アイヌ語Universal Dependencies再考

安岡孝一*

1　はじめに

アイヌ語 Universal Dependencies は、東京大学の瀬沼甫が提案[1]・開発[2]し、2017 年 3 月 31 日に公開[3]したもの (以下「瀬沼版」と呼ぶ) である。最初の公開は「ホテナオ」のみ (図 1) であり、その後にデータが追加されることもなく、4 年後にはリポジトリが閉鎖されてしまった。抱合語への Universal Dependencies の適用としては、ほぼ最初の試みだったので、筆者としては非常に残念だ。

とは言うものの、筆者の手元には瀬沼版のクローンが残されており、また、GitHub 上にもフォークがいくつか残っている。これらをもとに再構築を試みることも可能だろう。ただ、クローンを精査した限りでは、瀬沼版は「他者との共同作業を拒んでいる」ように、筆者には見える。アノテーションの基準が不明瞭な (他者と共有されていない) 上に、全体として「凝りすぎ」なのである。このままでは、「ホテナオ」以外の神謡[4]残り 12 編を、追加することすら侭ならない。

本稿では「共同作業」という視点から、瀬沼版アイヌ語 Universal Dependencies が抱えていた問題点を洗い出し、今後のアイヌ語コーパス設計への礎としたい。

2　Universal Dependencies と CoNLL-U の概要

Universal Dependencies[5](以下「UD」と呼ぶ) 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[6]に源を発し、Мельчук の有向グラフ記述[7]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

*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

[1]Hajime Senuma, Akiko Aizawa: Toward Universal Dependencies for Ainu, Proceedings of the 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]`https://github.com/hajimes/ud-ainu`

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

[5]Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, Daniel Zeman: Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection, Proceedings of the 12th Language Resources and Evaluation Conference (May 2020), pp.4034-4043.

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

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

- "sine-an-to-ta nismu=as kusu, pis ta sap=as.
- sinot=as kor okay=as awa, sine pon rupne aynu ek kor an.
- wakusu, hepasi san ko, hepasi ci=etusmak.
- heperay ek ko, heperay ci=etusmak.
- ikici=as awa, hepasi iwan suy heperay iwan suy ne i ta pon rupne aynu kor wenpuri enan tuyka eparsere.
- ene itak i,
- ’piy tun tun, piy tun tun!
- tan hekaci, wen hekaci, e=iki ciki, tan esannot teeta réhe tane réhe ukaepita e=ki kus ne na!’
- háwas ciki, ci=emina kor, itak=as hawe ene okay,
- ’nennamora tan esannot teeta réhe tane réhe erampewtek a!
- teeta anak sinnupur kusu, tapan esannot Kamuyesannot ari a=ye a korka, tane anakne sirpan kusu, Inawesannot ari a=ye ruwe tasi an ne!’
- itak=as awa, pon rupne aynu ene itak i,
- ’piy tun tun, piy tun tun!
- tan hekaci, sonno he tap e=hawan ciki, tapan petpo teeta réhe tane réhe ukaepita e=ki kus ne na.’
- háwas ciki, itak=as hawe ene okay,
- ’nennamora tapan petpo teeta réhe tane réhe erampewtek a!
- teeta kane sinnupur i ta tapan petpo Kanciwetunas ari a=ye a korka, tane sirpan kus, Kanciwemoyre ari a=ye ruwe tasi an ne!’
- itak=as awa, pon rupne aynu ene itak i,
- ’piy tun tun, piy tun tun!
- sonno he tap ne e=hawan ciki, usinritpita a=ki kus ne na!’
- háwas ciki, itak=as hawe ene okay,
- ’nennamora e=sinricihi erampewtek a!
- otteeta Okikirmuy kim ta oman wa kuca kar i ta kene inunpe kar a ike, ne inunpe apekar wa sat tek okere.
- Okikirmuy oar arkehe oterke ko, oar arkehe hotari.
- ne wa an pe Okikirmuy ruska kusu, ne inunpe pet or ta kor wa san wa osura wa, isam ruwe ne.
- oro wano ne inunpe petesoro mom ayne no, atuy oro osma.
- tu atuy penrur re atuy penrur ciesirkik siri kamuy utar nukar wa, a=eoripak Okikirmuy tekekar pe néno yayerampewtek wa mom ayne no, atuykomunin a=enunuke kusu, kamuy utar oro wa ne inunpe ceppo ne a=kar wa inunpepeceppo ari a=rékore ruwe ne.
- awa ne inunpepeceppo yaysinrit’erampewtek wa aynu ne yaykar wa iki kor an.
- ne inunpepeceppo e=ne ruwe tasi an ne.’
- itak=as awa, pon rupne aynu iporoho ka wen a wen a.
- ikokanu wa an ayne,
- ’piy tun tun, piy tun tun!
- eani anak pon horkew sani e=ne ruwe tasi an ne.’
- itak kese ta atuy or un terke humi copkosanu.
- osi inkar=as awa, sine húre ceppo honoyanoya wa too herepasi oman wa isam,"
- ari pon horkew kamuy isoytak.

図 1: 瀬沼版「ホテナオ」本文

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

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[8]。
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。

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

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[9]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。

本稿執筆時点で最新の UD は、2021 年 5 月 16 日発表の UD 2.8.1[10]である。UD 2.8.1 は、114 の言語にまたがるツリーバンクだが、筆者の見る限り、抱合語はチュクチ語 UD[11]とユピック語 UD (セントローレンス島)[12]の 2 つだけのようである。

3　瀬沼版アイヌ語 Universal Dependencies

筆者の手元に残された瀬沼版 CoNLL-U のクローンは、GitHub 上に残っている Francis Tyers のフォーク[13]と同様、36 文 516 語から成る「ホテナオ」である。最初のコミットは 2017 年 3 月 31 日で、それ以後のコミットは見当たらない。36 文には、それぞれ `ky-6-1` から `ky-6-36` の `sent_id`(文の ID) が振られている。XPOS・DEPS を除く 8 フィールドが使用されており、516 語を UPOS で分類すると、VERB が 107 語 (異なり FORM 数 54・異なり LEMMA 数 52)、PUNCT が 89 語 (異なり数 5)、NOUN が 86 語 (異なり FORM 数 37・異なり LEMMA 数 36)、PART が 47 語 (異なり数 9)、ADV が 43 語 (異なり数 19)、SCONJ が 29 語 (異なり FORM 数 8・異なり LEMMA 数 6)、CCONJ が 24 語 (異なり数 4)、INTJ が 24 語 (異なり数 2)、ADP が 17 語 (異なり数 6)、DET が 17 語 (異なり数 4)、AUX が 15 語 (異なり数 4)、PROPN が 8 語 (異なり数 5)、NUM が 6 語 (異なり数 4)、PRON が 4 語 (異なり数 2) である。なお、ライセンスは CC-BY-4.0 である。

3.1　底本の問題

瀬沼版「ホテナオ」本文 (図 1) は、知里版[4]とも切替版[14]とも片山版[15]とも異なっている。瀬沼版は 36 文だが、知里版・切替版・片山版は 66 行である。瀬沼版の表記は、どちらかと言えば片山版に近いが、「a ike」を分離している点は切替版を思わせる。一方で、瀬沼版のレポジトリには「Currently the text is based on “Ainu Shin’yōshū” by Yukie Chiri,

[9]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.
[10]`http://hdl.handle.net/11234/1-3687`
[11]`https://github.com/UniversalDependencies/UD_Chukchi-HSE`
[12]`https://github.com/UniversalDependencies/UD_Yupik-SLI`
[13]`https://github.com/ftyers/ud-ainu/blob/master/ain-ud.conllu`
[14]切替英雄:『アイヌ神謡集』辞典, 北大言語学研究報告, 第 2 号 (1989 年 6 月).
[15]片山龍峯, Julie Kaizawa:「アイヌ神謡集」を読みとく, 武蔵野: 片山言語文化研究所 (2003 年 5 月).

which is under public domain.」との記述があった[16]ことから、著作権切れの知里版を底本にしているはずなのである。どうもよくわからない。

瀬沼版には、たとえば「a=rékore」が1ヶ所あらわれるが、この部分に対応する知里版は「arekore」、切替版は「a rekore」、片山版は「a=rekore」であり、いずれも瀬沼版とは違ってアクセント記号が無い。瀬沼版 CoNLL-U の当該部分 (図 2) を見てみよう。

```
44  a=  a=  PART    _   Number=Plur|Person=4|PronType=Prs|Valency=Poly  45  aux _   SpaceAfter=No
45  rékore  rékore  VERB    _   Valency=2   40  conj    _   Tamura1996=571|Gloss=name
```

図 2: 瀬沼版 CoNLL-U における「a=rékore」

「rékore」の MISC フィールドに「Tamura1996=571」というアノテーションが見える。『アイヌ語沙流方言辞典』[17]の 571 ページには「reka」「reki(hi)」「rekkisar」「rekkisara(ha)」「rekkurpo」「réko」「rekomatu」「rékopa」「rékor」「rékore」が掲載されており、ここから「rékore」という表記を持ってきたと推測される。

瀬沼版 CoNLL-U の 427 語 (PUNCT を除く) のうち、MISC に「Tamura1996=数」を含むものは 322 語、「Tamura=502」を含むものは 2 語あり、これらの単語表記は『アイヌ語沙流方言辞典』に依拠している可能性が高い。ただ、辞書を引いて得た表記ならば、見出し語形を示す LEMMA にだけ入れておくのが妥当であり、表層形を示す FORM にまで及ぼすのは、筆者には全く理解できない。原テキストとの連絡性を失う、という点で、非常に危うい設計思想だと感じる。

3.2 グロスの問題

瀬沼版 CoNLL-U「rékore」(図 2) の MISC フィールドには、「Gloss=name」というアノテーションも見える。グロス (近似的な逐語英訳) だと思われるが、これは、何を典拠としているのだろう。『アイヌ語沙流方言辞典』には各単語の「英訳」が示されているが、「rékore」に対しては「英訳」が無い。『An Ainu-English-Japanese Dictionary』第 4 版[18]の「Rei-kore」には「名ヲ附ケル. *v.i.* To name.」と示されている。Peterson 訳[19]は、この部分を「And it was called *Inunpepecheppo*, the fire-fish.」としている。Kaizawa 訳[15]は、この部分を「and was named a 'hearth-frame fish.'」としている。どれも「Gloss=name」とは、微妙に違う。

英語の「name」には名詞と動詞があることから、グロスに「name」を用いる際は、それぞれ「a name」「to name」と示す場合が多い。もちろん、図 2 の「rékore」においては、UPOS に VERB が記されているので疑義は無いのだが、そうすると、ますます、この「Gloss=name」が何を典拠としているのか不明である。

瀬沼版 CoNLL-U では、VERB・NOUN・ADV・SCONJ・CCONJ・ADP・DET・AUX・NUM の全てを含む 363 語[20]にグロスが付与されている。しかし、これらのグロスの典拠

[16] http://web.archive.org/web/20200905085035/github.com/hajimes/ud-ainu
[17] 田村すず子: アイヌ語沙流方言辞典, 東京: 草風館 (1996 年 9 月).
[18] John Batchelor: An Ainu-English-Japanese Dictionary, 4th Edition, 東京: 岩波書店 (1938 年 10 月).
[19] Benjamin Peterson: The Song of the Owl God Sang, BJS Books (2013).
[20] 「hepasi」「heperay」を含む。これらの語は、瀬沼版 CoNLL-U において MISC フィールドが壊れており、「Gloss=」なしにいきなりグロスが記されている。

は、筆者には突き止められなかった。典拠がわからないと、元々の作業者以外はデータをいじることが出来なくなり、メンテナンス性が極端に下がる。正直かなり残念である。

3.3 結合価の問題

瀬沼版 CoNLL-U「rékore」(図 2) の FEATS フィールドには、「Valency=2」という形態素属性が示されている。「Valency」(結合価) は、動詞に繋ぐことのできる項 (主語と目的語) の個数である。UD で規定された言語普遍な形態素属性には含まれていないものの、バンバラ語 UD[21]やモクシャ語 UD[22]などに、言語固有な形態素属性として導入されている。

瀬沼版は、全ての VERB に「Valency」を付与している[1]。アイヌ語の品詞分類[17]に当てはめると、「Valency=0」は完全動詞、「Valency=1」は自動詞、「Valency=2」は単他動詞、「Valency=3」は複他動詞にあたる。「Valency=4」の動詞も存在するらしい[23]が、瀬沼版には見当たらない。

動詞以外にも、人称接辞の「=as」に「Valency=1」を、「a=」と「ci=」に「Valency=Poly」を、瀬沼版は付与している。しかしながら、なぜ、この 3 つの人称接辞にだけ結合価が付与されていて、他の人称接辞には付与されていないのか、筆者には全く理解できない。また「Valency=Poly」の意味するところも、皆目わからない。他の言語 UD では、接辞に対する結合価など導入されていない。

一方、連他動詞 (連動詞) における結合価をどう扱うべきか、という点も問題となるはずだが、瀬沼版は何も示していない。「ホテナオ」に連他動詞が現れない、という事情もあるのだが、筆者としては困惑せざるを得ない。

3.4 形態素属性の問題

瀬沼版 CoNLL-U の FEATS フィールドには、以下に示す 17 種類の形態素属性が使用されている。

Degree=Dim				
NumType=Card				
Number=Sing	Number=Plur	Number=Pluract		
Person=1	Person=2		Person=4	
Place=Yes				
Possessed=Yes				
PronType=Int	PronType=Prs			
Valency=0	Valency=1	Valency=2	Valency=3	Valency=Poly

「Valency」(結合価) に関しては前節で述べたので、残る 12 種類の形態素属性のうち、言語固有な 4 種類を見ていこう。

[21] https://github.com/UniversalDependencies/UD_Bambara-CRB
[22] https://github.com/UniversalDependencies/UD_Moksha-JR
[23] Anna Bugaeva: Valency Classes in Ainu, Andrej Malchukov and Bernard Comrie (Eds.): Valency Classes in the World's Languages, Vol.1 (2015), pp.807-854.

「Degree=Dim」は指小辞 (diminutive) である。瀬沼版では「ceppo」「inunpepeceppo」「petpo」に付与されている。言語固有な形態素属性だが、アフリカーンス語 UD[24]においても、同様に指小辞として使用されている。

「Number=Pluract」は複数行為形 (pluractional)[1]である。瀬沼版は「okay」「sap」に「Number=Pluract」を付与している。その一方で「an」(okay の単数形) や「san」(sap の単数形) に対し、瀬沼版は「Number=Sing」を付与しておらず、基準が謎である。なお「Number=Pluract」は、他の言語 UD では全く使われていない。

「Place=Yes」は位置名詞である。瀬沼版では「arkehe」「ka」「kese」「kim」「or」「oro」「pis」に付与されている。UD には、言語普遍な形態素属性「Case=Loc」が準備されているのに、なぜ瀬沼版が言語固有の「Place=Yes」を用いているのか、筆者には理解できない。

「Possessed=Yes」は名詞の所属形である。「arkehe」「humi」「iporoho」「kese」「oro」「réhe」「sani」「sinricihi」「siri」に付与されている。言語固有な形態素属性だが、アプリニャ語 UD[25]においても、同様に名詞の所属形として使用されている。

ここまで見た限り、瀬沼版の FEATS フィールドは、アイヌ語特有の事象を記述するのに熱心なあまり、言語普遍な形態素属性がなおざりにされている。結果として、否定を表す「Polarity=Neg」すら使用されていない。アノテーション設計のバランスを欠いている、と言ってもいいだろう。

3.5 係り受けの問題

瀬沼版「ホテナオ」の係り受けは、かなり良く書けているものの、疑問に思える点も少なくない。付録 A をもとに、ざっと見ていくことにしよう。

「nismu=as」「sap=as」「okay=as」など、人称接辞のリンクに aux が使用されているが、これらは nsubj の方が適切[2]である。動詞につく他の人称接辞も、nsubj もしくは obj を付与すべきである。

ky-6-1 の「kusu」－ nmod →「pis」は、「pis」← obl －「sap」の誤りである。また、以降の nmod のうち、リンク元が動詞であるものは、obl の方が適切である。

ky-6-2 の「pon rupne aynu」に acl:relcl が 2 つも使われているが、これは「aynu」に体言修飾型関係節 (adjective relative clause) が 2 つもぶら下がっている、というアノテーションになっていて、非常に気持ち悪い。いずれも amod の方が適切だと考えられる。同様に他の acl:relcl も、amod もしくは acl が適切である。

ky-6-5 の「hepasi」← advmod －「suy」は、筆者の感覚としては修飾の方向が逆である。「hepasi」－ nummod →「iwan」－ clf →「suy」とするか、あるいは「iwan suy」を 1 語にして「hepasi」－ nummod →「iwansuy」が適切である。「heperay iwan suy」も同様に、「heparay」－ nummod →「iwansuy」が適切である。

ky-6-8 と ky-6-14 の「ukaepita」← xcomp －「ki」は、ccomp の方が適切である。ky-6-20 の「usiritpita」← xcomp －「ki」も、ccomp の方が適切である。

ky-6-11 の「sirpan」← advcl －「ari」は、「sirpan」← advcl －「ye」の誤りである。「ruwe」－ advmod →「tasi」は case の誤りだが、これはそもそも、副助詞「tasi」を ADP

[24]https://github.com/UniversalDependencies/UD_Afrikaans-AfriBooms
[25]https://github.com/UniversalDependencies/UD_Apurina-UFPA

にしていない点が誤りの根源である。副助詞「anak」「anakne」「kane」についても同様である。

ky-6-22 の「e=sinricihi」は nmod:poss を用いているが、筆者の考えでは、それは英語 UD[26]に引きずられ過ぎている。この場合は det の方が適切だろう。

ky-6-23 の「ta」－ case →「ta」は、「i」－ case →「ta」の誤りである。ky-6-24 の「arkehe」－ mark →「ko」は、「oterke」－ mark →「ko」の誤りである。ky-6-25 の「kor」←ccomp －「ne」は、「ruwe」－ cop →「ne」の誤りである。ky-6-27 の「kar」←ccomp －「ne」も、「ruwe」－ cop →「ne」の誤りである。

4　アイヌ語 Universal Dependencies の再構築

ここまでの議論をもとに、アイヌ語 UD の再構築をどのようにおこなうか、筆者なりに考えてみた。対象は、とりあえず「ホテナオ」とする。

本文の底本は、知里版を使用する。誤植が多いのは承知の上で、知里版をそのまま FORM に使用する。文の切れ目は、知里版における文の切れ目 (ピリオドと感嘆符と「:–」) に従うが、1 行目のサケへ「Hotenao」だけは、独立した文とする。27 文の sent_id は、行番号 (切替版や片山版で示されている) と対応が付くよう、SYOS_6_1・SYOS_6_2-5・SYOS_6_6-8・SYOS_6_9・SYOS_6_10-12・SYOS_6_13-14・SYOS_6_15-16・SYOS_6_17-20・SYOS_6_21・SYOS_6_22・SYOS_6_23-25・SYOS_6_26・SYOS_6_27-28・SYOS_6_29-32・SYOS_6_33・SYOS_6_34・SYOS_6_35-36・SYOS_6_37・SYOS_6_38・SYOS_6_39-43・SYOS_6_43-45・SYOS_6_46-53・SYOS_6_54-56・SYOS_6_57-59・SYOS_6_60-61・SYOS_6_62-65・SYOS_6_66 とする。

LEMMA は『アイヌ語沙流方言辞典』に従う。作業上、アイヌ民族博物館アイヌ語アーカイブ[27]のオンライン版を使用してもよい。あるいは、単語の切れ目に疑義がある場合は、切替版や片山版を参照してもよい。

UPOS と XPOS は、表 2 に従って付与する。XPOS は『アイヌ語沙流方言辞典』に立脚しつつ、固有名詞を名詞から分離し、数詞を連体詞から分離し、さらに記号を加えたも

表 2: アイヌ語 UD 再構築のための UPOS と XPOS

UPOS	XPOS
NOUN	名詞
	位置名詞
	形式名詞
PROPN	固有名詞
PRON	代名詞
VERB	完全動詞
	自動詞
	他動詞

UPOS	XPOS
SCONJ	後置副詞
	接続助詞
	接続詞
CCONJ	
ADP	格助詞
	副助詞
PART	終助詞
	人称接辞

UPOS	XPOS
ADV	副詞
NUM	数詞
DET	連体詞
AUX	助動詞
	デアル動詞
INTJ	間投詞
PUNCT	記号

[26]https://github.com/UniversalDependencies/UD_English-EWT
[27]https://ainugo.nam.go.jp

のである。ただし、単他動詞と複他動詞は、まとめて他動詞としている。連他動詞に関しては、(位置) 名詞と他動詞の 2 語に分けて、それぞれに品詞を付与する。UPOS は XPOS から半自動的に決定できるが、接続詞のうち CCONJ とすべき語 (「awa」と「korka」) があるので、その点は注意されたい。人称接辞以外の接頭辞や接尾辞は、独立の語とせず、語幹とまとめる (品詞は語幹に従う) が、どうしても独立させざるを得ない場合 (たとえば図 4) は、UPOS を PART とすべきだろう。

FEATS は付与しない。グロスも付与しない。これらについては、LEMMA と XPOS から自動生成するようなシステムを、CoNLL-U とは別の形で準備し、それによって自動付与をおこなった方が効率的である。ただ、そのようなシステムに力を割くより、品詞と係り受けをちゃんと整備することが、現状としては必要な方策だと考えられる。

係り受けタグについては、表 1 の言語普遍なタグ 37 種類に限定し、言語固有のタグは使用しない。瀬沼版では、言語固有のタグとして acl:relcl と nmod:poss が使われている (付録 A) が、前節での議論の通り acl:relcl に対しては、言語普遍なタグ (amod あるいは acl) に置き換える。nmod:poss に対しても、言語普遍なタグ (det あるいは nmod) に置き換える。瀬沼が懸念[2]する「eani anak horkew e=ne」という例に対しては、筆者としては expl を用いる案を、図 3 に示しておく。また、連他動詞 (第三類の動詞) の例[28]に対しては、あくまで動詞中心主義を堅持しつつ fixed を用いる案を、図 4 に示しておく。

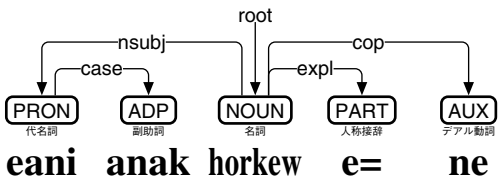


図 3: 「eani anak horkew e=ne」係り受け案

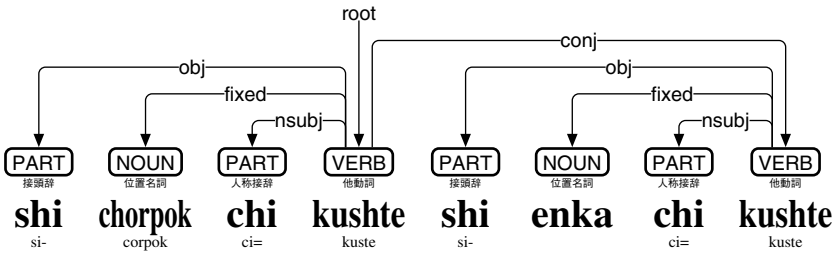


図 4: 「shichorpok chikushte shienka chikushte」係り受け案

ここまで述べた基準にもとづいて、実際に「ホテナオ」を再構築してみた。図 5 に SYOS_6_33 の CoNLL-U を示す。また、全データを GitHub で公開[29]すると同時に、付録 B に依存構造グラフを示す。この CoNLL-U のテキスト部分は、知里版を底本として、手作業で入力した。データ部分は、筆者作成のアイヌ語 UD エディター[30]で、品詞と係り受けリンクを付与し、その後に LEMMA を手作業で直した。

[28]佐藤知己: 知里幸恵『アイヌ神謡集』の難読個所と特異な言語事例をめぐって, 北海道立アイヌ民族文化センター研究紀要, 第 10 号 (2004 年 3 月), pp.1-32.

[29]https://github.com/KoichiYasuoka/ud-ainu/blob/master/ain_syos-ud-test.conllu

[30]https://koichiyasuoka.github.io/deplacy/demo/2021-07-30/editor-ainu.html

```
# sent_id = SYOS_6_33
# text = itskash awa ponrupneainu ene itaki : ―
# text_ja = 云ふと小男の云ふことには
1   itsk    itak    VERB    自動詞    _   0   root    _   SpaceAfter=No
2   ash     =as     PART    人称接辞  _   1   nsubj   _   _
3   awa     awa     CCONJ   接続詞    _   1   cc      _   _
4   pon     pon     VERB    自動詞    _   6   amod    _   SpaceAfter=No
5   rupne   rupne   VERB    自動詞    _   6   amod    _   SpaceAfter=No
6   ainu    aynu    NOUN    名詞      _   8   nsubj   _   _
7   ene     ene     ADV     副詞      _   8   advmod  _   _
8   itak    itak    VERB    自動詞    _   9   acl     _   SpaceAfter=No
9   i       -i      PART    接尾辞    _   1   conj    _   _
10  :       :       PUNCT   記号      _   9   punct   _   _
11  ―       ―       PUNCT   記号      _   9   punct   _   _
```

図 5: SYOS_6_33 の CoNLL-U(筆者作成)

5 おわりに

筆者とアイヌ語表記の関わりは、JIS X 0213 へのアイヌ語表記用カタカナ追加[31]に始まる。その後、ISO/IEC 10646 への Katakana Phonetic Extensions 追加[32]では、結果として、小書きの「ㇷ゚」など半濁点が使いにくくなってしまったものの、それでもアイヌ語表記に困らない程度にはなっている。また、ウポポイ (民族共生象徴空間) では、アイヌ語表記にカタカナを用いており (図 6)、今後もアイヌ語のカタカナ表記は増えていくだろう。

では、本稿で対象とした「ホテナオ」は、カタカナ表記とすべきだろうか。これに関しては、たとえば片山版のカタカナ表記を底本とするならば、FORM はカタカナにすべき[33]である。知里版を底本とするならば、図 5 に示した通りである。つまり、UD におけるアイヌ語表記をどうするかという問題は、筆者としては棚上げにして、底本の表記を優先するというポリシーを貫くことにした。その一方で、LEMMA(および XPOS) は『アイヌ語沙流方言辞典』に接地した。『アイヌ語沙流方言辞典』にはカタカナ表記も併用されているが、あくまでラテンアルファベット順の辞典なので、LEMMA はラテンアルファベット表記となる。

ただ、こういうポリシーを、今後のアイヌ語 UD でも貫けるかどうかは、書写言語としてのアイヌ語を今後どう扱っていくべきか、という問題に関わってくるのだろう。なかなか難しいところである。

[31]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 符号化文字集合調査研究委員会第 2 分科会 (WG2) 資料, JCS-2-8-02 (1996 年 11 月 25 日).

[32]Addition of forty eight characters, JTC1/SC2/WG2 N2092 (September 13, 1999).

[33]ただし、カタカナ表記の場合、縮約語の分解の問題が表面化する。たとえば、表記上「イタカシ」と書かれている語を、「イタㇰ」と「アシ」に分けるか否かが問題となる。筆者としては、ID フィールドに単語の範囲を示すやり方 (図 7) で、縮約語を分解すべきだと考えるが、実作業上は困難を伴うことが予想される。

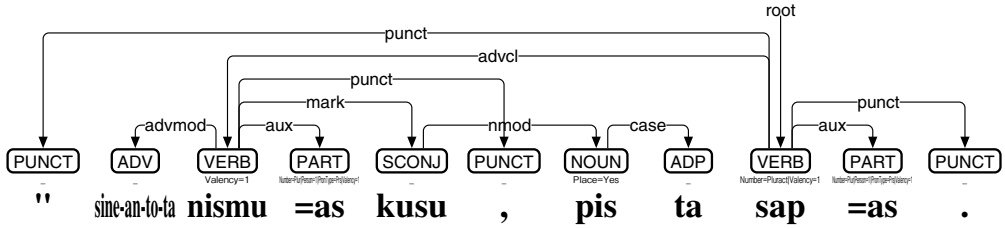
図 6: ウポポイにおけるアイヌ語表記 (2020 年 8 月 19 日筆者撮影)

```
# text = イタカシ アワ ポンルプネアイヌ エネ イタキ：—
1-2   イタカシ   _       _       _       _   _   _       _   _
1     イタク     itak    VERB    自動詞   _   0   root    _   _
2     アシ       =as     PART    人称接辞 _   1   nsubj   _   _
3     アワ       awa     CCONJ   接続詞   _   1   cc      _   _
4     ポン       pon     VERB    自動詞   _   6   amod    _   SpaceAfter=No
5     ルプネ     rupne   VERB    自動詞   _   6   amod    _   SpaceAfter=No
6     アイヌ     aynu    NOUN    名詞     _   8   nsubj   _   _
7     エネ       ene     ADV     副詞     _   8   advmod  _   _
8-9   イタキ     _       _       _       _   _   _       _   SpaceAfter=No
8     イタク     itak    VERB    自動詞   _   9   acl     _   _
9     イ         -i      PART    接尾辞   _   1   conj    _   _
10    ：         :       PUNCT   記号     _   9   punct   _   SpaceAfter=No
11    —          —       PUNCT   記号     _   9   punct   _   _
```

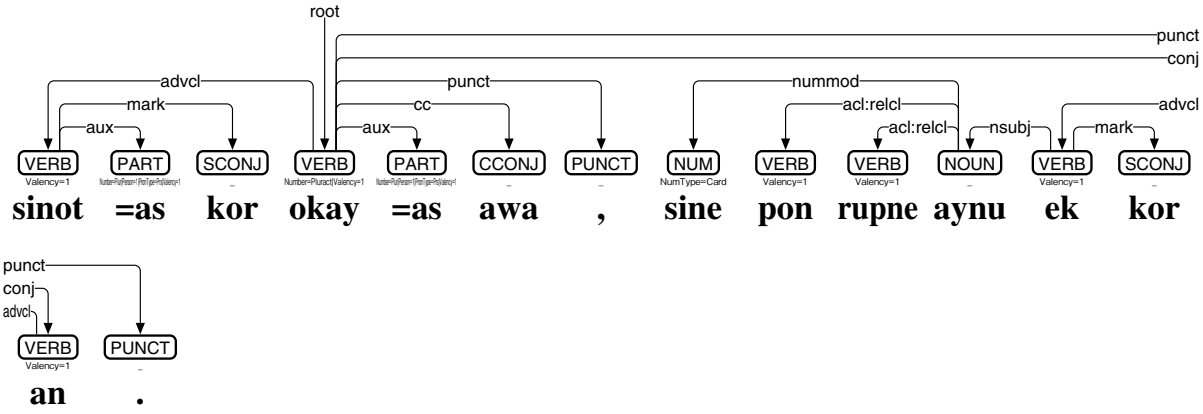
図 7: カタカナ表記アイヌ語 UD の CoNLL-U(筆者作成)

付録A　瀬沼版「ホテナオ」依存構造グラフ

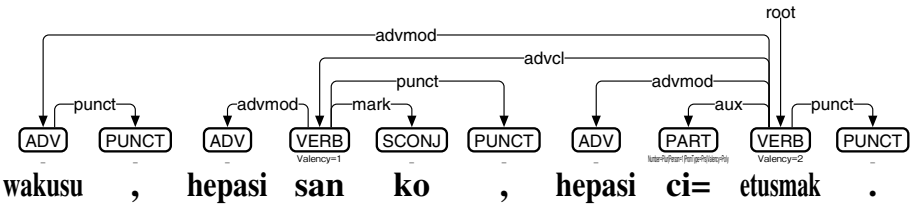
```
# sent_id = ky-6-1
```



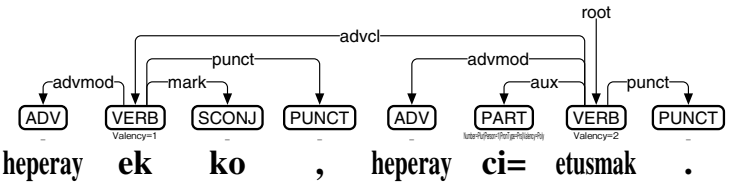

```
# sent_id = ky-6-2
```



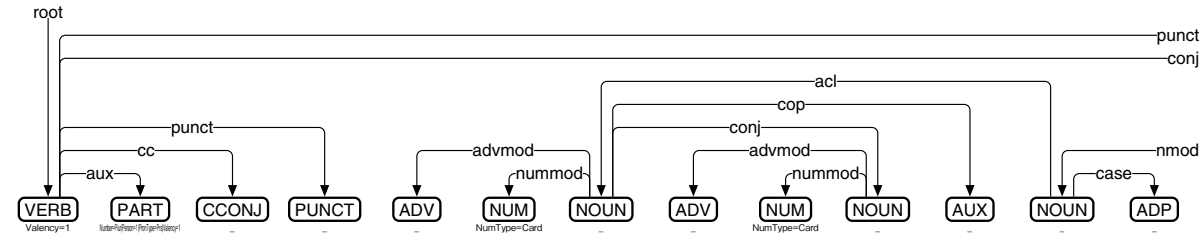

```
# sent_id = ky-6-3
```

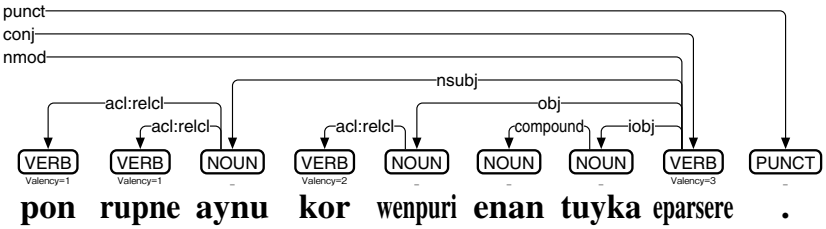



```
# sent_id = ky-6-4
```

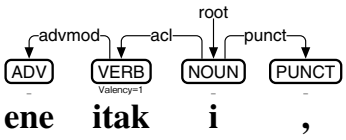


sent_id = ky-6-5

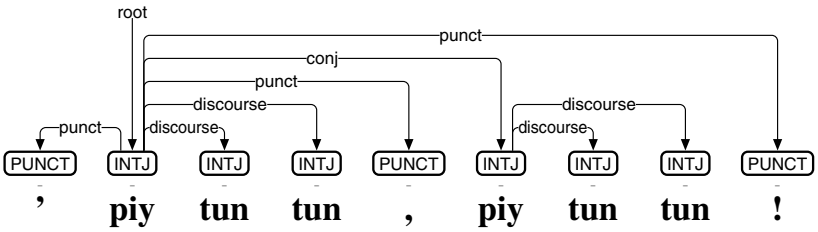




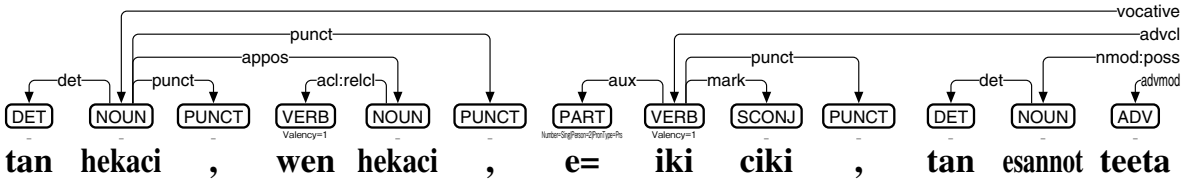
sent_id = ky-6-6

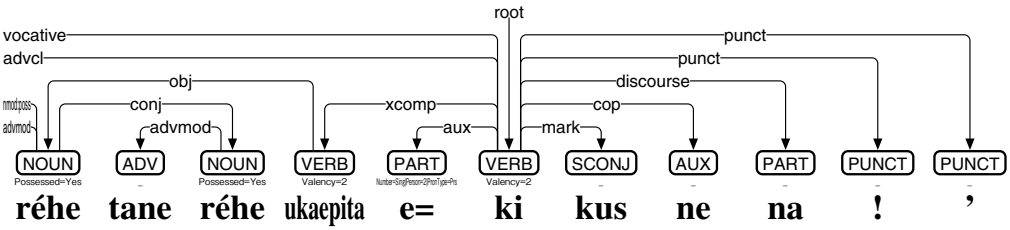


sent_id = ky-6-7

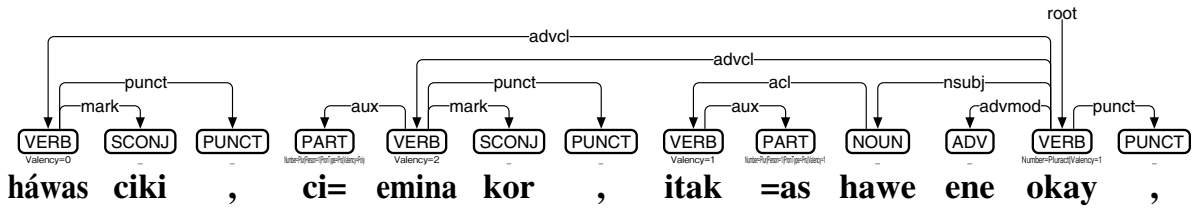


sent_id = ky-6-8

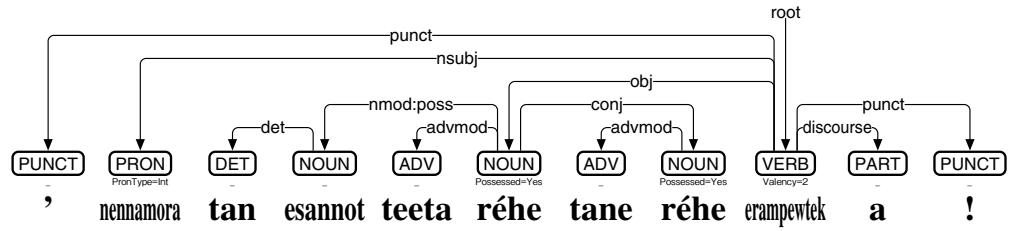




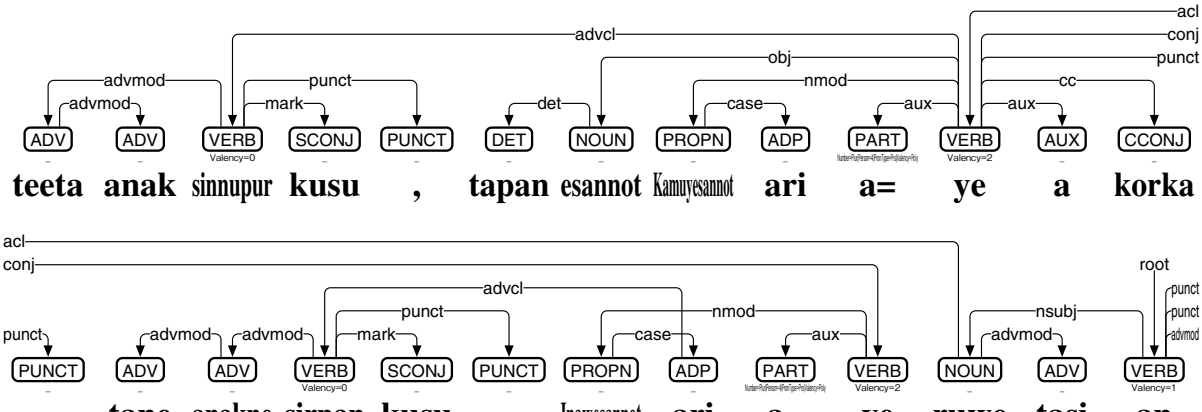
sent_id = ky-6-9

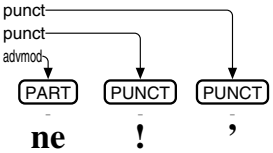


sent_id = ky-6-10

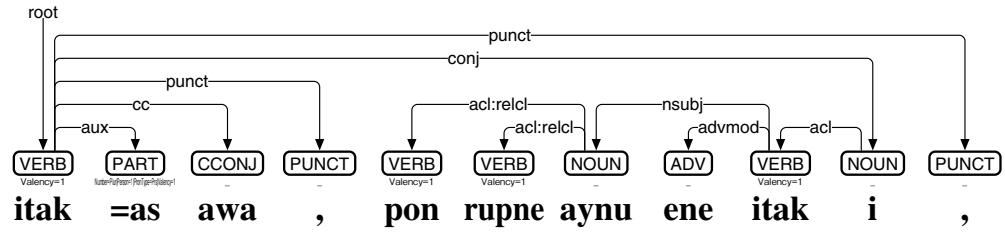


sent_id = ky-6-11

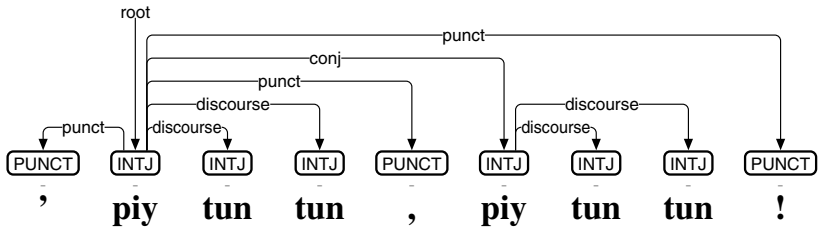




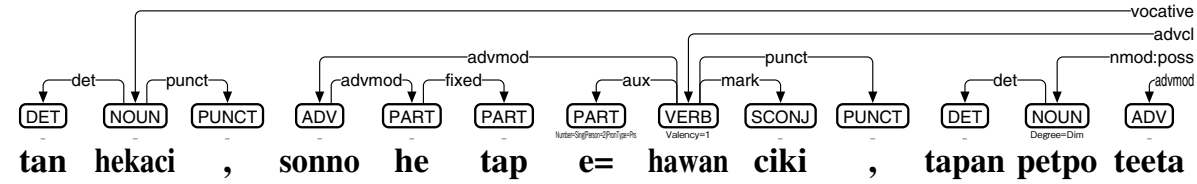
sent_id = ky-6-12

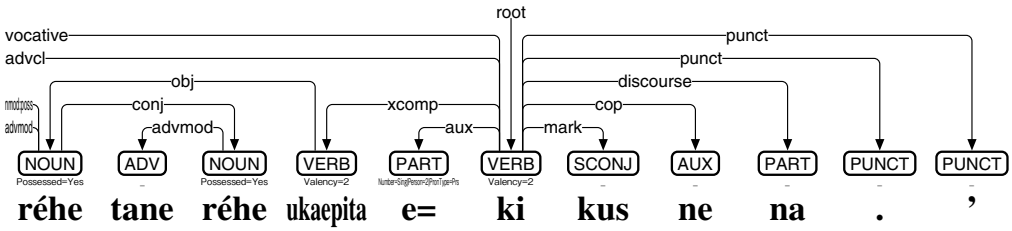


sent_id = ky-6-13

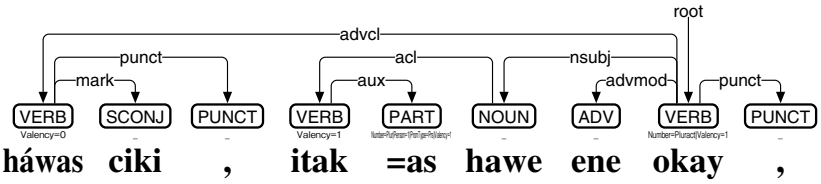


sent_id = ky-6-14

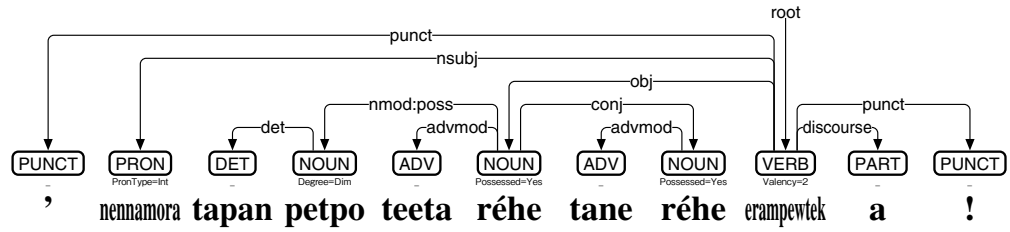




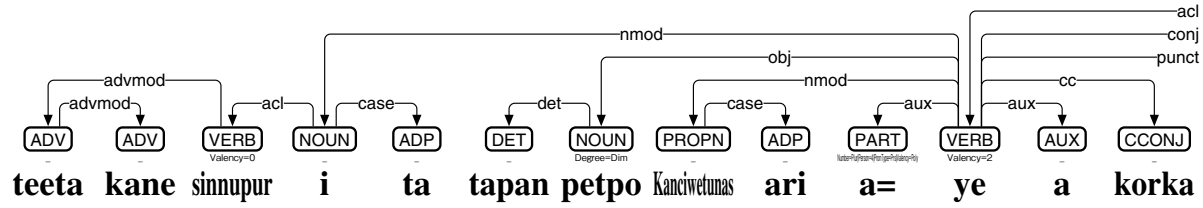
sent_id = ky-6-15

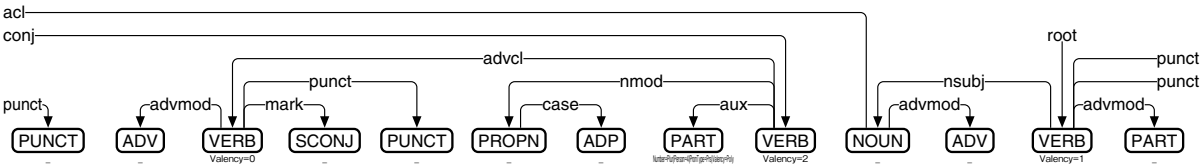


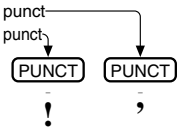
sent_id = ky-6-16



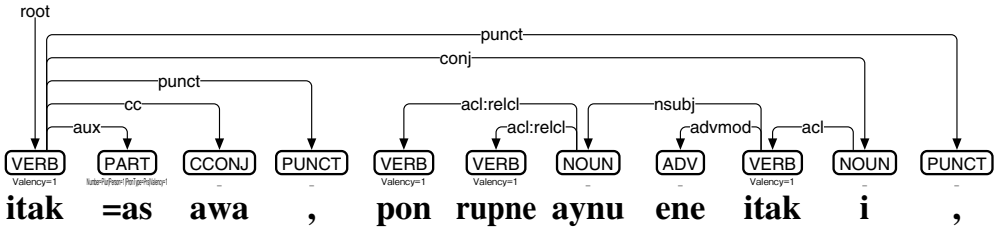
sent_id = ky-6-17



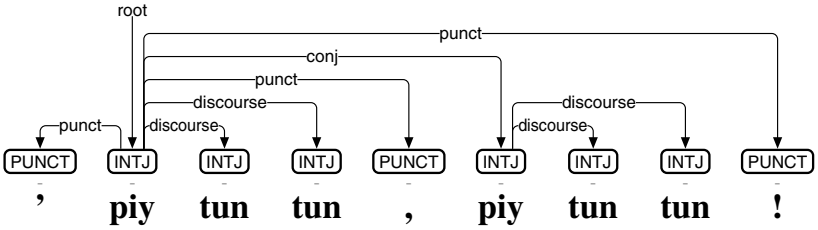




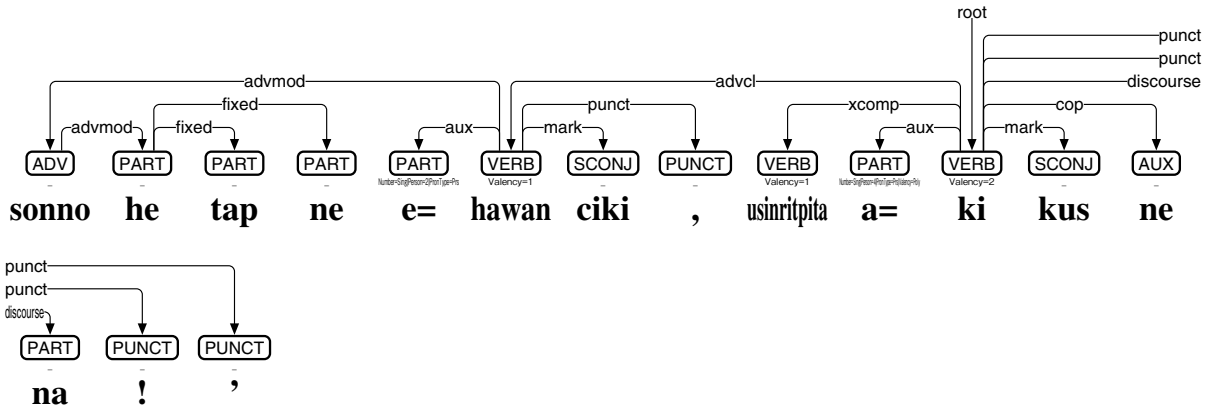
sent_id = ky-6-18



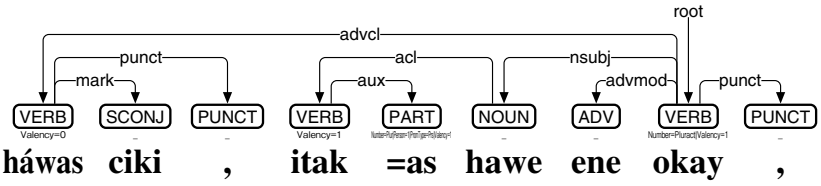
sent_id = ky-6-19



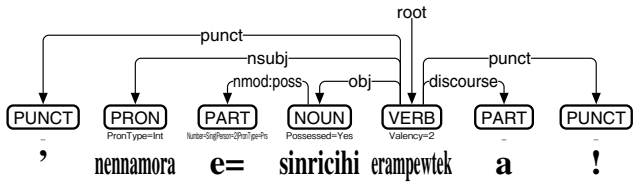
sent_id = ky-6-20



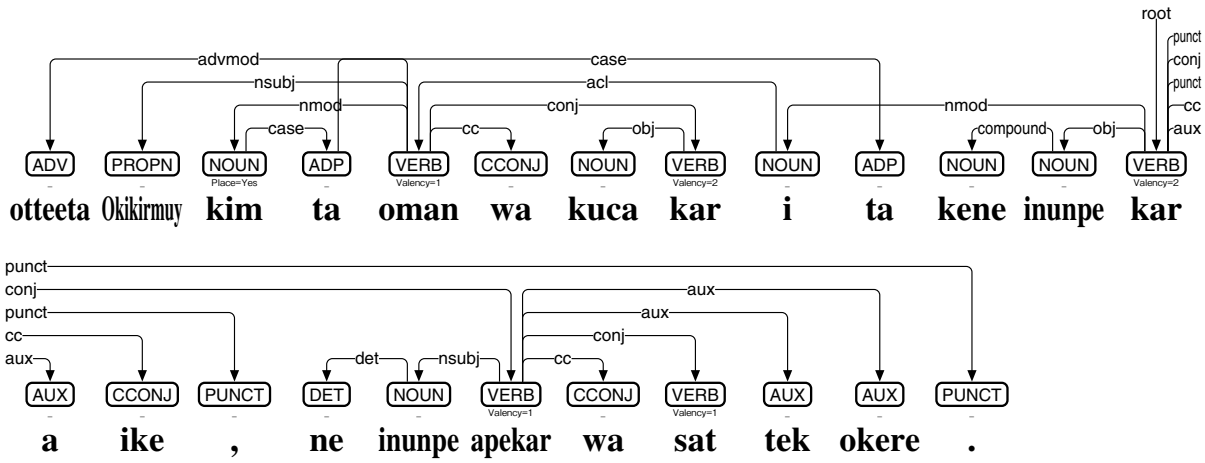
sent_id = ky-6-21



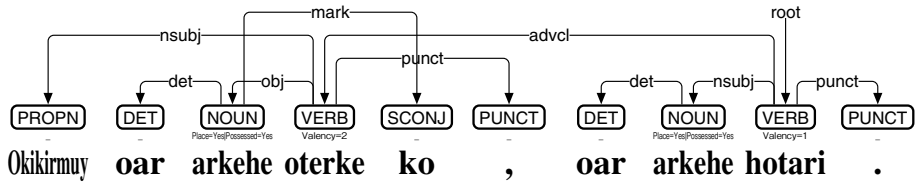
sent_id = ky-6-22



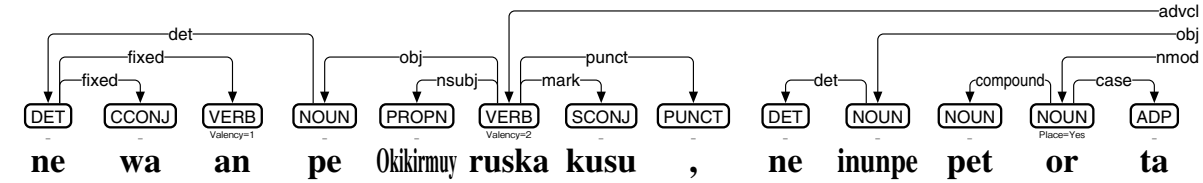
sent_id = ky-6-23

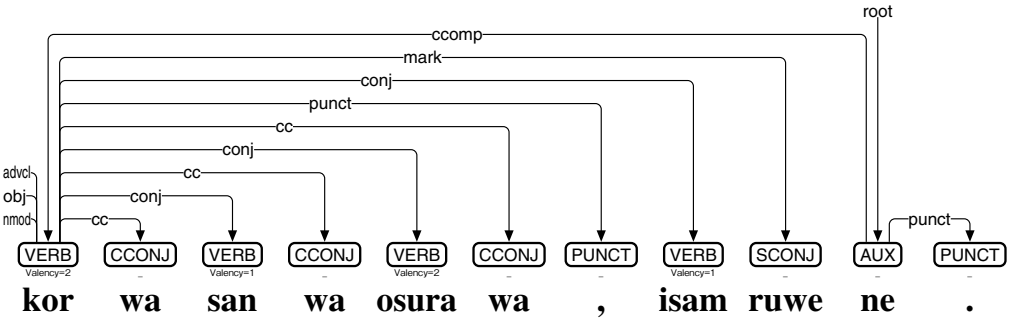


sent_id = ky-6-24

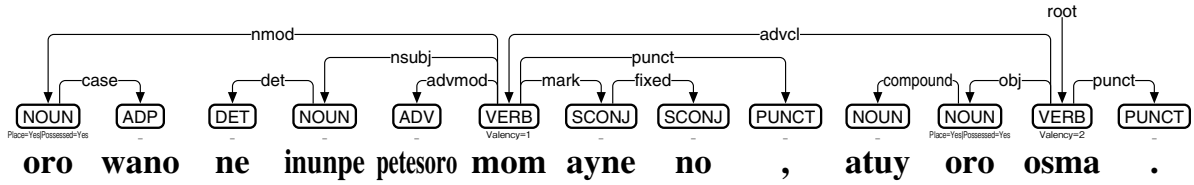


sent_id = ky-6-25

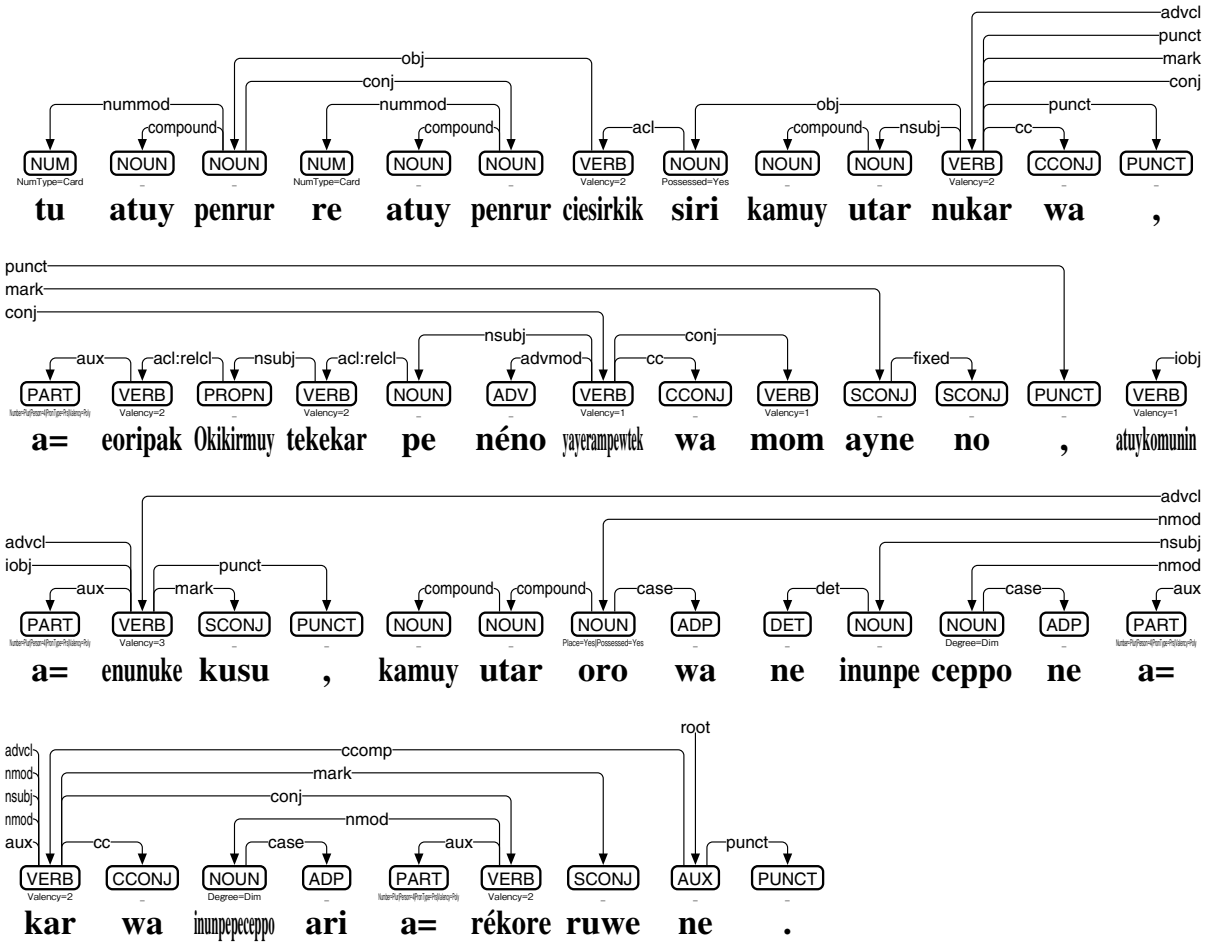




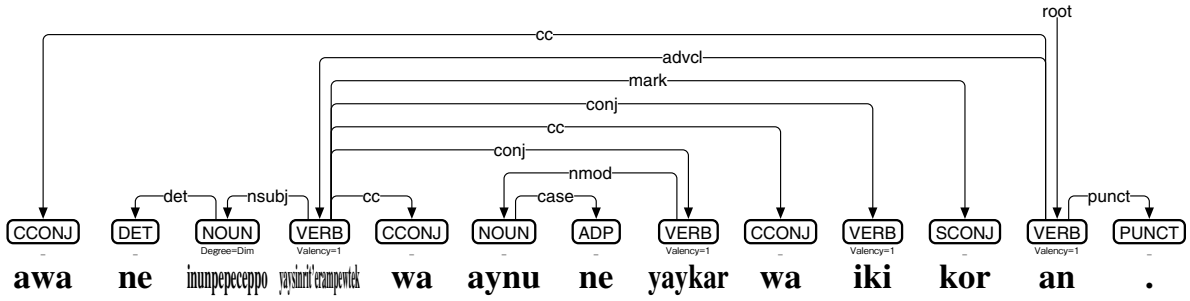
sent_id = ky-6-26



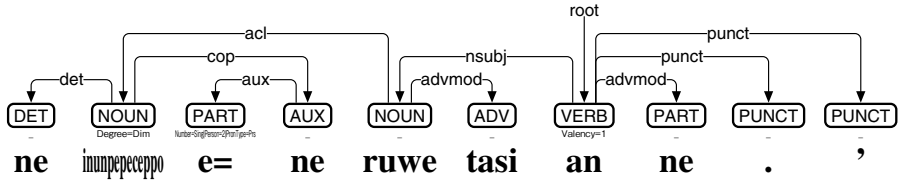
sent_id = ky-6-27



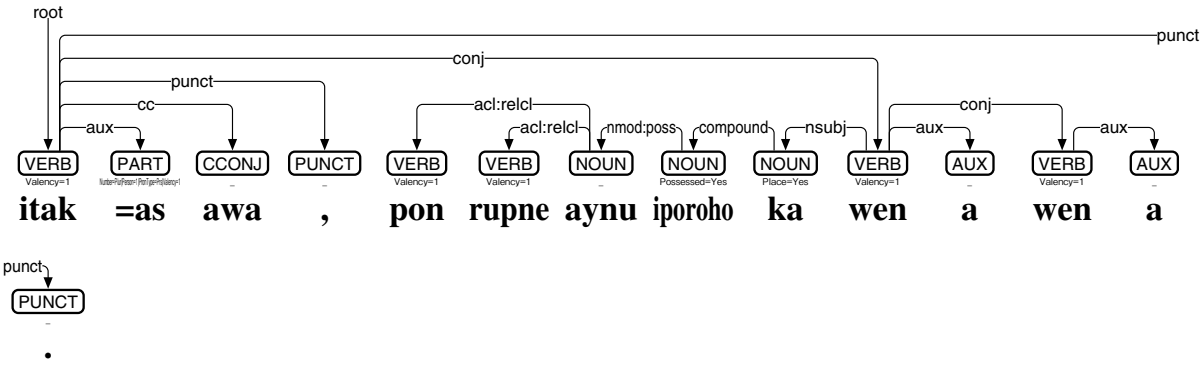
sent_id = ky-6-28



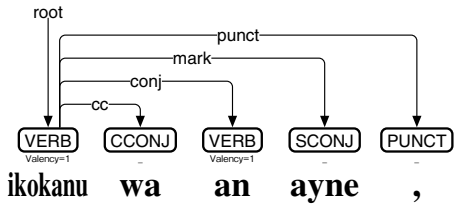
sent_id = ky-6-29



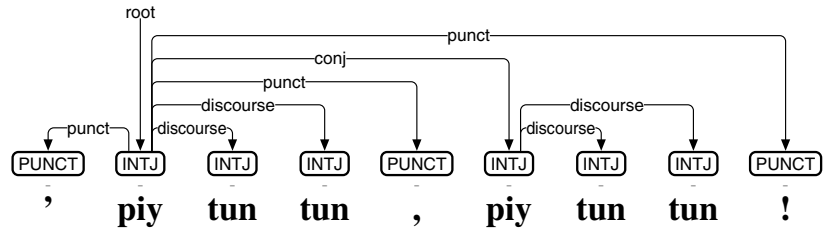
sent_id = ky-6-30



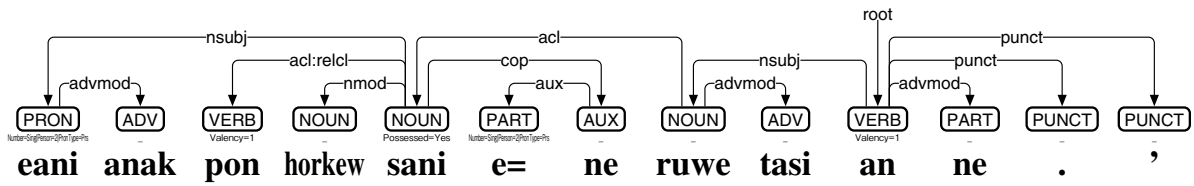
sent_id = ky-6-31



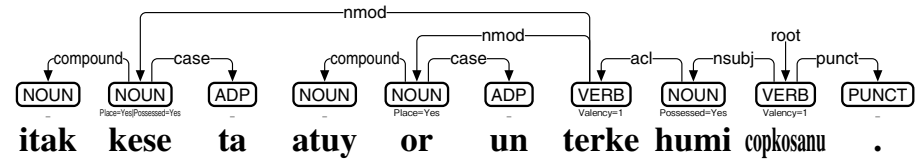
sent_id = ky-6-32



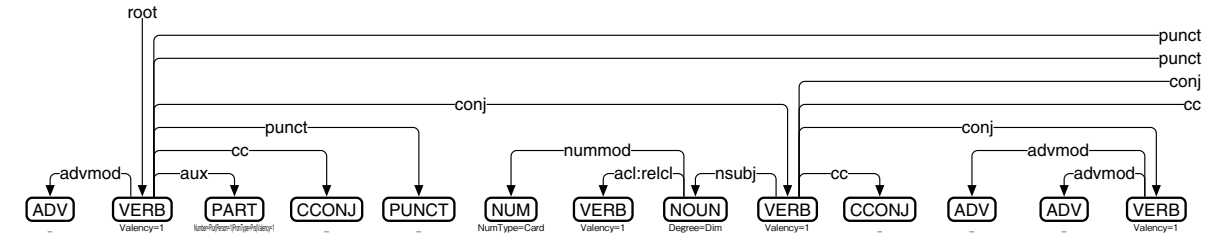
sent_id = ky-6-33

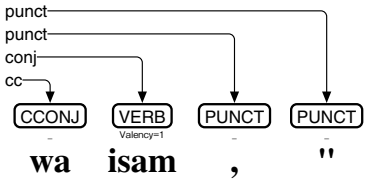


sent_id = ky-6-34

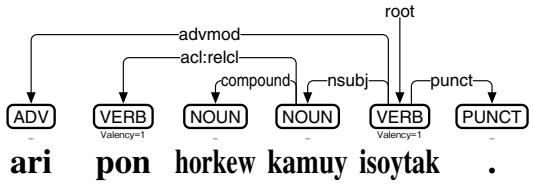


sent_id = ky-6-35





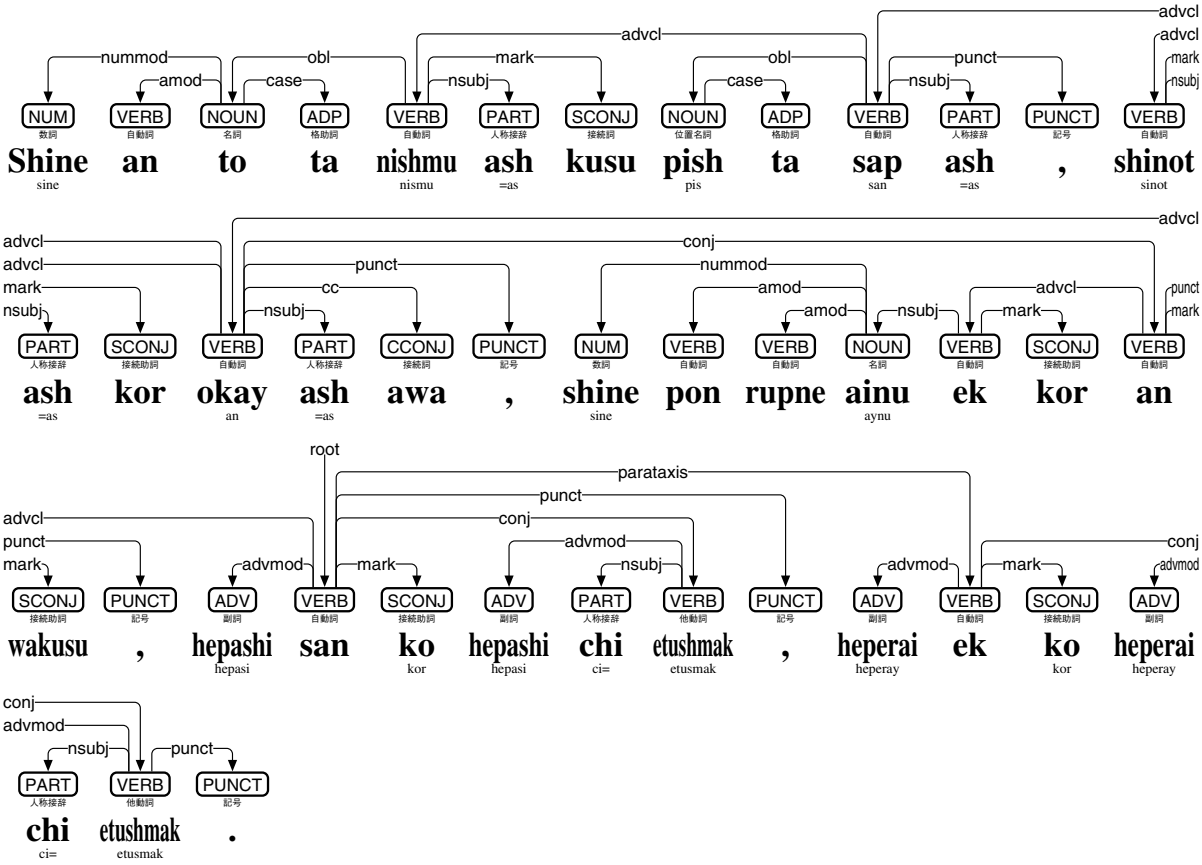
sent_id = ky-6-36



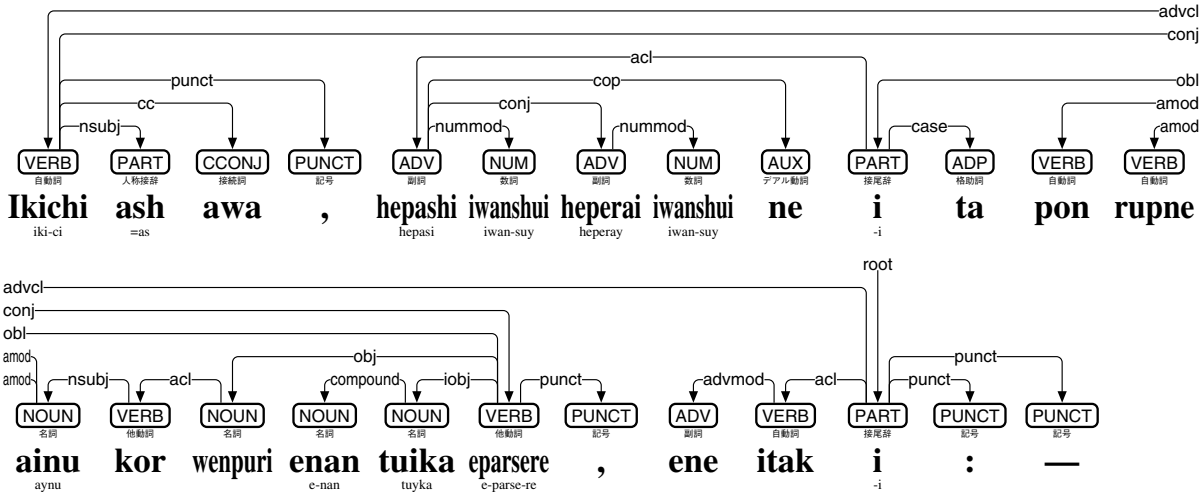
付録B　安岡版「ホテナオ」依存構造グラフ

```
# sent_id = SYOS_6_1
```

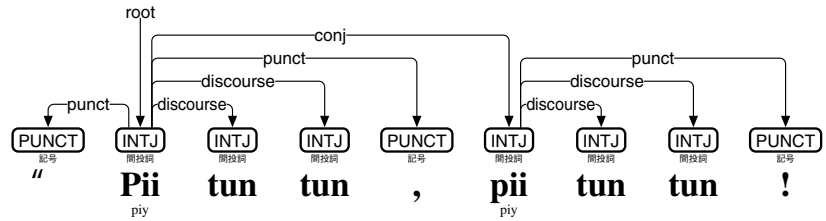

```
# sent_id = SYOS_6_2-5
```



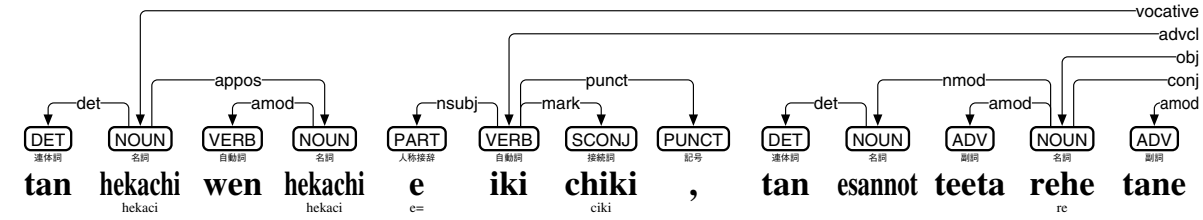

```
# sent_id = SYOS_6_6-8
```

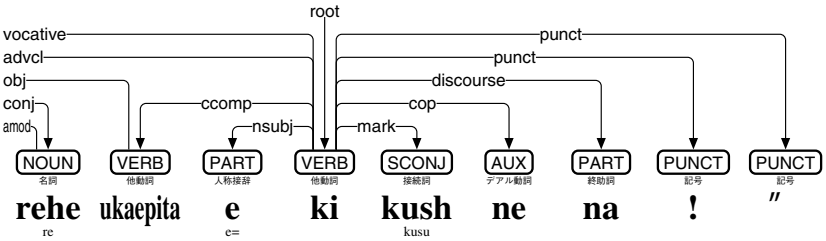


sent_id = SYOS_6_9

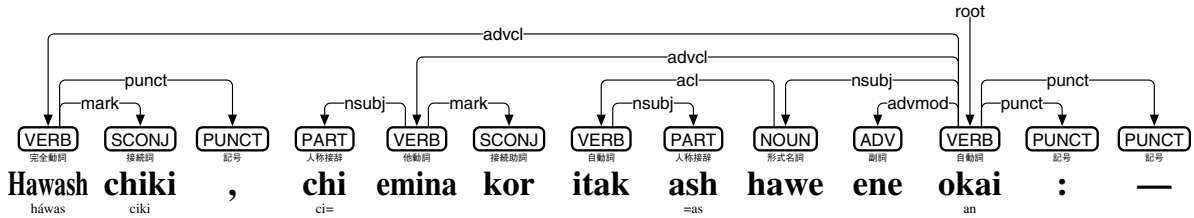


sent_id = SYOS_6_10-12

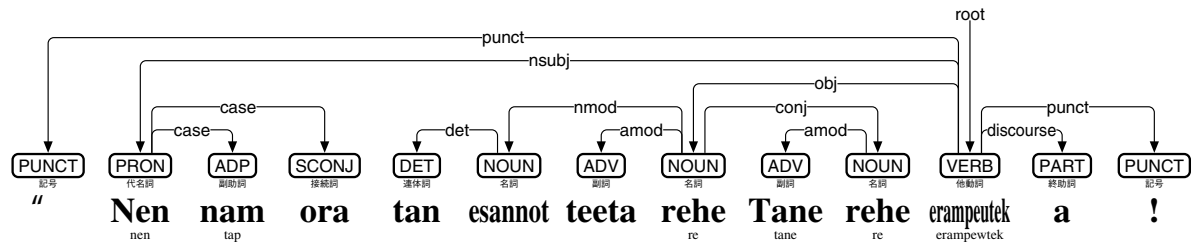




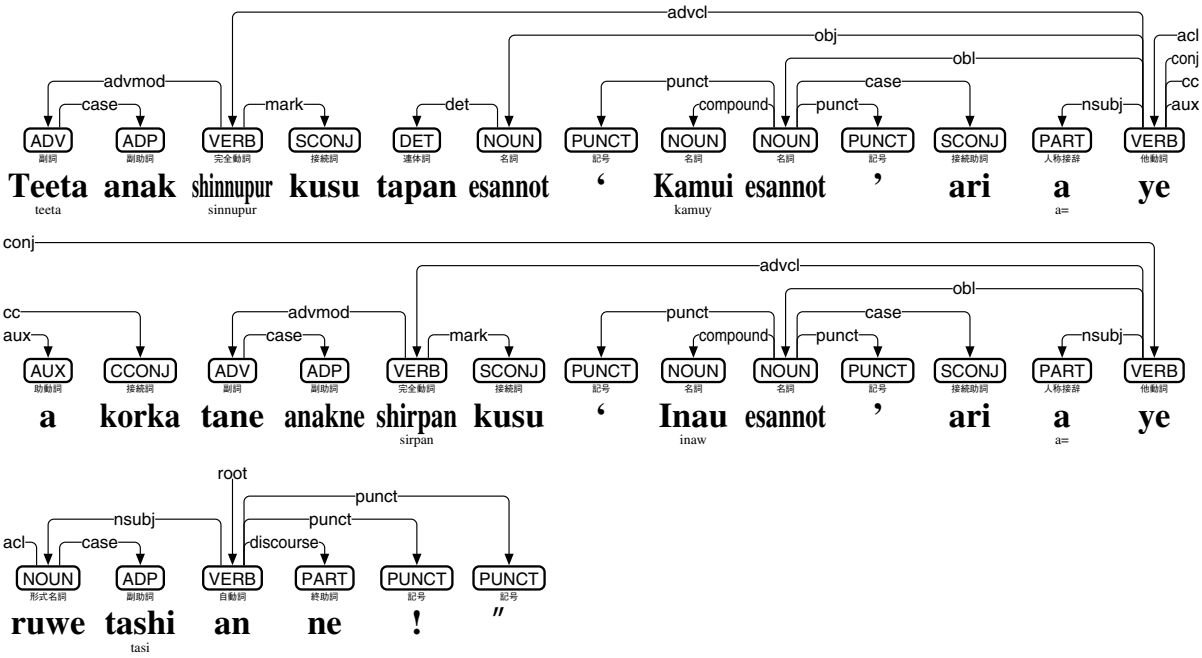
sent_id = SYOS_6_13-14



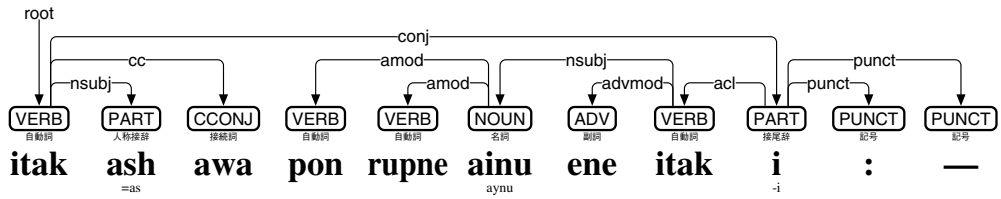
sent_id = SYOS_6_15-16



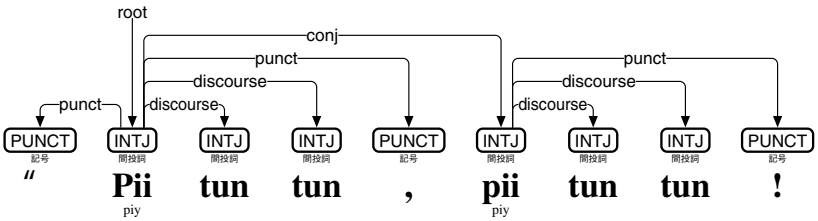
sent_id = SYOS_6_17-20



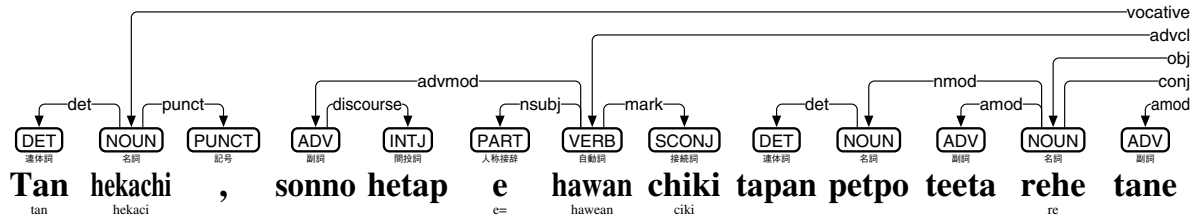
sent_id = SYOS_6_21

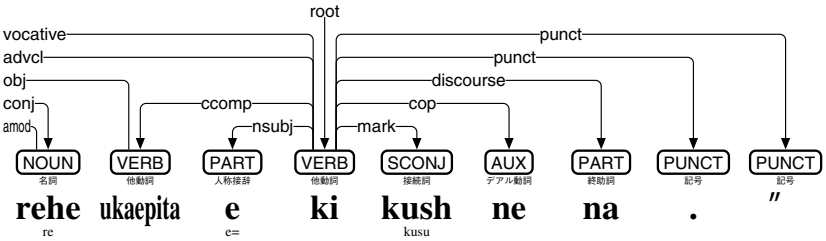


sent_id = SYOS_6_22

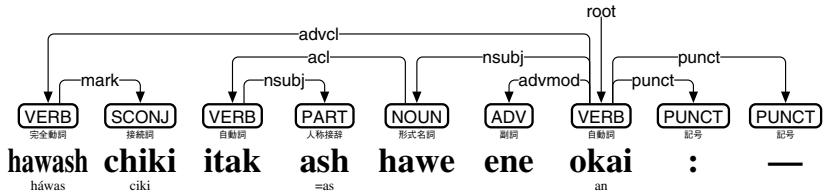


sent_id = SYOS_6_23-25

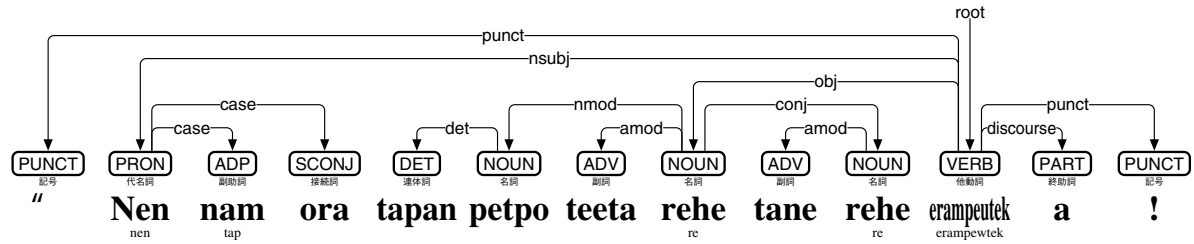




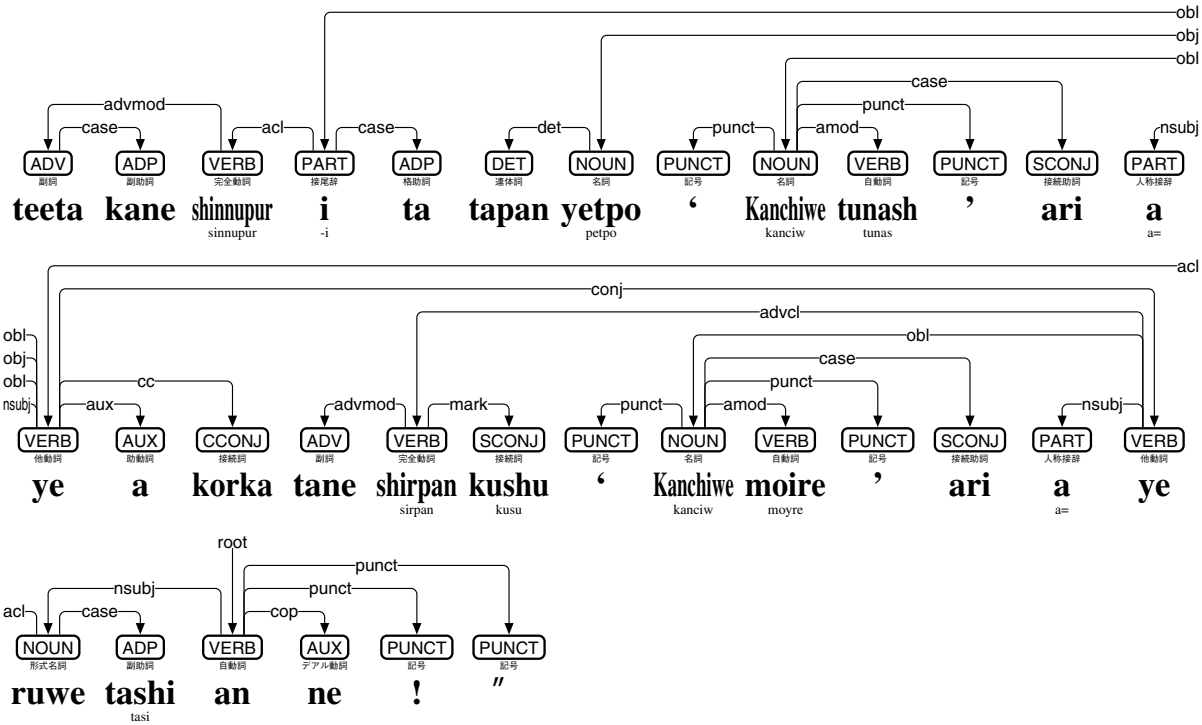
sent_id = SYOS_6_26



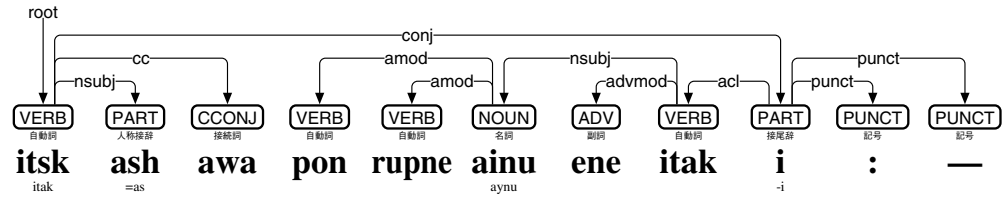
sent_id = SYOS_6_27-28



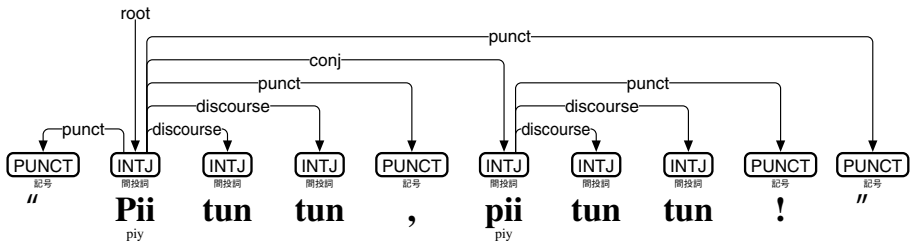
sent_id = SYOS_6_29-32



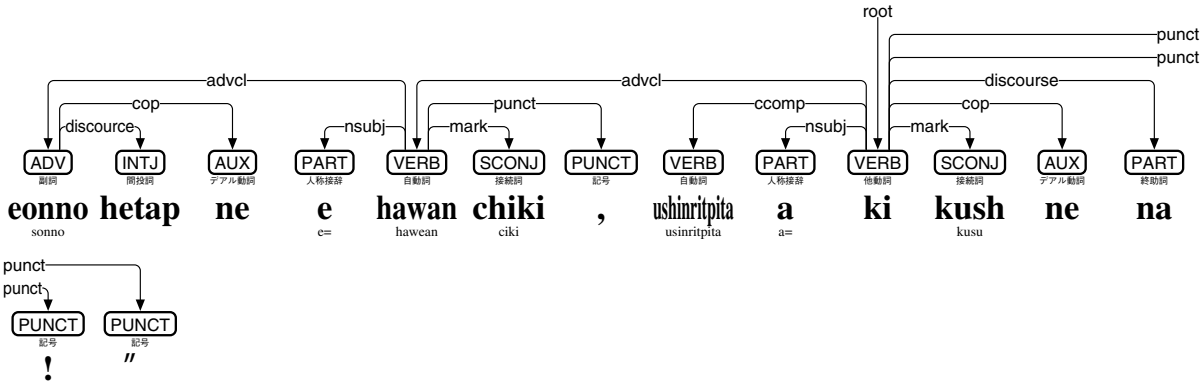
sent_id = SYOS_6_33



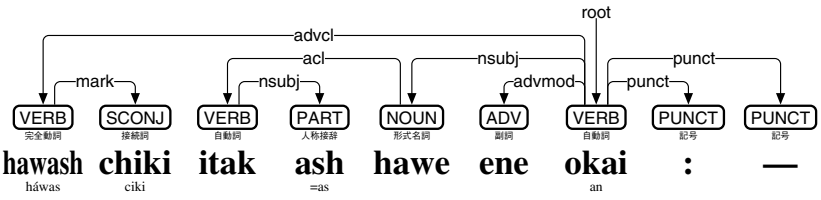
sent_id = SYOS_6_34



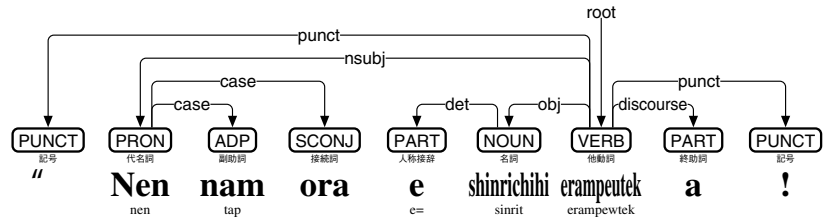
sent_id = SYOS_6_35-36



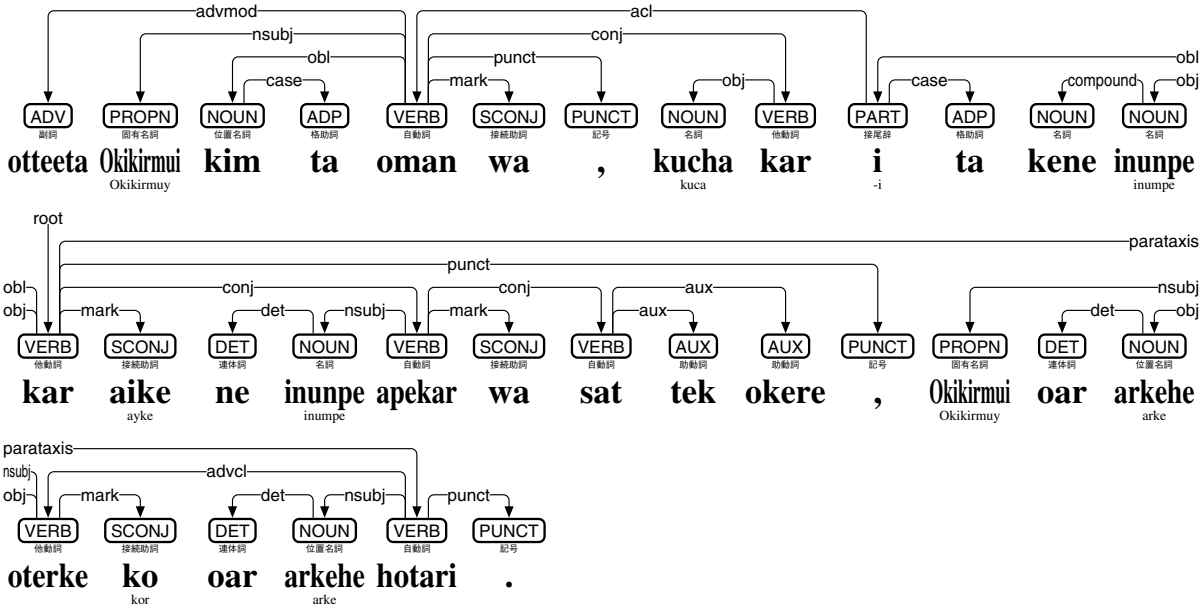
sent_id = SYOS_6_37



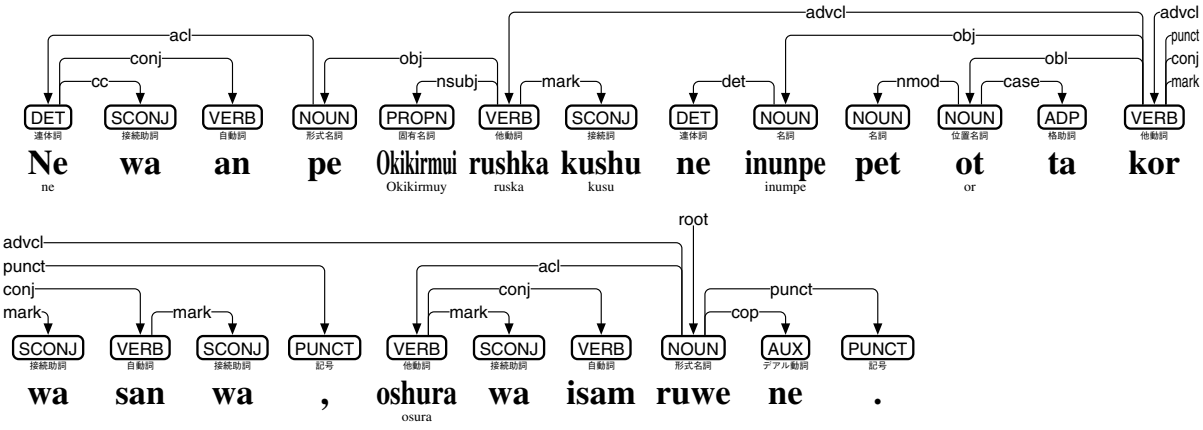
sent_id = SYOS_6_38



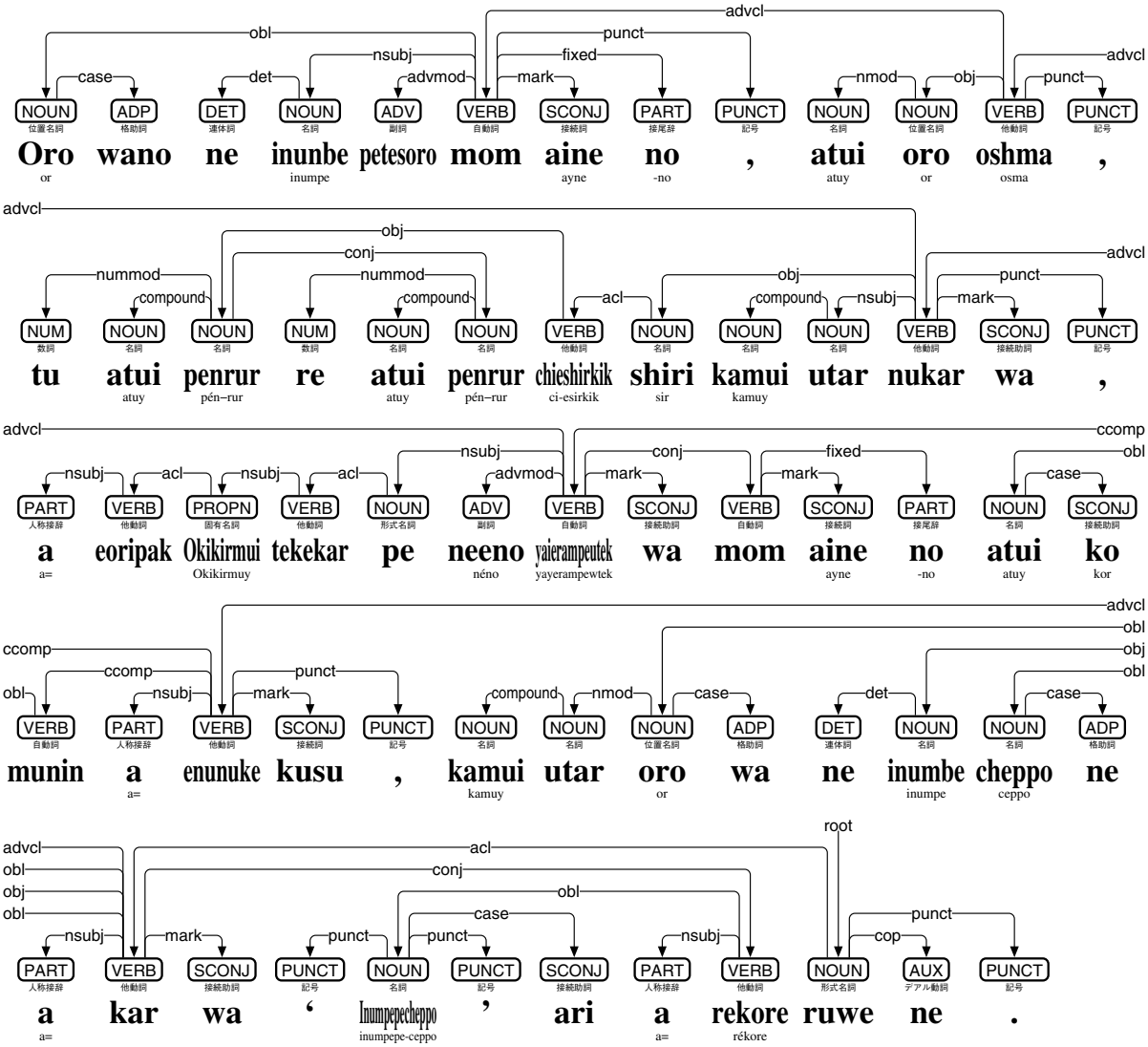
sent_id = SYOS_6_39-43



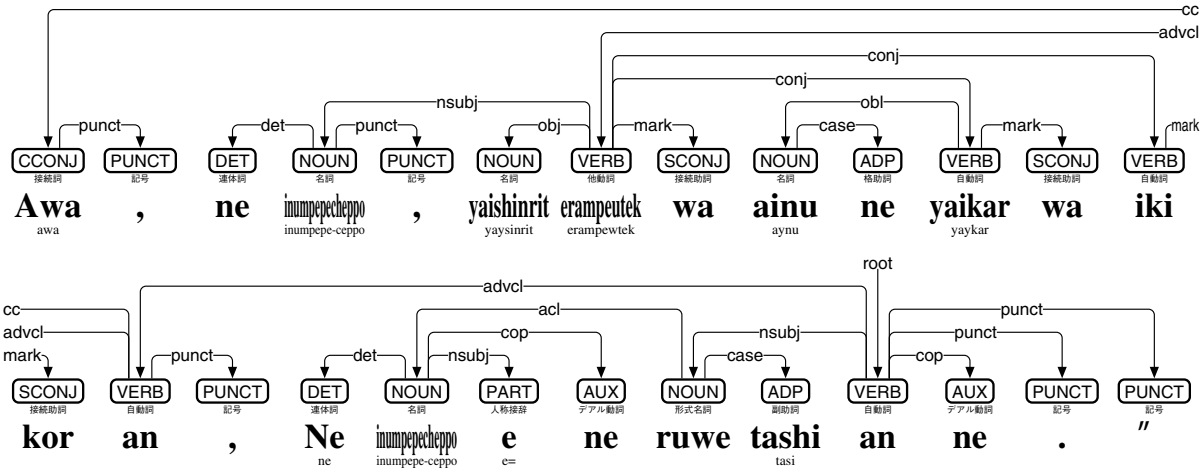
sent_id = SYOS_6_43-45



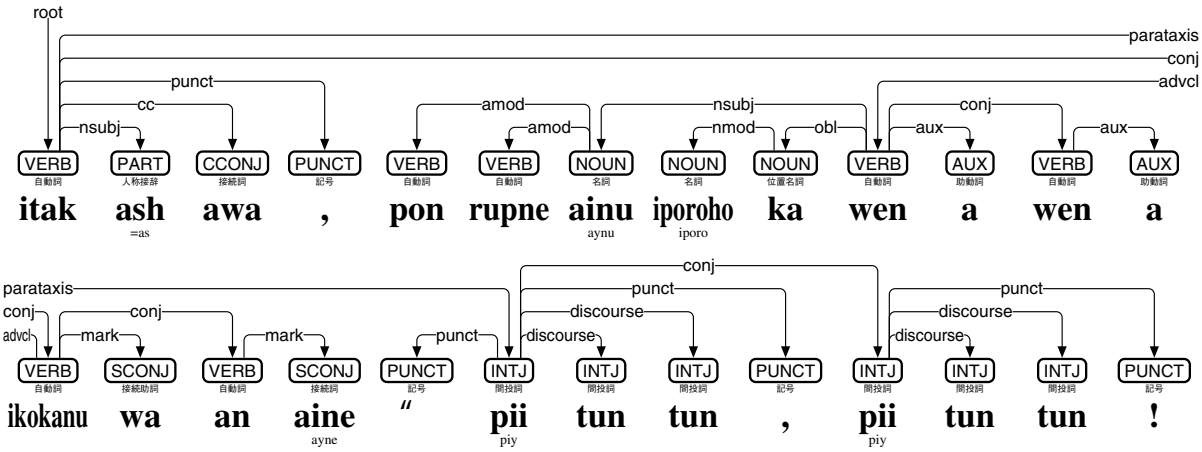
sent_id = SYOS_6_46-53



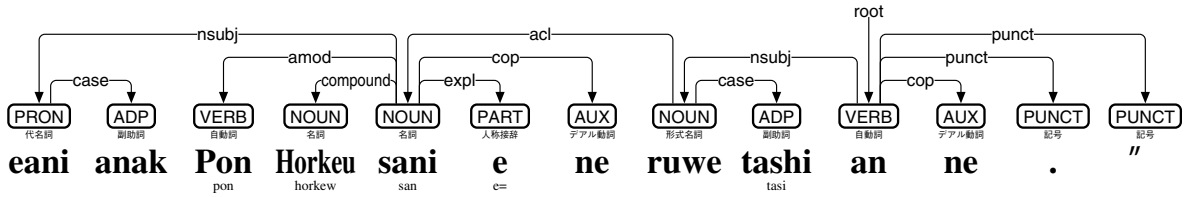
sent_id = SYOS_6_54-56



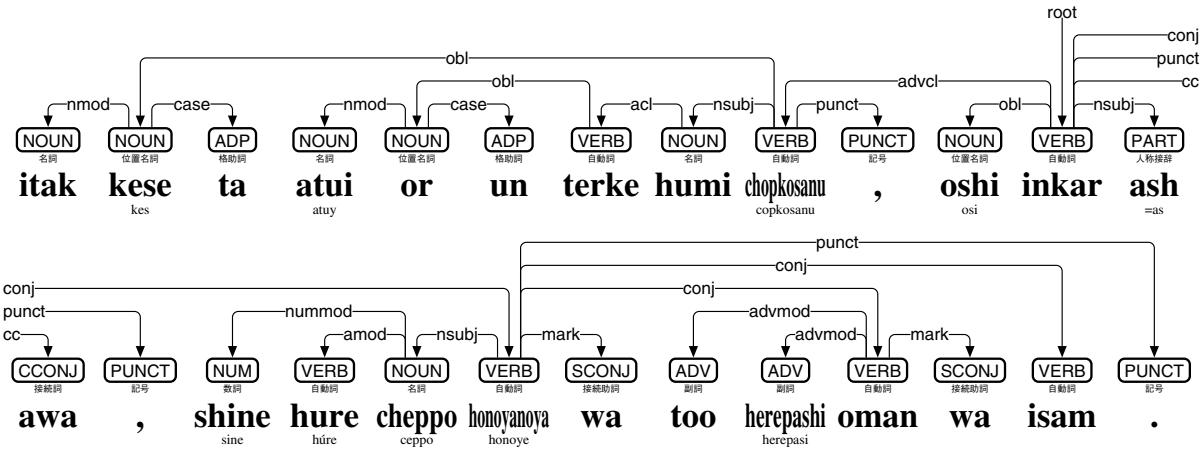
sent_id = SYOS_6_57-59



sent_id = SYOS_6_60-61



sent_id = SYOS_6_62-65



sent_id = SYOS_6_66