Building Okinawan Lexicon Resource for Language Reclamation/Revitalization and Natural Language Processing Tasks such as Universal Dependencies Treebanking

So Miyagawa
NINJAL, Tokyo
runa.uei@gmail.com

Kanji Kato
CODH, Tokyo
kanji_kato@nii.ac.jp

Miho Zlazli
SOAS, London
miho.zlazli@gmail.com

Salvatore Carlino
Daito Bunka University, Tokyo
nanajuu@gmail.com

Seira Machida
University of Hawaiʻi at Mānoa
machida.seira07@gmail.com

Abstract

The Open Multilingual Online Lexicon of Okinawan (OMOLO) project aims to create an accessible, user-friendly digital lexicon for the endangered Okinawan language using digital humanities tools and methodologies. The multilingual web application, available in Japanese, English, Portuguese, and Spanish, will benefit language learners, researchers, and the Okinawan community in Japan and diaspora countries such as the U.S., Brazil, and Peru. The project lays the foundation for an Okinawan UD Treebank, which will support computational analysis and the development of language technology tools such as parsers, machine translation systems, and speech recognition software. The OMOLO project demonstrates the potential of computational linguistics in preserving and revitalizing endangered languages and can serve as a blueprint for similar initiatives.

1 Introduction

This study introduces our ongoing project to create a learner-friendly dictionary of Okinawan, a Ryukyuan language, based on an existing printed dictionary.[1] We created two versions of the online dictionary, which are useful for the learning and revitalization of Okinawan.

Okinawan (*Uchinaaguchi*) is one of the indigenous Ryukyuan languages spoken in and around Okinawa Island in the Ryukyu Archipelago and across Okinawan diasporas worldwide. However, since mainly elderly speakers can speak it but not many younger speakers, according to (UNESCO, 2010), this language is endangered. However, various lexicographical works on the Okinawan language have been done so far. Among others, an Okinawan Shuri dialect speaker, Seibin Shimabukuro, created a 1,856-page manuscript of the *Okinawago Jiten* (Okinawan Dictionary) in 1951, with the headwords written in classical orthography with a *katakana* syllabary which did not faithfully represent the actual pronunciation. The National Institute for Japanese Language and Linguistics (NINJAL) extensively revised the headwords and other example sentences using a Latin alphabet supplemented with diacritical marks and some IPA faithful to the pronunciations in a unique way original to this dictionary. It published a revised version in 1963 (NINJAL, 1963). Its ninth version was digitized in XSLX format and published in NINJAL's repository under a CC BY 4.0 license in 2001.[2] Although helpful for researchers, it poses difficulties for language learners due to its use of alphabetic phonological notation and special supplementary symbols, requiring familiarity with both the language and the International Phonetic Alphabet (IPA). In Japan, using IPA-based letters and diacritical marks on the Latin alphabet is unknown to ordinary users usually. This study utilizes the digital *Okinawago Jiten* dataset to create a user-friendly multilingual web application, Open Multilingual Online Lexicon of Okinawan (OMOLO), in Japanese, English, Portuguese, and Spanish for language learners and contributes to language revitalization and Okinawans' "language reclamation" (Leonard, 2017) by collaborating with learners from Okinawan communities, including Okinawan diaspora communities in countries outside

[1] This work was supported by the NIHU Knowledge Co-creation Project "Digital Library for Humanities," the Institutional Core Research Projects "Co-creation of Research Infrastructure Through the Integration of Diverse Lexical Resources" and "Research on the Conservation of Endangered Languages," the JSPS Kakenhi Grant Number JP19H01265 and JP21K18376, and ROIS-DS-JOINT 2022 "LAJaR: Language Atlas of Japanese-Ryukyuan."

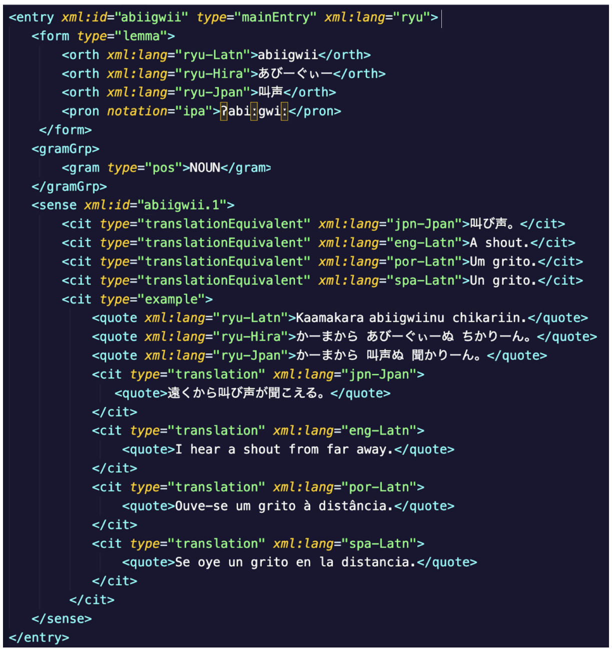
[2] https://mmsrv.ninjal.ac.jp/okinawago/, accessed on March 28, 2023.

of Japan such as the United States of America, Brazil, and Peru.

2 Orthographical and multi-lingual challenges in TEI Lex-0 and its visualization

Language revitalization has recently been flourishing in the Ryukyus. Although the Ryukyuan languages have been unwritten for a long time until recently, except for Okinawan, which has its classical literature and writing system, contemporary writing systems that are more conforming to the actual pronunciation have been established, most of which employ only phonetic *kana* characters. However, as a result of our survey of existing dictionaries, it was clarified that many Okinawan speakers prefer writing systems employing both phonetic *kana* and ideographic characters, i.e., Chinese (kanji) characters, and orthographies using only phonetic characters are not in line with speaker demand. Thus, there is a demand amongst learners for a *kanji-hiragana* combination in addition to *hiragana*-only text. If *Okinawago Jiten*, which until now has been written only in phonetic characters, were to be expressed in *kanji-kana* script, it could meet the demands of a larger number of learners. Therefore, we transcribed the headwords and example sentences in the *Okinawago Jiten* with our provisional orthography consisting of both *hiragana*-only and *kanji-hiragana* texts, which are designed to enable easy input with the default settings on ordinary computers or smartphones. The headwords and example sentences in the NINJAL's spreadsheet version of the *Okinawago Jiten* are written in alphanumeric characters in the ASCII range.

In this study, we conducted a survey of the existing Okinawan orthographies in eleven textbooks and dictionaries, such as Okinawa Prefectural Shimakutuba Orthography Council's (Shimakutuba Seishohō Kentō Iinkai, 2022), Nishioka et al. (2006), Uchima and Nohara (2006), Hanazono et al. (2020), Nakamatsu (1999), Fija (2015), Miyara (2021), Carlino (2022), an orthography for the Shuri dialect of Okinawan and another for the Tsuken dialect by Ogawa et al. (2015), and partial *katakana* renditions seen in the introductory chapter of NINJAL (1963). We created a database of these existing Okinawan orthographies and a Python program converting one orthography to another (see Miyagawa and Carlino, submitted). 14,549 headwords were converted into major orthographies. We chose the *hiragana* rendition of Nishioka et al. (2006)'s orthography as the standard for the headword but put the other orthographies in sub-layers including our original *kanji-hiragana* notation, in which the *hiragana* part is based on Nishioka et al., 2006. Using XSLT, we converted the data into TEI Lex-0,[3] a TEI XML subset for dictionary data (Fig. 1). TEI XML is a de facto standard of text mark-up in Digital Humanities. Currently, we are also translating the meanings of each word and example sentence written in Japanese into Portuguese, Spanish, and English for Okinawan diaspora communities in countries such as the United States of America, Brazil, and Peru. Each language and writing system is written according to BCP47[4] in the xml:lang attribute.



```
<entry xml:id="abiigwii" type="mainEntry" xml:lang="ryu">
   <form type="lemma">
       <orth xml:lang="ryu-Latn">abiigwii</orth>
       <orth xml:lang="ryu-Hira">あびーぐぃー</orth>
       <orth xml:lang="ryu-Jpan">叫声</orth>
       <pron notation="ipa">ʔabiːgwiː</pron>
    </form>
   <gramGrp>
       <gram type="pos">NOUN</gram>
   </gramGrp>
   <sense xml:id="abiigwii.1">
       <cit type="translationEquivalent" xml:lang="jpn-Jpan">叫び声。</cit>
       <cit type="translationEquivalent" xml:lang="eng-Latn">A shout.</cit>
       <cit type="translationEquivalent" xml:lang="por-Latn">Um grito.</cit>
       <cit type="translationEquivalent" xml:lang="spa-Latn">Un grito.</cit>
       <cit type="example">
           <quote xml:lang="ryu-Latn">Kaamakara abiigwiinu chikariin.</quote>
           <quote xml:lang="ryu-Hira">かーまから あびーぐぃーぬ ちかりーん。</quote>
           <quote xml:lang="ryu-Jpan">かーまから 叫声ぬ 聞かりーん。</quote>
           <cit type="translation" xml:lang="jpn-Jpan">
              <quote>遠くから叫び声が聞こえる。</quote>
           </cit>
           <cit type="translation" xml:lang="eng-Latn">
               <quote>I hear a shout from far away.</quote>
           </cit>
           <cit type="translation" xml:lang="por-Latn">
               <quote>Ouve-se um grito à distância.</quote>
           </cit>
           <cit type="translation" xml:lang="spa-Latn">
               <quote>Se oye un grito en la distancia.</quote>
           </cit>
        </cit>
   </sense>
</entry>
```

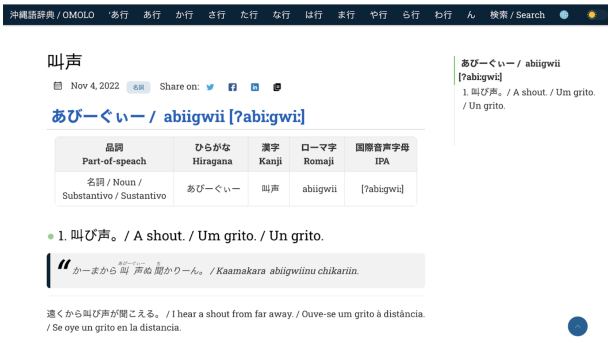
Figure 1: XML data compliant with TEI Lex-0

From this XML file, we created the prototype website using XSLT and Hugo, a static site generator developed in the Go language, with the theme Hugo Curious[5], which is searchable and has easy-to-read headwords on each page (Fig. 2).

[3] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html, accessed on March 28, 2023.

[4] See https://www.w3.org/International/articles/language-tags/index.en, accessed on March 28, 2023.

[5] https://github.com/vietanhdev/hugo-curious, accessed on March 28, 2023.



叫声

Nov 4, 2022

あびーぐぃー / abiigwii [?abi:gwi:]

品詞 Part-of-speach	ひらがな Hiragana	漢字 Kanji	ローマ字 Romaji	国際音声字母 IPA
名詞 / Noun / Substantivo / Sustantivo	あびーぐぃー	叫声	abiigwii	[?abi:gwi:]

- 1. 叫び声。/ A shout. / Um grito. / Un grito.

> かーまから叫声ぬ聞かりーん。 / *Kaamakara abiigwiinu chikariin.*

遠くから叫び声が聞こえる。 / I hear a shout from far away. / Ouve-se um grito à distância. / Se oye un grito en la distancia.

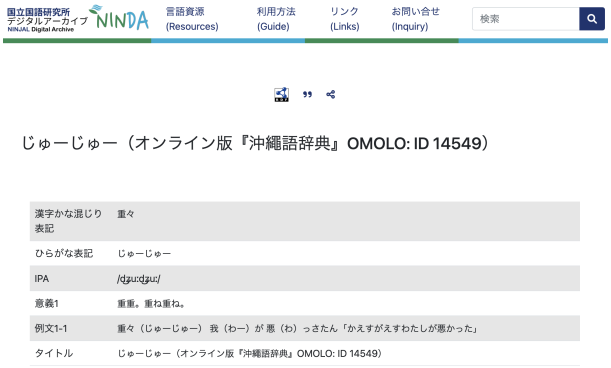
あびーぐぃー / abiigwii [?abi:gwi:]

1. 叫び声。/ A shout. / Um grito. / Un grito.

Figure 2: Example of visualization of a TEI Lex-0 dictionary entry of OMOLO in Hugo

3 Omeka S and Linked Open Data

In addition to the prototype website created using XSLT and Hugo, we are working on an alternative platform for presenting the Open Multilingual Online Lexicon of Okinawan (OMOLO) using Omeka S,[6] an open-source web publishing platform designed for sharing digital collections and creating media-rich online exhibits (Fig. 3). Omeka S is well-suited for digital humanities projects, offering a user-friendly interface and advanced features for organizing and presenting collections of various digital assets.

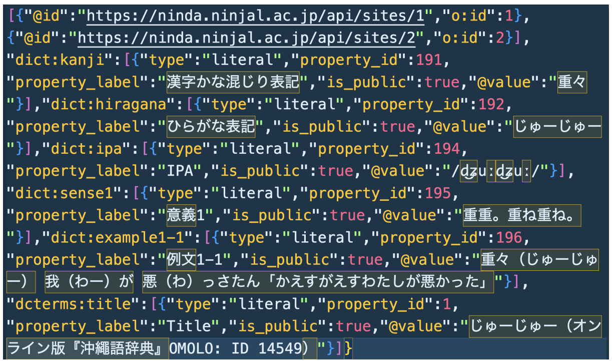


じゅーじゅー（オンライン版『沖縄語辞典』OMOLO: ID 14549）

漢字かな混じり表記	重々
ひらがな表記	じゅーじゅー
IPA	/dʑuːdʑuː/
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）っさたん「かえすがえすわたしが悪かった」
タイトル	じゅーじゅー（オンライン版『沖縄語辞典』OMOLO: ID 14549）

Figure 3: Omeka S visualization of the data

By utilizing the flexibility and extensibility of Omeka S, we plan to create a visually engaging and interactive experience for users to explore OMOLO. We import the TEI Lex-0 XML data into Omeka S, converting it into compatible metadata for items and item sets within the platform, providing JSON-LD data (Fig. 4) as Linked Open Data (LOD)/Resource Description Format (RDF). This enables us to present the lexicon entries in a more structured and organized manner, facilitating easy navigation and browsing for language learners and providing the data to other external services easily, following the standard of LOD/RDF.



```
[{"@id":"https://ninda.ninjal.ac.jp/api/sites/1","o:id":1},
{"@id":"https://ninda.ninjal.ac.jp/api/sites/2","o:id":2}],
"dict:kanji":[{"type":"literal","property_id":191,
"property_label":"漢字かな混じり表記","is_public":true,"@value":"重々
"}],"dict:hiragana":[{"type":"literal","property_id":192,
"property_label":"ひらがな表記","is_public":true,"@value":"じゅーじゅー
"}],"dict:ipa":[{"type":"literal","property_id":194,
"property_label":"IPA","is_public":true,"@value":"/dʑuːdʑuː/"}],
"dict:sense1":[{"type":"literal","property_id":195,
"property_label":"意義1","is_public":true,"@value":"重重。重ね重ね。
"}],"dict:example1-1":[{"type":"literal","property_id":196,
"property_label":"例文1-1","is_public":true,"@value":"重々（じゅーじゅ
ー） 我（わー）が 悪（わ）っさたん「かえすがえすわたしが悪かった」"}],
"dcterms:title":[{"type":"literal","property_id":1,
"property_label":"Title","is_public":true,"@value":"じゅーじゅー（オン
ライン版『沖縄語辞典』OMOLO: ID 14549）"}]}
```

Figure 4: JSON-LD output file of a lexicon entry following the LOD/RDF standard

Furthermore, Omeka S allows the incorporation of multimedia assets, such as audio recordings and images, which can be attached to individual lexicon entries. This feature will significantly enhance the learning experience, providing users access to native-speaker pronunciations and visual aids to support language acquisition.

4 Application of the data to Okinawan Universal Dependencies treebank

The data generated from the OMOLO project is to build an Okinawan Universal Dependencies Treebank (Fig. 5), which will be a valuable resource for researchers such as descriptive linguists and computational linguists working with the Okinawan language. Universal Dependencies (UD) is a framework for cross-linguistically consistent treebank annotation that aims to create a comprehensive and multilingual resource for natural language processing and linguistic research[7]. It was created by integrating three dependency grammar projects: (universal) Stanford dependencies (de Marneffe et al., 2006; de Marneffe and Manning, 2008; de Marneffe et al., 2014), Google universal part-of-speech tags (Petrov et al., 2012), and the Interset interlingua for morphosyntactic tagsets (Zeman, 2008). Currently, the latest version of UD is ver. 2.11, including 243 treebanks and 138 languages. More minority and endangered language treebanks such as Amazonian indigenous languages and Australian aboriginal lan-

[6] https://omeka.org/s/, accessed on March 28, 2023.

[7] https://universaldependencies.org/, accessed on March 28, 2023.

guages (Miyagawa et al., 2023). So far, there is one Japonic language, namely Standard Japanese, which has treebanks in UD. Okinawan UD (Miyagawa et al., 2023), using OMOLO's example sentences and other text corpora (NINJAL, 1978, 1985, 1986, 1987), will contribute to the diversity of UD so that we can execute more diverse typological research using UD.

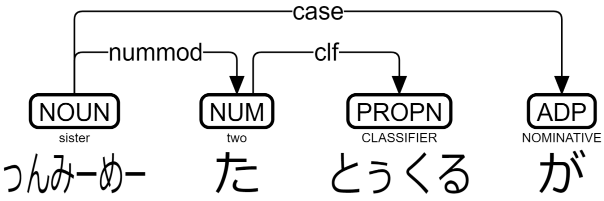


Figure 5: Visualization of *ʔmmiimee ta=tukuru=ga* (sister two=CLF=NOM) "two sisters" in Okinawan UD Treebank (Miyagawa et al., 2023) using deplacy (Yasuoka, 2020)

The Okinawan UD Treebank is constructed using the linguistic data available in the OMOLO, including the headwords, example sentences, and translations.[8] These data points are annotated in the CoNLL-U format following the UD guidelines to create dependency trees that capture the Okinawan syntactic structure.

After this phase, using this data as the training data, we train models included in the HuggingFace Transformers library[9], such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and T5 (Raffel et al., 2020), with this training data, and create a model to parse Okinawan text into CoNLL-U with dependency relation (DepRel) and universal parts-of-speech (UPOS) tags automatically. Thus, by developing the Okinawan UD Treebank, OMOLO, as its lexicon data and supplier of sample texts, supports the computational analysis of Okinawan, which can help develop language technology tools, such as parsers, machine translation systems, and speech recognition software. These tools can contribute to revitalizing the Okinawan language by making it more accessible to a broader audience and promoting its use in digital communication.

5 Conclusion

In conclusion, the OMOLO project aims to provide a user-friendly, multilingual, and accessible resource for language learners and researchers interested in Okinawan. By utilizing NLP and digital humanities methodologies, we have created a digital lexicon that can be used for language reclamation and revitalization efforts. Future developments will include creating an Omeka S version for a more visually engaging presentation with the output function of LOD/RDF and constructing an Okinawan UD Treebank to support computational analysis and language technology tools.

The foundation laid for the Okinawan UD Treebank is an essential aspect of this project, as it provides computational linguists with a valuable resource for working with Okinawan. The treebank facilitates the development of language technology tools, such as parsers, machine translation systems, and speech recognition software, which can significantly increase the accessibility and use of the Okinawan language in digital communication and contribute to its revitalization. At present, there are no large-language models (LLMs), such as GPT-4 (Bubeck et al., 2023), that can effectively handle the Okinawan language. This is primarily due to the limited availability of high-quality text corpora essential for training these models. However, the methodologies presented in this paper can create more comprehensive and accessible online resources for Okinawan.

As these Okinawan text corpora grow in size and quality, LLMs will be better equipped to learn and understand the language. With sufficient training data, future iterations of LLMs, like GPT-4, may be able to process and generate Okinawan text effectively, thereby contributing to the language's revitalization and making it more accessible to a broader audience. Additionally, the availability of high-quality Okinawan resources can help facilitate the development of advanced language technology tools, such as machine translation systems, parsers, and speech recognition software, further promoting the use and preservation of the Okinawan language in the digital age.

In summary, the OMOLO project showcases the immense potential of computational linguistics in preserving and revitalizing endangered languages. The methodologies and approaches employed in this project can serve as a blueprint for other similar initiatives, ultimately fostering linguistic diversity and preserving cultural heritage through the innovative use of digital technology.

[8] For more details, see Miyagawa et al. (2023).

[9] https://huggingface.co/docs/transformers/index, accessed on March 28, 2023.

References

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. http://arxiv.org/abs/2303.12712 Sparks of Artificial General Intelligence: Early experiments with GPT-4.

Salvatore Carlino. 2022. 'Nichiryū Shogo Online Jisho' no Shōkai [Japanese: Introduction to 'Japano-Ryukyuan Online Dictionary']. *Nihongo no Kenkū [Japanese: Studies of Japanese Language]*, 18:52–59.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. https://doi.org/10.18653/v1/N19-1423 BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Byron Fija. 2015. *Kimochi ga Tsutawaru! Okinawago Real Phrase Book: Pirin Paran Uchināguchi [Japanese: A Book of Real Phrases in Okinawan to Convey Your Feelings: Pirin Paran Uchinaaguchi]*. Kenkyūsha, Tokyo.

Satoru Hanazono, Satoshi Nishioka, Jō Nakahara, and Tomomasa Kuniyoshi. 2020. *Shokyū Okinawago [Japanese: Introductory Okinawan]*. Kenkyūsha, Tokyo.

Wesley Y. Leonard. 2017. `https://lddjournal.org/articles/10.25894/ldd146` Producing language reclamation by decolonising 'language'. In Wesley Y. Leonard and Haley De Korne, editors, *Language Documentation and Description*, volume 14, page 15–36. EL Publishing.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. `http://www.lrec-conf.org/proceedings/lrec2014/pdf/1062_Paper.pdf` Universal Stanford dependencies: A Cross-Linguistic Typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592, Reykjavik, Iceland. European Language Resources Association (ELRA).

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. `http://www.lrec-conf.org/proceedings/lrec2006/pdf/440_pdf.pdf` Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. https://aclanthology.org/W08-1301 The Stanford Typed Dependencies Representation. In *COLING 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, Manchester, UK. COLING 2008 Organizing Committee.

So Miyagawa and Salvatore Carlino. submitted. Database of Writing Systems and Orthographies for Okinawan Language: Toward Preservation of Okinawan Linguistic Cultural Heritage. *Proceedings of Japanese Association for Digital Humanities (JADH) 2023*.

Sō (So) Miyagawa, Hiroshi Kanayama, Chihiro Taguchi, and Nana Tōyama (Tohyama). 2023. Okinawago no Universal Dependencies Treebank Corpus no Kōchiku [Japanese: Building Okinawan Universal Dependencies Treebank Corpus]. In Gengo Shori Gakkai Jimukyoku, editor, *Gengo Shori Gakkai Dai-29 Kai Nenji Taikai (NLP2023) Happyō Ronbunshū [Japanese: Proceedings of NLP2023]*, pages 743–748.

Shinshō Miyara. 2021. *Uchināguchi Katsuyō Jiten [Japanese: Okinawan Practical Dictionary]*. National Institute for Japanese Language and Linguistics, Tachikawa.

Takeo Nakamatsu. 1999. *Okinawaken no Kotoba [JPN: Languages in Okinawa Prefecture]*. Okinawa Gengo Bunka Kenkyūjo [JPN: Institute of Okinawan Language and Culture], Naha.

NINJAL. 1963. *Okinawago Jiten [Japanese: Okinawan Dictionary]*. Zaimushō Insatsukyoku, Tokyo.

NINJAL. 1978. *Hōgen Danwa Shiryō 6 [Japanese: Dialect Discourse Resource 6]: Tottori, Ehime, Miyazaki, Okinawa*. National Insitute for Japanese Language and Linguistics, Tokyo.

NINJAL. 1985. *Hōgen Danwa Shiryō 8: Rōnensō to Jakunensō to no Kaiwa [Japanese: Dialect Discourse Resource 8: Dialogue between Elderly and Younger People]: Gumma, Nara, Tottori, Shimane, Ehime, Kōchi, Nagasaki, Okinawa*. National Insitute for Japanese Language and Linguistics, Tokyo.

NINJAL. 1986. *Hōgen Danwa Shiryō 9: Bamen Settei no Taiwa [Japanese: Dialect Discourse Resource 8: Dialogue with Specific Scenes]: Aomori, Gumma, Chiba, Niigata, Nagano, Shizuoka, Aichi, Fukui,* *Nara, Tottori, Shimane, Ehime, Kōchi, Nagasaki, Okinawa*. National Insitute for Japanese Language and Linguistics, Tokyo.

NINJAL. 1987. *Hōgen Danwa Shiryō 10: Bamen Settei no Taiwa, Sono 2 [Japanese: Dialect Discourse Resource 8: Dialogue with Specific Scenes II]: Aomori, Gumma, Chiba, Niigata, Nagano, Shizuoka, Aichi, Fukui, Nara, Tottori, Shimane, Ehime, Kōchi, Nagasaki, Okinawa*. National Insitute for Japanese Language and Linguistics, Tokyo.

Satoshi Nishioka, Jō Nakahara, Noriko Ikari, and Yumi Nakajima. 2006. *Okinawago no Nyūmon: Tanoshii Uchināguchi [Japanese: Introduction to Okinawan: Enjoyable Uchinaaguchi*, 2nd edition. Hakusuisha, Tokyo.

Shinji Ogawa, Hiromi Shigeno, Yūto Niinaga, Satomi Matayoshi, Nana Tōyama (Tohyama), Thomas Pellard, Yuka Hayashi, Michinori Shimoji, Kayoko Shimoji, Natsuko Nakagawa, Christopher Davis, Reiko Asō, and Masahiro Yamada. 2015. *Ryūkyū no Kotoba no Kakikata: Ryūkyū Shogo Tōitsuteki Hyōkihō [Japanese: Writing Ryukyuan Languages : A Unified Orthography of Ryukyuan Languages]*. Kuroshio Shuppan, Tokyo.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. `http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf` A Universal Part-of-Speech Tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. http://arxiv.org/abs/1910.10683 Exploring the limits of transfer learning with a unified text-to-text transformer.

Shimakutuba Seishohō Kentō Iinkai. 2022. *Okinawaken ni okeru 'Shimakutuba' no Hyōki ni tsuite [Japanese: On Orthography of 'Shimakutuba' in Okinawa]*. Okinawaken Bunka Kankō Sports-bu [Japanese: Department of Culture, Tourism, and Sports, Okinawa Prefecture], Naha.

Chokujin Uchima and Mitsuyoshi Nohara. 2006. *Okinawago Jiten: Naha Hōgen wo Chūshin ni [Okinawan Dictionary: Centering on Naha Dialect]*. Kenkyūsha, Tokyo.

UNESCO. 2010. *Atlas of the World's Languages in Danger*, third edition. UNESCO, Paris.

Kōichi Yasuoka. 2020. Universal Dependencies ni motozuku Tagengo Kakariuke Kaiseki Tool deplacy [Japanese: Multilingual Dependency Parsing Tool deplacy based on Universal Dependencies]. In *Jimbun Kagaku to Computer "Jimmonkon 2020" Symposium Ronbunshū [Japanese: Proceedings of Humanities and Computer "Jimmonkon 2020" Symposium]*, pages 95–100.

Daniel Zeman. 2008. `http://www.lrec-conf.org/proceedings/lrec2008/pdf/66_paper.pdf` Reusable Tagset Conversion Using Tagset Drivers. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).