## P-7　ケラビット語のオンライン辞書構築についての経過報告:
## フィールドワークのデータから、話者コミュニティと研究者が使用できる辞書を目指して*

深谷 康佳（広島大学）

**要旨：**

本発表では、構築中であるケラビット語のオンライン辞書についての経過報告を行う。ケラビット語はオーストロネシア語族に属するマレーシア、ボルネオ島のサラワク州で話されている話者数 6,000 人以下の言語である。ケラビット語については、これまでにいくつかの辞書ないし語彙集が作成されてきたが、それらは話者コミュニティで使用されていない。本発表ではケラビット語の辞書構築に関わる背景や現状、問題点をまとめ、辞書作成のためのデータ収集方法と、それらを用いて辞書を構築する過程について述べる。そして、現状のデータで仮作成したケラビット語オンライン辞書に含まれる項目と使用方法について紹介し、話者コミュニティへの還元を兼ねた研究である辞書構築を行うに当たり、現在問題として生じている点について議論を行う。

### 1.　はじめに

第 1 節では前提として、本研究の対象言語であるケラビット語の概要と、本稿の構成について述べる。

ケラビット語はオーストロネシア語族、西マライ・ポリネシア語派に属する、マレーシアのボルネオ島、サラワク州北部のバリオ村を中心に居住するケラビット人により話されている言語である(図 1 )。

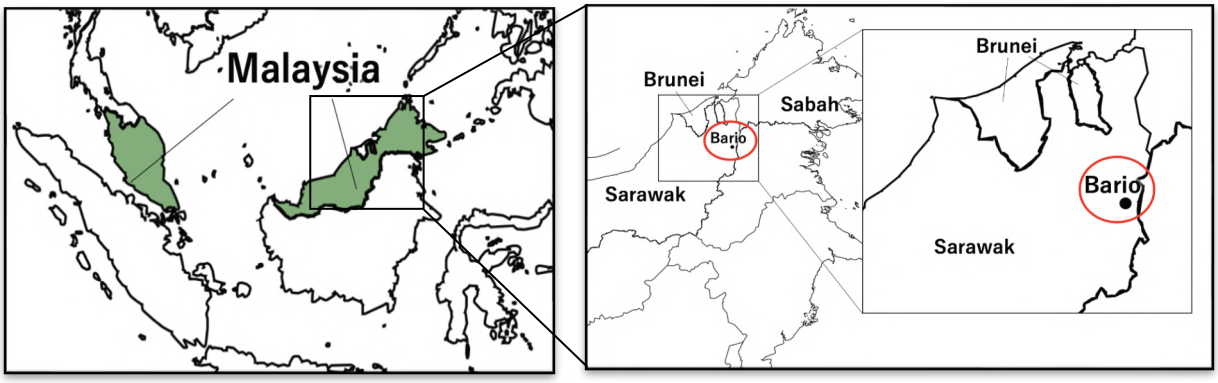


図 1：マレーシア、ボルネオ島、バリオ村の地図

ケラビット人の人口は 5000～6000 人程度と言われているが(Amster 1995)、大多数は都市部に居住している。現在、ケラビット人では、マレー語や英語への言語推移が進行しており、話者の多くは都市部に移住して日常的に英語等を話している。3,40 代はケラビット語を理解できない人が多く，理解できる人も運用能力に自信がないという人が多い。UNESCO はケラビット語を「消滅の危機が差し迫った（definitely endangered）言語」として認定している。

本稿の構成は以下の通りである。第 2 節では、これまでに作成された語彙集の概要と問題点について述べる。第 3 節では、これまでに行なった調査の概要と収集したデータを提示し、そのデータを用いて辞書を構築する過程について述べる。第 4 節では現状のデータから仮作成した辞書を示し、第 5 節ではそれまでの内容をまとめ、今後の展開についても触れる。

* 本研究は三島海雲記念財団の助成を受けた。

## 2. 先行研究とその問題点

第2節では、まず、これまでに作成された辞書ないし語彙集の概要と問題点について述べ、次にケラビット語がどの程度記述されているのかという状況を示すために、これまでに行われたケラビット語を対象とする研究についてまとめる。

### 2.1. 既存の語彙集とその問題点

ケラビット語の語彙集として代表的なものに、Amster(1995)がある。しかし、この語彙集は危機言語を半永久的に残していくための言語資料として問題がある。1点目は、この語彙集にだけ使われる特殊な記号で音声表記がなされていることである。言語保存のための辞書は、通言語的に標準の表記である国際音声字母(IPA)での音声表記が望ましい。2点目は、この語彙集では類義語に対する語義が大雑把に記述され、微細な使い分けがわからなくなっていることである。3点目に、多くの語に例文がなく、例文がある場合もその内容に対して構成要素の注釈はなく、対訳のみが記されている。よって、実際に使用にあたって、正確な使い方を知ることが難しいという問題がある。他にも、Janowski(2014)はケラビットの文化を記した *Tuked Rini* の巻末に語彙集を付けて出版している。Blust(1993)は祖語再建の手がかりとするため、ケラビット語–英語の辞書を作成している。深谷(2022)には記述文法の巻末に400語程度の語彙集が付いている。これらの語彙集には例文が少なく、類義語の微細な使い分けも明らかでない。アクセス・入手も困難であり、上記のうちウェブ上や書店で誰でも入手可能なものはJanowskiによる書籍ぐらいである。他の語彙集は著者や著者の協力者と知り合いでもない限り入手することは困難である。

### 2.2. ケラビット語を対象とした先行研究の状況

ケラビット語を対象とした主要な研究は、多くない。Blust (2006)はケラビット語の有声帯気音（Voiced Aspirates）について主に音韻論的観点からその存在を論じている。Hemmings (2016)は、Blust (2006) に基づきケラビット語の文法スケッチを行い、その上でヴォイス体系を明らかにすることを焦点として博士論文を書いている。また、Hemmingsはケラビット語の情報構造についても他のボルネオ諸語と比較し研究しており、その資料となるケラビット語の音声、動画をウェブサイト上で公開している。深谷(2022)では、ケラビット語の言語内的一貫性と言語個別性に基づき、ケラビット語の音韻体系から形態、基本語順や単文の構成など、体系を一通り記述した。このように、複数の研究が存在しているが、根拠となるデータが公開されているものはHemmingsの情報構造のもののみであり、これにはケラビット語の表記にグロスなどの文法を説明する要素は付けられていない。また、ケラビット語には「有声帯気音」と呼ばれる音声があるか否かについて議論があるが、音声の公開はされていない。音声やグロスつき例文を含むオンライン辞書を公開する事により、このような音声、音韻に関わる特徴や、その他の記述文法の項目についても検証することが可能となる。

このように、ケラビット語を対象とした研究は少なく、言語の保存、記述が十分に達成されないまま、1節で述べたように、話者が急速に減少し、消滅の危機に瀕している。現状では、既存の辞書の修正箇所について、作成者へ話者がコメントをしても返事がなく、改訂版も作成できないと言う話者もいる。研究成果を還元するということは辞書を作って終わりなのではないはずである。言語保存は話者が望んでこそ成し得る目標であり協働は欠かせない。本研究の長期的な最終目標は、話者コミュニティが持続的に修正・更新していけるオンライン辞書を作成することである。

## 3. 研究方法

第 3 節では、まず、これまでに発表者が行ってきた調査のうち、辞書作成に関わりのあるものの概要について述べ、収集したデータを例として示す。そして、それらのデータを用いて辞書を構築する過程について述べる。

### 3.1. 調査概要

発表者はこれまで、2014 年から 2023 年にかけて、通算で約 4 ヶ月間程度、バリオ村での調査を行ってきた。特に、2014 年に行った初回、第 2 回の調査では、ケラビット語バリオ方言の音素体系を明らかにするための基礎語彙収集調査を行い、Swadesh list (Swadesh 1955)の 206 語に生活の中で使用されていた語彙を加え、全部で 400 語の基礎語彙の音声を収集した。その後は 2015 年から 2019 年まで、文法の全体像を記述するための調査を行っており、数編の歌や語彙の用例等を音声、フィールドノートに記録している。2019 年以降は 1 度 Covid-19 の感染拡大を受け、調査をすることはできなかったが、2023 年に再度調査を行うことができた。その際には、身体語彙の一部とその例文のデータと、生活に関わる野菜や、文化に関わるビーズワークについての語彙や例文、モノローグのデータを、音声やフィールドノートに記録している。

本発表では上記の調査により収集したデータで辞書を仮作成するよう努めた。しかし、第 2 節で述べたような、語彙の IPA 表記、音声、グロス付き例文等の要素を含む辞書を作成するためには、上記の調査により収集したデータでは不十分である。今後、例文・音声付き辞書作成のために、2014 年から 2019 年の間に収集した基礎語彙の音声、用例や歌などのデータの他にも、2023 年に収集したような語彙と例文、モノローグをより多く収集する必要があるだろう。さらに、複数人による自然談話を収集し、辞書に例文として掲載することで、語彙の自然な使用例や類義語の微細な使い分けを辞書から検索することが可能になるだろう。

### 3.2. 辞書構築方法

3.1 のデータを用いて、辞書を仮作成する方法について述べる。本研究における辞書構築には大きく分けて 2 つの段階がある。 1 つ目は 3.2.1 で触れる TEI(Text Encoding Initiative)を用いた語彙の情報のコード化である。 2 つ目は 3.2.2 で述べる TEI を出力するためのスタイルを決める XSLT である。

#### 3.2.1. TEI による語彙情報のコード化

TEI は上で述べたように Text Encoding Initiative の略語である。これは TEI 協会が策定する、人文学を中心とするテキスト構造化のためのデータ形式であり、現在は XML(Extensible Markup Language)技術に基づき実装されている（石田ほか編 2022: 418)。TEI にはデジタル資料の構造的記述の標準的な形式を提案する「TEI ガイドライン」(https://tei-c.org/guidelines/）が存在している（永崎 2022)。TEI ガイドラインは第 1 章から第 23 章まであり、1 章から 5 章までは様々な種類の資料に関わる、全体の構造やヘッダ、タグ、文字などの基本的な情報を示している。第 6 章からは、デジタル資料の種類に合わせた構造的記述の方法が提示されている。第 9 章は辞書のデータの構造的記述の方法が提示されている。

第 9 章の構成は次に示す通りである。

9 Dictionaries

Table of contents



図 2: TEI Guidelines 9 章の構成

TEI Guidelines の 9 章において、9.1 では辞書の全体の構成について、標準的な記述方法が示されている。9.2 では辞書に含まれる語彙のエントリーの記述方法の例を複数提示している。具体的には、表記、音声、品詞、意味などをどのように記述したら良いか、例を挙げている。9.3 では 9.2 で扱った基本的な項目に更に情報を加える場合、どのように記述すべきか例を示している。9.4 では見出し語への参照をどのようにコード化するのかについて述べられており、9.5 はタイポグラフィックの情報をどのようにコード化するのかについて述べられている。9.6 では、これまでに示した方法では記述できない構造の辞書をコード化するためにどのように<entryFree>や<dicrScrap>といったコードを用いるかについて示している。9.7 では辞書モジュールとして、辞書のテキストデータ化に関するコードをまとめている。

本研究においては、すでに存在している辞書を TEI 化するのではなく、書籍になっていないケラビット語のデータを辞書としてアウトプットできるように語彙のデータを構築していく。そのため、現時点では 9.1〜9.4 までを参考にして基準に沿った辞書のためのデータを記述する。次節では、TEI 化したデータを出力するための形式に変換する方法として XSLT について概説する。

### 3.2.2. XSLT スタイルシートによる TEI データの出力

XSLT(Extensible Stylesheet Language Transformations)は XML 文書を処理・成形する有力な手法の一つである。XML データを変換するための XSL(Extensible Stylesheet Language)という言語を用いて XML を任意の形式に変換する（石田ほか編 2022: 418)。TEI を用いた XML 文書に辞書内の項目の情報をまとめることで、XSLT により HTML や PDF に変換し、複数の形式にアウトプットできる。辞書の項目に修正を加える場合は、XML 文書を修正することで、複数のアウトプットを効率的に修正することができる。

## 4. 現状のデータによるオンライン辞書

本節では、現状のデータでオンライン辞書を仮作成した結果を過程とともに示す。4.1 では、3.1 で示した語彙の情報を、3.2 で紹介した TEI 形式を用いて整形したものの一部を例として示す。次に、4.2 では、XSLT を用いて TEI 形式の XML 文書を HTML に出力したものの一部を示す。

### 4.1. TEI 形式を用いた語彙の情報のコード化

本研究では、これまでに収集した語彙のデータをもって、TEI 形式の XML 文書を作成する。先述の TEI Guidelines の第 9 章を参考として作成した XML 文書の一部を下に示す。

```
<entry n="0" xml:id="AAlistentry1">
   <form>
      <orth>表記1</orth>
      <pron notation="ipa">発音記号</pron>
   </form>
   <gramGrp>
      <pos>品詞</pos>
      <pos>part of speech</pos>
   </gramGrp>
   <sense>
      <sense n="">
         <def xml:lang="jp">意味</def>
         <def xml:lang="en">meaning</def>
      </sense>
      <sense>
         <cit type="example">
            <def style="ex">
               <quote>例文</quote>
            </def>
            <cit type="trans" xml:lang="jp">
               <quote>例文の日本語訳, </quote>
            </cit>
            <cit type="trans" xml:lang="en">
               <quote>example</quote>
            </cit>
         </cit>
      </sense>
   </sense>
   <cit type="note">
      <re>関連語彙, related words in Kelabit</re>
      <note type="usage">note for usage, related word in other languages</note>
   </cit>
</entry>
```

図 3: ケラビット語辞書の XML 文書の一部（例の部分）

<entry>の内部には、<form>、<gramGrp>、<sense>、<cit>が存在している。<form>内に表記、発音記号、<gramGrp>には品詞についての情報が日本語と英語で示され、<sense>内には意味と例文を示す。<cit>内には関連する語彙や語彙の使用例、状況などについての補足情報を示す。

### 4.2. XSLT を用いた TEI 形式の出力

4.1 の XML 文書を辞書の形式で出力するために、XSLT ファイルを用いる。XSLT については GitHub サイトから、Michal Měchura により作成された、TEI Guidelines の第 9 章を基準として様々な辞書の形式を表出するためのスタイルシートである tei-dictionary.xsl（https://github.com/michmech/tei-dictionary.xsl）を参考に作成した。XSLT ファイルの内容の一部を以下に示す。

```
<!--top-level entry element: basically just a paragraph-->
<xsl:template match="tei:superEntry | tei:entry | tei:entryFree">
  <span style="display: block; line-height: 1.5em; margin: 1.5em 0 1.5em -1em; text-indent: -1em;" xml:space="preserve">
          <xsl:apply-templates/>
      </span>
</xsl:template>

<!--subentry with grey left border and indent-->
<xsl:template match="tei:superEntry/tei:entry | tei:superEntry/tei:entryFree | tei:hom">
  <span style="display: block; line-height: 1.5em; margin: 0.35em 0 0 0; border-left: 5px solid #dddddd; padding: 0.25em 0 0.25em 1em; text-indent: 0em;">
    <xsl:apply-templates/>
  </span>
</xsl:template>

<!--subentry with bullet and indent-->
<xsl:template match="tei:re">
  <span style="display: block; margin: 0.5em 0 0.5em 0.5em; min-height: 1.5em; text-indent: 0em;">
    <span style="float: left; font-weight: bold; color: #666666">»</span>
    <span style="display: block; margin-left: 1.5em;"><xsl:apply-templates/></span>
  </span>
</xsl:template>
```

図 4: XSLT ファイルの例の一部

図 4 では、XML 文書の各項目をどのようなフォーマットで出力するのか指示をしている。この指示に従って出力された HTML の形式の一部を示す。

```
<span style="font-family: sans-serif;">表記1</span>
<span style="font-family: sans-serif; color: #006600">[発音記号]</span>

<span style="font-family: sans-serif; color: #000066; background-color: #eeeeee; padding: 1px 5px; border-radius: 2px;">品詞</span>
<span style="font-family: sans-serif; color: #000066; background-color: #eeeeee; padding: 1px 5px; border-radius: 2px;">part of speech</span>

<span style="display: block; margin: 0.5em 0 0.5em 0.5em; min-height: 1.5em; text-indent: 0em;"><span style="float: left; font-weight: bold; color: #999999">▪</span><span style="display: block; margin-left: 1.5em;">
  <span style="display: block; margin: 0.5em 0 0.5em 0.5em; min-height: 1.5em; text-indent: 0em;"><span style="float: left; font-weight: bold; color: #999999"></span><span style="display: block; margin-left: 1.5em;">
    意味 <span style="color: #999999">|</span>
    meaning
  </span></span>
  <span style="display: block; margin: 0.5em 0 0.5em 0.5em; min-height: 1.5em; text-indent: 0em;"><span style="float: left; font-weight: bold; color: #999999">▪</span><span style="display: block; margin-left: 1.5em;">
    <span style="display: block; margin: 0.5em 0 0.5em 0.5em; min-height: 1.5em; text-indent: 0em;"><span style="float: left; font-weight: bold; color: #666666">◆</span><span style="display: block; margin-left: 1.5em;">
      例文
      <span style="color: #999999">|</span>
      例文の日本語訳,
      example
    </span></span>
  </span></span>
</span></span>

<span style="display: block; margin: 0.5em 0 0.5em 0.5em; min-height: 1.5em; text-indent: 0em;"><span style="float: left; font-weight: bold; color: #666666">»</span><span style="display: block; margin-left: 1.5em;">関連語彙, related words in Kelabit</span></span>
note for usage, related word in other languages
```

図 5: 出力された HTML 形式の文書の一部

以上のような流れで、XSLT を用いて XML 文書を HTML に変換している。実際に画面に表示されるのは、以下の図に示すようなものとなる。

表記1 [発音記号] 品詞 part of speech

意味 | meaning

◆ 例文 | 例文の日本語訳, example

» 関連語彙, related words in Kelabit

note for usage, related word in other languages

図 6: 実際に出力される冒頭の例の画面の一部

以上、図 3 から図 6 のように、語彙のデータを XML 文書から辞書として出力する過程を示した。

## 5. まとめ

現状のデータでは、上記のように仮作成したケラビット語オンライン辞書に含まれる項目はまだ約 50 語である。現時点では音声や画像、例文の構成など修正が必要な箇所はあるものの、TEI を用いて更新を進めていく予定である。音声、例文が掲載された辞書が完成した際には、これまで議論されていた「有声帯気音」やヴォイス体系などの先行研究の妥当性についても検証することが可能となるだろう。

文法記述を行う際には、循環的プロセスとして、テキストの収集と注釈付け、語彙の収集作業も並行して行われる(下地 2013)。本辞書を作成する過程でも文法記述を進め、また、記述した文法の根拠として参照できるテキストから、辞書の例文を抽出する。このような循環的なプロセスを繰り返すことにより、言語記述・保存にとって理想的な情報量を備えた辞書となるだろう。

現時点における辞書作成の大きな問題点は、表記体系と例文の提示方法である。表記体系については、現地コミュニティにおいて主流の 2 つの表記体系が存在し、採用しなかった片方の表記を使用する話者とどのようにコミュニケーションをとるかが問題である。例文の提示方法についてはグロス付き例文のグロスの基準をどのように定めるのか、また、その例文をどのようにコード化し出力するのが最適であるかが問題である。

## 参考文献・参考 Web サイト


Amster, Matthew H. (1995). *Kelabit/English, English/Kelabit glossary: a concise guide to the Kelabit language*: Rurum Kelabit Sarawak.

Blust, R. A. (2006). “The Origin of the Kelabit Voiced Aspirates: A Historical Hypothesis Revised,” Vol. 2, No. 45, pp. 311–338.

Blust, R. A. (1993). Kelabit-English vocabulary. Sarawak Museum Journal, 44(65), 141-226.

深谷 康佳. (2022). 「ケラビット語バリオ方言の記述文法」博士論文. 広島大学.

Hemmings, C. (2016). The Kelabit language, Austronesian voice and syntactic typology. (Unpublished doctoral dissertation). SOAS, University of London, England.

石田 友梨, 大向 一輝, 小野 綾乃, 永崎 研宣, 宮川 創, 渡邉 要一郎 (編) (2022) 『人文学のためのテキストデータ構築入門 TEI ガイドラインに準拠した取り組みにむけて』, 文学通信

Janowski, Monica (2014). *Tuked Rini: cosmic traveller: life & legend in the heart of Borneo*: NIAS Press

永崎 研宣. (2022). 「利活用演習：TEI 準拠テキストの活用方法」, 石田 友梨, 大向 一輝, 小野 綾乃, 永崎 研宣, 宮川 創, 渡邉 要一郎 (編) 『人文学のためのテキストデータ構築入門 TEI ガイドラインに準拠した取り組みにむけて』, 180-198, 文学通信.

下地 理則 (2013)「フィールドワークと辞書」『日本語学』臨時増刊号 32 (14): 1-15.


### 参考ウェブサイト


TEI Consortium (2023). “Guidelines — TEI”, <Text Encoding Initiative>. 2023-04, https://tei-c.org/guidelines/, (最終アクセス 2023.05.10).

Michal Měchura (2017). “An XSLT stylesheet for TEI-encoded dictionaries”, 2017-12, https://github.com/michmech/tei-dictionary.xsl, (最終アクセス 2023.5.10).