

Building Web Corpus of Old Nubian with Interlinear Glossing as Digital Cultural Heritage for Modern-Day Nubians¹

So Miyagawa², Vincent W.J. van Gerven Oei³

1. Old Nubian

Old Nubian is a Nubian language recorded in the kingdoms of Nobadia and Makuria in the Middle Nile Valley, modern-day southern Egypt and northern Sudan, between the 8th and 15th centuries CE⁴. Besides Meroitic, the Old Nubian language is the oldest written language from the Nilo-Saharan language phylum⁵. It was written using the Coptic-derived Nubian alphabet, and includes three characters from the Meroitic alphasyllabary. The Old Nubian corpus consists of both literary material, such as Bible translations and sermons, and burial texts, and documentary materials, such as contracts, land sales, as well as numerous wall inscriptions of various kinds.

2. Goal of the project

Despite recent advances in the description of Old Nubian language and the publication of numerous new texts, there is no central digital corpus of Old Nubian texts. A team consisting of So Miyagawa, a digital humanist and Coptologist, and Vincent van Gerven Oei, the current expert in the Old Nubian language, have started to compile a digital corpus of Old Nubian. The underlying goal is to make Old Nubian texts available via a linguistically and philologically tagged corpus of Old Nubian that can be accessed via a user-friendly, highly visual online digital platform. Data sustainability and interoperability will be realized using *de facto* digital humanities standards; for instance, TEI XML for the linguistic and philological tag mark-up of Old Nubian texts. A visualization will also be created that is helpful for both linguistic experts and Nubian heritage holders. Specifically, the team has opted to show interlinear glosses following Leipzig Glossing Rules⁶ (LGR;

¹ This work was supported by JSPS KAKENHI Grant Numbers JP20K21975, JP21K00537.

² Center for Studies of Cultural Heritage and Inter Humanities (CESCHI), Kyoto University

³ Punctum Books and Community-led Open Publication Infrastructures for Monographs (COPIM), Coventry University.

⁴ For more information on the genealogy and history of Old Nubian and Nubian languages, see Chapter 1 of van Gerven Oei 2021 [1].

⁵ Though it is doubtful whether Nilo-Saharan forms a coherent genetic unity, the genealogical relations among Eastern Sudanic languages within Nilo-Saharan language phylum, which include Old Nubian and Meroitic, are well-established. See Rilly 2010 [2].

⁶ The *de facto* standard for interlinear glossing in linguistics papers, created by the Department of Linguistics at the Max Planck Institute for Evolutionary Anthropology, Leipzig [3].

modified to suit Old Nubian) under each word and morpheme on the web corpus for professional linguists. In addition, the corpus will in the future also incorporate Arabic and plain English grammatical annotations for Nubians, heritage holders of Old Nubian, especially those without a linguistic educational background.

3. Corpus-building methodology

The Old Nubian corpus data were primarily provided by Vincent van Gerven Oei. The data were originally composed in XeLaTeX, and the interlinear glossing were realized in a gb4e.sty format that consists of four lines: the first line is the Old Nubian text in the Old Nubian alphabet, the second line is the Romanized version of the same Old Nubian text with the morphemes parsed by hyphens, the third line provides interlinear glossing, and the fourth line is the English translation of the Old Nubian texts.

The entire Old Nubian corpus is not as big as that of Coptic and far smaller than that of Ancient Greek; however, it contains large amounts of epigraphic materials. The most sizable text is Pseudo-Chrysostomus's *In venerabilem crucem sermo*, a ~3,200-word Old Nubian translation of a Greek homily. The second most sizable text is the *Miracle of Saint Mina*, a Nubian miracle story of ~950 words. As the project is currently in the commencement phase, we are creating a pipeline to automatically convert the interlinear glossed corpus of Old Nubian from LaTeX or CSV into an adaptation of the TEI XML format developed for the Coptic SCRIPTORIUM⁷ and subsequently web pages with the interlinear glossing remaining under the text.

4. Pilot project

First, we chose the *Stauros Text* as the prototype. At ~800 words, the *Stauros Text* is a relatively long text in comparison to the other Old Nubian texts. Vincent van Gerven Oei developed the interlinear glossed text file in XeLaTeX which he is currently rendering into a CSV spreadsheet. To convert this into a simple interlinearly glossed text with an English translation in this LaTeX file, So Miyagawa created an XSLT program that converted the LaTeX file into TEI XML format, which was shared on GitHub to support data interoperability. An example of the TEI XML file that resulted from the conversion between LaTeX and gb4e.sty is below (Fig. 1).

⁷ The first multi-layered corpus project of Coptic texts [4]. For the project, see Schroeder and Zeldes 2016 [5].

```

<ab xml:id="SC4">
  <s type="orig">Γαειᾶ οὐκ ὀκιδᾶρρε·</s>
  <s type="parse">Γαει-ᾶ οὐ-κ ὀκ-ιδ-αρ-ρ-ε</s>
  <s type="roman">ηaei-a ou-k ok-ij-ar-r-e</s>
  <s type="gloss">who-QUOT 2PL-ACC call-PLACT-INTEN-PRS-1SG.PRED</s>
  <s type="trans" xml:lang="en">‘What shall I call you?’</s>
  <note>Notes The following affirmative forms in -μα are all dependent on the verb
  ὀκιδᾶρρε.</note>
</ab>

```

Figure 1: Tiers tagged by <s> </s> are the original text, parsed text, Romanized parsed text, interlinear glossing and English translation in TEI XML.

He then created a further XSLT program to transform the TEI XML into an HTML file that was combined with a JavaScript file to enable visualization of our interlinear glosses in the form of LGR on any Internet browsers (Fig. 2). This pipeline produced the first online Old Nubian corpus.

A large part of the documentary material is yet to be digitized and is unavailable in the XeLaTeX format described above. We will work on a pipeline that will allow efficient data entry for the remaining materials.

Γαειᾶ οὐκ ὀκιδᾶρρε·		
Γαει-ᾶ	οὐ-κ	ὀκ-ιδ-αρ-ρ-ε
ηaei-a	ou-k	ok-ij-ar-r-e
who- <u>QUOT</u>	<u>2PL-ACC</u>	call- <u>PLACT-INTEN-PRS-1SG.PRED</u>
‘What shall I call you?’ ^[2]		

Figure 2: Interlinear corpus rendition of Fig. 1 as a web page transformed by XSLT and leipzig.js [6].

5. Future perspectives for digital cultural heritage

Using the pipeline described above, we will produce more Old Nubian literary texts with LGR-styled interlinear glossing. To facilitate an online search, we will create a search function of each lemmata and glosses on the homepage using XQuery and also provide photos of Old Nubian manuscripts in IIF if their affiliations permit us. Moreover, as a side-product from the interlinear glosses, we intend to create the first digitized lexicon data of Old Nubian for further NLP development for Old Nubian.

Our work is primarily aimed at Old Nubian philology experts or linguistics, or historians in Medieval Nubia. However, we are also planning to embed plain English and Arabic translations, easy-to-read explanations for each gloss, a basic grammar of Old Nubian, and lecture videos on our corpus homepage. Through doing so, we hope to contribute to the cultural preservation and heritage education of Nubians and spread knowledge of the Nubian culture to a wider global audience.

Reference

- [1]. Vincent W.J. van Gerven Oei, *A Reference Grammar of Old Nubian*, Leuven: Peeters, 2021.
- [2]. Claude Rilly, *Le méroïtique et sa famille linguistique*, Leuven: Peeters, 2010.
- [3]. Max Planck Institute for Evolutionary Anthropology - Department of Linguistics, Leipzig Glossing Rules: Conventions for Interlinear Morpheme-by-Morpheme Glosses, <https://www.eva.mpg.de/lingua/resources/glossing-rules.php>, accessed on 2021-05-05.
- [4]. Caroline T. Schroeder, Amir Zeldes, et al., Coptic SCRIPTORIUM: Digital Research in Coptic Language and Literature, <https://copticcriptorium.org/>, accessed on 2021-05-05.
- [5]. Caroline T. Schroeder and Amir Zeldes, Raiders of the Lost Corpus, *Digital Humanities Quarterly* 10 (2), 2016, <http://www.digitalhumanities.org/dhq/vol/10/2/000247/000247.html>, accessed on 2021-06-21.
- [6]. Benjamin Chauvette, Leipzig.js: Interlinear Glossing for the Browser, <https://bdchauvette.net/leipzig.js/>, accessed on 2021-06-21.

JADH 2021

“Digital Humanities and COVID-19”

September 6-8, 2021

The University of Tokyo, JAPAN



<https://www.hi.u-tokyo.ac.jp/JADH/2021/>

The 11th Conference of Japanese Association for Digital Humanities

Proceedings of JADH conference, vol. 2021

Organized by

Organizing Committee, Japanese Association for Digital Humanities

Hosted by

Historiographical Institute, The University of Tokyo

Co-organized by

International Institute for Digital Humanities

Supported by

IPSJ SIG Computers and the Humanities

Japan Art Documentation Society

Japan Association for English Corpus Studies

Japan Society for Digital Archives

Japan Society for Information and Media Studies

Japan Society of Information and Knowledge

The Mathematical Linguistic Society of Japan

Proceedings of JADH conference, vol. 2021

Edited by Historiographical Institute, The University of Tokyo

Copyright © 2021 by the Japanese Association for Digital Humanities

Published by Historiographical Institute, The University of Tokyo

3-1, Hongo 7-chome, Bunkyo-ku, Tokyo 113-0033, JAPAN

<https://www.hi.u-tokyo.ac.jp/>

Online edition: ISSN 2432-3144

Print edition: ISSN 2432-3187

Table of Contents

Table of Contents.....	3
JADH 2021 Committees	9
Time Table.....	11
[Plenary – 1]	12
Keener than Connoisseurs’ Eyes: Analysis and Experience of Ancient Art through Virtual Reality (VR)	12
<i>Kyoko Haga</i>	
[Plenary – 2]	14
(Dis)Connections in Digital Japanese Studies	14
<i>Paula R. Curtis</i>	
Style Comparative study of Japanese medieval picture scrolls focusing on landscapes using GM Method with IIF Curation Platform	16
<i>Chikahiko Suzuki, Akira Takagishi, Asanobu Kitamoto</i>	
Book Barcoding for Differential Reading -Application to Woodblock-printed Books in the Bukan Complete Collection-	22
<i>Asanobu Kitamoto</i>	
Digital technologies and the spatial organisation of exhibitions: Interactive art as reflective experience	28
<i>Marianna Charitonidou</i>	
<i>The Re-centered Text: Digital restructuring in Amira Hanafi’s A Dictionary of the Revolution.....</i>	31
<i>David Thomas Henry Wright</i>	
Experimental LDA Topic Modelling of Tennyson’s Epic Poems.....	34
<i>Iku FUJITA</i>	
A study on the readerly aspects of Electronic Poetry through Cognitive Poetics	45
<i>Mariyam Nancy J, David Arputharaj</i>	

Architectural Drawings Exposed and the Effect of Digitization: The Rise of Artefactual Value vs the Democratization of Knowledge.....	47
<i>Marianna Charitonidou</i>	
Intersectionality and Digital Humanities in the Teaching of Architectural History: Diversity in the Dissemination of Knowledge	48
<i>Marianna Charitonidou</i>	
Multilingual word embeddings and low resources: identifying influence in Antiquity	51
<i>Marianne Reboul, École Normale Supérieure de Lyon</i>	
Skin Deep: Exploring Ideals of Japanese Beauty through Social Media	55
<i>Amy Grace Metcalfe, Emily Ohman</i>	
Analyzing “Mechanisms” in the British National Corpus	59
<i>Yuki Sugawara</i>	
One Challenge, Not Two Problems: Regular Expressions for Researching a Single-Author Corpus.....	62
<i>Dr. Robert W. Williams</i>	
Picking out Arabian Names from <i>Fahrasa</i> by Ja‘far b. Idrīs al-Kattānī without Reading Arabic.....	66
<i>Yuri Ishida, Kensuke Baba</i>	
POS tagging for Vedic Sanskrit using deep learning.....	69
<i>Yuzuki Tsukagoshi</i>	
Spectral analysis for identifying octave playing in piano works.....	72
<i>Mai Takahashi, Michikazu Kobayashi, Ikki Ohmukai</i>	
Token-based semantic vector space model for classic poetic Japanese.....	77
<i>Xudong Chen, Hilofumi Yamamoto, and Bor Hodošček</i>	
Open source datasets of the Hachidaishū for the research of classical Japanese poetic vocabulary.....	82
<i>Hilofumi Yamamoto, Bor Hodošček</i>	

Exploring Metadata Quality Issues in Non-English Corpora: Preliminary Assessments of HathiTrust Records of Late Imperial Chinese Books	88
<i>Wenyi Shang, Jacob Jett, J. Stephen Downie</i>	
Dataset Construction for Cross-genre Plot Structure Extraction.....	93
<i>Hajime Murai, Shuuhei Toyosawa, Takayuki Shiratori, Takumi Yoshida, Shougo Nakamura, Yuuri Saito, Kazuki Ishikawa, Sakura Nemoto, Junya Iwasaki, Akiko Uda, Shoki Ohta, Arisa Ohba, Takaki Fukumoto</i>	
Basic Plot Structure in the Adventure and Battle Genres	97
<i>Yuuri Saito, Takumi Yoshida, Shougo Nakamura, Kazuki Ishikawa, Shoki Ohta, Arisa Ohba, Takaki Fukumoto, Hajime Murai</i>	
Construction of ShiJi Spatiotemporal Information Platform on the Framework of Research-oriented Knowledge Bases	101
<i>Jung-Yi Tsai, Pi-Ling Pai, Hsiung-Ming Liao, You-Jun Chen, Richard Tzong-Han Tsai*, I-Chun Fan</i>	
Cross-genre Plot Analysis of Detective and Horror Genres	106
<i>Junya Iwasaki, Shuuhei Toyosawa, Kazuki Ishikawa, Shoki Ohta, Hajime Murai</i>	
Using Moodle as a Multi-Modal Tool for Ainu Language Education	111
<i>Matthew Cotter, Takayuki Okazaki, Jennifer Teeter</i>	
An Attempt at Creating Integrated Retrieval for Chinese Excavated Materials: An Implementation of a Search Function across Interpretations of Ancient Characters	114
<i>Shumpei Katakura</i>	
Collecting Canons: Comparing Guodian and Mawangdui Laozi Texts with the Dead Sea Scrolls.....	117
<i>Janelle Peters</i>	
Development of Database for Japanese Conversation Patterns: an observation from noun phrases ending with focus particle "mo (also)"	120
<i>Mika Ebara, Hilofumi Yamamoto</i>	

Drug-focused text summarization of coronavirus-related articles for the discovery of COVID-19 therapies	124
<i>Setsuro Matsuda</i>	
Reconstruction and Utilization of Text Data Using TEI: Case study of the Shibusawa Eiichi Denki Shiryo	126
<i>Boyoung Kim, Satoru Nakamura, Yuta Hashimoto, Naoki Kokaze, Sayaka Inoue, Toru Shigehara, Kiyonori Nagasaki</i>	
Development of a support system for extracting mentioned bibliographical data from the Encyclopédie entries	130
<i>Satoru Nakamura, Ayano Kokaze, Yoshiho Iida, Naoki Kokaze, Tatsuo Hemmi</i>	
Platformed reflections on the Pandemic: Covid-19 and Electronic Literature.....	134
<i>Anna Nacher, Søren Bro Pold, Scott Rettberg</i>	
Digital Humanities and the way forward for ethnographic research: What we learned from Covid-19?	139
<i>Deepika Kashyap</i>	
Virtual Communities and Post-Pandemic Possibilities: Animal Crossing New Digital Humanities	141
<i>Quinn Dombrowski, Elizabeth Grumbach, Merve Tekgürler</i>	
Building Web Corpus of Old Nubian with Interlinear Glossing as Digital Cultural Heritage for Modern-Day Nubians.....	144
<i>So Miyagawa, Vincent W.J. van Gerven Oei</i>	
Development of data-driven historical information research infrastructure at the Historiographical Institute in the University of Tokyo.....	148
<i>Satoru Nakamura, Taizo Yamada</i>	
Compilation of Semantic Data Archive: A New Method of Learning “Local Culture”	152
<i>Kwangwoo Kim, Soohyeon Kim</i>	
Towards a Structured Description of the Contents of the Taisho Tripitaka.....	161

Yoichiro Watanabe, Kiyonori Nagasaki, Hyunjin Park, Yifán Wáng, Tomohiro Murase, Masayoshi Watanabe, Norimichi Yajima, Yoshihiro Sato, Yūi Sakuma, Xinxing Yu, Masahiro Shimoda, Ikki Ohmukai

Classification of face images in the frontispiece paintings of Sutra copies in gold ink on indigo paper by deep convolutional neural networks 164

Toshiaki Aida, Tomomi Kobayashi, Aiko Aida

The difference in transitional process between Western instrumental and vocal music 169

Daisuke Miki, Akihiro Kawase, Kenji Hatano

e-Sukhāvati: An Innovative Digital Platform for Studying the *Smaller Sukhāvavīyūha* 172

SIU Sai-yau

New Possibilities of Digital Publishing and Online Exhibition— A Case Study of the Website “Reflections on COVID-19” 178

Lin, Wen Jiun

Sonifying the pandemic – innovative approaches towards data interaction and engagement formats for scientific, educational and artistic purposes 192

Michael Stark, Amelie Dorn, Renato Rocha Souza

Thailand Towards Digitization– the past, the present, the future and gray digital gap 196

Saiyud Moolphate, Nadila Mulati, Thin Nyein Nyein Aung, Motoyuki Yuasa, Myo Nyein Aung

Wikidata as a Low-tech Solution to Leverage Semantic Technologies and A Case Study of CBDB ID’s Reconciliation with Wikidata 199

Fudie Zhao

[Workshop – 1] 203

歴史学におけるデータ共有, 統合化, 多角的協働 203

[Workshop – 2] 205

海外 DH 教育動向調査 205