Acknowledgement: This work was supported by Grant-in-Aid for JSPS Fellows Grant Number 24KJ1814, Grant of Mishima Kaiun Memorial Fund and Hiroshima University's Career Advancement Project for Women Researchers (CAPWR) for Academic Year 2023.

Compiling an Online Dictionary based on Field Data:

The Case of Kelabit Utilizing TEI/XML, XSLT and ChatGPT

Yasuka Fukaya

Kyushu University/JSPS (Japan Society for the Promotion of Science)

(yasuka.fukaya@gmail.com)

Abstract

Introduce Kelabit and ODK project

*Online Dictionary of Kelabit

How NLP assists in this project

ChatGPT helped the author create...

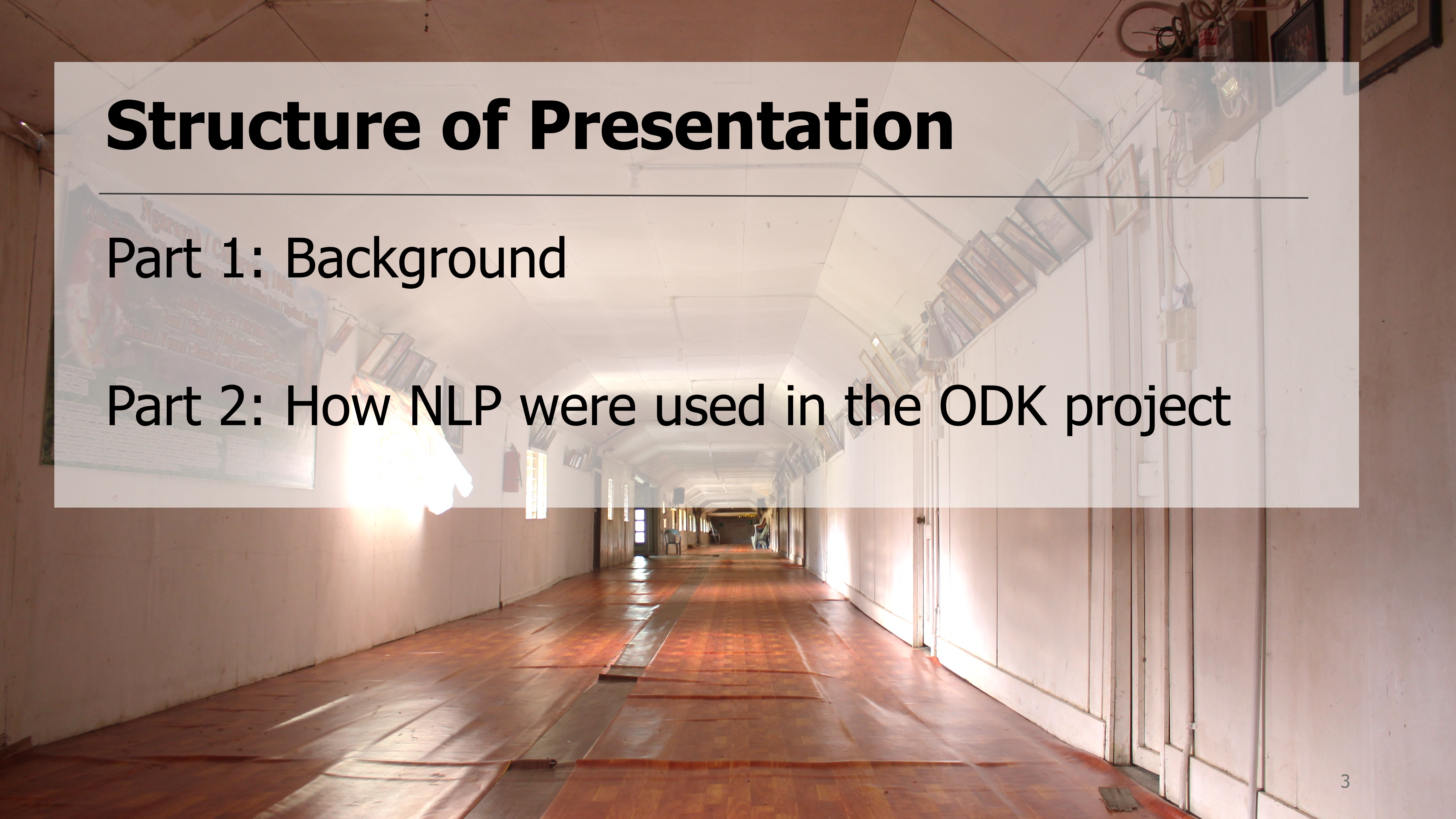
XSLT (to transform XML to HTML)

JavaScript (to add research function)

Structure of Presentation

Part 1: Background

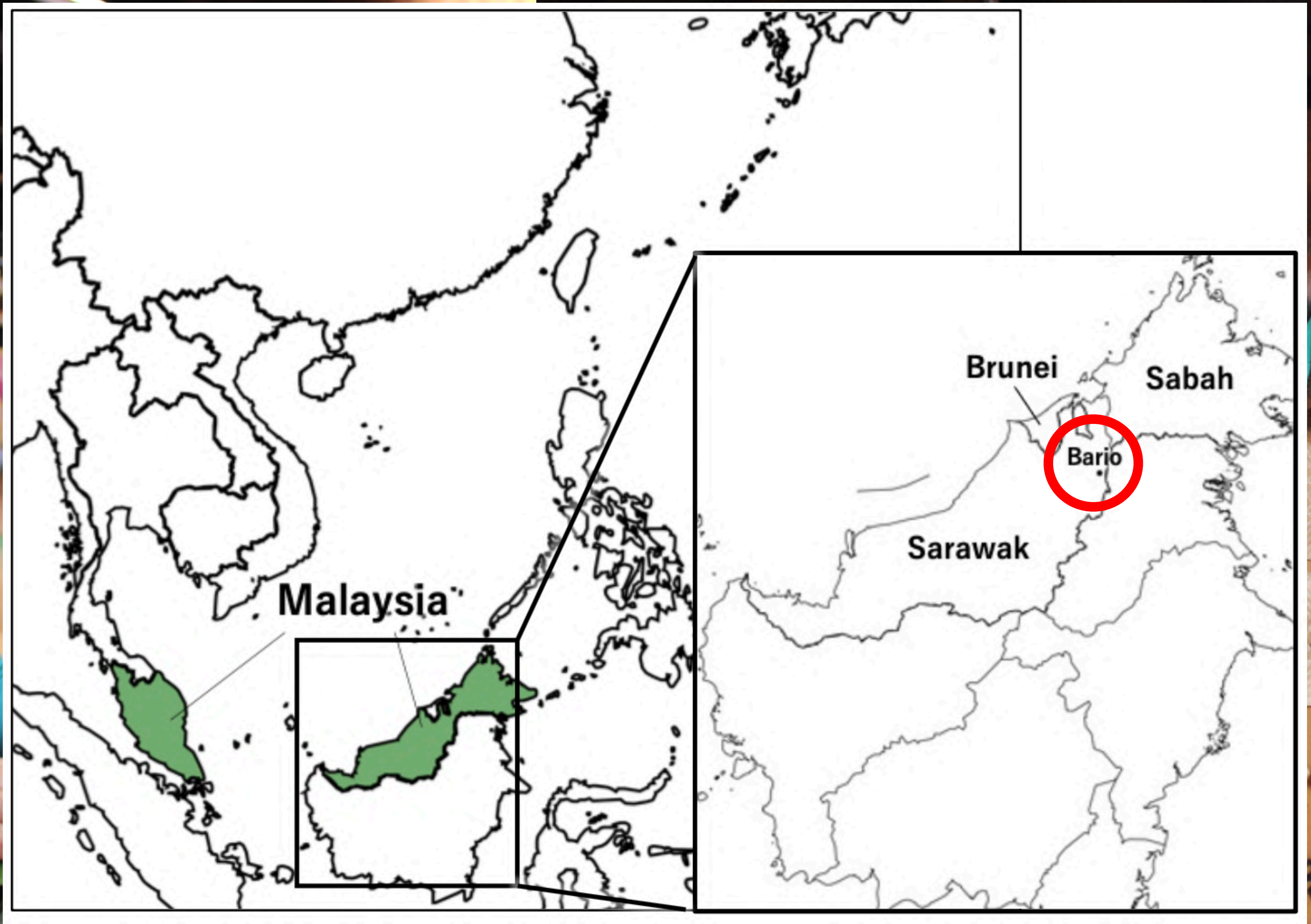
Part 2: How NLP were used in the ODK project

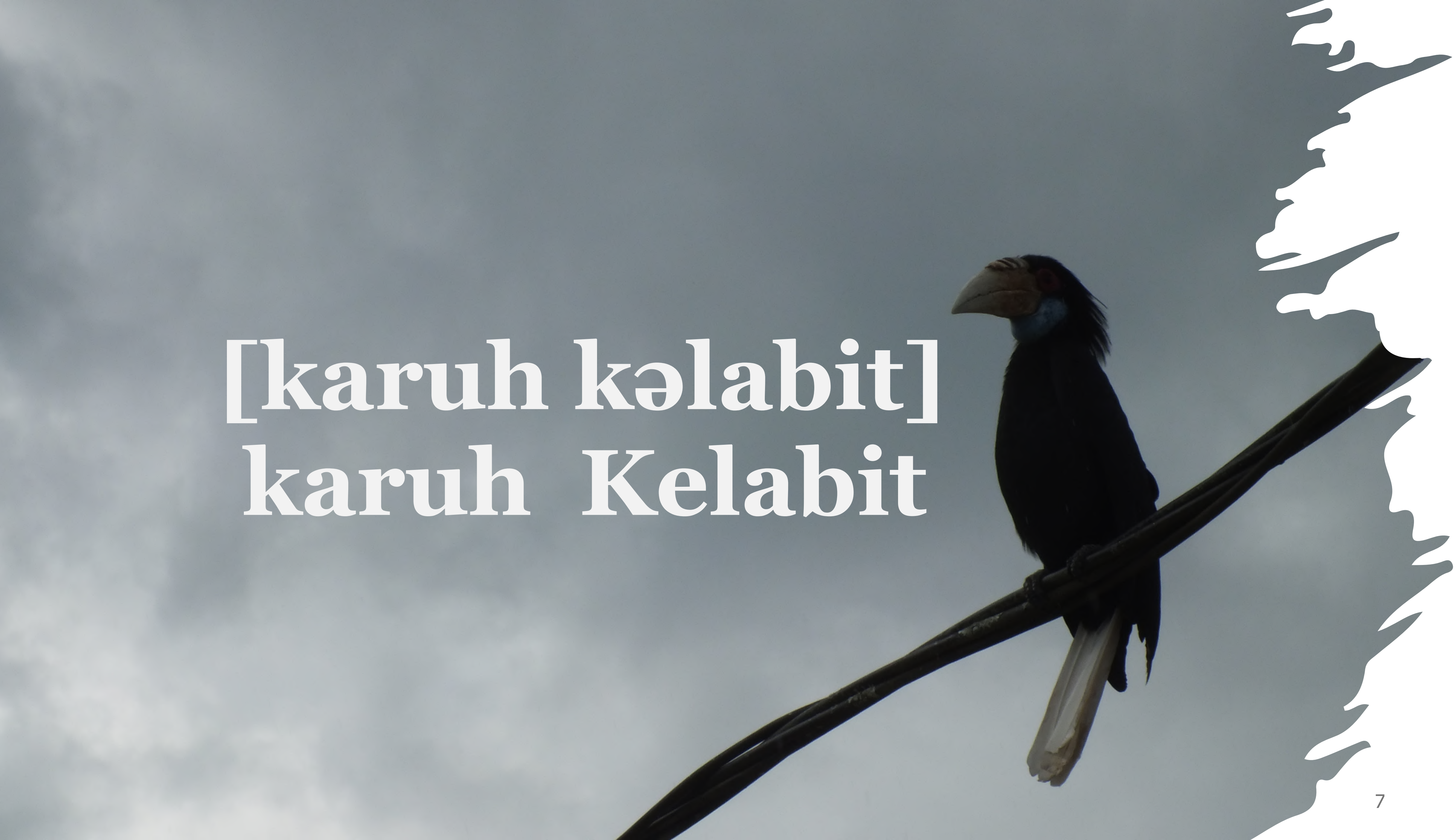


Part 1: Background

Bario, Sarawak, Malaysia

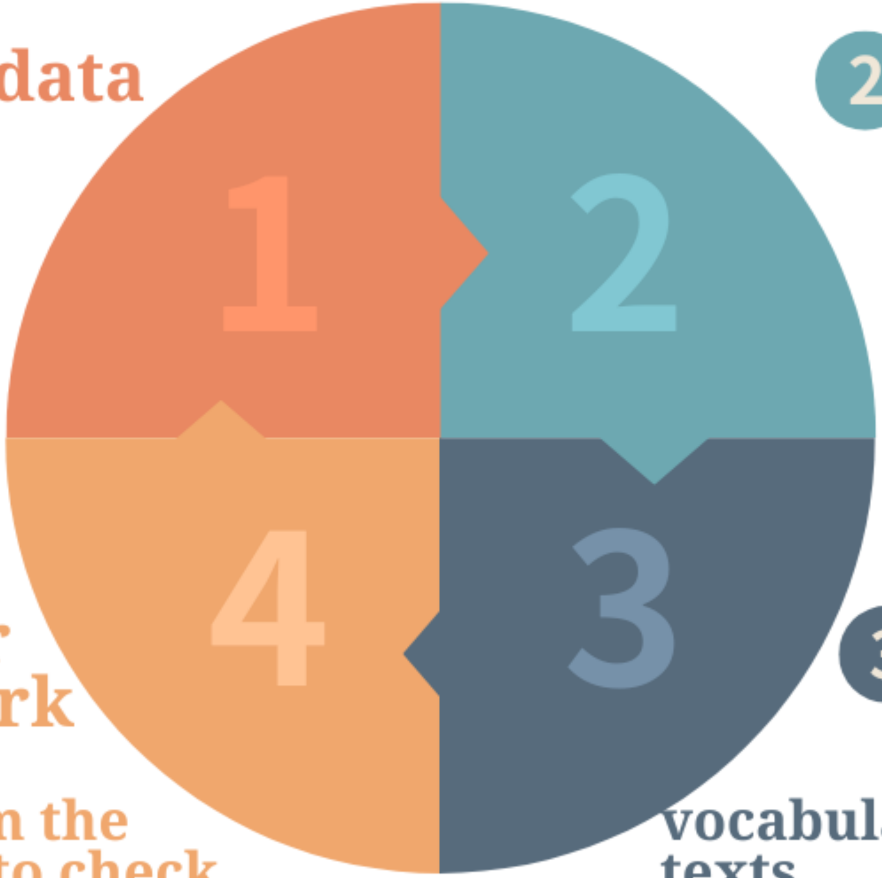
Kelabit





[karuh kəlabit]
karuh Kelabit

Process of grammar description



1 Collect data

2 annotation
analysis
(ELAN)

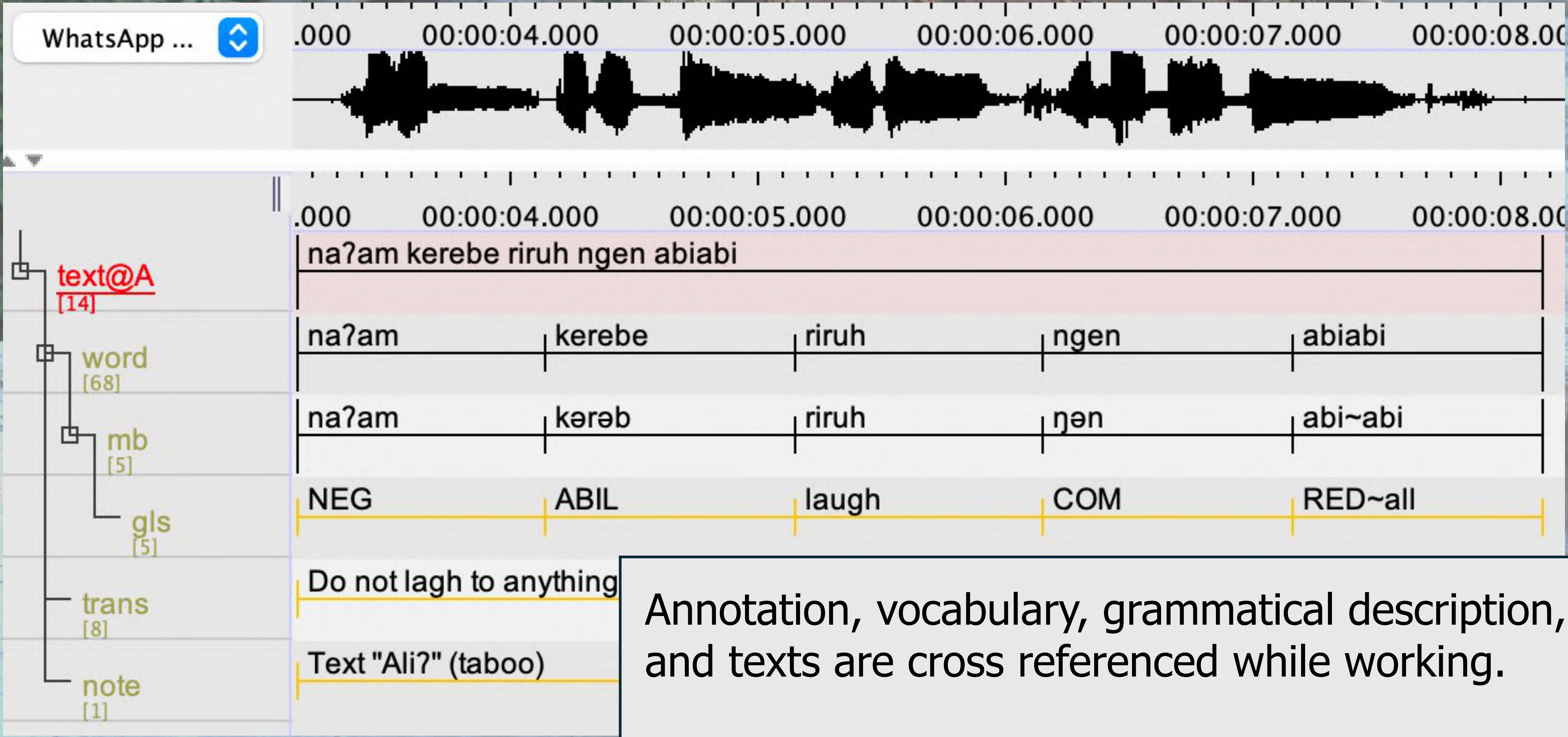
3 cross
referencing

vocabulary
texts
grammatical description

4 another
fieldwork

to confirm the
analysis, to check
another grammar…

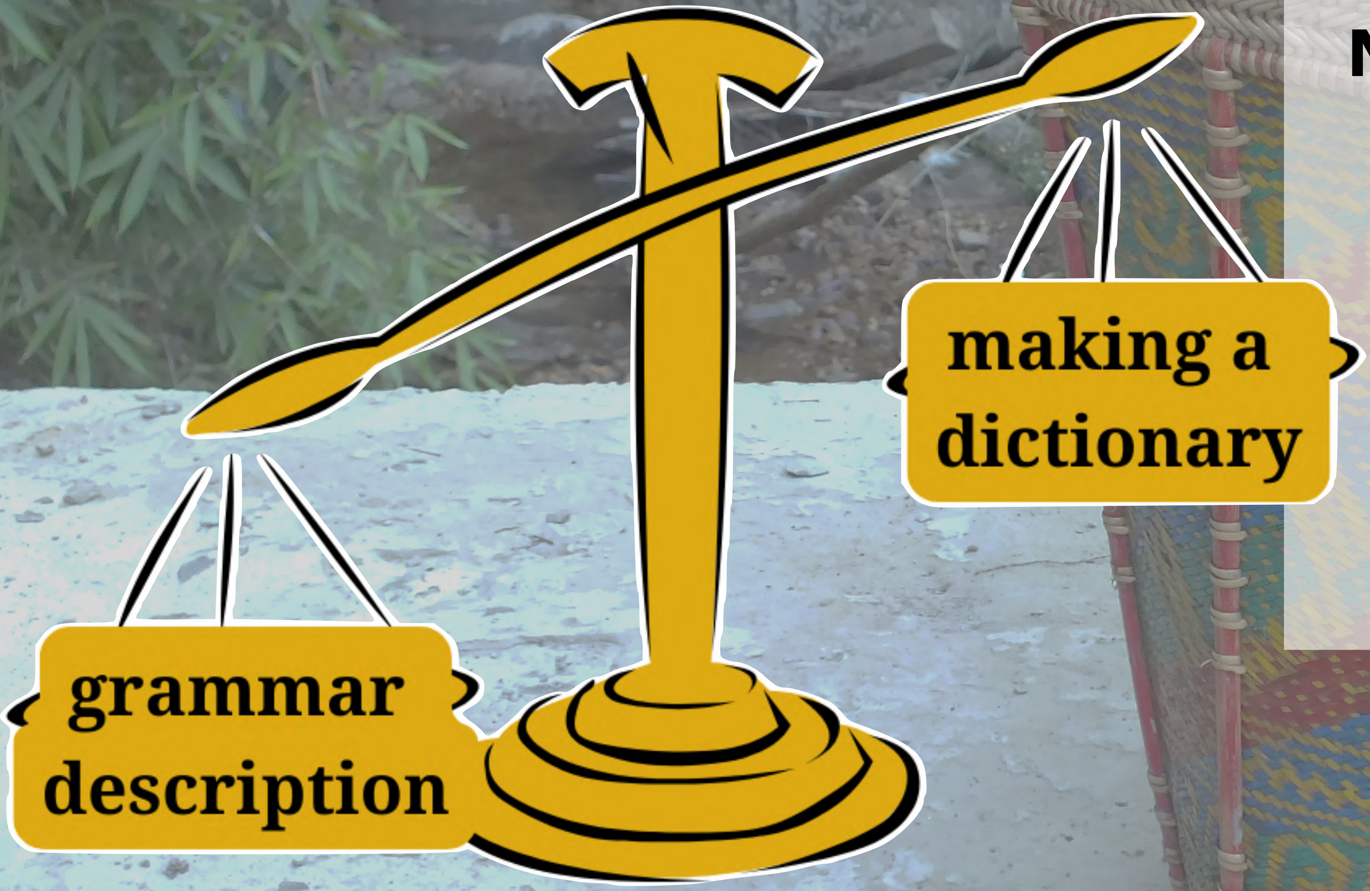
Annotation work using ELAN



text@A [14]	naʔam kerebe riruh ngen abiabi				
word [68]	naʔam	kerebe	riruh	ngen	abiabi
mb [5]	naʔam	kərəb	riruh	ŋən	abi~abi
gls [5]	NEG	ABIL	laugh	COM	RED~all
trans [8]	Do not lagh to anything				
note [1]	Text "Aliʔ" (taboo)				

Annotation, vocabulary, grammatical description, and texts are cross referenced while working.

We have lexical data for a dictionary…



Needs of speakers' community

- **language inheritance**

  dictionary
  children's book
  textbook for grown-ups
  …

Part 2:
How NLP were used in the ODK project?

The process of creating ODK

- **The process of making lexical data for the dictionary**

  Raw data → transcription (ELAN) **Done while grammar writing**

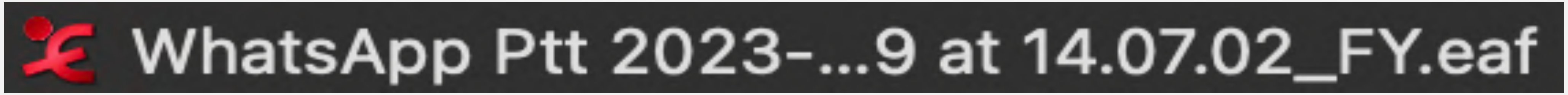
  → **transformation to TEI/XML (Oxygen)**

- **The process of getting the output in different formats**

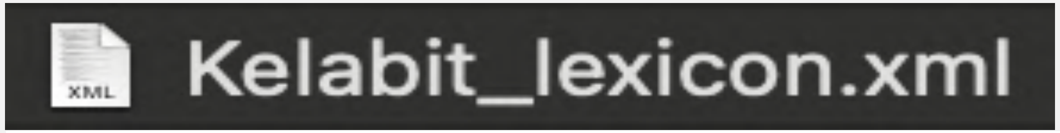
  XSLT(ChatGPT) → XML to HTML (Oxygen)

  → search function in JavaScript (ChatGPT)

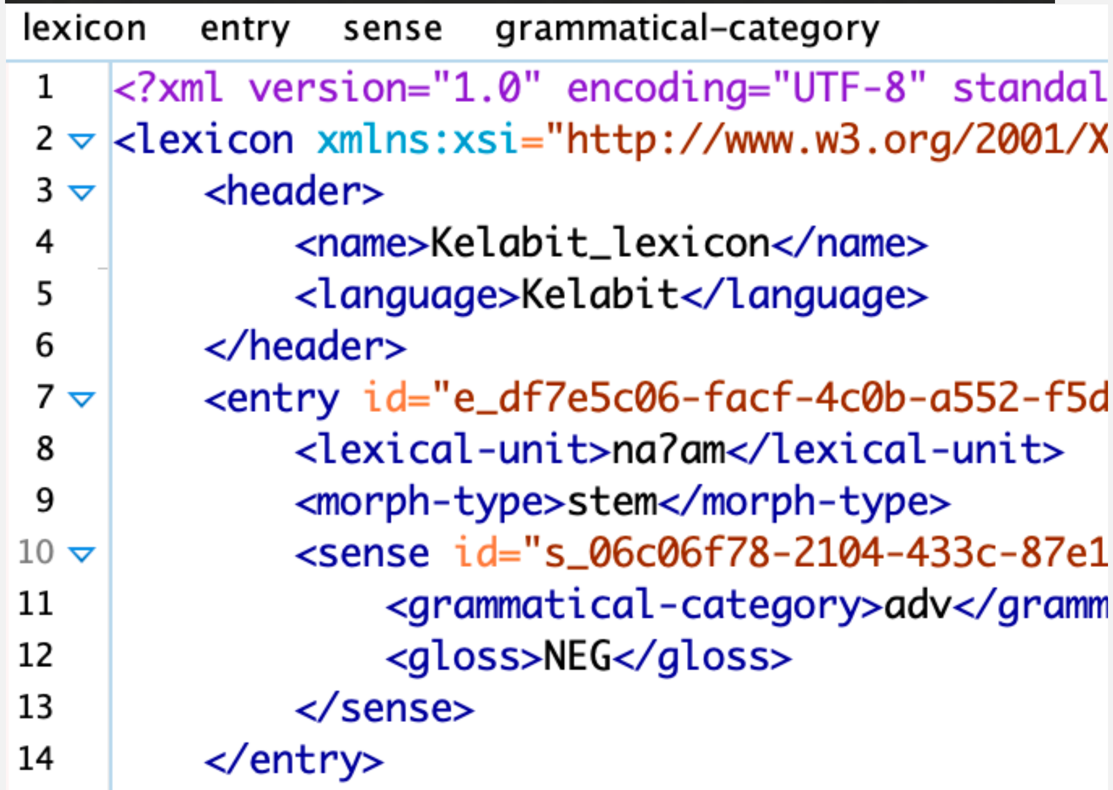
ELAN --> TEI/XML (OxygenXMLEditor)



WhatsApp Ptt 2023-...9 at 14.07.02_FY.eaf



Kelabit_lexicon.xml



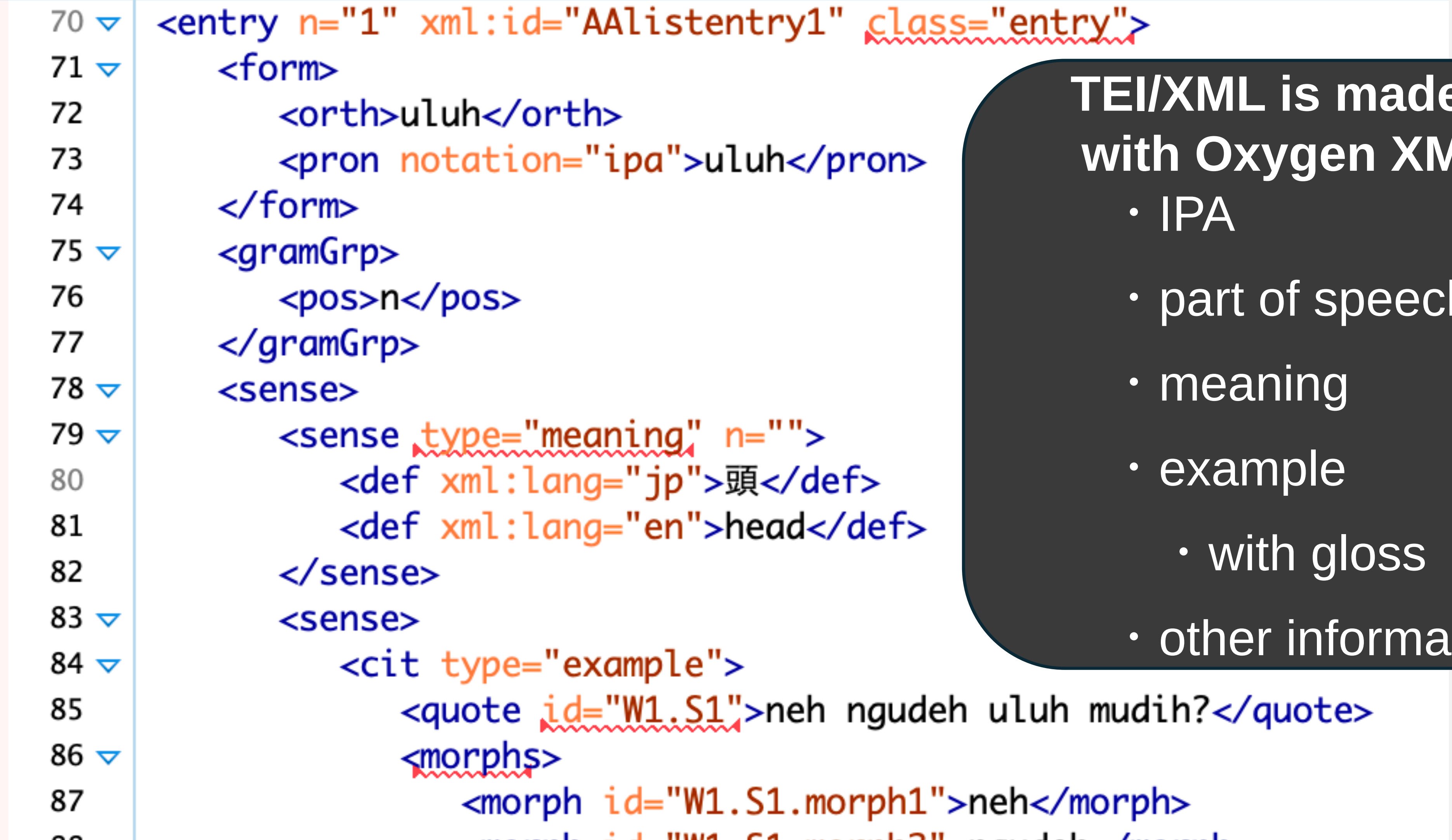
```
<?xml version="1.0" encoding="UTF-8" standal
<lexicon xmlns:xsi="http://www.w3.org/2001/X
    <header>
        <name>Kelabit_lexicon</name>
        <language>Kelabit</language>
    </header>
    <entry id="e_df7e5c06-facf-4c0b-a552-f5d
        <lexical-unit>na?am</lexical-unit>
        <morph-type>stem</morph-type>
        <sense id="s_06c06f78-2104-433c-87e1
            <grammatical-category>adv</gramm
            <gloss>NEG</gloss>
        </sense>
    </entry>
```

```
<entry>
    <lexical-unit>...
    <morph-type>...
    <sense>
        <grammatical-category>...
        <gloss>...
    </sense>
</entry>
```

NOT suitable for ODK

TEI/XML for ODK (TEI Guideline Ch. 10 Dictionaries)

https://tei-c.org/release/doc/tei-p5-doc/en/html/DI.html



```
<entry n="1" xml:id="AAlistentry1" class="entry">
    <form>
        <orth>uluh</orth>
        <pron notation="ipa">uluh</pron>
    </form>
    <gramGrp>
        <pos>n</pos>
    </gramGrp>
    <sense>
        <sense type="meaning" n="">
            <def xml:lang="jp">頭</def>
            <def xml:lang="en">head</def>
        </sense>
        <sense>
            <cit type="example">
                <quote id="W1.S1">neh ngudeh uluh mudih?</quote>
                <morphs>
                    <morph id="W1.S1.morph1">neh</morph>
                    <morph id="W1.S1.morph2">ngudeh</morph>
```

TEI/XML is made by hand with Oxygen XML editor.

- IPA
- part of speech
- meaning
- example
    - with gloss
- other information

The process of creating ODK

- **The process of making lexical data for the dictionary**

  Raw data → transcription (ELAN)

  → transformation to TEI/XML (Oxygen)

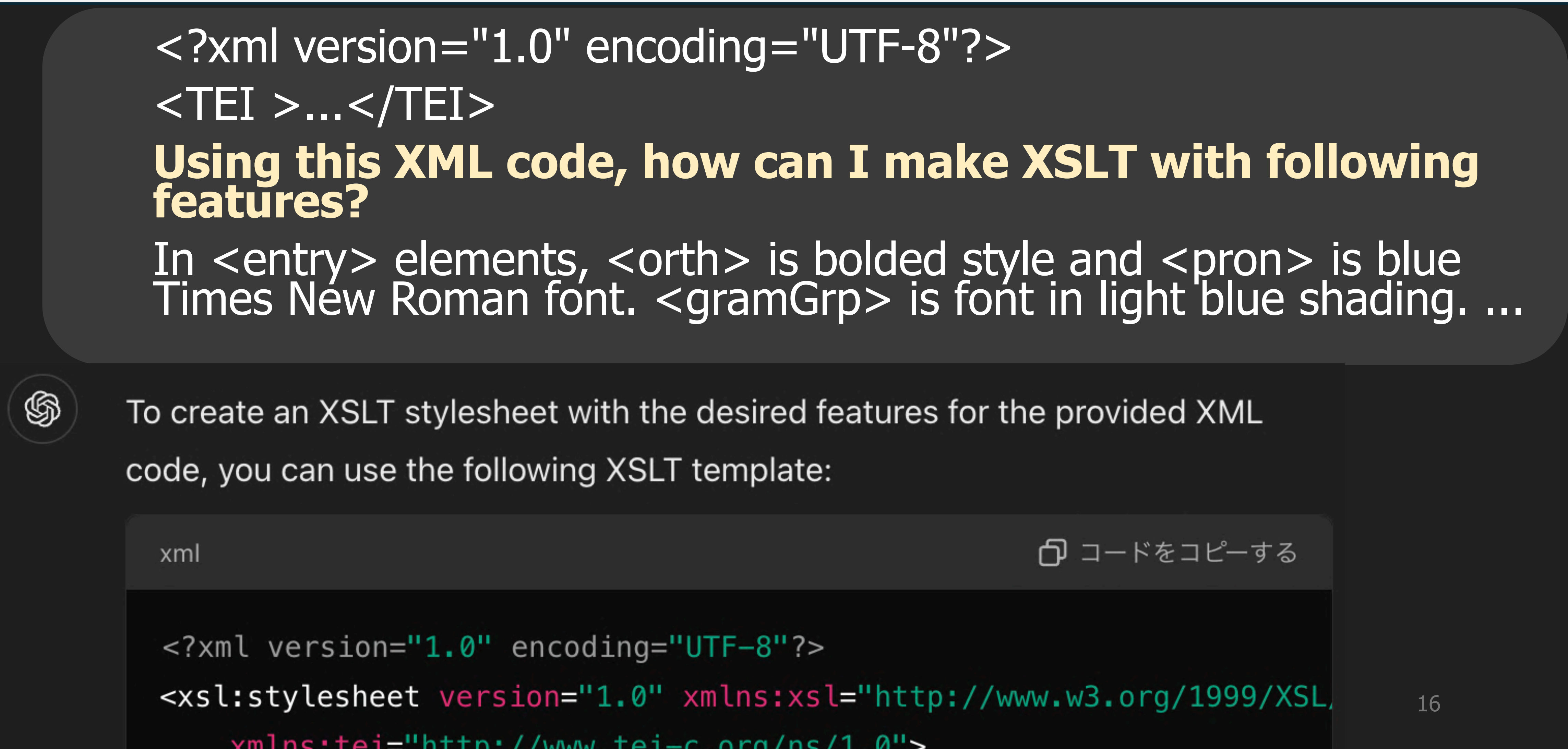
- **The process of getting the output in different formats**

  XSLT(ChatGPT) → XML to HTML (Oxygen)

  → search function in JavaScript (ChatGPT)

Creating XSLT with ChatGPT

ChatGPT recommended XSLT over CSS.



<?xml version="1.0" encoding="UTF-8"?>
<TEI >...</TEI>

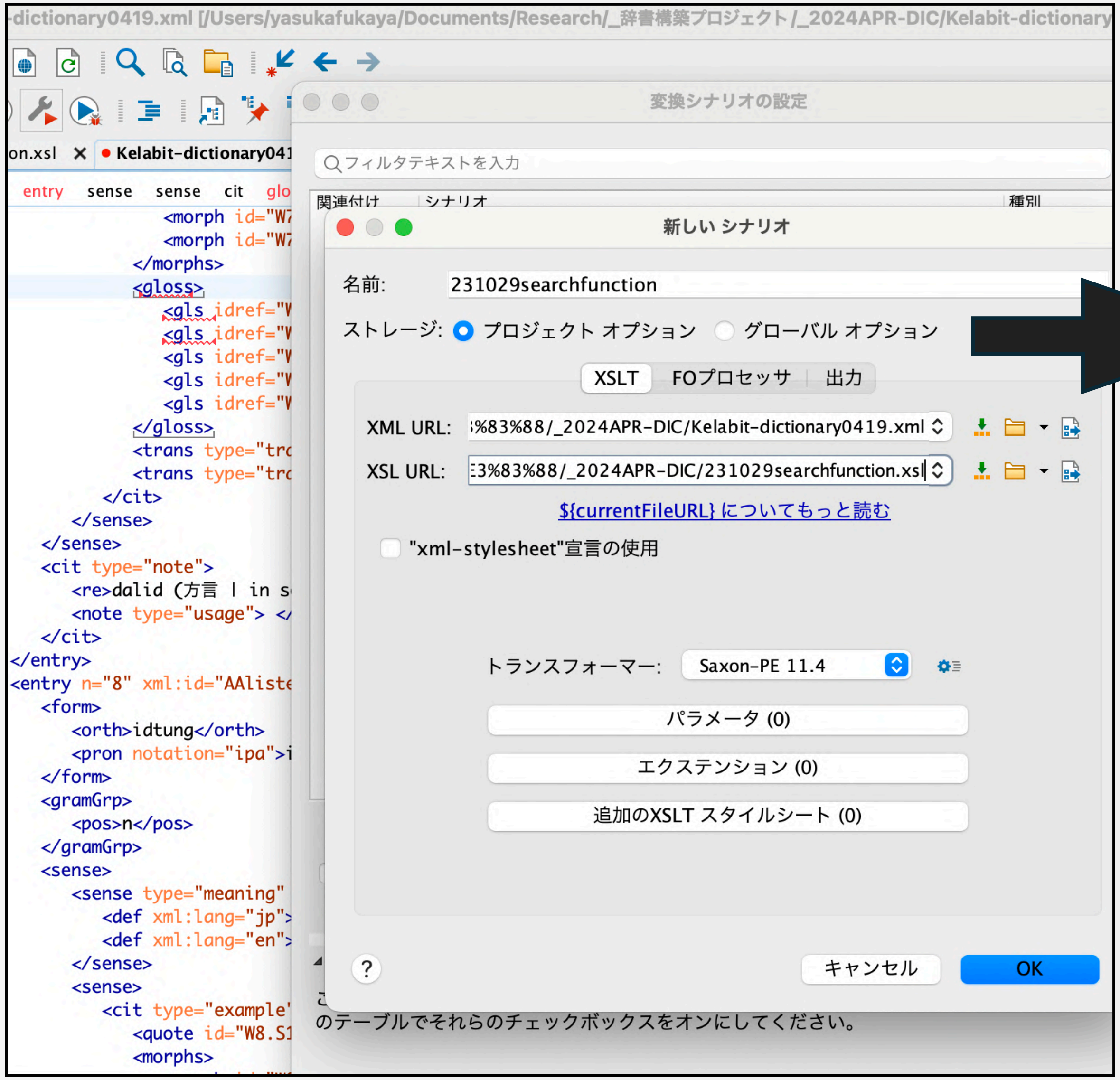
Using this XML code, how can I make XSLT with following features?

In <entry> elements, <orth> is bolded style and <pron> is blue Times New Roman font. <gramGrp> is font in light blue shading. ...

To create an XSLT stylesheet with the desired features for the provided XML code, you can use the following XSLT template:

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/
    xmlns:tei="http://www.tei-c.org/ns/1.0">
```

XML to HTML



-dictionary0419.xml [/Users/yasukafukaya/Documents/Research/_辞書構築プロジェクト/_2024APR-DIC/Kelabit-dictionary

```
            <morph id="W7
            <morph id="W7
         </morphs>
         <gloss>
            <gls idref="V
            <gls idref="V
            <gls idref="V
            <gls idref="V
            <gls idref="V
         </gloss>
         <trans type="tro
         <trans type="tro
      </cit>
    </sense>
  </sense>
  <cit type="note">
    <re>dalid (方言 | in s
    <note type="usage"> </
  </cit>
</entry>
<entry n="8" xml:id="AAliste
  <form>
    <orth>idtung</orth>
    <pron notation="ipa">i
  </form>
  <gramGrp>
    <pos>n</pos>
  </gramGrp>
  <sense>
    <sense type="meaning"
      <def xml:lang="jp">
      <def xml:lang="en">
    </sense>
    <sense>
      <cit type="example"
        <quote id="W8.S1
        <morphs>
```

変換シナリオの設定

フィルタテキストを入力

新しい シナリオ

名前: 231029searchfunction

ストレージ: ● プロジェクト オプション ○ グローバル オプション

XML URL: %83%88/_2024APR-DIC/Kelabit-dictionary0419.xml

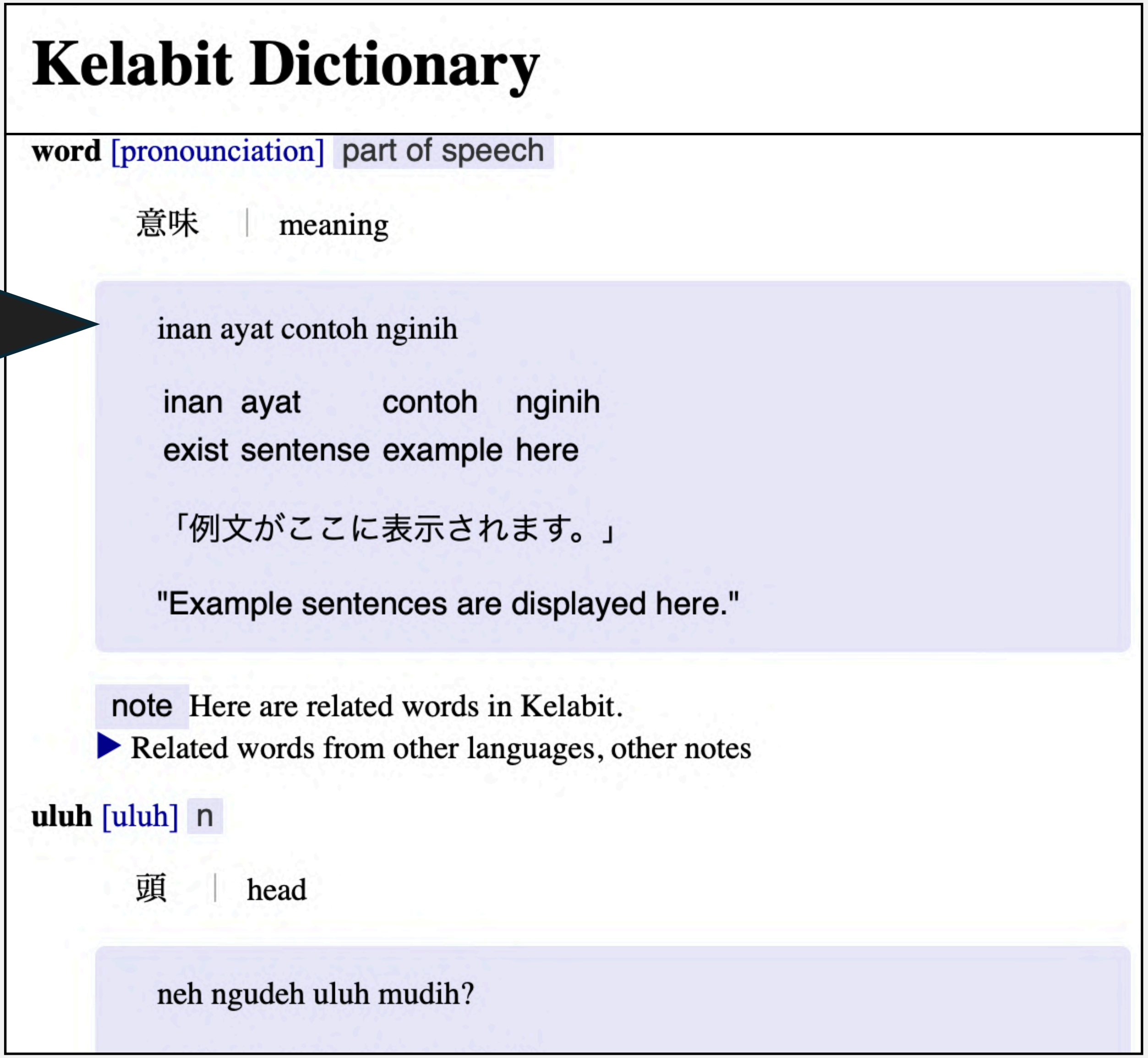
XSL URL: 3%83%88/_2024APR-DIC/231029searchfunction.xsl

${currentFileURL} についてもっと読む

□ "xml-stylesheet"宣言の使用

トランスフォーマー: Saxon-PE 11.4

のテーブルでそれらのチェックボックスをオンにしてください。



Kelabit Dictionary

word [pronounciation] part of speech

意味 | meaning

inan ayat contoh nginih

inan	ayat	contoh	nginih
exist	sentense	example	here

「例文がここに表示されます。」

"Example sentences are displayed here."

note Here are related words in Kelabit.

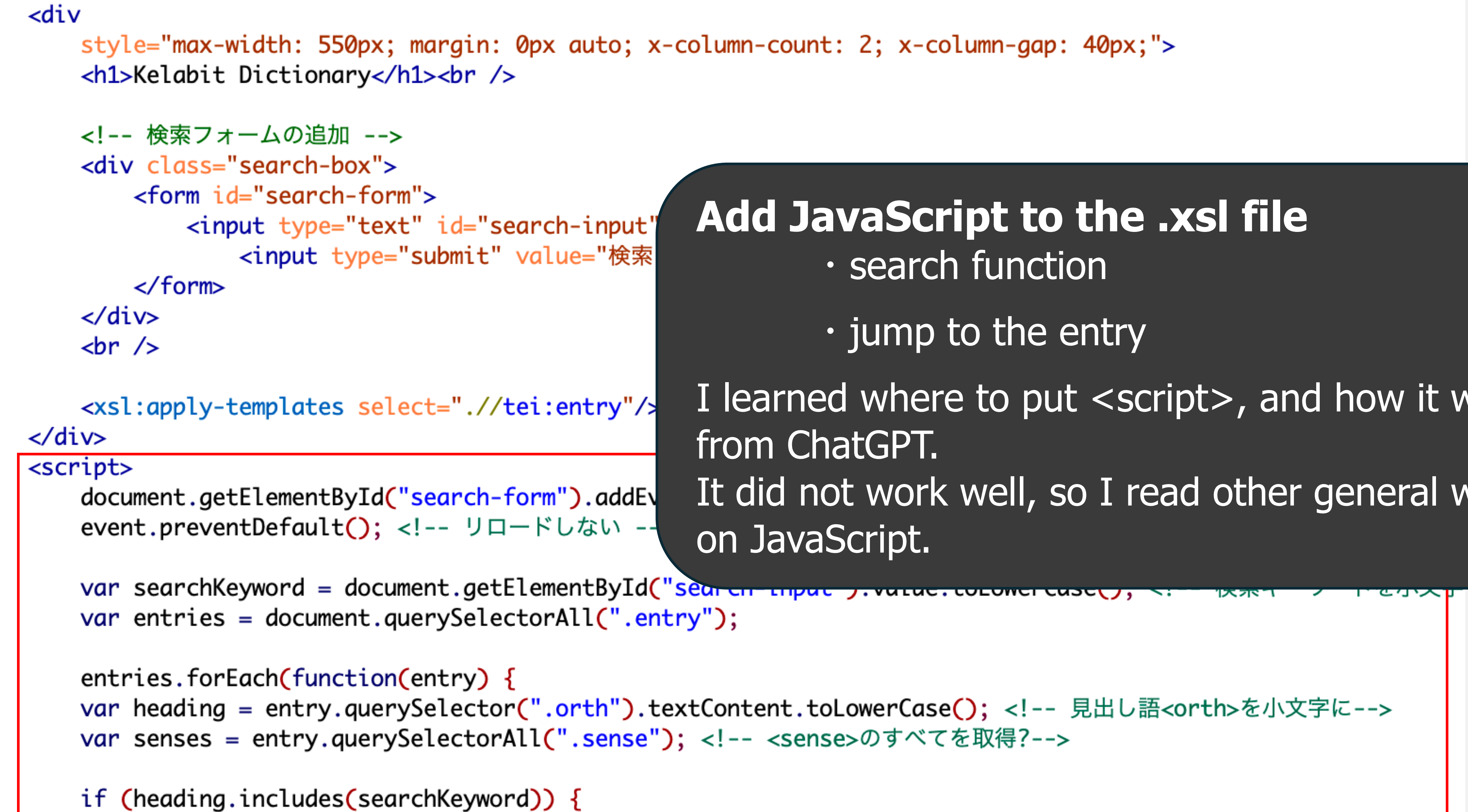
▶ Related words from other languages, other notes

uluh [uluh] n

頭 | head

neh ngudeh uluh mudih?

Add a search function made by ChatGPT

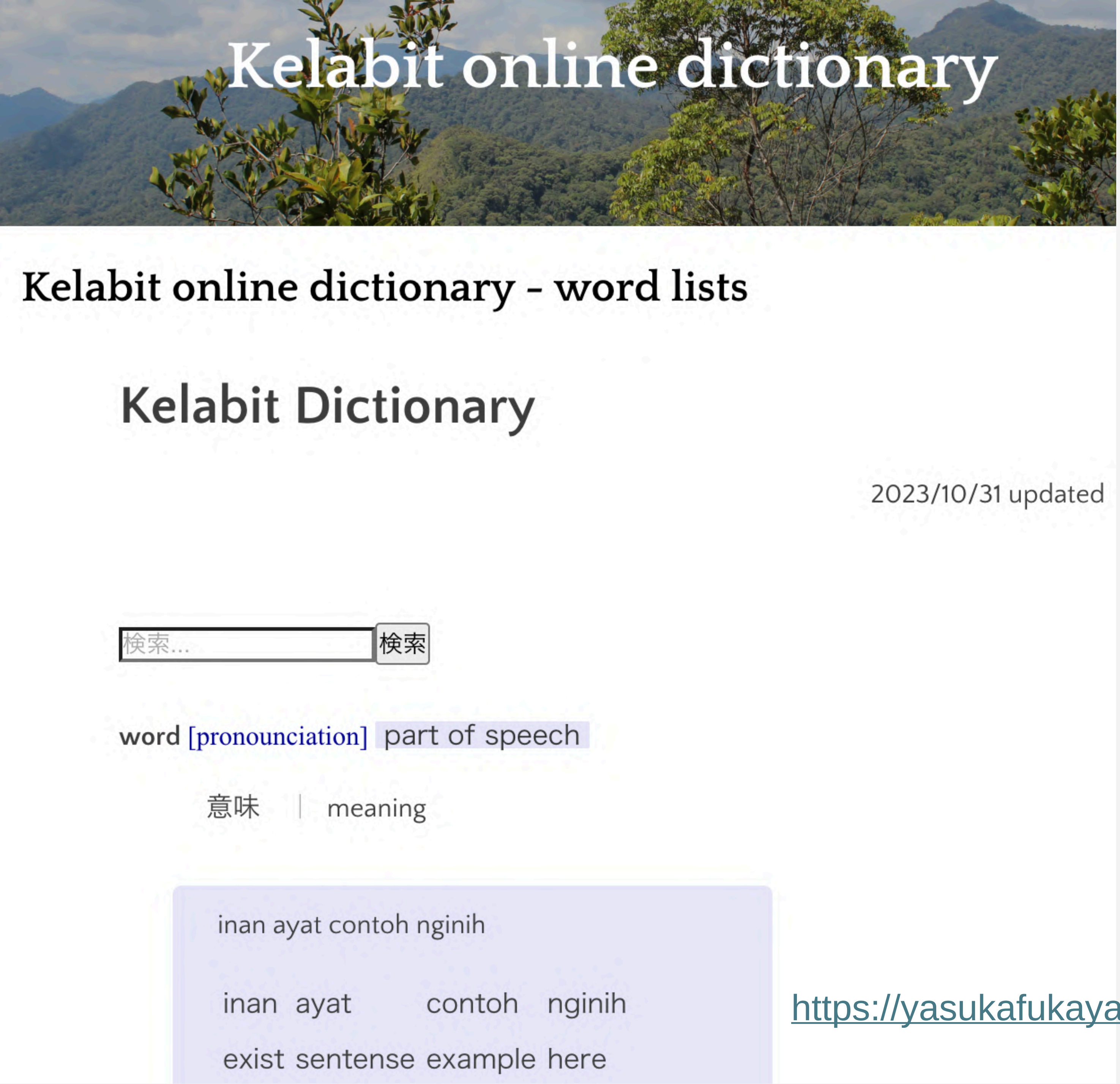


```
<div
    style="max-width: 550px; margin: 0px auto; x-column-count: 2; x-column-gap: 40px;">
    <h1>Kelabit Dictionary</h1><br />

    <!-- 検索フォームの追加 -->
    <div class="search-box">
        <form id="search-form">
            <input type="text" id="search-input"
                <input type="submit" value="検索
        </form>
    </div>
    <br />

    <xsl:apply-templates select=".//tei:entry"/>
</div>
<script>
    document.getElementById("search-form").addEv
    event.preventDefault(); <!-- リロードしない --

    var searchKeyword = document.getElementById("search-input").value.toLowerCase(); <!-- 検索キーワードを小文字
    var entries = document.querySelectorAll(".entry");

    entries.forEach(function(entry) {
    var heading = entry.querySelector(".orth").textContent.toLowerCase(); <!-- 見出し語<orth>を小文字に-->
    var senses = entry.querySelectorAll(".sense"); <!-- <sense>のすべてを取得?-->

    if (heading.includes(searchKeyword)) {
```

Add JavaScript to the .xsl file

- search function
- jump to the entry

I learned where to put <script>, and how it works from ChatGPT.
It did not work well, so I read other general websites on JavaScript.



Kelabit online dictionary

Kelabit online dictionary - word lists

Kelabit Dictionary

2023/10/31 updated

word [pronounciation] part of speech

意味 | meaning

inan ayat contoh nginih

inan ayat contoh nginih
exist sentense example here

Present version

limited words and function

Future goal

Better way (xml, xsl…)

Photos and sounds

https://yasukafukaya.weebly.com/kelabit-dictionary.html

I would like to express my special thanks of gratitude to my language consultants and Kelabit community who supports me.

Terima kasih mula'-mula'!