

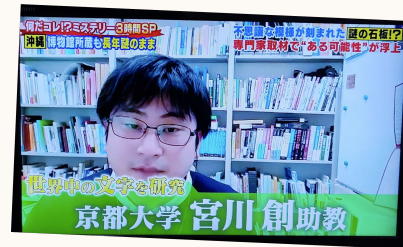
2023年3月17日 言語処理学会第29回年次大会ワークショップ『深層学習時代の計算言語学』

消滅の危機にある言語・方言のために 深層学習を応用する

みやがわ そう
宮川 創

国立国語研究所 テニユアトラック助教

自己紹介・宮川創（みやがわそう）



フジテレビ 世界の何だコレ！？ミステリー 2022年1月5日

- 国立国語研究所テニュアトラック助教
- 学振DC1（京大）→ ドイツ研究振興協会研究員（ゲッティンゲン大）→ 関西大学PDフェロー → 京都大学助教 → 国語研
- ドイツでは、コプト語や古ヌビア語など、復興運動のある、アフリカのローリソース文献言語のコーパス・語彙資源の作成、デジタルアーカイブの作成
- 現在は、国語研で日本語・琉球諸語・アイヌ語を中心としたコーパス・語彙資源、デジタルアーカイブを作成中

コプト語での自然言語処理とコーパス

COPTIC
SCRIPTORIUM

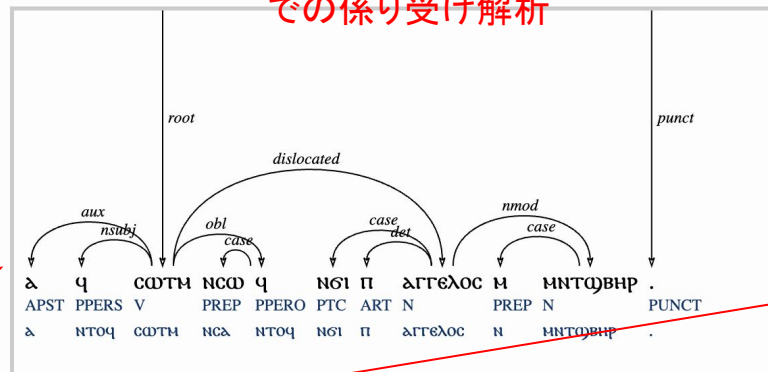
Result: **XML / SGML での形態素分析**

α<hi rend="red">ϣ</hi>ϣⲓⲛⲁⲩⲛⲁⲩ ⲛⲥⲱⲩ ϣⲛⲥ
 ⲓⲛⲁⲩⲉⲗⲟⲥ ⲙⲓⲛⲧⲱⲃⲏⲣ .
 ⲱⲧⲉⲧⲏⲣⲱⲩⲉ ⲛⲧⲱⲧⲏ ⲛⲉⲧⲕⲁⲙⲁⲛⲁⲩ
 ⲉⲧⲣⲉⲩⲉⲓ ⲉⲣⲟⲩⲛ ⲛⲧⲱⲩⲛ ⲛⲓⲛⲁ ⲉⲧⲟⲩⲁⲁⲃ
 ⲙⲓⲛⲟⲩⲧⲉ ⲙⲁⲩⲱ ⲛⲧⲉⲧⲏⲣⲱⲩⲱⲩ ⲙⲓⲛⲉⲩⲣⲓⲉ
 ⲉⲧⲉⲛⲧⲱⲧⲏ ⲛⲉ ⲙⲓⲛⲓⲛⲙⲱⲧⲏ ⲙⲓⲛⲉⲧⲏⲣⲓⲛⲙⲉⲩⲉ
 ⲛⲧⲁⲛⲟⲩⲁⲥⲓⲥ ⲙⲓⲛⲁⲛⲟⲥⲧⲟⲗⲟⲥ
 ⲩⲉⲛⲉⲧⲏⲁⲥⲱⲩⲱⲩ ⲙⲓⲛⲉⲣⲓⲉ ⲙⲓⲛⲟⲩⲧⲉ .

コプト語テキスト入力

Syntax:

Universal Dependencies
 での係り受け解析



```
<norm_group orig_group="αϣϣⲓⲛⲁⲩⲛⲁⲩ" norm_group="αϣϣⲓⲛⲁⲩⲛⲁⲩ">
  <norm xml:id="u1" pos="APST" lemma="α" func="aux" head="#u3" new_sent="true"
    norm="α">
    α
  </norm>
  <norm xml:id="u2" pos="PPERS" lemma="ρⲩⲱⲩ" func="nsubj" head="#u3" orig="q" n
    <hi rend="red">
    ϣ
  </hi>
  <norm>
  <multiword multiword="ϣⲓⲛⲁⲩⲛⲁⲩ">
  <norm xml:id="u3" pos="V" lemma="ϣⲓⲛⲁⲩⲛⲁⲩ" func="root" orig="ϣⲓⲛⲁⲩⲛⲁⲩ" norm="ϣⲓⲛⲁⲩⲛⲁⲩ">
    ϣⲓⲛⲁⲩⲛⲁⲩ
  </norm>
</norm_group>
```

Entities:

エンティティ認識



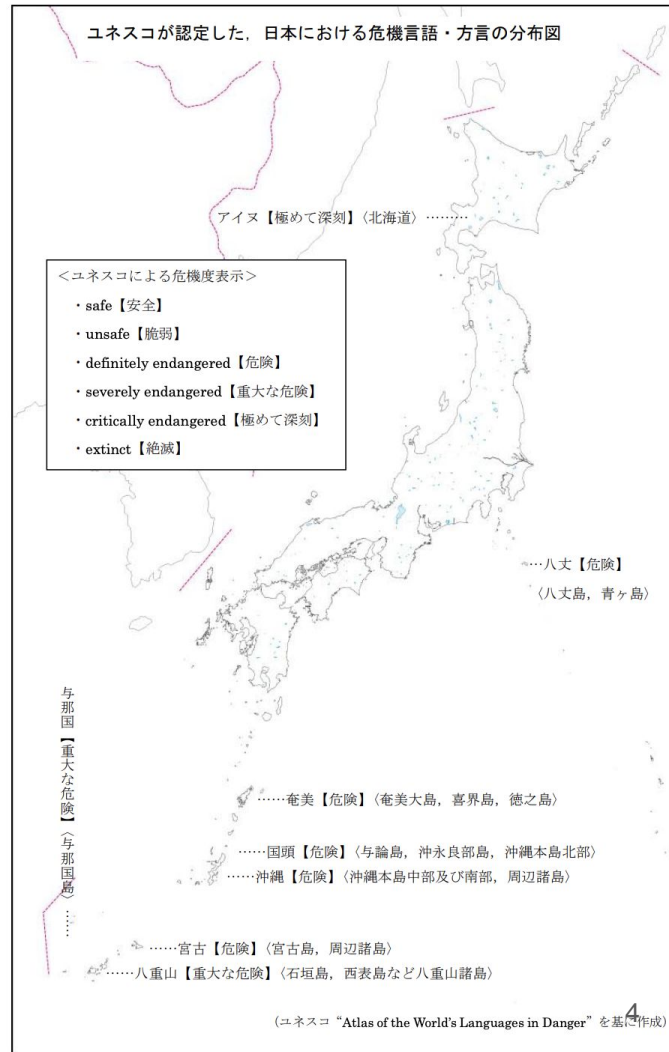
α ϣ ϣⲓⲛⲁⲩⲛⲁⲩ ϣⲛⲥ ϣⲛⲥ ⲓⲛⲁⲩⲉⲗⲟⲥ ⲙⲓⲛⲧⲱⲃⲏⲣ . ⲱⲧⲉⲧⲏⲣⲱⲩⲉ ⲛⲧⲱⲧⲏ
 ⲛ ⲉⲧ ⲕⲁ ⲙⲁⲛ ⲙⲓ ⲛ ⲥⲁⲧⲁⲛⲁⲥ ⲉ ⲧⲣⲉ ⲩⲉⲓ ⲉⲣⲟⲩⲛ ⲛⲧⲱⲩⲛ ⲛⲓⲛⲁ ⲉⲧⲟⲩⲁⲁⲃ ⲙⲓ ⲛ ⲟⲩⲧⲉ . ⲁⲩⲱ ⲛ
 ⲧⲉⲧⲏⲣⲱⲩⲱⲩ ⲙⲓ ⲛ ⲉⲧ ⲉⲧⲉ ⲛⲧⲱⲧⲏ ⲛⲉ ⲙⲓⲛⲓⲛⲙⲱⲧⲏ ⲙⲓⲛⲉⲧⲏⲣⲓⲛⲙⲉⲩⲉ ⲛ
 ⲧⲁⲛⲟⲩⲁⲥⲓⲥ ⲙⲓ ⲛ ⲁⲛⲟⲥⲧⲟⲗⲟⲥ ⲩⲉ ⲛ ⲉⲧ ⲏⲁ ⲥⲱⲩⲱⲩ ⲙⲓ ⲛ ⲉⲣⲓⲉ ⲙⲓ ⲛ ⲟⲩⲧⲉ .

[Coptic NLP Service](#)

消滅の危機にある言語とは

文化庁「消滅の危機にある言語・方言」より→

- 消滅＝母語話者・学習者がいなくなること
- 日本の消滅危機言語
 - アイヌ語、八丈語、奄美語、国頭語、沖縄語、宮古語、八重山語、与那国語
 - 方言レベルでは、ほとんどの日本語方言（関西方言を除く）
- 世界では、約7000言語のうち、約2500言語（UNESCO 2009）
 - 母語話者も学習者もなくなった言語
 - タスマニア諸語、エフタル語、ピクト語、クリミア・ゴート語など
 - 母語話者がいなくなったが学習者がいる言語（古典言語含む）
 - ラテン語、シュメール語、アッカド語、ウラルトゥ語、古典日本語など
 - そこからさらに復興して母語話者が生じた言語
 - ヘブライ語、サンスクリット語、コーンウォール語、マン島語、コプト語など
 - 危機的な状況から母語話者・学習者ともに増えた言語
 - ハワイ語、マオリ語など



消滅の危機にある言語・方言

日本の言語・方言の中には、消滅の危機にあるものがあります。

それには、ユネスコ（国連教育科学文化機関）が平成21年2月に発表した“Atlas of the World's Languages in Danger”（第3版）に掲載された8言語・方言や東日本大震災の被災地の方言が該当します。

文化庁では、消滅の危機にある言語・方言の実態や保存・継承の取組状況に関する調査研究をはじめ保存・継承に資する様々な取組を行っています。

“Atlas of the World's Languages in Danger”（第3版）には、世界で約2,500に上る言語が消滅の危機にあるとして掲載されています。日本国内では、8言語・方言が消滅の危機にあるとされており、掲載されている8言語・方言とそれぞれの危機の度合いは次のとおりです。

【極めて深刻】 アイヌ語

【重大な危機】 やえやま 八重山語（八重山方言）， よなぐに 与那国語（与那国方言）

【危険】 はちじょう 八丈語（八丈方言）， あまみ 奄美語（奄美方言）， くにながみ 国頭語（国頭方言）， おきなわ 沖縄語（沖縄方言）， みやこ 宮古語（宮古方言）

※ユネスコでは「言語」と「方言」を区別せず、全て「言語」で統一しています。

沖縄「方言」、宮古「方言」ではないのか？ →世界的に見たら、日本語と沖縄語と宮古語は別言語

- スペイン語の「月」 luna
 - イタリア語の「月」 luna
 - フランス語の「月」 lune
- これらは祖語(ロマンス祖語 / 俗ラテン語)は同じだが別言語
-
- 日本語共通語の「月」 tsuki
 - 沖縄語那覇方言の「月」 chichi
 - 宮古語大神方言の「月」 kskks
- これらは祖語(日琉祖語)は同じだが別言語
-
- アイヌ語の「月」 kunnecup → 上の6言語とは完全に異なる言語

消滅危機言語継承・保全の問題点

- 政治・経済的な威信言語（日本の場合は日本語共通語）の使用域の拡大
←一番の要因
- 教育の少なさ（←→ニュージーランドやハワイの先進的な取り組み）
- 学習書・教科書・教室・勉強グループが少数

ここに、深層学習系の自然言語処理の技術を使えないか？

→例えば、ChatGPTにその言語を覚えてもらう・添削してもらう、など

ChatGPTの語学力

- ChatGPTは、琉球諸語やアイヌ語をどれだけ理解するか
- 右の沖縄語に関するChatGPTの回答のほとんどが誤り
- ←そもそも沖縄語でデジタルテキストとしてオンラインに流通している言語資源が少ないのが原因か

so

沖縄語で「こんにちは」はなんと言いますか？



沖縄語で「こんにちは」は「はいさい」(haisai)と言います。

so

沖縄語で「お腹がすいた」はなんと言いますか？



沖縄語で「お腹がすいた」は「ちゅらめー (churaamee)」と言います。

so

沖縄語で「今日は気分が悪い」はなんと言いますか？



沖縄語で「今日は気分が悪い」は「ちょーひとしゅん (chou hitoshun)」と言います。

so

沖縄語で「NLP2023に来たけど、有名な先生がいっぱいて緊張する。」はなんと言いますか？



沖縄語では、次のように言います。「NLP2023に来たけど、有名な先生がいっぱいおって糸張りぐわさるー。」（「NLP2023 ni kita kedo, yuumeina sensei ga ippai ottote itobari guwasaru.」）

消滅危機言語のデジタルテキストリソースを増やす

- 国立アイヌ博物館「[アイヌ語アーカイブ](#)」
- 国立国語研究所「[アイヌ語口承文芸コーパス—音声・グロス付き—](#)」
- 国立国語研究所「[ことばのミュージアム](#)」
- 国立国語研究所「[日本語諸方言コーパス \(COJADS\)](#)」
- 国立国語研究所「[危機言語データベース](#)」
- えらぶむに「[しまむに宝箱](#)」（国頭語沖永良部島諸方言）
- ウィキメディア・インキュベーター「[エキペディヤ](#)」（沖縄語）
- 読谷村「[読谷村史編集室](#)」（沖縄語）
- 沖縄県「[しまくとうばナビ](#)」（沖縄県内の琉球諸語）

などなど・・・でも（多分）現状はNLPするには足りない

国立国語研究所「アイヌ語口承文芸コーパス—音声・グロス付き—」

アイヌ語口承文芸コーパス 音声・グロス付き

このコーパスについて ユーザガイド トップ English Version

作品別 全文検索 ダウンロード

作品

パンペン鬼の手から逃れる

K7708241UP 沙流 ウエベケレ

悪い赤犬に毒を飲ませた少女

K7708242UP 沙流 ウエベケレ

笛の神様に育てられた少年

K7803231UP 沙流 ウエベケレ

村長の奥さんの墓を掘り返した貧乏人

K7803232UP 沙流 ウエベケレ

ぶどうづるの輪がトバットウミを退けてくれた話

K7803233UP 沙流 ウエベケレ

ケムカチカッポになった女

K7807151KY 沙流 カムイユカナ

チセコリカムイが育ててくれた

K7807151UP 沙流 ウエベケレ

火の女神が地の果ての魔神と戦った

K7807152KY 沙流 カムイユカナ

ウエベケレ：パンペン鬼の手から逃れる

あらすじ 一話の音声のダウンロード

001

パノ ニシバ イサム ニシバ ア= ネ ヒネ アン =アン(アナン) ペ ネ ヒケ,

pakno nispa isam nispa a= ne hine an =an pe ne hike,

pak-no nispa isam nispa a= ne hine an =an pe ne hike,

～まで-副 ニシバ ない ニシバ 4.他主= ～である して ある.単 =4.自主 の ～である したが

私は並ぶもののない長者であった。

002

エキムネ =アン コロ ユク ネ チキ カムイ ネ チキ ヌウェ ア= コアン.

ekimne =an kor yuk ne ciki kamuy ne ciki nuwe a= koan.

ekimne =an kor yuk ne ciki kamuy ne ciki nuwe a= ko-an

山に行く =4.自主 すると シカ ～である したら クマ ～である したら たくさんの獲物 4.他主= ～に対して-ある.単

山へ行けばシカでもクマでもたくさん捕ってくる。

003

ウトウル タ アトウイ テササム タ サン =アン(サナン) ワ ペライ =アン ネ ヤ

uturu ta atuy teksam ta san =an wa peray =an ne ya

utur-u ta atuy teksam ta sa-n =an wa peray =an ne ya

間-への ～に 海 そば ～に 前-自形成.単 =4.自主 して 釣りをする =4.自主 ～である か

その合間に海岸に下りては釣りをしたりして、

004

レボッチコイキヤ ヤオッチコイキヤ ヌウェ ア= コアン ペ ネ コロ,

rebottchikoya yaottchikoya nuwe a= koan pe ne kor,

rebottchikoya yaottchikoya nuwe a= koan pe ne kor,

◀ ▶ ⏮ ⏭ ⏯ ▶▶ 🔊

簡易表示 標準表示 詳細表示

アイヌ語口承文芸コーパス 音声・グロス付き Copyright © 2016-2021 National Institute for Japanese Language and Linguistics. All rights reserved.

読谷村史編集室

ホーム > 読谷村しまくとぅば「むんがたい」 > 勝連バーマー

勝連バーマー

かっちんばーまー

話者 伊波蒲戸 (1894・M27)

地域 伊良皆

時間 03 : 03

かっちん
勝連バーマーは、な一、あれ一、^{んかし}昔えよ、^{すい}首里ぬ
^{なかぐしくどうん}中 城御殿ぬ^{めー}前なかい^{いゆぐむい}魚小堀んち^あ在たしえ。^{いけふ}あぬ池掘
^{おおむかし}いによ、大 昔てーうれ一、あぬ池掘い^{いけふ}にありすたんで
い。

^{かっちん}「勝連ぬ夫や時間^{ぶー}おちやー^{じかの}遅りてい、^{うく}あんしえーな
^{かっちん}く ま とぅー

首里の中城御殿の前に魚小堀（^{りゅうたんいけ}龍 潭 池）というの
があるでしょう。勝連バーマーの話はね、あの池を造
ったときの話だよ、大昔のこと。

「勝連の人夫はいつも時間に遅れて、こんなことで
はいけない」と役人が言うと、「いやもう、勝連はこ



ウィキメディア・インキュベーター「オキペディア」



オキペディヤんかい、いめんしえーびり！

くれー、^{うちなーぐち} ^{すい} ^{くとうば} ぬオキペディヤやいびーん。
 オキペディヤー誰やていん 作 ^{ちやう} ^く いゆーするいちゃんだ百 物書物やいびーん。
 近頃一、^{ちかごろ} ^{ふん} ^ち ^じ 2,549 本ぬ記事ぬあいびーん。
 キャッシュ ^{ちゆー} ^{あーだ} ^{てい} ^{きぐとう} 清みーんじ ぐぬ間一ぬ出い来 事 じ

- [いめんしえーびり！](#)
- [ずー考え](#)
- [玄関](#)
- [免責事項](#)
- [助け](#)
- [Help for non-Okinawan-speakers](#)
- [沖縄口ぬ漢字表](#)
- [沖縄口うとーていぬ表記法方針](#)


とらめー 探 ゆん

ちゆくゆん / 作成

ち じ
記事



玉陵 (たまうどうん)、^{うちなーちん} ^{うどうん} ^{ていー} 沖縄県ぬ 陵ぬ一ちや
 ん。「玉」や「玉黄金」んでい言る言葉んある通
 い、「宝物ぬ如、美らさん」でいぬちむえーぬあ
 ん。^{すいでしく} ^{いの} 首里城ぬ西んかいあん。去じやる^{なま} 沖縄戦ぬば
 す、^{すいでしく} ^{あー} ^{ぐとう} ^{くー} 首里城とうぬめ如、壊さてい、今あるむぬ
 お一、戦 後、造 い直ちええるむぬやん。...[な一ふいん見じゆん](#)

むぬがたい
うむさる物語

龍とう 鶏

る一 にわとうい
 龍とうよ、^{にわとうい} 鶏 でいーねーあぬハートウヤーて、うりが話やしがよ。うぬ
 龍ぬ耳ぬ中んかいよ、あぬンカジよ、ムカデ、ありがへーりんちえーるふ
 ーじ、^{る一} ^{みみ} ^{なか} 龍ぬ耳ぬ中んかい。へーりんちやぐとう、な一痛りふしがらん、^{あと} ^{とう} 後
 お、あれ一^{くー} 噛いんすしえーや痛りふしがらん。な一逃ぎーん、^{うん} ^ま 其処から出
 じーぬ所 お道ばっペーていねーらんよ、あまなくまん全部 噛ちきやーち。
 あんさぐとう、^{くんど} ^{いしや} ^{やー} 今度おまた医者ぬ家かい、^{にんじんめ} ^ば な一人間んかい化きてい行だえ

型

琉球諸語とアイヌ語ではどちらの方がNLPしやすいか

- おそらく、アイヌ語
 - アイヌ語の表記は、ラテン文字、修正カタカナ、キリル文字で比較的安定
 - 琉球諸語は、表記法が乱立している
 - 例：沖縄語
 - 『おもろさうし』などの伝統的かな表記法
 - しまくとぅば正書法検討委員会（しまくとぅば県民運動）
 - 琉球諸語統一的表記法（小川+2015）
 - 新沖縄文字（船津好明氏）
 - 漢字を混ぜるか、ひらがなののかカタカナなのかもバラバラ
 - ……調べると20以上の変種があった；Miyagawa and Carlino 2023→
 - 他の琉球諸語は表記法はばらばら、かつ、使えるテキストが少ない

[ʔɰa:]  [ɰa:]
 “pig” “I”

ʔワー



ワー

ッワー

‘わー



わー

あー



ワー

ッワー



ワー

ウァー



ワー

対訳コーパス構築

「[国立アイヌ民族博物館](#)
[アイヌ語アーカイブ](#)」からスクレイピングしたほか、アイヌ語の教科書のPDFからコピーしたテキストをChatGPT (GPT-3.5 / -4.0) を用いてテキスト-和訳の表形式に

so

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。

表現

◎ ワ/wa ～(し) て、～(し) た

「私が〇〇〇して、私が〇〇〇した」のように、

出来事が続いて起こることを表す単語に「ワ/wa」や「イネ /ine」を用います。

* 例文

私は 走ってき て 私は 疲れた

ku=hoyupu wa ku=sinki

ク ホユプ ワ ク シンキ

私は 食べ て 私は 寝た

ku=e wa ku=mokor

ク エ ワ ク モコロ

私は 料理をし て 私は 食事した

ku=suke wa ku=ipe

ク スケ ワ ク イペ

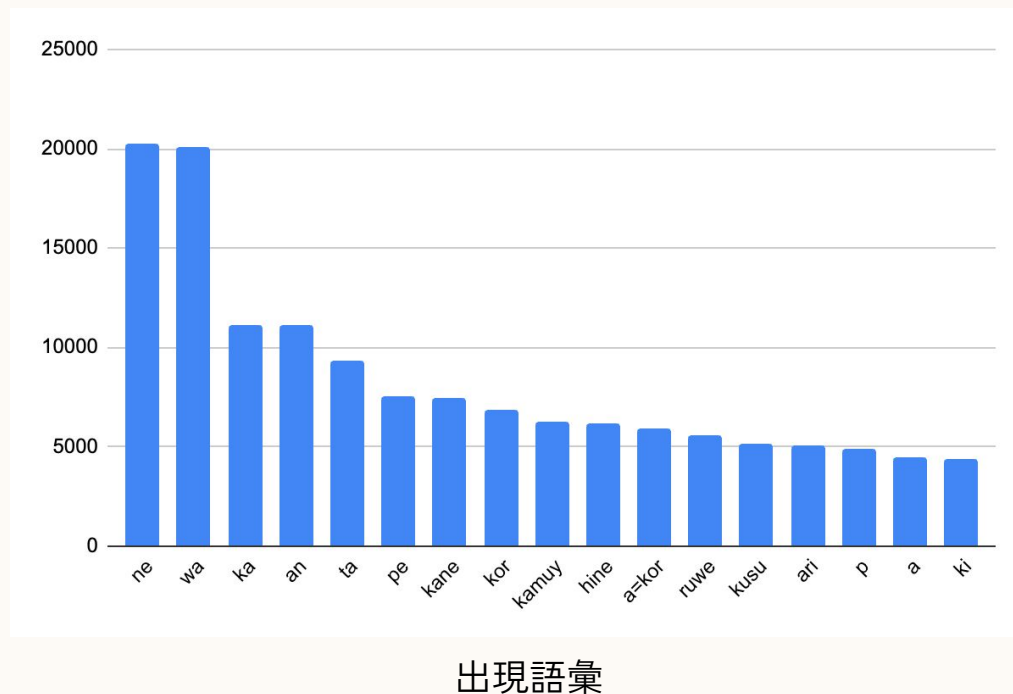


アイヌ語	日本語訳
ku=hoyupu wa ku=sinki	私は走ってきて、私は疲れた
ku=e wa ku=mokor	私は食べて、私は寝た
ku=suke wa ku=ipe	私は料理をして、私は食事した



アイヌ語-日本語自動翻訳モデルの開発：方法

- HuggingFace Transformers,
MarianMT
 (BartForConditionalGenerationの
 修正版) を使用
- 「[国立アイヌ民族博物館アイヌ語
 アーカイブ](#)」などのアイヌ語資料
 のテキストと和訳を抽出したアイ
 ヌ語の文字数2,488,264, 単語数
459,827(重複を除き18,156)と日本語
 対訳の教師データで深層学習



実験結果 ainu-2-japanese (HuggingFaceで公開中)

訓練済みモデル

- SacreBLEU: 10.448
- 試しに「アイヌ語口承文芸コーパス ー音声・グロス付き」(教師データにない)の文をHuggingFace上で入力すると、ほぼ完璧に和訳されることがあった

010



キラ	=アン	カ	エアイカプ	ペ	ネ	クス,
kira	=an	ka	eaykap	pe	ne	kusu,
kira	=an	ka	e-aykap	pe	ne	kusu
逃げる	=4.自主	も	〜について-できない	の	〜である	ので

逃げることもできないので、

Hosted inference API ⓘ

Translation

kira=an ka eaykap pe ne kusu,

Compute

⌘+Enter

1.6

Computation time on Intel Xeon 3rd Gen Scalable cpu: 1.402 s

私は逃げることもできないので



</> JSON Output

Maximize

沖縄語の語彙資源・コーパス開発

オンライン版沖縄語辞典

宮川・金山・田口・當山 2023

沖縄語UDコーパス

国立国語研究所
デジタルアーカイブ
NINJAL Digital Archive

言語資源(Resources)
利用方法(Guide)
リンク(Links)
お問い合わせ(Inquiry)

じゅーじゅー (オンライン版『沖縄語辞典』 OMOLO: ID 14549)

漢字かな混じり表記	重々
ひらがな表記	じゅーじゅー
IPA	/dʒuːdʒuː/
意義1	重重。重ね重ね。
例文1-1	重々(じゅーじゅー) 我(わー) が 悪(わ) っ さん「かえすがえすわしが悪かった」
タイトル	じゅーじゅー (オンライン版『沖縄語辞典』 OMOLO: ID 14549)



190-0014 東京都立川市緑町10-2
大学共同利用機関法人 人間文化研究機構 国立国語研究所
060-8943-2222

沖縄語の Universal Dependencies ツリーバンクコーパスの構築

宮川 創¹ 金山 博² 田口 智大³ 當山 奈那⁴
¹ 国立国語研究所 研究系 ² 日本アイ・ビー・エム株式会社 東京基礎研究所
³ University of Notre Dame ⁴ 琉球大学 人文社会学部
so-miyagawa@ninjal.ac.jp hkana@jp.ibm.com
ctaguchi@nd.edu tohyama@hs.u-ryukyu.ac.jp

概要

沖縄語は日琉語族のうちの琉球諸語に属する言語で、消滅の危機に瀕しているが、ローリソース言語の中では比較的テキスト資料は多い。そこで、国立国語研究所の語彙資源『沖縄語辞典』および、談話テキスト資源『方言談話資料』をもとに、沖縄語の Universal Dependencies ツリーバンクを構築している。本稿ではその過程で検討した、表記法、語分割、品詞や係り受けタグの選択について報告する。

国語研究所の『沖縄語辞典』[3]に基づき、コーパスのテキストは同研究所の『方言談話資料』の第6・8・9・10巻[4, 5, 6, 7]に収録されている沖縄語首里方言の談話テキスト資料を用いている。いずれの資料も、国立国語研究所のリポジトリ上でCC BY 4.0ライセンスで配布されている。琉球諸語のタグ付きコーパスは現在存在せず、本稿が述べる沖縄語 UD ツリーバンクコーパスが琉球諸語で初の言語学的タグ付きコーパスとなる見込みである。

2 沖縄語の言語学的特徴

1 はじめに

沖縄語は、日琉語族のうちの琉球諸語の中の北琉球諸語に属し、沖縄島中南部を中心に使用されてきた言語である。しかし、その使用は急速に衰退しており、UNESCOによって、消滅の危機にある言語と認定された[1]。それにもかかわらず、沖縄語の中でも、琉球国時代の政治の中心地であった首里の言語は、日琉諸語のなかでは、日本語のつぎに古い書記記録をもち、書かれたテキストは比較的多い。Universal Dependencies [2] (以下、UD) は、世界の諸言語の係り受け構造を記述するオープンコミュニティのプロジェクトである。世界中の消滅危機言語や少数言語や古典言語のツリーバンク開発も盛んであり、UDの最新版 ver. 2.11 は、世界の138の言語のツリーバンクを有している。消滅の危機に瀕する言語ならびに少数言語としては、ブラジルの少数民族の言語や、オーストラリアの先住民の言語などが含まれている。しかし、日本語派の諸語と琉球語派の諸語からなる日琉語族の諸言語で、UD ツ

沖縄語は日琉語族に属しており、日本語とは親縁関係にあり、統語論・形態論の特徴もその多くを日本語と共有している。例えば、次の図1は、「私は(わんねー)彼(あり)に(んかい)酒(さき)を与えた(くいたん)」という意味の例文であるが、基本語順はSOVであり、格助詞(んかい)が名詞の後に来ている。この図では、例文の上には単語の依存先(HEAD)を表す矢印と5節で述べるそれぞれの依存/係り受け関係(DEPREL)タグが、例文の下には単語毎に4節で述べる Universal Part-of-Speech (UPOS) とグロス(語釈)が示されている。

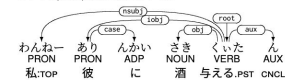


図1 「私は彼に酒を与えた」を意味する例文¹⁾

形態素同士の融合が、沖縄語では日本語よりも頻繁に見られる。例えば、図1の「わんねー」(代名詞

JSON-LD / RDF, TEI XMLでも配布予定 (夏までに公開)

国語研『方言談話資料』を基に開発中

国立国語研究所デジタルアーカイブNINDA

- デジタルアーカイブ専用のCMSのOmeka S
- Linked Open Data
(RDFでメタデータを出力)
- IIF (国際的画像相互運用枠組み)
- 映像・音声・語彙・コーパス



国立国語研究所 デジタルアーカイブ NINDA 言語資源(Resources) 利用方法(Guide) リンク(Links) お問い合わせ(Inquiry) 検索

方言録音資料シリーズ1「鹿児島市方言」1965年

MORE INFORMATION

00:00 00:00

方言録音資料シリーズ1「鹿児島市方言」1965年

タイトル	方言録音資料シリーズ1「鹿児島市方言」1965年
権利管理/Rights	CC0
タイプ	方言音声
作成者/Creator	国立国語研究所
日付	1965年

国語研「危機言語データベース」

- 琉球諸語，日本語方言の音声つき語彙，例文，談話テキスト（拡充予定）

日本の危機言語
 日本の消滅危機方言の音声データを紹介するサイト

TOP イベント データ

語彙データ

単語検索 ②

入力例

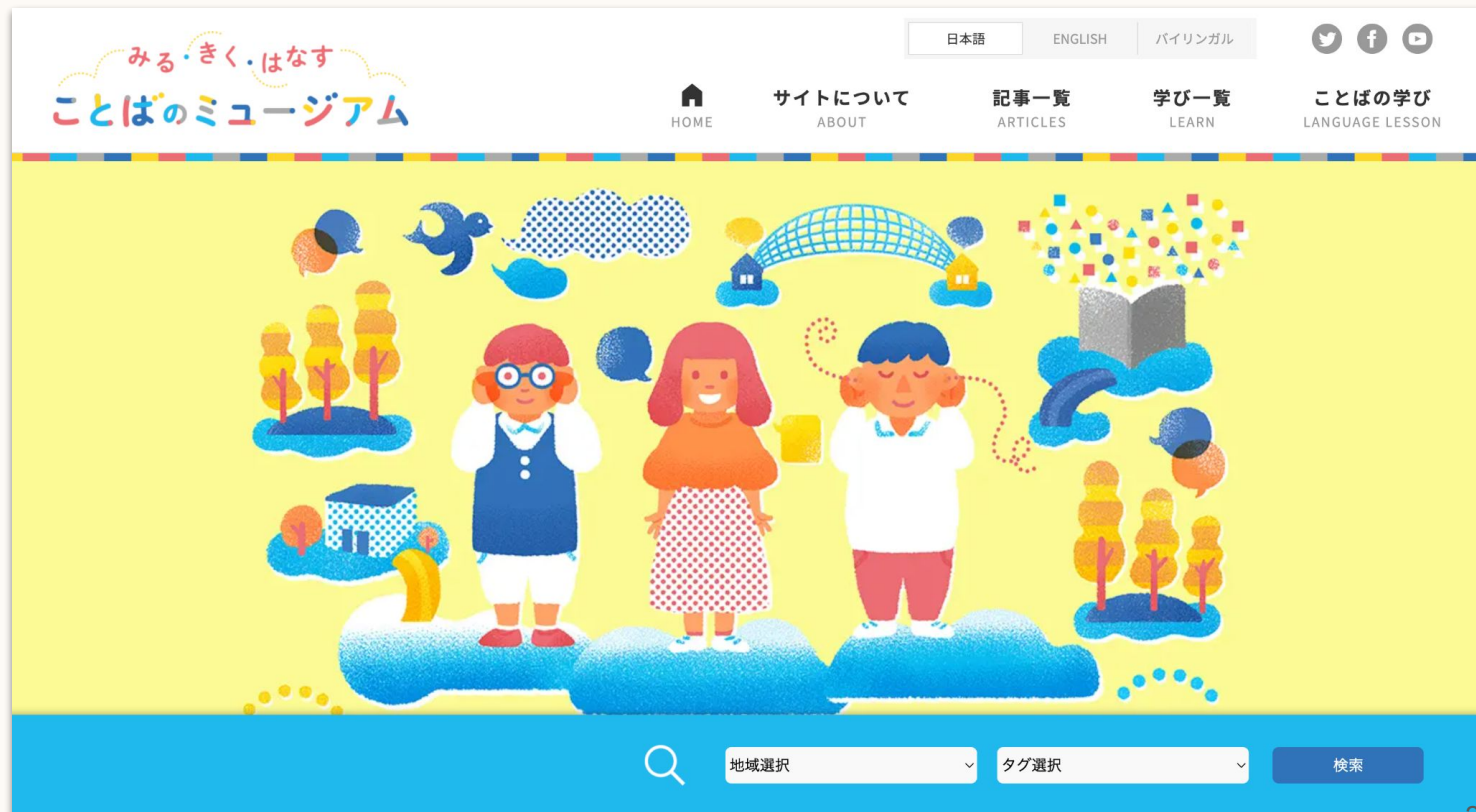
☐ 正規表現

地域・地点絞り込み ②

☐ 全選択	☐ 池田町西角間
☐ 福井県	☐ 三根 ☐ 末吉
☐ 八丈島	☐ 甕島里
☐ 鹿児島県	☐ 喜界 ☐ 荒木 ☐ 阿伝 ☐ 小野津 ☐ 上嘉鉄 ☐ 湾
☐ 喜界	☐ 上平川 ☐ 国頭 ☐ 芦清良
☐ 沖永良部島	☐ 東区
☐ 与論	☐ 伊平屋村田名 ☐ 首里 ☐ 本部瀬底島
☐ 沖縄	☐ 砂川 ☐ 池間西原 ☐ 城辺仲原 ☐ 水納
☐ 宮古	☐ 川平
☐ 八重山	☐ 祖納
☐ 与那国	

国語研「ことばのミュージアム」

- 日本の危機言語に関する記事や教材など



まとめ

- 日本の消滅危機にある言語のために深層学習系NLPの技術を使いたい
- しかし、メジャーな言語に比べて、危機言語のインターネット上のデジタルテキストは極端に少ない
 - そもそもデジタルテキストで書き起こされたorボーンデジタルな資料が少ない
 - それに、表記が安定していない
- アイヌ語では、今現在インターネットにある資料を使っただけで、まあまあの自動翻訳モデルができた
- → **琉球諸語や日本語の危機方言でもデジタル言語資源を増やしていくべき**