Building Web Corpus of Old Nubian with Interlinear Glossing as Digital Cultural Heritage for Modern-Day Nubians[1]

So Miyagawa[2], Vincent W.J. van Gerven Oei[3]

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 CE[4]. Besides Meroitic, the Old Nubian language is the oldest written language from the Nilo-Saharan language phylum[5]. 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 Rules[6] (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 SCRIPTORIUM[7] 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">ⲥⲁⲉⲓⲁ̄ ⲟⲩⲕ ⲟ̄ⲕⲓⲇ̅ⲁⲣⲣⲉ·</s>
  <s type="parse">ⲥⲁⲉⲓ-ⲁ̄ ⲟⲩ-ⲕ ⲟ̄ⲕ-ⲓⲇ̅-ⲁⲣ-ⲣ-ⲉ</s>
  <s type="roman">ŋaei-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 -ⲛⲁ are all dependent on the verb
    ⲟ̄ⲕⲓⲇ̅ⲁⲣⲣⲉ.</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.

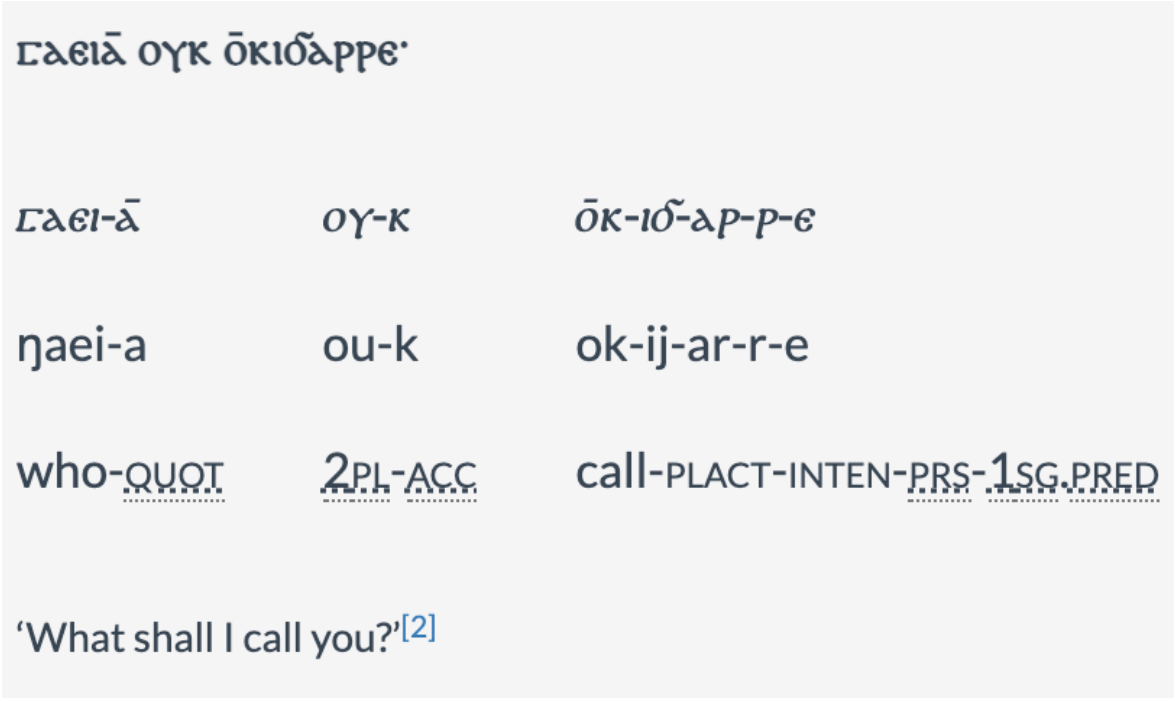

Figure 2: Interlinear corpus rendition of Fig. 1 as a web page transformed by XSLT and leipzig.js [6].

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 IIIF 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 preservationand 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://copticscriptorium.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

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