

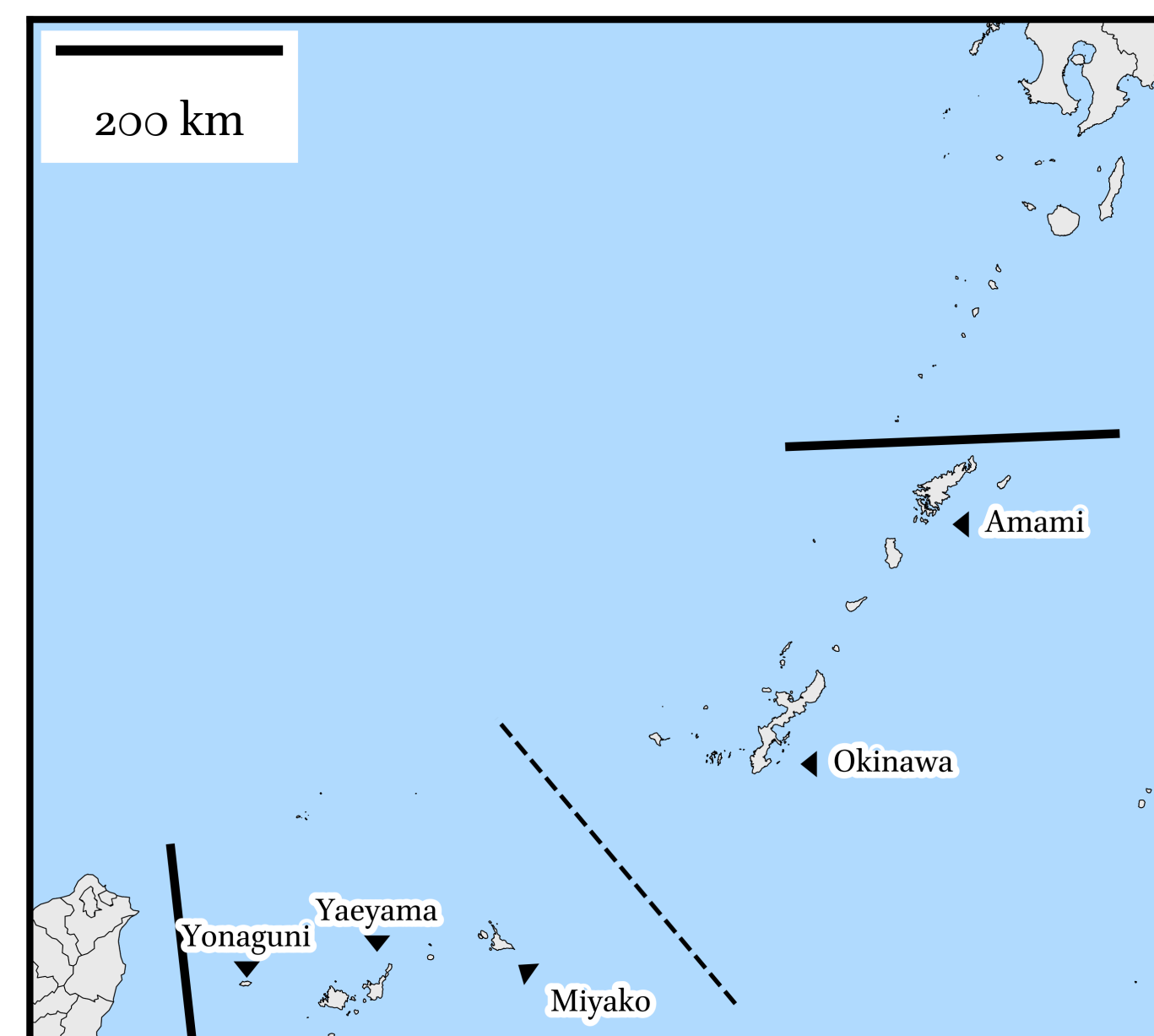
# Okinawan Lexicography in TEI: Challenges for Multiple Writing Systems

1 So Miyagawa, 2 Kanji Kato, 3 Miho Zlazli, 4 Seira Machida, and 5 Salvatore Carlino

1: National Institute for Japanese and Linguistics (NINJAL), Japan; 2: Tokyo University of Foreign Studies/JSPS/NINJAL, Japan; 3: SOAS University of London, UK; 4: University of Hawai'i at Hilo, US; 5: Kyushu University/Hitotsubashi University, Japan

## Okinawan and Multiple Spelling Problems

- Okinawan
  - ▶ Classified as one of the Northern Ryukyuan languages in the Japonic language family
  - ▶ Primarily spoken in the south and central parts of the Okinawa Island
  - ▶ Official lingua franca of the Ryukyu Kingdom and a literary vehicle, e.g., the Omoro Soshi poetry collection
  - ▶ Currently an endangered language
- Various written systems used
  - ▶ A combination of Kanji logograms and Hiragana syllabary with archaic spellings
  - ▶ Modern spelling variations to approximate actual pronunciation
  - ▶ Pure Katakana syllabary (e.g., Bettelheim's Bible translation)
  - ▶ Latin alphabet (mostly by linguists)
  - ▶ Pure Hiragana (popular)



```
<entry>
  <cit>
    <bibl>『沖縄語辞典』国立国語研究所資料集5第9刷（2009），p.100</bibl>
  </cit>
  <form>
    <orth xml:lang="ryu-Hira">あびーぐいー</orth>
    <orth xml:lang="ryu-Jpan" n="1">叫び一声</orth>
    <orth xml:lang="ryu-Jpan" n="2">叫声</orth>
    <orth xml:lang="ryu-Latn" n="1">?abiigwii</orth>
    <orth xml:lang="ryu-Latn" n="2">abiigwii</orth>
    <pron notation="ipa">?abiigwii</pron>
    <pron notation="accent">0</pron>
  </form>
  <gramGrp>
    <pos>NOUN</pos>
    <subc>名</subc>
  </gramGrp>
  <sense xml:lang="jp-Jpan" n="1">
    <def>叫び声。kaamakara~nu cikariiN.遠くから叫び声が聞こえる。</def>
  </sense>
  <usg></usg>
</entry>
```

## Digitization of NINJAL's Okinawan Dictionary

- The *Okinawago Jiten* (Okinawan Dictionary; OD): published by the National Institute for Japanese Language and Linguistics (NINJAL) in 1963 and revised in 2001; the Latin alphabet for each lexeme
- Addition of the possible writing forms listed above to the data in CSV format
- Conversion of the CSV into TEI XML using Python
- Solution of the multiple writing forms with <orth> tags with corresponding writing systems in @xml:lang attribute following BCP 47
  - ▶ E.g., xml:lang="ryu-Hira" for Okinawan words written in Hiragana
- International Phonetic Alphabet (IPA) and the accent type to make the pronunciation clearer with the <pron> tags

## User-friendly dictionary website

Using XSLT, we transformed this TEI file into a static webpage with a user-friendly GUI, as shown in the figure below. It is anticipated that this digitization of OD and its publication under the open license will benefit key stakeholders, such as Okinawan heritage learners and worldwide Okinawan learners, being the largest Okinawan dictionary available online.

	A	B	C	D	E	F	G	H	I	J
辞書	見出し語	アクセント型	品詞	文語などの種別	補足	意味 1.	意味 2.	意味 3.	意味 4.	
1	101	?abu=cuN	①	自 =kaN,=ci		暑くて蒸し蒸しする。蒸すように暑い。				
52	101	?abui	①	名		あぶみ。				
53	101	?abuikuu	①	名		魚や餅などを焼く金網。				
54	101	?abu=juN	①	他=raN,=ti		あぶる。食物などを焼く。また、火にかざし暖める。muci~。餅を焼く。tiinusaci~。司危い。元来は?ukaasjaNという。?abunee tukuru.危い所。				
56	101	?abusi	①	名		あぜ。田のあぜ。				
57	101	?abusibaree	①	名		あぜ祓いの意。旧暦4月に行なう田の祭り。仕事を休み、ごちそうを作っ。日を				
58	101	?abusimaku]ra	①	名	文	稲の穂がみのり。?abusi(あぜ)を枕にすること。豊作を形容している語。				
59	101	?abusimici	①	名		あぜ道。				
60	101	?aca	①	名		あした。あす。				
61	101	?aca]asa	①	名		あしたの朝。				
62	101	?acagai	①	名		時期が終わること。				
63	101	?acaga=juN	①	自 =raN,=ti		1.時期が去る。時期2.飽きが来る。仕事・人などに対する熱がさめる。tageeni ?i				
64	101	?aca]juru	①	名		あしたの夜。				
65	101	?aca]jusaNdi	①	名		あしたの晩。				
66	101	?aca]sutumiti	①	名		あしたの朝。				
67	101	?aci	①	名	文、古	1.[文]秋。秋という季2.[古]収穫の時期。刈り入れ時。				
68	101	?aCibee	①	名		熱灰。まだ火気のある灰。~Nkai hwisja ?iQtaNnee.熱い灰に足を入れたように。				
69	101	?aCibii	①	名		柔らかい御飯。おかゆと御飯との中間ぐらいの子供・病人などに食べさせる御飯				
70	101	?aCibiiraci	①	名		はれものなどが熱をもって痛むこと。また、やけどが痛むこと。				
71	101	?aCiguni	①	名		暑い地方。熱帯地方。hwiiguniの対。				
72	101	?acihaisieutu	①	名		納きてしまうような仕事。きりのないようないやいやながらの仕事。~nu ?asihaisi				

The screenshot shows the website interface for the word '叫びゆん' (Abiyun). It displays the word in various scripts: Kanji (叫びゆん), Hiragana (あびゆん), and Latin Alphabet (abiyun). The entry includes the date 'Jan 1, 0001', a share button, and a list of three senses: 1. 叫ぶ。大声で呼ぶ。また、わめく。どなる。 (Abiyun), 2. 大声で泣く。泣き叫ぶ。 (Abiyun), and 3. ほえる。犬・猫・豚などが鳴く。 (Abiyun). The interface also shows a 'Conjugation type' section with 'Alphabet' and 'IPA' options, and a 'Sense' section with 'Sense 1', 'Sense 2', and 'Sense 3' buttons.