Digitization of the Concise Sociopolitical and Mass-Media Dictionary (English – Amharic)

Hewan Semon Marye

Electronic version

URL: <https://journals.openedition.org/jtei/3726>

DOI: 10.4000/jtei.3726

ISSN: 2162-5603

Publisher

TEI Consortium

Electronic reference

Hewan Semon Marye, "Digitization of the Concise Sociopolitical and Mass-Media Dictionary (English – Amharic)", *Journal of the Text Encoding Initiative* [Online], Rolling Issue, Online since 22 January 2022, connection on 14 January 2023. URL: <http://journals.openedition.org/jtei/3726> ; DOI: <https://doi.org/10.4000/jtei.3726>

For this publication a Creative Commons Attribution 4.0 International license has been granted by the author(s) who retain full copyright.

Digitization of the Concise Sociopolitical and Mass-Media Dictionary (English — Amharic)

ABSTRACT

The Concise Sociopolitical and Mass-Media Dictionary is an English-to-Amharic dictionary that was prepared primarily by Professor Zdeněk Poláček and digitized at the Hiob Ludolf Center for Ethiopian and Eritrean Studies (HLCEES) under the supervision of Professor Alessandro Bausi. The edition and digitization were carried out initially on Microsoft Word and then through XML using similar technology to other production work at the HLCEES. By using a lightly modified TEI Publisher application within eXist-db, this project successfully produced a 1,600-page PDF that may also be published as a book. This project note highlights the key steps taken in the process of digitizing the dictionary. It gives a succinct exploration of how TEI Publisher simplified

the digitization process in comparison to Word. It also presents some of the extra revision steps that had to be taken using eXide after a dictionary entry was entered into the application. We supplement this note with figures to demonstrate these points.

INDEX

Keywords: dictionary, edition, XML tool, eXist-db, Ethiopia, Ethiopian studies, Amharic

- 1 The Concise Sociopolitical and Mass-Media Dictionary is an English-to-Amharic dictionary that was prepared primarily by Professor Zdeněk Poláček. It was edited and digitized at the Hiob Ludolf Center for Ethiopian and Eritrean Studies (HLCEES) from 2017 until 2020. In this brief note, I describe the project that led to a forthcoming print publication of the dictionary.

1. Dictionary Description

- 2 The dictionary consists of words that are associated with the sociopolitical and mass media. Before digitization, each dictionary entry was written on a small index card with definitions and examples given as necessary. These index cards were delivered to the HLCEES in 2017 in eight boxes. In addition to these boxes, the English words beginning with A were delivered in a bound document separately from the boxes with index cards. This bound document also included the introduction to the project, the appendices, linguistic explorations by the original author, and the sources that were consulted. The HLCEES, under the guidance of Professor Alessandro Bausi and with funding from Universität Hamburg, took the manuscript to edit and create a digital version that would then be used for publication.
- 3 At first, it seemed that the project would be undertaken using Microsoft Word. I worked through the introduction and the English words beginning with A from the bound document by typing the entries into Word. The dictionary entries were inserted into a two-columned table. The left-hand column included the words, in bold format, with the part of speech (noun, adjective, adverb, and so on) in italics, as well as a definition of the word if the definition was supplied in the original hand-written document. This same column gave some English-language examples, as complete

sentences or phrases, of the word in use. The right-hand column included the Amharic equivalent term or the meaning of the English word in Amharic. It also included translations of the examples listed in the left-hand column.

Figure 1. Entry for *abdicate* from the Word document.

<ul style="list-style-type: none"> • abdicate <i>v.t.t., -cates, -cating, -cated:</i> <ul style="list-style-type: none"> - <to> abdicate from power - <to> abdicate one's rights - The King decided (<i>hon. 3rd</i>) to abdicate. - The Queen abdicated/gave up, renounced/ (<i>hon. 3rd</i>) the throne in favour of the Crown Prince. - They have abdicated/have abandoned/all responsibility for the project. 	<p>ሥልጣንን ለቀቀ (መልቀቅ)</p> <p>መብቱን በራሱ ፈቃድ ተወ (መተወ)</p> <p>ንጉሡ ሥልጣናቸውን ለመልቀቅ ወሰኑ። ንጉሡ ከሥልጣናቸው ለመውረድ ወሰኑ።</p> <p>ንግሥቲቱ ዙፋናቸውን /አልጋቸውን/ በገዛ ፈቃዳቸው ለአልጋ ወራሹ ለቀቁ።</p> <p>ፕሮጀክቱን በሚመለከት የነበራቸውን ኃላፊነት ሁሉ ለቀቁ /ተወ/።</p>
---	--

- In general, working with Word seemed simple enough. However, it was difficult and time-consuming to change the style selection that varied with each word entry, its constituent part of speech, and the usage examples. These elements had to be manually updated for each entry of the dictionary.
- In total, 125 pages of the dictionary were typed into Word before Professor Bausi suggested using XML instead because it would be an interesting experiment. More out of curiosity than out of necessity, we decided that some brief training on XML and its tools would be helpful to accomplish this. Professor Bausi predicted that a tool that was developed for digitizing the dictionary would also minimize the time taken to mark the different parts of speech of each word entry. At this point, we turned to the technology used by the *Online Lexicon Linguae Aethiopiae* (Bausi and Sokolinski 2016)¹ at the HLCEES to see what methods and tools I would be able to use to generate a digital version of the dictionary. Some of the technology used by the *Online Lexicon Linguae Aethiopiae* could have been reused, but we eventually opted for a different workflow, which only had the use of TEI and *exist-db*² in common with that resource.

2. XML Encoding

- 6 The dictionary was digitized using a lightly modified version of an early release of the TEI Publisher application³ within eXist-db, with a simple ODD that was dedicated to the task that used the TEI Dictionary module (see [TEI Consortium 2020](#), chap. 9, “Dictionaries”).⁴ The application in itself was never an intended output and was used simply as a tool to obtain the PDF version of the entire dictionary for the printable book. In comparison with other dictionary projects such as the digitization and updating of Dillmann’s Lexicon (see [Bausi and Sokolinski 2016](#), [Liuzzo 2019](#), 143–159), this digitization work could not be automated, as it was started from scratch and needed a solution to support data entry as well as a professional layout of the final product. This PDF was produced with an XQuery⁵ script that transforms the data in XSL-FO⁶ and renders it with APACHE FOP,⁷ which is common practice for producing professional output ([Liuzzo 2019](#), 219–32).
- 7 As shown in [figure 2](#), instead of entering every word entry into eXide,⁸ I entered the term, definition (in English) if available, the part of speech, and the usage examples, and then the Amharic term, definition, and examples. After each term was entered, a link to eXide for further editing was produced.

Figure 2. The application data entry interface.

Concise Socio-Political and Mass Media Dictionary

Great! So you are entering something new in the dictionary!

Enter here the basic informations about your new entry.
A form will never be able to give you the flexibility that you can have in encoding your data, so here there are only a few things you can enter to start your entry.
You will then edit it in eXide, but here it will get a file name, a unique id, a date of creation.
Actually you can just enter letter and the term

General data

Term *

type here the title of the new file (will go to /node5/mc/title) and in the main form, where you can then change it.

Definition

the definition (not the examples!)

Main grammatical information

PoS

Part of Speech

Usage domain

usage domain

Amharic

Amharic Term

type here the Amharic lemma

Definition

the definition in Amharic (not the examples!)

Examples

ADD EXAMPLE

CREATE NEW ENTRY

Please enter the data and hit the button to create a new entry.

- 8 After entering the data in this TEI Publisher application with the interface shown in figure 2, I returned to eXide to check if the data were appropriately marked (e.g., using italics or boldface); and if not, then I selected the particular term. For example, if a part-of-speech term (for example, *plural* [pl.]) in the definition or the examples provided did not appear in italics, I returned to eXide and entered <emph> around the parts of speech so the final result would be italicized. Another example is when a part of speech, or the entries themselves, needed to appear in bold. I entered <orth> around those phrases.
- 9 Example 1 gives a sample record that presents the XML encoding for the word *abdicate*. It also demonstrates how the English and Amharic equivalents are encoded through parallel numbering.

Example 1. XML encoding for the word *abdicate*.

```

<text>
  <body>
    <div>
      <entry xml:id="CSPMDb389f92c-9942-46d9-89f9-2c994236d9fc" n="5122">
        <form>
          <orth>abdicate </orth>

          <pos>
            <emph>v.t.i.</emph>,<orth> -cates, -cating, -cated</orth>
          </pos>

        </form>
        <sense xml:lang="en">
          <cit n="1" type="example">
            <quote>&lt;to&gt; abdicate from power</quote>
          </cit>
          <cit n="2" type="example">
            <quote>&lt;to&gt; abdicate one's rights</quote>
          </cit>
          <cit n="3" type="example">
            <quote>The King decided (<emph>hon. 3rd</emph>) to abdicate.</quote>
          </cit>
          <cit n="4" type="example">
            <quote>The Queen abdicated/gave up, renounced/ (<emph>hon. 3rd</emph>) the
            throne in favour of the Crown Prince.</quote>

```

```

</cit>
<cit n="5" type="example">
  <quote>They have abdicated/have abandoned/all responsibility for the
project.</quote>
</cit>
</sense>
<sense xml:lang="am">
  <cit n="1" type="example">
    <quote>ሥልጣንን ለቀቀ (መልቀቅ)</quote>
  </cit>
  <cit n="2" type="example">
    <quote>መብትን በራሱ ፈቃድ ተወ (መተው)</quote>
  </cit>
  <cit n="3" type="example">
    <quote>ንጉሡ ሥልጣናቸውን ለመልቀቅ ወሰኑ። ንጉሡ ከሥልጣናቸው ለመውረድ ወሰኑ።</quote>
  </cit>
  <cit n="4" type="example">
    <quote>ንግሥቲቱ ዙፋናቸውን /አልጋቸውን/ በገዛ ፈቃዳቸው ለአልጋ ወራሽ ለቀቁ።</quote>
  </cit>
  <cit n="5" type="example">
    <quote>ፕሮጀክቱን በሚመለከት የነበራቸውን ኃላፊነት ሁሉ ለቀቁ /ተዉ/። </quote>
  </cit>
</sense>
</entry>
</div>
</body>
</text>

```

- 10 The rendering was not checked every time for the entire PDF, but it was checked within TEI Publisher using `<modelSequence>` and `<model>`s to define the rendering. The alignment of the English and Amharic was simply achieved by styling a table of the two `<sense>` elements with `@xml:lang`.

Example 2. Using `<elementSpec>` to define the table of `<sense>` elements.

```

<elementSpec ident="sense" mode="change">
  <modelSequence predicate="@xml:lang='en' ">
    <model behaviour="block">

```

```

<outputRendition>display: table-cell; min-width:
300px; padding-right: 30px</outputRendition>
</model>
</modelSequence>
<modelSequence predicate="@xml:lang='am' ">
  <model behaviour="block">
    <outputRendition>display: table-cell; width: 100%; vertical-align:
top;</outputRendition>
  </model>
</modelSequence>
</elementSpec>

```

- 11 Figure 3 shows a sample image for the entry *abdicate* when generated as a PDF. We observe that this particular entry does not have an Amharic equivalent for the word, it only provides some examples to demonstrate how the word can be used.

Figure 3. Image of the dictionary entry *abdicate* in the output PDF.

<ul style="list-style-type: none"> • <i>abdicate v.t.i., -cates, -cating, -cated</i> • <to> abdicate from power • <to> abdicate one's rights • The King decided (<i>hon. 3rd</i>) to abdicate. • The Queen abdicated/gave up, renounced/ (<i>hon. 3rd</i>) the throne in favour of the Crown Prince. • They have abdicated/have abandoned/all responsibility for the project. 	<ul style="list-style-type: none"> • ሥልጣንን ለቀቀ (መልቀቅ) • መብትን በራሱ ፈቃድ ተወ (መተው) • ንጉሡ ሥልጣናቸውን ለመልቀቅ ወሰኑ። ንጉሡ ከሥልጣናቸው ለመውረድ ወሰኑ። • ንግሥቲቱ ዙፋናቸውን /ክልጋቸውን/ በገዛ ፈቃዳቸው ለአልጋ ወራሹ ለቀቁ። • ፕሮጀክቱን በሚመለከት የነበራቸውን ኃላፊነት ሁሉ ለቀቁ /ተዉ/።
---	---

- 12 Note that in the fourth example, there is no space between *abdicated* and the slash (/); we had to add one in eXide. The fifth example also presents a similar error, and neither of these errors is present in the Amharic translation of the examples. Checking for particular entries is done in the database directly using the behaviors specified in the ODD.
- 13 For a user accustomed to Word, the transition to XML was not easy at first. However, it was much easier after the application was developed and after learning the most important codes for the dictionary. The application is efficient and it minimized the amount of time needed to digitize the dictionary. In comparison, Word was a great drain of time and energy. There are, of course, some

things to keep in mind when using XML. One is regarding the parts of speech that are given in the definition and examples section. These have to be marked in eXide. Another is to check the spacing between words and symbols for each entry.

- 14 Another point that has not been addressed thus far concerns tables. In the introduction of the dictionary, there are instances where short tables needed to be inserted. It was easier to create tables in Word than in eXide, but this could generally be a limit of our knowledge regarding XML. It would be useful if some of the tools readily available in Word were also accessible in eXide. But it is evident that dictionaries and works of lexicography can be edited with the tools we used in the future.
- 15 TEI Publisher and ODD sufficed as a working setting for this project and, after three years of work, the dictionary was ready for publication. The output is an almost 1,600-page-long searchable PDF version of the dictionary, which will also be published in print.

BIBLIOGRAPHY

- Bausi, Alessandro, and Eugenia Sokolinski, eds. *150 Years After Dillmann's Lexicon Perspectives and Challenges of Gə'əz Studies*. Supplement to Aethiopica, 5. Wiesbaden: Harrassowitz Verlag, 2016.
- Liuzzo, Pietro Maria. *Digital Approaches to Ethiopian and Eritrean Studies*. Supplement to Aethiopica, 8 Wiesbaden: Harrassowitz Verlag, 2019. doi:10.2307/j.ctvrnfr3q.
- TEI Consortium. 2020. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 4.1.0. Last updated August 19. <https://tei-c.org/Vault/P5/4.1.0/doc/tei-p5-doc/en/html/index.html>.

NOTES

- 1 Accessed January 22, 2022, <https://betamasaheft.eu/Dillmann/>.
- 2 Accessed January 18, 2022, <http://exist-db.org/>.
- 3 “TEI Publisher: The Instant Publishing Toolbox,” accessed November 10, 2020, <https://teipublisher.com/index.html>.

4 <https://tei-c.org/Vault/P5/4.1.0/doc/tei-p5-doc/en/html/DI.html>. TEI Lex-0 would have been too restrictive and specialized a customization for this workflow, while the Dictionary module allows enough freedom. The aim of this project was not to exchange data, but only to prepare to publish a book.

5 See <https://www.w3.org/TR/xquery-31/>, accessed January 21, 2022.

6 See <https://www.w3.org/TR/xsl11>, accessed January 21, 2022.

7 See <https://xmlgraphics.apache.org/fop/>, accessed January 21, 2022.

8 <https://exist-db.org/exist/apps/eXide/index.html>, accessed January 18, 2022.

AUTHOR

Hewan Semon is a student of modern Ethiopian history, interested in urban social history as well as the traditions and history of the Ethiopian Orthodox Church. Hewan has just completed her PhD, “The Rhetorics and Polemics of Humor in Addis Abäba Ethiopia, 2005—2019” at Universität Hamburg. Besides working on the Concise Sociopolitical and Mass-Media Dictionary, she also has short essays, translation works, and articles mostly relating to Addis Abeba and its contemporary realities.