

ウポポイ園内マップによる 12 種の Universal Dependencies

安岡孝一 (京都大学)・安岡素子 (京都外国語大学)

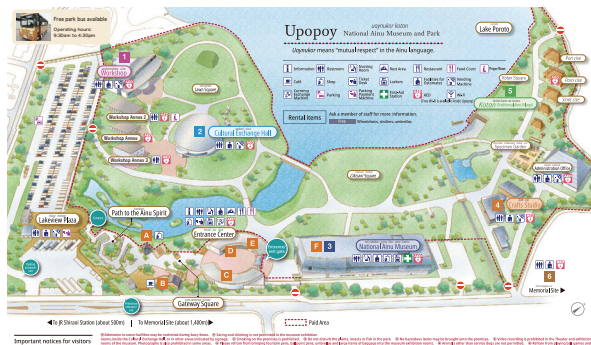
はじめに

ウポポイ園内マップは、日本語・英語・繁体中文・簡体中文・韓国語・タイ語・ロシア語の 7 種類が配布されている。これら 7 種類のマップにはアイヌ語が付記されており、それぞれカタカナ・ローマ字・ローマ字・ローマ字・ハングル・タイ文字・キリル文字で書かれている (図 1)。文字種から見ると、ひらがなが日本語に、カタカナが日本語とアイヌ語に、漢字が日本語と繁体中文と簡体中文に、ローマ字が英語とアイヌ語に、ハングルが韓国語とアイヌ語に、タ

<https://ainu-upopoy.jp/wp-content/uploads/2023/03/957bd95b932d869814b026c1dfbd7769.pdf>



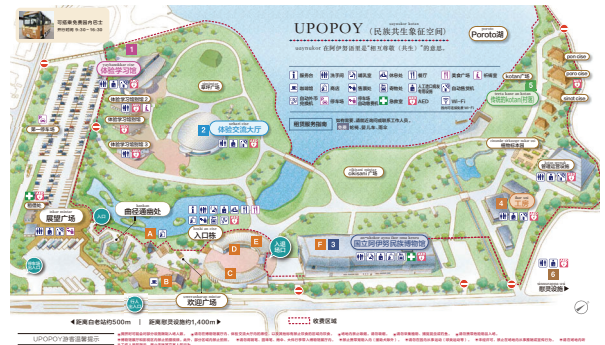
<https://ainu-upopoy.jp/wp-content/uploads/sites/2/2022/03/4475832502cf0e6c3928ee5fa826c04.pdf>



<https://ainu-upopoy.jp/wp-content/uploads/sites/3/2022/03/3c8b15c1d7b1fe38534bf0a2643d5423.pdf>



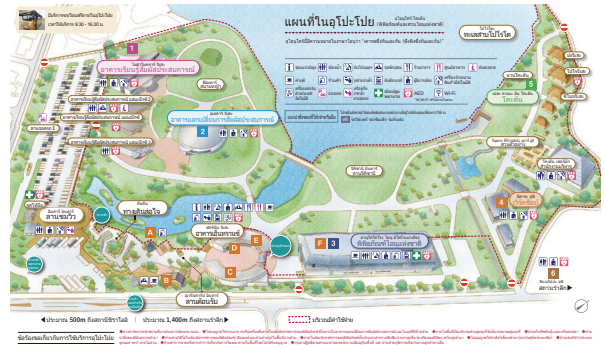
<https://ainu-upopoy.jp/wp-content/uploads/sites/4/2020/08/73da1712e654238dc7e5b2c650b6e03-1.pdf>



<https://ainu-upopoy.jp/wp-content/uploads/sites/5/2020/08/259c91220319e56c5e782c3ce4699f2f.pdf>



<https://ainu-upopoy.jp/wp-content/uploads/sites/6/2020/08/26ea70c0ca695ed072236b61de80.pdf>



<https://ainu-upopoy.jp/wp-content/uploads/sites/7/2022/03/5c94388414668be87a6a0beefb59ed37.pdf>

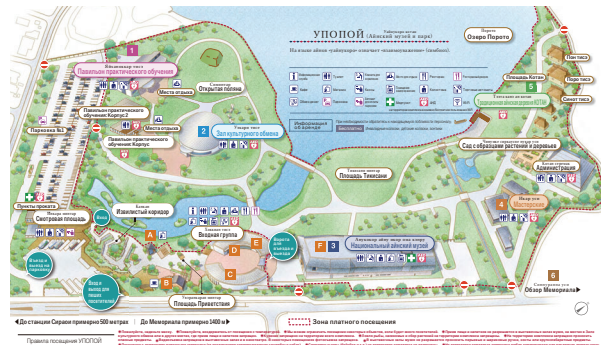


図 1: ウポポイ園内マップ

表 1: esupar の多言語サポート (一部)

日本語	https://huggingface.co/KoichiYasuoka/bert-base-japanese-upos
英語	https://huggingface.co/KoichiYasuoka/roberta-base-english-upos
中国語	https://huggingface.co/KoichiYasuoka/chinese-bert-wwm-ext-upos
韓国語	https://huggingface.co/KoichiYasuoka/roberta-base-korean-upos
タイ語	https://huggingface.co/KoichiYasuoka/roberta-base-thai-spm-upos
ロシア語	https://huggingface.co/KoichiYasuoka/bert-base-russian-upos
アイヌ語	https://huggingface.co/KoichiYasuoka/deberta-base-ainu-upos

イ文字がタイ語とアイヌ語に、キリル文字がロシア語とアイヌ語に、それぞれ使われている。すなわち、ウポボイ園内マップには、12種類の書写言語が並行して記述されているとみなしてよい。

安岡孝一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・Christian Wittern・池田巧・守岡知彦・鈴木慎吾・李媛・劉冠偉)では、多言語係り受け解析エンジン esupar^{a)}を開発中である[1]。表1に示した各言語モデルを用いることで、日本語[2]・英語[3]・中国語(繁体中文・簡体中文)[4]・韓国語[5]・タイ語[6]・ロシア語[7]・アイヌ語(カタカナ・ローマ字・キリル文字)[8]については、Universal Dependencies (UD)[9]にもとづく品詞付与・係り受け解析が可能となっている。しかしながら、ハングルやタイ文字で書かれたアイヌ語の解析は、われわれには経験がない。

では、ウポボイ園内マップにおける12種類の書写言語を、パラレルUDコーパスとして記述するには、どのような方策が必要なのか。本稿では、これについて述べる。

Universal Dependencies の概要

UD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[10]に源を発し、Мельчукの有向グラフ記述[11]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が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的品詞タグ^{b)}。
5. XPOS: 言語固有の品詞タグ。
6. FEATS: UDで規定された言語普遍的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。係り受け元が無い場合は0とする。
8. DEPREL: UDで規定された言語普遍的係り受けタグ(表2)。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・

^{b)}ADJ・ADP・ADV・AUX・CCONJ・DET・INTJ・NOUN・NUM・PART・PRON・PROPN・PUNCT・SCONJ・SYM・VERB・Xの17種類。

^{a)}<https://pypi.org/project/esupar>

表 2: 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 繫辞 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

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 [12] が、Мельчук とは異なっている。ちなみに、コンピュータ文においては、補語をリンク元として、主語へとリンクする。

12 種 UD コーパスの製作

図 1 のウポポイ園内マップ中、以下に示す 16 地点の名称に対して、12 種の書写言語による UD コーパスを製作した。

- 1 体験学習館 (yayhanokkar cise)
- 2 体験交流ホール (uekari cise)
- 3 国立アイヌ民族博物館 (an=ukokor aynu ikor oma kenru)

- 4 工房 (ikar usi)
- 5 伝統的コタン (teeta kane an kotan)
- 6 慰霊施設 (sinnurappa usi)
 - 芝生広場 (simimtar)
 - 展望広場 (inkar mintar)
 - いざないの回廊 (kankan)
 - 歓迎の広場 (uwerankarap mintar)
 - エントランス棟 (hoski an cise)
 - チキサニ広場 (cikisani mintar)
 - 管理運営施設 (kotan sermak)
 - 草木の見本園 (cinumke sirkaspe nukar usi)
 - ポロト湖 (poroto)
 - 民族共生象徴空間 (uaynukor kotan)

日本語・英語・中国語 (繁体中文・簡体中文)・韓国語・タイ語・ロシア語・アイヌ語 (ローマ字) に対しては、表 1 の言語モデルを用いて、Google Colaboratory 上

```
!pip install esupar
import esupar,deplacy
mdl="KoichiYasuoka/deberta-base-ainu-upos"
nlp=esupar.load(mdl)
doc=""
for t in ["yayhanokkar cise","uekari cise",
"an=ukokor aynu ikor oma kenru",
"ikar usi","teeta kane an kotan",
"sinnurappa usi","simimtar",
"inkar mintar","kankan",
"uwerankarap mintar","hoski an cise",
"cikisani mintar","kotan sermak",
"cinumke sirkaspe nukar usi","poroto",
"uaynukor kotan"]:
doc+=f"# text = {t}\n{str(nlp(t))}\n"
deplacy.serve(doc,port=None)
```

図 2: アイヌ語 (ローマ字) UD 作成プログラム

text = 国立アイヌ民族博物館

1	国立	-	NOUN	-	-	5	compound	-	SpaceAfter=No
2	アイヌ	-	PROPN	-	-	5	compound	-	SpaceAfter=No
3	民族	-	NOUN	-	-	5	compound	-	SpaceAfter=No
4	博物	-	NOUN	-	-	5	compound	-	SpaceAfter=No
5	館	-	NOUN	-	-	0	root	-	SpaceAfter=No

text = National Ainu Museum

1	National	-	ADJ	-	-	3	amod	-	-
2	Ainu	-	PROPN	-	-	3	compound	-	-
3	Museum	-	NOUN	-	-	0	root	-	SpaceAfter=No

text = 國立愛努民族博物館

1	國	國	NOUN	-	-	2	nsubj	-	SpaceAfter=No
2	立	立	VERB	-	-	6	acl	-	SpaceAfter=No
3	愛努	愛努	PROPN	-	-	6	nmod	-	SpaceAfter=No
4	民族	民族	NOUN	-	-	6	nmod	-	SpaceAfter=No
5	博物	博物	NOUN	-	-	6	compound	-	SpaceAfter=No
6	館	館	PART	-	-	0	root	-	SpaceAfter=No

text = 国立阿伊努民族博物館

1	国	国	NOUN	-	-	2	nsubj	-	SpaceAfter=No
2	立	立	VERB	-	-	6	acl	-	SpaceAfter=No
3	阿伊努	阿伊努	PROPN	-	-	6	nmod	-	SpaceAfter=No
4	民族	民族	NOUN	-	-	6	nmod	-	SpaceAfter=No
5	博物	博物	NOUN	-	-	6	compound	-	SpaceAfter=No
6	館	館	PART	-	-	0	root	-	SpaceAfter=No

text = 국립아이누민족박물관

1	국립	-	NOUN	-	-	4	compound	-	SpaceAfter=No
2	아이누	-	PROPN	-	-	4	compound	-	SpaceAfter=No
3	민족	-	NOUN	-	-	4	compound	-	SpaceAfter=No
4	박물관	-	NOUN	-	-	0	root	-	SpaceAfter=No

text = พิพิธภัณฑ์ไอนุแห่งชาติ

1	พิพิธภัณฑ์	พิพิธภัณฑ์	NOUN	-	-	0	root	-	SpaceAfter=No
2	ไอนุ	ไอนุ	PROPN	-	-	1	compound	-	SpaceAfter=No
3	แห่ง	แห่ง	NOUN	-	-	1	nmod	-	SpaceAfter=No
4	ชาติ	ชาติ	NOUN	-	-	3	compound	-	SpaceAfter=No

text = Национальный айнский музей

1	Национальный	-	ADJ	-	-	3	amod	-	-
2	айнский	-	ADJ	-	-	3	amod	-	-
3	музей	-	NOUN	-	-	0	root	-	SpaceAfter=No

text = an=ukokor aynu ikor oma kenru

1	an=	an=	PART	人称接辞	-	2	nsubj	-	SpaceAfter=No
2	ukokor	ukokor	VERB	他動詞	-	6	acl	-	-
3	aynu	aynu	NOUN	名詞	-	4	nmod	-	-
4	ikor	ikor	NOUN	名詞	-	5	nsubj	-	-
5	oma	oma	VERB	他動詞	-	6	acl	-	-
6	kenru	kenru	NOUN	名詞	-	0	root	-	SpaceAfter=No

text = アヌココロ アイヌ イコロマケンル

1-2	アヌココロ	-	-	-	-	-	-	-	-
1	アン	an=	PART	人称接辞	-	2	nsubj	-	-
2	ウココロ	ukokor	VERB	他動詞	-	6	acl	-	-
3	アイヌ	aynu	NOUN	名詞	-	4	nmod	-	-
4-5	イコロマ	-	-	-	-	-	-	-	SpaceAfter=No
4	イコロ	ikor	NOUN	名詞	-	5	nsubj	-	-
5	オマ	oma	VERB	他動詞	-	6	acl	-	-
6	ケンル	kenru	NOUN	名詞	-	0	root	-	SpaceAfter=No

text = 아누코콜 아이누 이콜 오마 켄루

1-2	아누코콜	-	-	-	-	-	-	-	-
1	안	an=	PART	人称接辞	-	2	nsubj	-	-
2	우코콜	ukokor	VERB	他動詞	-	6	acl	-	-
3	아이누	aynu	NOUN	名詞	-	4	nmod	-	-
4	이콜	ikor	NOUN	名詞	-	5	nsubj	-	-
5	오마	oma	VERB	他動詞	-	6	acl	-	-
6	켄루	kenru	NOUN	名詞	-	0	root	-	SpaceAfter=No

text = อานูโคโคโร ไอนุ อีโครโอมาคังรุ

1-2	อานูโคโคโร	-	-	-	-	-	-	-
1	อัน	an=	PART	人称接辞	-	2	nsubj	-
2	อุโคคอร์	ukokor	VERB	他動詞	-	6	acl	-
3	ไอนุ	aynu	NOUN	名詞	-	4	nmod	-
4-5	อีโครโอม	-	-	-	-	-	-	SpaceAfter=No
4	อีโคร	ikor	NOUN	名詞	-	5	nsubj	-
5	โอม	oma	VERB	他動詞	-	6	acl	-
6	คังรุ	kenru	NOUN	名詞	-	0	root	SpaceAfter=No

text = Анукокор айну икор ома кэнру

1	Ан	an=	PART	人称接辞	-	2	nsubj	SpaceAfter=No
2	укокор	ukokor	VERB	他動詞	-	6	acl	-
3	айну	aynu	NOUN	名詞	-	4	nmod	-
4	икор	ikor	NOUN	名詞	-	5	nsubj	-
5	ома	oma	VERB	他動詞	-	6	acl	-
6	кэнру	kenru	NOUN	名詞	-	0	root	SpaceAfter=No

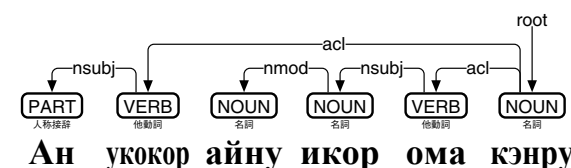
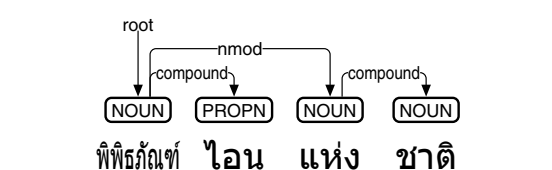
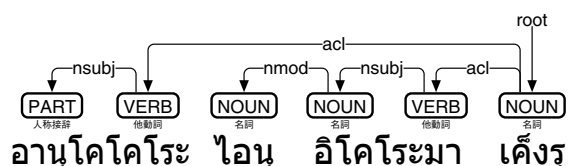
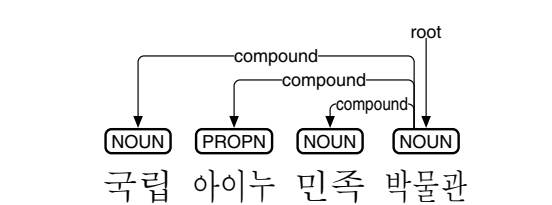
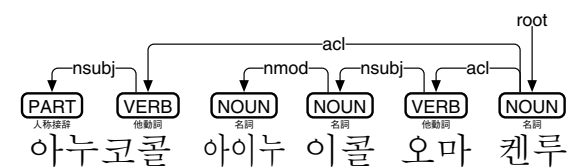
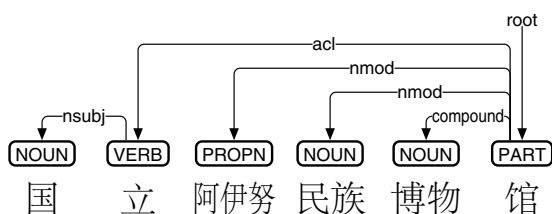
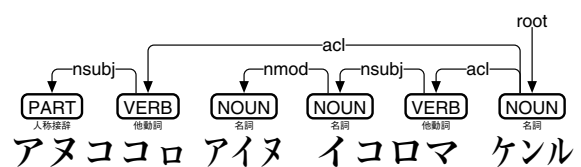
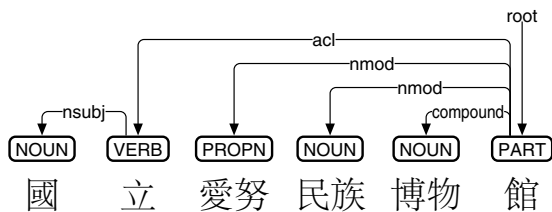
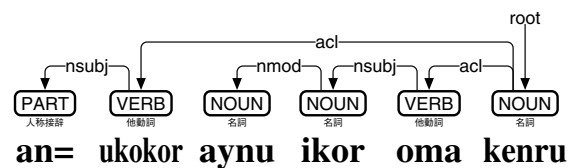
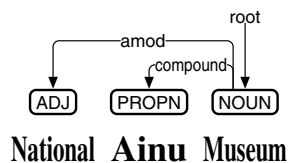
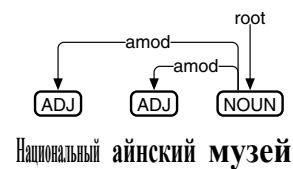
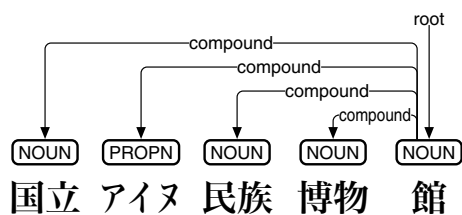


図 3: 「国立アイヌ民族博物館」の 12 種 UD コーパス

の esupar で品詞付与・係り受け解析をおこない、さらに deplacy [13] で可視化・編集をおこなった。Google Colaboratory 上のアイヌ語 (ローマ字) UD 作成プログラムを、図 2 に示す。ただし、アイヌ語 UD の編集に関しては、アイヌ語 UD エディター [8]⁹⁾ も適宜併用した。

アイヌ語 (カタカナ・ハングル・タイ文字・キリル文字) に対しては、アイヌ語 (ローマ字) の CoNLL-U データを、アイヌ語 UD エディターで編集する方法を採用した。ただし、以下の 5 つではアンシェヌマン (末子音と母音の融合) が起こっているため、これらは縮約語とみなし、内部的には 2 語に分割して処理することにした。

- アヌココロ → アン ウココロ
- イコロマ → イコロ オマ
- 아누코콜 → 안 우코콜
- อานุโคโคโร → อัน อุโคโคโร
- อีโคโรมา → อีโคโร โอมมา

なお「이칼 민탈」の이칼は、인칼の誤植だと考えられるが、あえて訂正しないこととした。「ເອຕະ ດານະ ອັນ ໂຄະດັນ」のເອຕະは、ເທຕະの誤植だと考えられるが、あえて訂正しないこととした。綴りのゆれも散見される (たとえばອຸຸທິとອຸສິ) が、特に統一していない。

12 種 UD コーパスの例として、「国立アイヌ民族博物館」を図 3 に示す。タイ語 UD が、いわゆる後置修飾であり、他の 11 種とは異なる依存構造になっているのが、特徴的である。

おわりに

ウポボイ園内マップ中 16 地点の名称に対し、12 種の書写言語すなわち日本語・英語・繁体中文・簡体中文・韓国語・タイ語・ロシア語・アイヌ語 (ローマ字)・アイヌ語 (カタカナ)・アイヌ語 (ハングル)・アイヌ語 (タイ文字)・アイヌ語 (キリル文字) による UD コーパスを製作した。付録に全データを可視化するとともに、WWW で公開¹⁰⁾している。ただし、全てが体言であるため、係り受けコーパスとしては不十分だと考えられる。

正直なところを言えば、タイ文字でアイヌ語を表記するのは、無理があるように思える。タイ語の末子音には r が立たないため、無理矢理 r で表記しようとしているのだが、他の末子音との間で表記が統一

されておらず、かなり読みにくい。一方、ハングルで書かれたアイヌ語は、韓国の外来語表記法 [14] を守ろうとすると、末子音の s を시で表記せざるを得ず、微妙に悲しい。タイ文字やハングルをアイヌ語表記に用いるより、むしろローマ字で表記した方が、中国語 (繁体中文・簡体中文) マップとも整合するのではないかと考える次第である。

参考文献

- [1] 安岡孝一: Universal Dependencies と BERT/RoBERTa/DeBERTa/GPT モデルによる言語情報処理 (2025 年 3 月版), 京都大学人文科学研究所・未踏科学研究ユニット・データサイエンスで切り拓く総合地域研究ユニット (2025 年 3 月).
- [2] 浅原正幸, 金山博, 宮尾祐介, 田中貴秋, 大村舞, 村脇有吾, 松本裕治: Universal Dependencies 日本語コーパス, 自然言語処理, Vol.26, No.1 (2019 年 3 月), pp.3-36.
- [3] Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel R. Bowman, Miriam Connor, John Bauer and Christopher D. Manning: A Gold Standard Dependency Corpus for English, LREC 2014: 9th International Conference on Language Resources and Evaluation (May 2014), pp.2897-2904.
- [4] Herman Leung, Rafaël Poirer, Tak-sum Wong, Xinying Chen, Kim Gerdes and John Lee: Developing Universal Dependencies for Mandarin Chinese, 12th Workshop on Asian Language Resources (December 2016), pp.20-29.
- [5] Jayeol Chun, Na-Rae Han, Jena D. Hwang, Jinho D. Choi: Building Universal Dependency Treebanks in Korean, LREC 2018: 11th International Conference on Language Resources and Evaluation (May 2018), pp.2194-2202.
- [6] Koichi Yasuoka: Sequence-Labeling RoBERTa Model for Dependency-Parsing in Classical Chinese and Its Application to Vietnamese and Thai, ICBIR 2023: 8th International Conference on Business and Industrial Research (May 2023), pp.169-173.
- [7] Kira Drogonova, Olga Lyashevskaya, Daniel Zeman: Data Conversion and Consistency of Monolingual Corpora: Russian UD Treebanks, TLT 2018: 17th International Workshop on Treebanks and Linguistic Theories (December 2018), pp.52-65.
- [8] 安岡孝一, 安岡素子: ローマ字・カタカナ・キリル文字によるアイヌ語 Universal Dependencies の可能性, Evidence-based Linguistics Workshop 2023 発表論文集 (2023 年 9 月), pp.47-60.
- [9] Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, Daniel Zeman: Universal Dependencies, Computational Linguistics, Vol.47, No.2 (June 2021), pp.255-308.
- [10] Lucien Tesnière: Éléments de Syntaxe Structurale, Paris: C. Klincksieck (1959).
- [11] Igor A. Mel'čuk: Dependency Syntax: Theory and Practice, New York: State University of New York Press (1988).
- [12] 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.
- [13] 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.
- [14] 외래어 표기법, 문화체육관광부 고시, 제2017-14호 (2017 年 3 月 28 日).

⁹⁾<https://koichiyasuoka.github.io/UD-Ainu/editor/editor-ain.html>

¹⁰⁾<https://koichiyasuoka.github.io/deplacy/demo/2025-05-17>

 体験	 学習	 館	 体験	 交流	 ホール	 国立	 アイヌ	 民族	 博物館	 工房	 伝統	 的	 コットン	 慰霊	 施設	 芝生	 広場	 展望		
 広場	 いざない	 の	 回廊	 歓迎	 の	 広場	 エントランス	 棟	 チカサニ	 広場	 管理	 運営	 施設	 草木	 の	 見本	 園	 ポルト	 湖	
 民族	 共生	 象徴	 空間																	
 Workshop	 Cultural Exchange	 Hall	 National	 Ainu	 Museum	 Crafts Studio	 Kotan	 (	 Traditional	 Ainu	 Village	)	 Memorial	 Site	 Lawn	 Square	 Lakeview			
 Plaza	 Path	 to	 Ainu	 Spirit	 Gateway	 Square	 Entrance	 Center	 Cikisani	 Square	 Administration	 Office	 Specimen	 Garden	 Lake	 Poroto	 National	 Ainu	 Museum	
 and	 Park																			
 体験	 学習	 館	 体験	 交流	 大廳	 國	 立	 愛努	 民族	 博物館	 工房	 傳統	 的	 kotan	 (	 村落	)	 慰霊		
 設施	 草皮	 廣場	 展望	 廣場	 曲徑	 通幽	 處													

