歴史—』セコロ チレコ ワ オロ タ エチエカノㇰ。」(展示は縦書き) である．アイヌ語に日本語が混在している文であり，われわれも一旦は，アイヌ語 UD と国語研短単位 UD を混在させる形で CoNLL-U を準備した (図 4)．ただ，この文をそのままアイヌ語 RoBERTa・DeBERTa モデルの評価に用いるのは，さすがに難があることから，「『ウアイヌコロ コタン アカㇻ』セコロ チレコ ワ オロ タ エチエカノㇰ。」へと抜粋した (図 3)．

本稿では，ローマ字・カタカナ・キリル文字を併用したアイヌ語 RoBERTa・DeBERTa モデルの製作手法を示した．現時点での解析精度は不十分だが，言語モデル開発の端緒は示せたと思う．ただ，現代におけるカタカナ書きのアイヌ語は，日本語との混在に向いており，そのようなアイヌ語と日本語の混在は，今後は増えていくのではないかと考えられる．な

[e] 語頭の Metaspace (U+2581) は，付与しないことにした．
[f] https://ainugo.nam.go.jp/siror/contents/Library1_3.html
[g] https://ainu.ninjal.ac.jp/folklore
[h] http://www.town.biratori.hokkaido.jp/biratori/nibutani/culture/language/story
[i] https://ainugo.aa-ken.jp
[j] RoBERTa トークナイザは SentencePiece との相性が悪い．
[k] 「2020パ」は，挨拶文の読み上げでは「ホッネ パ イカシマトゥアシクネワンホッネ パ」だったが，「2020パ」のままにした．

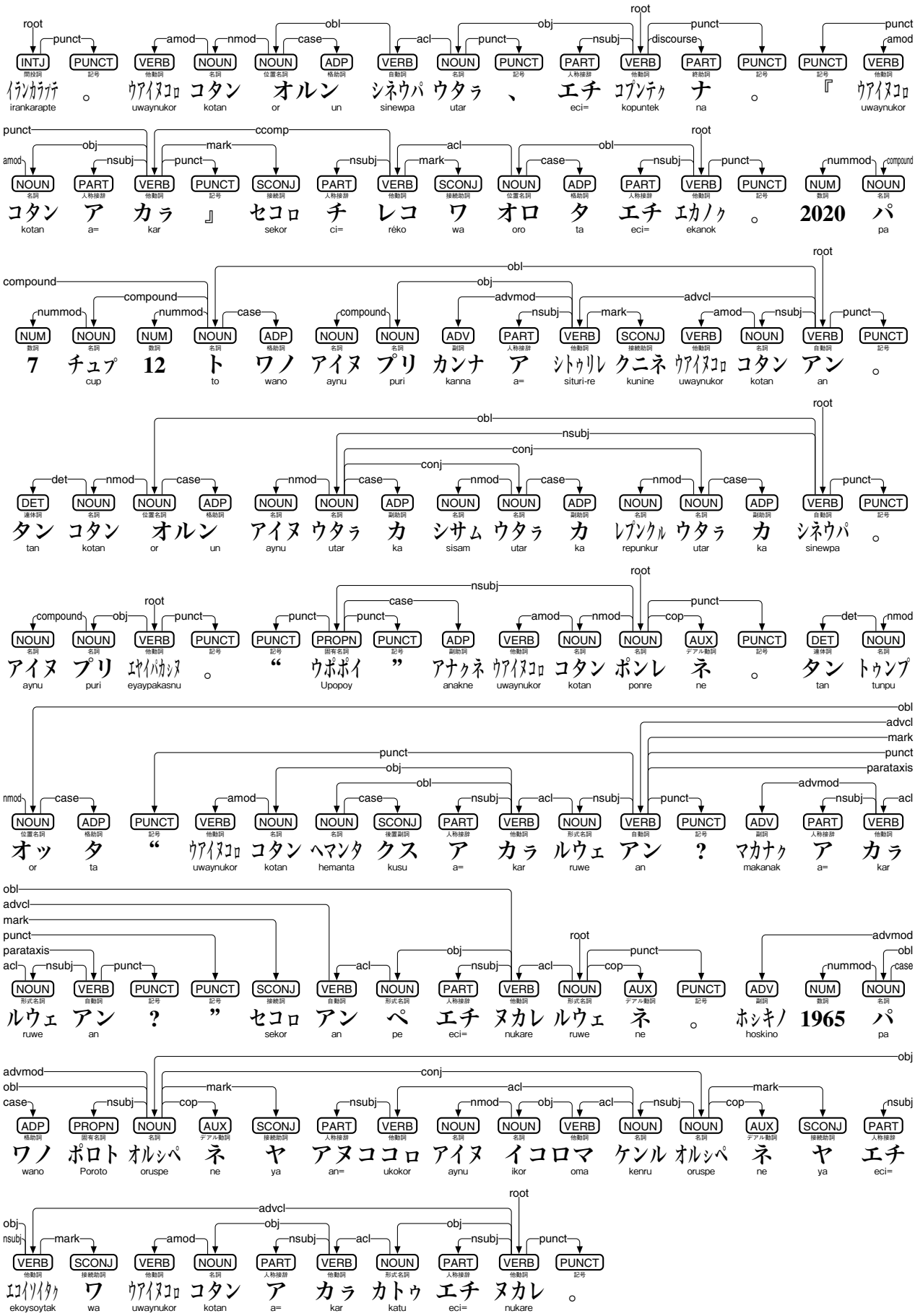

図 3: 『ウアイヌコロ コタン アカラ』挨拶文 (前半抜粋) CoNLL-U

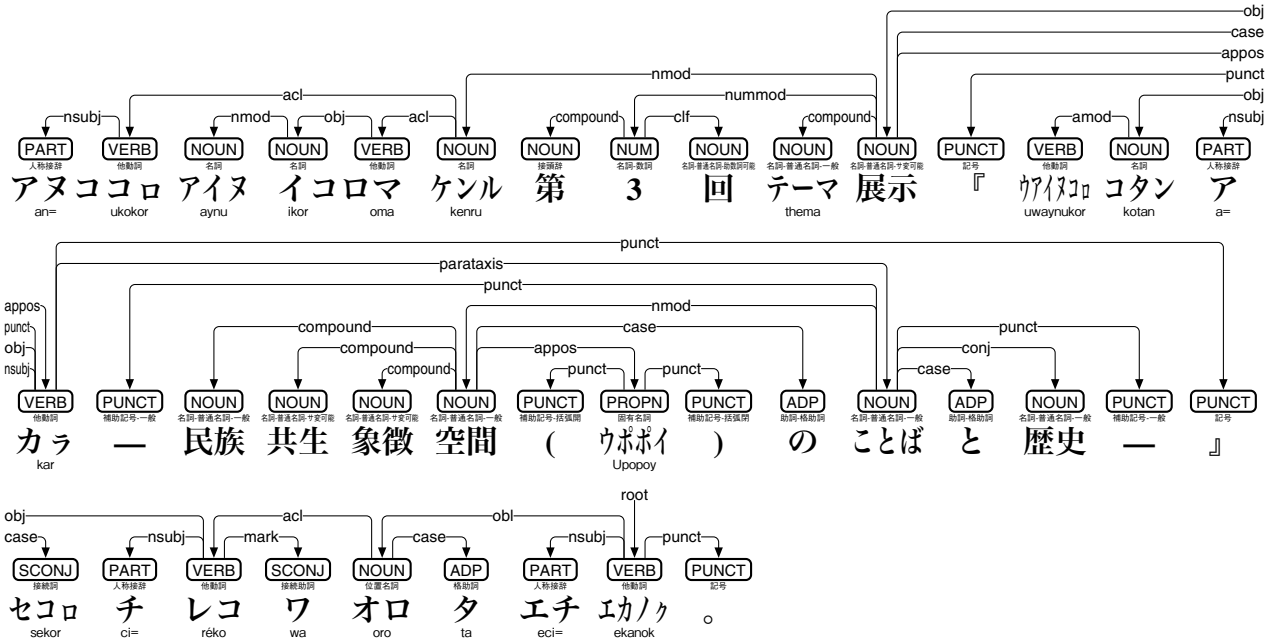

図 4: 『ウアイヌコロ コタン アカラ』挨拶文 3〜5 行目 CoNLL-U

らば，単純にアイヌ語のカタカナ表記を処理するだけでなく，日本語との混在を視野に入れていく必要がある．アイヌ語と日本語では文法が全く異なっており，頑張って UD で混在させても，それを自動で解析するのは至難の業である．まだまだ道程は長い．今後のわれわれの研究の進展に期待されたい．

参考文献

[1] Hajime Senuma, Akiko Aizawa: Toward Universal Dependencies for Ainu, NoDaLiDa 2017 Workshop on Universal Dependencies (May 2017), pp.133-139.

[2] Hajime Senuma, Akiko Aizawa: Universal Dependencies for Ainu, LREC 2018: Eleventh International Conference on Language Resources and Evaluation (May 2018), pp.2354-2358.

[3] 安岡孝一: アイヌ語 Universal Dependencies 再考, 東洋学へのコンピュータ利用, 第 34 回研究セミナー (2021 年 7 月 30 日), pp.25-53.

[4] 安岡孝一: Universal Dependencies によるアイヌ語テキストコーパス, 情報処理学会研究報告, Vol.2021-CH-127『人文科学とコンピュータ』, No.5 (2021 年 8 月 28 日), pp.1-8.

[5] 知里幸恵: アイヌ神謡集, 東京: 郷土研究社 (1923 年 8 月).

[6] 神保小虎, 金澤庄三郎: アイヌ會話字典, 東京: 金港堂書籍 (1898 年 4 月).

[7] 国立アイヌ民族博物館ガイドブック, 白老: 国立アイヌ民族博物館 (2020 年 3 月).

[8] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov: RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv:1907.11692 (July 2019).

[9] Pencheng He, Xiaodong Liu, Jianfeng Gao, Weizhu Chen: DeBERTa: Decoding-enhanced Bert With Disentangled Attention, 9th International Conference on Learning Representations (May 2021), Poster 03-2562.

[10] Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, Daniel Zeman: Universal Dependencies, Computational Linguistics, Vol.47, No.2 (June 2021), pp.255-308.

[11] Lucien Tesnière: Éléments de Syntaxe Structurale, Paris: C. Klincksieck (1959).

[12] Igor A. Mel'čuk: Dependency Syntax: Theory and Practice, New York: State University of New York Press (1988).

[13] Joakim Nivre: Towards a Universal Grammar for Natural Language Processing, CICLing 2015: 16th International Conference on Intelligent Text

Processing and Computational Linguistics (April 2015), pp.3-16.

[14] 田村すず子: アイヌ語沙流方言辞典, 東京: 草風館 (1996 年 9 月).

[15] 安岡孝一: Universal Dependencies にもとづく多言語係り受け可視化ツール deplacy, 人文科学とコンピュータシンポジウム「じんもんこん 2020」論文集 (2020 年 12 月), pp.95-100.

[16] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, Alexander M. Rush: Transformers: State-of-the-Art Natural Language Processing, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (October 2020): System Demonstrations, pp.38-45.

[17] Taku Kudo, John Richardson: SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (November 2018): System Demonstrations, pp.66-71.

[18] ローマ字のつづり方, 昭和 29 年内閣告示第 1 号, 官報, 第 8382 号 (1954 年 12 月 9 日), p.189.

[19] ISO 9:1995 Information and documentation — Transliteration of Cyrillic characters into Latin characters — Slavic and non-Slavic languages, 2nd Edition, Geneva: ISO (February 15, 1995).

[20] 安岡孝一: Transformers と国語研長単位による日本語係り受け解析モデルの製作, 情報処理学会研究報告, Vol.2022-CH-128『人文科学とコンピュータ』, No.7 (2022 年 2 月 19 日), pp.1-8.

[21] Timothy Dozat, Christopher D. Manning: Deep Biaffine Attention for Neural Dependency Parsing, 5th International Conference on Learning Representations (April 2017), C25.

[22] 安岡孝一, 安岡素子: 古典中国語の形態素解析と係り受け解析, 근역한문학회 2022년 추계 기획학술대회: 디지털과 한문 고전 연구 (2022 年 11 月 26 日), pp.148-160.

[23] Chu Yoeng-jin (朱永津) and Liu Tseng-hong (刘振宏): On the Shortest Arborescence of a Directed Graph, Scientia Sinica, Vol.XIV, No.10 (October 1965), pp.1396-1400.

[24] Jack Edmonds: Optimum Branchings, Journal of Research of the National Bureau of Standards—B. Mathematics and Mathematical Physics, Vol.71B, No.4 (October-December 1967), pp.233-240.

[25] Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov: CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Proceedings of the CoNLL 2018 Shared Task (October 2018), pp.1-21.