Language Documentation & Conservation Special Publication No. 27 (2022)
Voices: Perspectives from the International Year of Indigenous Languages
ed. by Sarah Sandman, Shannon Bischoff, and Jens Clegg pp. 117-135
http://nflrc.hawaii.edu/ldc/
http://hdl.handle.net/10125/74693

Making Dictionaries of Lesser-Known Indigenous languages: Coding of Lexical Semantic Information

Umarani Pappuswamy
Central Institute of Indian Languages, India

1. Introduction

Several scholars predict that almost 90% of the world's languages will disappear by the end of the 21st century, thereby reducing the number of living languages from about 7000 to 700 or so. Interestingly, 96% of the world's languages are spoken by only 4% of its people with about 1500 languages having less than 1000 speakers (Turin 2005). The languages that are widely spoken by these 4% of people are usually the 'big' languages which are either official languages of a nation/State or a dominant language of a region that is socio-politically significant. The immediate threat is to the so-called 'small' languages that do not have any official status anywhere. The UNESCO Atlas of the World's Languages in Danger (Moseley 2010) shows that 2471 languages are endangered, out of which 197 of them are spoken in India making it the country with the most endangered languages of the world.[1] *Ethnologue* (Eberhard et al 2020) lists 2,895 languages as endangered whereas the *Catalogue of Endangered Languages* developed (Campbell & Belew 2018) has identified 3,394 as endangered. These numbers are disputable because there is no clear-cut distinction between language/dialect in many cases. Further, the criteria for endangerment varies from one study to the other. Thus, the number keeps changing. But the fact is languages are dying faster than before. This crisis affects both major and minor languages, the most vulnerable ones being the linguistic minorities or the 'small' languages.

The buzz phrase since 2019 heard everywhere was 'Year of Indigenous Languages' and now 'International Decade of Indigenous Languages'. Indigenous languages are those languages spoken by the Indigenous people,[2] the most disadvantaged and vulnerable groups of people in the world. Many of these languages are lesser-known and are the so-called 'smaller languages' spoken by 'smaller' communities that are at risk of disappearing at a faster rate. This has a direct impact on the intangible cultural heritage and their identity. Endangerment of these languages leads to language loss, which will eventually lead to an irreparable loss of Indigenous wisdom and knowledge.

In response to the language endangerment, efforts are being made all over the globe to protect, preserve and promote these Indigenous languages, societies, and cultures. One of the most common of them being language documentation, which is seen as a lasting, multipurpose record of a language which will not only be seen as data repositories for scientific inquiries, but also as important resources for supporting language maintenance (Himmelmann 2006). Many language documentation projects have been initiated by research institutions, government agencies, and individual researchers on language endangerment and revitalization that specifically refer to Indigenous language situations. This includes lexicographic work that may not necessarily result in dictionaries.

[1] The dynamics underlying the language endangerment in India are more complex than those prevailing in the rest of the world. See Pappuswamy (2019) for some discussion on this.
[2] The definition of Indigenous people is adopted from the United Nations document at https://www.un.org/development/desa/Indigenouspeoples/about-us.html.

Since the advancements in information and technology and with the arrival of machine-readable corpora, a fundamental shift of priorities has taken place in the lexicographic practices. Corpora have become some of the most reliable lexicographic resources for sense inventories. The number of corpus-based dictionaries is steadily growing. However, this change applies only to the making of dictionaries of well-resourced languages. Things are still grim when it comes to the compilation of dictionaries of Indigenous languages, which is still expensive and time-consuming. In most cases, there are no corpora, and many of these languages are still unwritten. Moreover, dictionaries are compiled usually by linguists or native-speakers or anthropologists with little or no knowledge in lexicography and often times end up as 'word lists' or 'vocabulary lists', which is indeed a serious concern to many of us.

This paper discusses the issues and challenges involved in creating dictionaries for Indigenous languages, drawing examples from Ruga, Atong, Garo and Khasi,[3] languages spoken in Meghalaya of North-East India that I worked on and have been associated with for several years. The main focus is on the design issues and the codification of lexical semantic information. §2 provides a brief overview of the lexicography and lesser-known Indigenous languages of India; §3 describes the design issues related to the macro- and microstructures and outlines the importance of making decisions about these double structures before compiling a dictionary; §4 focuses on the codification of lexical semantic information that includes providing definitions and explanations, sense inventory and coding sense relations, ordering of senses etc. and further demonstrates the need to incorporate new methods of data elicitation from lexicographic and community-focused perspectives. §5 presents conclusions.

2. Lexicography and lesser-known Indigenous languages of India

India is the 4th most linguistically diverse nation, with 453 living languages (Eberhard, Simons & Fening, 2019). There are 22 scheduled and 99 non-scheduled languages (Census of India 2011).[4] The Government of India has recognised 550 tribes as 'notified tribes' that constitute 8.6% of the total population. The languages spoken by them are referred to as 'tribal languages' in Article 342 of the Constitution of India. Except Bodo and Santali, none of them are scheduled languages. The tribes of India are predominantly concentrated in three distinct zones:

(i) North-East India
(ii) Central and East India comprising Madhya Pradesh, Bihar, Telangana, Andhra Pradesh, Jharkhand, Chattisgarh, Orissa and West Bengal
(iii) Southern India

Most of the tribes of North-East India are economically and socially well-placed, and are more thickly populated than the other two groups. All these communities speak a variety of languages that are under-resourced, less-explored, under-described, and have had very little

[3] Ruga, Atong and Garo are Bodo-Garo languages of the Tibeto-Burman family. Garo, the dominant language of the region, is the lingua franca among the Ruga and Atong speakers. Khasi is a Mon-Khmer language of the Austro-Asiatic family. All four languages are spoken in Meghalaya of the North-East of India. Ruga is a severely endangered language as it has only three fluent speakers, who are elderly, with one of them being hearing impaired. There is a large ethnic Ruga group living in Rugapura who are interested in reviving their ancestral tongue. As per Census of India (2011), Garo has 11,45,323 speakers while Khasi has 14,31,344. Population figures for Atong are not available in the Census data – there are no correct estimates of the number of speakers in other sources, but it is certainly less than 10000.

[4] The scheduled languages are the officially recognised languages included in the Eighth Schedule to the Constitution of India, and they are the major literary languages of the country. The non-scheduled languages are obviously less-privileged for a variety of reasons.

documentation done so far. These tribal languages are Indigenous languages in the sense that they are spoken by the most underprivileged and vulnerable groups of people.

In the modern era, though many documentation and revitalization projects have begun on these languages, societies, and cultures, there are few comprehensive works on dictionary making. The earliest dictionaries of some of these languages are those compiled by the missionaries. Some of these languages have just one or two dictionaries to represent their languages compared to the major scheduled languages of the country, while many others do not even have any writing system, and thus no dictionaries.

Dictionaries are one of the most important resources desired by these Indigenous language communities because they are, as rightly observed by Haviland (2006:129), the linguistic genre that is most familiar to the general public. Even speakers of unwritten languages see it as a matter of 'pride' and 'prestige' to have dictionaries in their languages in order to be on par with the others who possess such linguistic resources. Hence, in recent years, many communities approach linguists and other researchers to help them create dictionaries for their languages despite the fact that they may not find any significant role in the current educational system of India.

Making dictionaries of these Indigenous languages is more complex and often differs dramatically in scale and scope compared to those of major well-known languages of the country: firstly, because of a lack of existing dictionaries or corpora, and secondly, one has to start from scratch with limited resources of time, money and manpower. Oftentimes, dictionaries compiled in recent years in North-East India are off-shoots of the grammars and texts produced by field linguists and individual researchers. Besides that, since only a few people directly benefit from their publication, they have very limited production. The problem lies in the fact that people are paying attention to the instrumental value of these resources instead of seeing them as priceless treasures that document the worldview of a particular community or society.

Dictionary making is the area where linguists and the Indigenous community have the most divergent interests (Hinton & Weigel 2002). Native speakers are not just keepers of the language but are also culture bearers. The way they view dictionaries is different from that of the lexicographer or the linguist or the field worker compiling it. The community sees it as a storehouse of their Indigenous knowledge or wisdom to keep their cultural heritage in memory and transmit it to the younger generation. They are interested in seeing more encyclopedic information in their dictionaries than the linguists' dictionary. Thus, the focus of the dictionary user and usage is very important while making these dictionaries. The contemporary lexicography should consider the following aspects:

a) Anthropocentric aspect (for whom the dictionary is made);
b) Socio-pragmatic aspect (purpose of the dictionary); and
c) Cognitive aspect (essential skills necessary for the user).

These, of course, have certain idiosyncrasies when we speak of endangered, lesser-known Indigenous languages, especially the cognitive part. One cannot expect all the users to possess or acquire the essential skills to effectively use the dictionary compiled. Nevertheless, it is important to integrate the above three factors and produce reasonably good dictionaries.[5]

Historically, dictionaries for endangered languages have been mainly concerned with language maintenance. The whole idea was to document the lexical knowledge of a particular

[5] I have created a general-purpose lexicon in such a way that it can cater to different users (academic, community people, and second language learners) with multiple purposes and the dictionaries are being made as user-friendly as possible.

language for linguists and researchers interested in undertaking further research on them (Schebeck 1983). Hence, earlier days of lexicography in these languages saw a more encyclopedic documentation to represent as much knowledge as possible in a written form (Corris et al 2002). Though documentation is still not done for many of the languages of North-East India, the literacy rates of these communities are relatively increasing due to urbanization and globalization, and thus many of them are interested in developing orthographies which in turn can help them write down their languages and prepare dictionaries so that their language can also have a 'status' and be on par with the lingua franca of the region (for example, Atong or Ruga on par with Garo). The symbolic function of the dictionary is gaining more significance than the instrumental purpose, perhaps, in the context of identity crisis.

Until recently, the structure and the usability of dictionaries of Indigenous, endangered and lesser-known languages were planned and executed primarily for the benefit of academics and much less focus was given to the needs of the speakers of the language. Further, single/multiple author or compiler issues also crop up, which makes the process even more complex. Types of data elicitation also matters. The traditional method of eliciting words using mere word lists will not suffice the purpose. As stated by Evans (2010:109), there are hidden treasures of knowledge in many areas that reflect the culture and way of living of the Indigenous people, for instance, archaeological information that might not be important enough for linguists to gather vocabulary on, say, an arcane type of grindstone, distinctive pattern of pottery, etc. Thus, it is important to use other devices to elicit words. I have used several devices[6] such as extended culturally-specific wordlists, picture prompts, silent video clips, natural texts (sentences and free narratives), family talk and talk in community gatherings that would yield ethnocultural vocabulary from a community focused perspective for compiling dictionaries for Ruga-Garo-English, Atong-English, Khasi-English and Garo-English.

It is essential for lexicographic projects to carry out documentation of "the human perception of the natural world and survival strategies" (Harrison 2007) as well as "the fringes of linguistic diversity" (Wohlgemuth & Cysouw 2010). The esoteric and exoteric Indigenous knowledge should also be collected and the community should be consulted if that information should be made public or should be only in a special edition of a dictionary that would be made available only to select few members of the community. This is very important because some of the communities, for instance, the Tewa community, as noted by Brandt (1980), are particular that the sensitive information regarding social and ceremonial organization (the esoteric knowledge) be shared with only those who have both a right and a need to know. Such concerns should be respected and the dictionary should be created that may include the exoteric knowledge of the society that can be circulated to all without any inhibition.

Dictionaries, thus, serve as a guardian of the purity of the language, of language standards and of moral and ideological values of a community. In that case, how to extract the Indigenous knowledge? How much is enough to make a 'good' dictionary? What should be the design and structure of these dictionaries? How to determine the senses? What kind of ordering of senses should be done? What types of meanings should be included to explain words? There are no simple, straightforward answers to these questions, but I will reflect on

[6] I used several new methods to elicit data: (i) picture prompts wherein I used the pictures we clicked in their natural environment and asked them to describe the picture (free text) and also elicited words by pointing at the objects and other contents, (ii) Pear story, and a few native silent documentary video clips were used with different language consultants to elicit a variety of data and (iii) culture specific active elicitation from native speakers on several topics. I have given several online talks on data elicitation techniques, for instance, https://www.youtube.com/watch?v=KQFNwfEz6yI .

some of what I have learned from my experiences in dictionary making in the subsequent sections.

3. Design of double structures

A dictionary may consist of several main components such as the front matter, the central list and the back matter. It is the central list that forms the core of the 'dictionary proper' and the one that is subject to lexicographic treatment. A dictionary is usually ordered alphabetically by main entry and has a double structure. That structure is usually referred to as the macrostructure and the microstructure, first introduced by Rey-Debove (1971:21) as cited in Gouws (2017), following which discussion of a variety of structures such as *access structure* began by various scholars, for instance, Wiegand (1989); Wiegand & Beer (2013) as cited in Gousw (2017). My discussion is limited only to the double structures which cannot function in isolation. By macrostructure, I mean the list of all the words in a dictionary i.e., headwords or lemmas – what constitutes a headword, how they are organized etc. Recent advancements in technology and lexicology have helped contemporary lexicographers to use many online lexical databases to build up their lexicon and auto-generate the intended dictionaries. In such cases, the macrostructure refers to the way the lexicon is set up so users can enter the lexicon and find the desired headword.

The microstructure, on the other hand, is all the information given about each headword. The dynamics can vary depending on the mode of the dictionary, for instance, CD/DVD-ROM/Portable dictionary. But then again, a definite choice has to be made. For on-line dictionaries there is certainly no limit on what to include. It is usually in terms of information categories ranging from phonological to semantic.

In the following subsections, I will describe these double structures with appropriate examples from Ruga-English, Garo-English, Atong-English, and Khasi-English dictionaries being compiled. These dictionaries are different from the existing dictionaries of Garo-English, Atong-English, and Khasi-English in that they adhere to the principles of contemporary lexicography for both print and digital editions and have additional information in both macro- and microstructures which are discussed in the subsequent sections.

3.1 Macro-structure

Two main macrostructural traditions exist in lexicography, i.e., the onomasiological and the semasiological traditions. *Semasiology* takes its starting-point in the individual word and goes on to identify the concepts associated with it (word-to-concept approach). *Onomasiology,* takes the opposite perspective – it starts with a concept and investigates which words may be associated with that concept (concept-to-word approach). I use a hybrid approach for elicitation of words i.e., start with concept-to-word and then when one comes across new forms, say for instance, change in tone, intonation, etc. also elicited words using the semasiological approach. However, when it comes to organization of the lemma list, I prefer the alphabetical macrostructure based on the spelling[7] of the lemmas (as shown in Figure 1) according to the Word-by-Word principle for print dictionaries for the best access possibilities. Figure 1 shows lemmas with straight alphabetical macrostructure without grouping where all the lemmas are equal and each lemma appears at the left-most of the column (Svensen 2009: 371).

[7] Khasi and Garo use roman script, while Ruga and Atong do not have any writing system. I use roman script based on the Garo orthography to compile the Ruga and Atong lemmas since the communities are familiar with it.

On the other hand, for digital dictionaries, both alphabetical macrostructure and systematic macrostructure – which is the arrangement of entries according to Semantic Domains (SD), derived mostly from Fieldworks Language Explorer with nine major SDs and SIL comparative African wordlist (SILCAWL[8]) with 12 major SDs (as depicted in Figure 2) – may be used. Each of the major semantic domains have at least five subdomains which may also have several sub-domains.

Ruga-English
babiia [babija] n father
babilshi [bablʃi] n kitchen
bahlak [baʔlak] n ladle
baibadaia [baibadaja] v to move
maiphring [maiphriŋ] n breakfast
maithama [maithama] n dinner

Atong-English
garu [garu] n mustard
gasam [gasam] n time
gasu [gaʔsu] adj splendid
gausu [gausu] n rib
gawang [gawaŋ] n spider

Figure 1: Straight alphabetical ordering without grouping

Example of a classified Ruga-English dictionary that follows systematic macrostructure[9]:

Human Body-parts:
jahkhinchokniia [ɟaʔkhinchoknija] n heel (SD: External organs, Wh: Leg)
jahphawa [ɟaʔphawa] n foot (SD: External organs Wh: Leg)
kereng [kereŋ] n skeleton (SD: Internal organs, Bone, joint.)
kerenga[1] [kereŋa] n skull (SD: Internal organs, Wh: Head.)
kerenga[2] [kereŋa] n bone (SD: Internal organs, Bone, joint.)
khahrongthiniia [khaʔrɔŋthinija] n kidney (SD: Internal organs)
khahsopniia [khaʔsɔpnija] n lung (SD: Internal organs)
khahthongniia [khaʔthɔŋnija] n heart (SD: Internal organs)

Kinship:
jikniia [ʤiknija] n father's younger sister's husband (SD: Related by marriage)
mathaniia [mathanija] n mother's elder sister (SD: Related by birth)
mathaniia abani jukjuwah [mathanija abani ʤukʤuwaʔ] *n* elder sister's husband (SD: Related by marriage)
nanijika [naniʤika] n husband (SD: Related by marriage)
nanijukjuwa [naniʤukɟuwa] n wife (SD: Related by marriage).

Figure 2: Systematic macrostructure of Ruga-English classified dictionary

Having decided upon the approach to be undertaken, the next step is to look into the contextual requirements of a headword. The criteria for the selection of lexical items to be included in the dictionary should be laid down. These items are entered into the macrostructure as lemma signs which become the guiding element of the dictionary article.[10] Earlier dictionaries have included only words i.e., single roots as lemmas and thus were

[8] https://www.sil.org/resources/archives/7882
[9] Though the discussion here is on the macrostructure, I have included the SD information as well to illustrate the use of semantic domains and subdomains. "Wh" in these entries indicate the "Whole" – the holonyms.
[10] 'Article' is the lemma or headword and all the information items related to it (basically the dictionary entry).

dominated by a word-bias, cf. Gouws (1991). A lexical-based approach to macrostructure emphasizes the need to include other types of lexical items as well. Thus, lexical items selected for inclusion as lemmata can be entered as either main lemmata or sublemmata or multiword lemmata. It is important to allow this flexibility so that languages like Khasi, where multi-word units such as nouns that also specify their gender, can be accommodated. This criterion is illustrated in Figure 3 and discussed in detail in §3.1.2.

Establishment of lemmas also includes the decisions to be made in case of lexical items with identical base forms. For instance, should separate lemmas be established for homonymous forms? How to deal with the lemma for polysemous forms? Homonymy is normally manifested in the macrostructure of the dictionary and polysemy in the microstructure (Svensen 2009: 96). Besides this, should the entry structure of the lemmas be flat or hierarchical? Thus, deciding on the types of entries the dictionary will include, and organizing the headword list, are macrostructure decisions.

Lemmata
Ruga
santok n bangle

Sublemmata
Khasi-English
jing- Pre nominalising prefix

Multiword lemmata
Khasi-English
u klong, u skaw n pumpkin
ka sngi n sun

Figure 3: Types of lemmata

In a monolingual lexicon, all the lemmas are entered according to semantic domains which can be sorted alphabetically or thematically to generate a monolingual dictionary. Sorting can be customized.

For a bi/multilingual dictionary, building the lemma is not easy. The macro structure of such dictionaries usually depends on a specific monolingual – sometimes a bilingual – dictionary or a dictionary file that is chosen as its basis. In the next subsection, I will discuss the lexicographic concerns of developing bilingual dictionaries of Indigenous languages.

3.1.1 Building bilingual dictionaries

Dictionaries of Indigenous languages are almost always bilingual and are usually monoscopal unidirectional dictionaries i.e., the headword will be in the Indigenous language and the definitions or explanations are provided either in English or in the lingua franca of the region, and the intended user is the native speaker of one language. Recently, many biscopal bidirectional dictionaries are also being created wherein the lemmas and the equivalents may be given in two languages if intended for native speakers of both languages. Sometimes, we come across trilingual dictionaries too with a link language used for interpretation (as in the case of the Ruga-Garo-English dictionary). Keeping in view that the dictionaries are primarily made for the community members, I have kept the Indigenous language (IL) headword first and not English-IL or English-other dominant language. This is echoed by many scholars. For instance, Corris et al. (2002), in making dictionaries of endangered languages (EL) state the following:

(i) *This arrangement is typically most useful for speakers of LWC (including the lexicographer) trying to learn, understand or explicate the EL, in other words for decoding EL.*
(ii) *It can also be put down to the symbolic function of the dictionary. Speakers sometimes feel that EL-LWC is the only direction that could truly be described as a dictionary of the EL. Having the EL first gives it a kind of primacy.*

There are many tools for developing digital lexicons. For instance, LEXUS[11] for the creation of multi-media encyclopedic dictionaries and lexica; ViCoS[12] and FLeX[13] for constructing and visualizing conceptual spaces, and Kirrkirr[14] for software exploration of Indigenous language dictionaries. I use FLeX for developing lexicons for these languages of North-East India, which could be used for multiple purposes including generating dictionaries. As far as the macrostructure is concerned, FLeX is designed in a user-friendly manner for headword lookup.[15]

3.1.2 Types of lemma

What constitutes a lemma in a bilingual dictionary is to a certain extent language-specific. It may contain single items, partial words and multi-word expressions (Atkins & Rundell 2008).

1. **Single item**
   These are the most common in any language and may be of three types: simple words, abbreviations, and partial words.
   (i) Simple words are those that have only a root, as illustrated in Figure 4:

Ruga-English
ganga [gaŋa] n village
dariiah [darijaʔ] adj sharp
joba [ʤɔba] n hibiscus

Garo-English
phringo [pʰrɨŋo] n morning
matchu [matcʰu] n cow
aringga [ariŋga] n a large water lizard

Figure 4: Simple words in bilingual dictionary

(ii) Abbreviations
Abbreviations are also included as headwords with two subclasses as shown with examples from Khasi:
a) *Alphabetisms* the initial letters of a group of words, pronounced as series of letters, as in:
KSU: Khasi Students Union
b) *Acronyms* the initial letters of a group of words, pronounced as a word as in:
NEHU: North Eastern Hill University

[11] https://www.mpi.nl/corpus/html/lexus2/ch01s01.html
[12] developed by Max Plank Institute for Psycholinguistics
[13] https://software.sil.org/fieldworks/
[14] https://nlp.stanford.edu/kirrkirr/
[15] Users should be able to access entries by simply searching for headwords matching a string they type in, or by searching for entries of a certain part of speech, or based on the semantic category of a particular word etc.

(iii) Partial Words

These are the sublemmata – usually affixes in a language. It has become common these days to include derivational prefixes, suffixes, infixes and circumfixes (if any) in dictionaries as shown in Figure 5. Naturally, this type of headword may cause confusion to the users with no linguistic background. Depending on for whom the print dictionary is prepared, the dictionary-maker may choose to include these as headwords or not. On the other hand, in online dictionaries, options could be provided for searching partial words too.

Atong-English
dot- Pre, classifier for cylindrical object
Garo-English
sak- Pre, classifier for humans
Khasi-English
jing- Pre, nominalising prefix
pyn- Pre, causative prefix

Figure 5: Partial words as lemma

2. Multiword expressions

Multiword expressions form a central part of vocabulary in most languages, be it major or minor, and they should be included in the dictionary. Phrases, idioms, collocations and compounds are treated as multiword items as illustrated in Figure 6.[16]

Khasi-English
Phrases
u bnai n moon (SD: Universe and Creation)
ka jingmap v pardon (SD: Verbs of forgiving)
Idioms
kum ka sim lai dieng n gossip-monger
dong shimat n a person who cannot keep secrets
Collocations
u briew uba khuid n holy man
u briew khlain n strong man

Atong-English
Compounds
mai rwmgaba [mai rəmgaba] n cook (SD: Occupation and related)
phiram gaba [phiram gaba] n beggar (SD: Occupation and related)
tangka botgaba [taŋka botgaba] n money lender (SD: Occupation and related)

Figure 6: Multi-word expressions

To summarize, headwords are given in bold typeface. Homonyms are marked with figures printed as superscripts after the word. Headwords give information about spelling and function as a key or guiding element to the information sought. In some cases, headwords may contain a cross-reference to another entry (see Figure 14). The selection of words is kept authentic and includes words that are representative of the language spoken by these

[16] This is not the right place to quibble about what counts as an idiom and what a collocation or fixed phrase is because the idea is to have some place in the dictionary to deal with them. See Atkins and Rundell (2008: 166) for detailed discussion on this concept. For our purposes, idioms are different from phrases in that the meaning of an idiom is the sum of the meaning of its parts whereas collocations are combinations of words which frequently occur together in a particular language.

communities even in the case of those that lack written records. There could be issues with, spellings especially when the languages are not written down. Developing an orthography for a language that is acceptable by all the stake holders is essential (cf. Pappuswamy 2017) for issues related to orthographies of unwritten languages.

3.2 Microstructure

Microstructure of dictionaries differs in the profile they present the information of a headword. The lemma may contain information on a number of categories such as orthography, phonetics, morphology, syntax, semantics, pragmatics, usage, etc. It should be recalled that these different levels of linguistic information are equally important and play a vital role in designing the microstructure. My focus in this paper is limited to the coding of lexical semantic information.

4. Coding of lexical semantic information

4.1 Definitions and explanations

Lexical semantic information is coded in definitions and explanations of the headword. Definition is the most important part in the microstructure, mainly because it is what many consult a dictionary for. The compilation of appropriate definitions or explanations for the lemmas is indeed a very challenging task for dictionary makers. This is one of the most debated areas and perhaps one of the features that needs improvement before a reasonably acceptable explanation for a lemma is arrived at. In Johnson's words, "The rigour of interpretive lexicography requires that *the explanation and the word explained should be reciprocal*" (Johnson, *Preface*, 1755 in Johnson 2012).

Most of the dictionaries prepared for Indigenous languages are bilingual in nature, but are different from those prepared for translation purposes. Data elicitation is either through English or another dominant language of the region (for instance, Ruga-Garo-English) in which the definitions are also given. Thus, providing lexical equivalents for the IL lemma will not be satisfactory because it is entirely understandable that the concepts need not be the same and the dictionary-maker is more likely to embark on totally new concepts unfamiliar to them. Classificatory issues related to semantic domains might crop up. For instance, a 'fruit' in one society might be perceived as a 'vegetable' in another.

Definitions should be as close as possible to the dominant values of the society. It should be intelligible in the sense that the familiar simple words are used so that the user does not have to consult another definition or worse even another dictionary to understand the word they are searching for. Thus, definitions should be simple, easy to understand, relevant, accurate, sufficient and coherent in depicting the ideological values of the society. Printed dictionaries usually suffer from space limitations, and therefore, many dictionary makers tend to provide shorter definitions or explanations for lemmas which merely help the user to differentiate one lemma from the other.

The definition explains the meaning of the headword. Hence, it is essential to avoid circularity. There are different explanatory techniques, the classic one being the 'genus-differentiae' model, which uses the hypernym-hyponym or holonym-meronym features. A dictionary compiled for Indigenous languages has multiple purposes. For instance, the Ruga-English dictionary is intended for native speakers, particularly for those who have completely lost their language and have switched over to Garo, the dominant language of the region. In such a scenario, it is important to use an analytical definition, which is the classical, most common, and perhaps one of the most important types of definitions (Ayto 1983:90; Hartmann & James 1998:6). This type of definition "consists of the word to be defined, the

definiendum, and the explaining part, the definiens" (Ayto 1983:90). This parallels the 'genus-differentiae' model. But sometimes, these types of definitions can be very vague and not helpful in explaining the complete meaning of a particular word. In such cases, one may have to rely on other lexical semantic devices such as synonymy or near-synonymy, collocations, usages, etc. As stated by Wittgenstein (1953), "The meaning of a word is its use in the language." Thus, it is important to determine the senses of a word depending on its context. For instance, the entry for Atong-English "**choki** [cho:ki] n chair" is not enough to get the appropriate meaning of *choki*. Here, it is important to specify which meaning of the English 'chair' is provided as definition, since *chair* is homonymous in English. This information may be included using a hypernym. In one sense, this English gloss will help in sense disambiguation as illustrated in Figure 7:

Atong-English:
choki [cho:ki] n chair (a kind of furniture)

Figure 7: Using glosses for explanations

Usages of the sense could be given with simple short examples. If no other mechanism is possible, collocations can be of use as in Figure 8:

Khasi-English
khuid [kʰuid] adj 1. clean (khuid karma 'clean room') (2. holy (u briew uba khuid 'holy man') 3. good (pyrkhat khuid 'good thought')

Figure 8: Use of collocations in definitions

Handling polysemy is not that easy either. In the examples in Figure 9, the word-forms that are similar but with related meanings are listed in one single entry as sub-senses with glosses when necessary within parentheses.

Khasi-English
khlieh [khlieʔ] n 1. head (body part) 2. head of an institution/organization
kjat [kʤat] n 1. leg (body part) 2. bottom part of something

Ruga-English
kujulla [kuʤulla] n 1. mouth (body part) 2. Mouth of something (river, bottle, etc.).
jahphawa [ɟaʔphawa] n 1. foot (body part) 2. bottom part of something

Figure 9: Definitions of polysemous lemmata

Handling homonymy is yet another tricky task. In a traditional historical approach, the lexemes with the same form but different origin were treated as homonyms and given separate entries in a dictionary. However, there are inconsistencies in this theory as pointed out by several scholars; for instance, Svenson (2009) and many have moved on to adopt a 'semantic approach' where lexemes with the same forms but unrelated meanings can be treated as homonyms. Each of the different senses are represented by unique lemmas. Meanings can be discriminated with the help of hypernyms, meronyms etc. using semantic domain names as in Figure 10:

Ruga-English
ningjah$_1$ [niŋʤaʔ] n squash (SD: plant)
ningjah$_2$ [niŋʤaʔ] n lizard (SD: reptile)
Atong-English

kuchuk$_1$ [kʊchuk] n beak (SD: birds and related)
kuchuk$_2$ [kʊchuk] n language (SD: education and related)
Khasi-English
san[1] [san] n five (SD: numerals)
san[2] [san] v grow (SD: verbs of growing)
thied[1] [t^{h}ied] n root (SD: part of plant)
thied[2] [t^{h}ied] n vein (SD: body part, internal organ)
thied[3] [t^{h}ied] v buy (SD: verbs of buying and selling)

Figure 10: Definitions for homonymous entries

A word can have *n* number of relationships with other words, and it can be both in vertical and horizontal directions (hierarchical and linear relations). The hierarchical relations, taxonomic (kind-of) and meronymic (part-of) relations, are asymmetric and hence are ideal to describe the definiendum without the problem of circularity. For instance, see Figure 11, especially the entries for **dala** which are self-explanatory.

Garo-English
bitchy n, rice beer (a kind of drink).
Atong-English
dala[1] n round bamboo mat (a kind of a mat)
dala[2] n branch of a tree (part of a tree)

Figure 11: Definitions using taxonomic and meronymic relations

Besides the above, the symmetrical relations of synonymy and antonymy provide ancillary information about the meaning of a headword: Synonymy is less used in definitions[17] mainly because there are no true synonyms (Nida 1949). The use of near-synonyms is valid in cases where one is not required to provide semantic provision. Sometimes cumulative synonym definitions may be provided, as in the Khasi example below, but certainly it cannot help the user to understand the subtle differences in the meanings until and unless some collocations are used in the entry: **buh** [buʔ] v set, put, keep, place (SD: Arrange).

The relation 'antonymy' is often associated with adjectives, and dictionaries usually present antonyms as 'run-on' information at sense entries (Murphy 2016: 453) but care should be taken to avoid circularity. For instance, the definition for *hot* = 'not cold' is fine but it is not desirable to define *cold* = 'not hot': **tengoh** [tɛŋoʔ] *adj* hot ANT **gisinna.** The question is whether to include **tengoh** as antonym in the entry for **gisinna** as well. Such information is redundant in a paper dictionary because one has to cross-refer to **gisinna** manually, unlike the machine-readable dictionaries where one could simply click on the lexical relation to get the meaning: **gisinna** [gisinna] *adj* cold.

Definitions of culture-specific items may be done by means of providing encyclopedic information, and in case that is still not sufficient to capture the meaning, appropriate graphic illustrations may be used to support the definition as shown in Figure 12.

[17] However, check Zgusta (1971: 262) for when it is allowed to use synonyms for defining lemmas.

Ruga-English
repena [repena] n body garment,
(a cloth used by men to cover their waist)

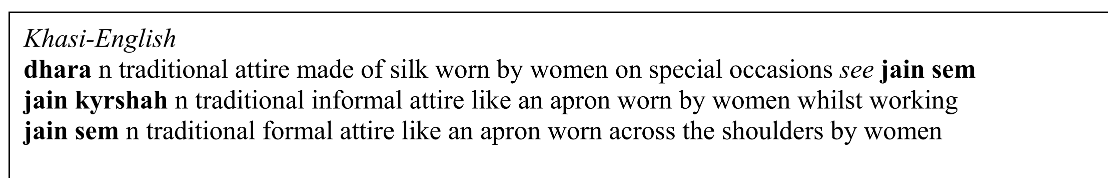
Figure 12: Explanation with encyclopedic and graphic information

Though the picture in Figure 12 is the right one, it is weak in the sense that it does not offer any clue as how Ruga men wear this garment. So, the most appropriate and adequate explanation should be provided. Many questions such as ‘How is this piece of garment worn by men?’ ‘How long is it?’ etc. might crop up in the mind of the user. Hence, the use of full-sentence explanation along with a picture will be ideal, as illustrated in Figure 13. Thus, when it is not possible to explain a headword in a satisfactory manner, use of pictures or diagrams help to clarify the meaning of the lemma.

Ruga-English
repena [repena] n men’s clothing,
(a knee-length cloth wrapped around their waist by men)

Figure 13: Appropriate explanation with graphic illustration

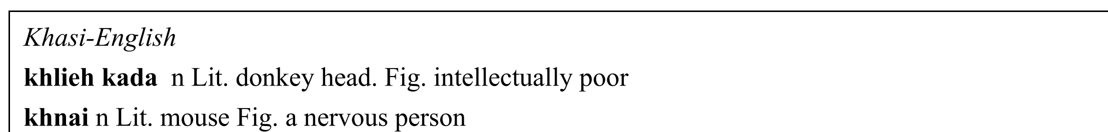
Cross-referencing may also be used not just to save space in paper dictionaries but also to indicate that the headword is related to another one in several aspects with the help of the label *cf* or *see* as in the first entry in Figure 14.



Khasi-English
dhara n traditional attire made of silk worn by women on special occasions *see* **jain sem**
jain kyrshah n traditional informal attire like an apron worn by women whilst working
jain sem n traditional formal attire like an apron worn across the shoulders by women

Figure 14: Example of cross-referencing[18]

Metalinguistic information may be used to indicate the style labels (*fml* for formal, *inf* for informal, *sl* for slang, *vulg* for vulgar, etc.). Metaphorical use of words and phrases are indicated by a label *fig* along with the literal meaning, as in Figure 15.



Khasi-English
khlieh kada n Lit. donkey head. Fig. intellectually poor
khnai n Lit. mouse Fig. a nervous person

Figure 15: Metaphoric senses

4.2 Determining word senses

[18] Appropriate pictures will certainly help the user to know the differences among these three types of women’s garments. For lack of space, I am not including them here.

Most words have only one meaning. The more common a word is, the more likely it is to have multiple meanings. It is these high-frequency words with multiple senses that generally form the 'core' vocabulary of the language. Moon (1987: 176) notes that Johnson's entry for *take* lists 134 senses, while the corresponding entry in the first edition of the *OED* has about 341.

Determining word senses is not an easy task, especially in an Indigenous language. Dictionary makers are said to be word-watchers. The question is, how does one watch a word in Indigenous languages that lack written or spoken corpora or other lexical resources? One possible solution to this issue is to move away from the traditional data elicitation just from word-lists and, as mentioned before, include other means of 'watching' and 'collecting' words, such as using picture prompts, silent video clips, narratives etc.

In bilingual dictionaries, the issue of sense division is more complex because it involves two lexical systems representing two world views. Words, in fact, do not exist in isolation. They are always related to each other in various ways: they may be synonyms, or antonyms, or simply belong to the same semantic domain (like kinship or body parts etc.). Some usages find their way into dictionaries and become established word senses. There are several strategies for identifying dictionary senses:

(i) Using lexical relations: Relatedness of Senses
   (a) Similarities in word-forms but with unrelated meanings (homonyms)
   (b) Similarities in word-forms but with related meanings (polysemes)
   (c) Similarities in word-forms but figurative meanings (metaphors)

(ii) Using different types of meanings
   (a) Denotative and Connotative
   (b) Words that carry specialised meanings
   (c) Changing provisional explanations into concrete definitions

(iii) Splitting and Clustering
   For common words, the dictionary makers have to struggle with several usages and try to group them into manageable 'clusters' of meaning, which is desirable. However, 'splitters' are also seen in use by lexicographers who tend to generate a large number of finely distinguished senses, as shown in Figure 16. Until and unless one has mastery over an Indigenous language, these strategies may be loosely used and oftentimes will be only a matter of convenience.

Atong-English
kha[1] [khaʔ] v to tie (SD: Stative or Position Verbs)
kha[2] [khaʔ] v to be bitter (SD: Stative or Position Verbs)
kha[3] [khaʔ] v to pour liquid (SD: Stative or Position Verbs)

Figure 16: Distinguishing senses using 'splitters'

4.3 Ordering the senses

Polysemous senses are normally listed as subentries inside one lemma either using a flat structure with sense numbers as 1, 2, 3 etc. or a tiered one. It is not easy to list the senses in that manner. Sometimes one might come across two or three meanings that are close to each other while a few others might be quite distinct. Lexicographers prefer to use a tiered structure to represent the same. A tiered structure has the provision of specifying the

sub-senses into the main senses (1a, 1b, 1c, 2a, 2b, 3a, 3b, 3c, 4, etc.).

Choice of a flat or tiered structure of organizing the polysemous senses depends entirely on the purpose and user of the dictionary. The most common one is to use a flat structure with the presentation part. However, the problem still remains in organizing the sub-senses. Traditionally, dictionaries have applied historical order of senses, starting with the oldest sense and completing it with the most recent. This is a very complex phenomenon for Indigenous languages. It is indeed very difficult or sometimes even impossible to decide whether a particular sense is older than another. Even the elderly native speakers of these languages will not be able to provide us with this information. Further, lack of written corpora in these languages adds to the difficulty in arriving at the frequency-based order of the senses of a word. Even more complex is to identify the 'core' or the 'prototypical' sense and the ones deviating away from it. In other words, semantic relatedness is not easy to figure out even with the help of the native speakers. Therefore, it is usually the linguist or the fieldworker who assigns the sense numbers, which are oftentimes done in an arbitrary fashion as illustrated in Figure10. One of the disadvantages in this method is that the most usual sense might occur somewhere in the fifth or sixth place. Thus, the sense and sub-sense numbers in these entries should be treated only numerically with an understanding that they basically indicate the division into n senses, and no logical claims should be made about their order of occurrence.

Ordering homographs is a complex phenomenon. It is not sufficient just to give homograph number (either in the superscript or subscript) to the n senses that have the same form and orthography and list them one after the other as unique lemmas. The issues are the same as that of the polysemous senses: which homographic sense gets the first placement? Which ones follow that? On what basis is this order of occurrence decided by the dictionary maker? The same issues crop up while ordering synonyms too for a lemma.

Thus, it is debatable which format of ordering of senses is desirable. There is no 'perfect way' to make a reasonably good dictionary. In some approaches, polysemy is manifested in the microstructure while homonymy is reflected in the macrostructure. As stated by Svenson (2009: 365), a dictionary prioritizing polysemy has a comparatively rich microstructure but a poor macrostructure (micro-structure oriented dictionary) and the opposite is also possible (macro-structure oriented dictionary). The nature of the dictionary obviously depends on the treatment of polysemy and homonymy by the dictionary makers. This is, of course, overlooked by many and perhaps may not even matter to the Indigenous languages' situation, especially if those are the first dictionaries that are being brought out. However, it is necessary to follow a standard procedure in creating dictionary entries for these two important sense relations.

5. Conclusion

The issues and challenges discussed in this paper highlight the complexities of making dictionaries for lesser-known Indigenous languages, right from designing the macro and microstructures. These languages are the main victims of India's slow but steady language erosion. Therefore, it is essential to document the Indigenous knowledge as early as possible because it will provide insights into human culture, especially the knowledge in the areas of ethnobotany, ethnozoology, ecosystems, medicinal and nutritional practices, which will certainly benefit mankind. One of the best deliverables of such a documentation is the production of dictionaries that rely on systematic elicitation, serendipitous discovery and employment of lexicographic principles, especially in the codification of lexical semantic information. Cross-disciplinary collaboration with anthropologists, biologists, geologists and

others along with the native speakers of a language will certainly help the dictionary makers for a reasonably good coverage of the lexical and cultural knowledge. Many community members are engaged in the development of dictionaries on their own. They may be provided training in easy-to-use software like WeSay[19] for dictionary-making so that their efforts can be utilized in a proper manner, not just to bring out lexicographic resources for the under-resourced languages but also for capacity building. To sum up, production of dictionaries and other lexical resources will certainly provide an instrument for social mobility and technological empowerment of the Indigenous communities.

[19] WeSay is a dictionary-building tool for use by native-speakers see http://www.wesay.org.

References[20]

Atkins, B.T. & M. Rundell. 2008. *The Oxford guide to practical lexicography*. Oxford: Oxford University Press.

Ayto, J. R. 1983. On specifying meaning. In R. R. K. Hartmann (ed.), *Lexicography: principles and practice*. 89–98. London: Academic Press.

Campbell, L. & A. Belew (eds.). 2018. *Cataloguing the world's endangered languages*. London and New York: Routledge.

Corris, M., C. Manning, S. Poetsch & J. Simpson. 2002. Dictionaries and endangered languages. In D. Bradley & M. Bradley (eds.), *Language endangerment and language maintenance*. 329-347. London: Routledge Curzon. http://nlp.stanford.edu/pubs/eldic.ps M. Corris 1999, *Dictionary use and useability*. Sydney: Department of Linguistics, University of Sydney. (BA Honours thesis).

Evans, N. 2010. *Dying words: endangered languages and what they have to tell us*. West Sussex: Wiley-Blackwell.

Gouws, R. H. 1991. Toward a lexicon-based lexicography. *Dictionaries,* 13, 75–90.

Gouws, R. H., U. Heid, W. Schweickard & H. E. Wiegand (eds.). 2013. *Dictionaries: an international encyclopedia of lexicography. Supplementary volume: recent developments with focus on electronic and computational lexicography.* Berlin/New York: Walter de Gruyter.

Harrison, D. 2007. *When languages die: the extinction of the world's languages and the erosion of human knowledge*. New York: Oxford University Press.

Hartmann, R. R. K. & Gregory J. 1998. *Dictionary of lexicography*. London and New York: Routledge.

Hausmann, F. J., O. Reichmann, H. E. Wiegand & E. Zgusta (eds.) (1989–1991) Wörterbücher: *Dictionaries: An International Encyclopedia of Lexicography*. Berlin/New York: de Gruyter.

Haviland, J. B. 2006. Documenting lexical knowledge. In J. Gippert, N. Himmelmann & U. Mosel (eds.), *Essentials of Language Documentation*, pp. 129-162. Berlin: Mouton de Gruyter.

Himmelmann, N. P. 2006. Language documentation: What is it and what is it good for? In J. Gippert, N. P. Himmelmann and U. Mosel (eds.) *Essentials of language documentation* (Trends in Linguistics. Studies and Monographs, 178), 1-30. Berlin: Mouton de Gruyter.

Hinton, L. & William W. 2002. A dictionary for whom? Tensions between academic and nonacademic functions of bilingual dictionaries. In W. Frawley, K. Hill & P. Munro (eds.),

[20] **Acknowledgements**

My sincere thanks are due to the following for their assistance in data elicitation: Egira Shadap (for Khasi); Triksimeda B. Sangma (for Garo and Atong); Sansuma Brahma, Saint Quill N. Sangma and Nazaretha D. Sangma (for Atong), Venial In Good Hope Marthong (for Ruga) and Gripseng Momin (for Garo interpretation with Ruga consultants) in the initial stages of the projects. I am grateful for various funding agencies and institutions for their financial and academic support: Khasi-English, Garo-English mini-interactive dictionary project was funded by NEICOD project, North-Eastern Hill University Shillong. Ruga-Garo-English dictionary was initially funded by UGC Major Research Project; and the Atong-English dictionary is part of my ongoing work supported by Scheme for Protection and Preservation of Endangered Languages Project, Central Institute of Indian Languages, Mysuru. I acknowledge the contribution made by many native speakers, without whose help I could not have expanded my lexical database over the last few years. I also thank the organizers of IYIL19 Perspectives conference, especially Shannon Bischoff and Shobhana Chelliah, for inviting me to present the earlier version of this paper in the conference. I also thank the reviewers for their valuable feedback and suggestions on my manuscript.

Making dictionaries: Preserving Indigenous languages of the Americas. 155–170. Berkeley: University of California Press

Moon, R. 1987. Monosemous words and the dictionary. In A. Cowie (ed). *The dictionary and the language learner*. 173–182. Darmstadt: De Gruyter.

Murphy, L. M. 2016. Meaning relations in dictionaries: hyponymy, meronymy, synonymy and contrast. In P. Durkin (ed.). *The Oxford handbook of Lexicography*. 439-457. Oxford: Oxford University Press.

Nida, E. 1949. *Morphology: the descriptive analysis of words*. Ann Arbor: University of Michigan Press.

Pappuswamy, U. 2017. Issues and challenges in the search of effective orthography for unwritten languages of North-East India. In A. Abbi (ed.), *Unwritten Languages of India.* 117-159. Delhi: Sahitya Akademi.

Pappuswamy, U. 2019. Language endangerment, ethnic minorities and identity crisis in the Indian subcontinent. *Science and Culture* [*'Ciência e Cultura'*] 71(4): 29-38, São Paulo out./dez. 2019, versão On-line ISSN 2317-6660.

Rey-Debove, J. 1971. *Étude linguistique et sémiotique des dictionnaires francais contemporains* [Linguistic and semiotic study of contemporary French dictionaries]. The Hague/Paris: Mouton.

Schebeck, B. 1983. Dictionaries for Australian languages: Some general remarks. In P. Austin, P., P. McConvell, R. Day, P. Black, R. Zorc, B. Schebeck, G. McKay, K. Hale, M. Laughren, D. Nash, M. Laughren, H. Koch, & A. Wierzbicka, (ed.) *Papers in Australian Linguistics No.15: Australian Aboriginal lexicography*. 41-55. Canberra, Australian National University: Pacific Linguistics.

Svenson, B. 2009. *A handbook of lexicography: The theory and practice of dictionary-making*. Cambridge: Cambridge University Press.

Turin, M. 2005. Language endangerment and linguistic rights in the Himalayas: A case study from Nepal. *Mountain Research and Development*. 25(1): 4-9.

Wohlgemuth, J. & Michael C. 2010. *Rara & Rarissima: Documenting the fringes of linguistic diversity*. Berlin: Mouton de Gruyter.

Wiegand, H. E. 1989. "Aspekte der Makrostruktur im allgemeinen einsprachigen Wörterbuch: alphabetische Anordnungsformen und ihre Probleme" [Aspects of macrostructure in the general monolingual dictionary: alphabetical arrangements and their problems],in *Wörterbücher: Ein internationales Handbuch zur Lexikographie* (1989–1991). 371–409. Berlin: De Gruyter Mouton.

Wiegand, H. E, & R. H. Gouws. 2013. Addressing and addressing structures in printed dictionaries" in R. H. Gouws, U. Heid, W. Schweickard and H. E. Wiegand (eds). *Supplementary Volume Dictionaries. An International Encyclopedia of Lexicography: Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*, pp. 273-314. Berlin, Boston: De Gruyter Mouton,

Wittgenstein, L. 1953. *Philosophical investigations*. Oxford: Basil Blackwell.

On-line Resources:

Census of India. 2011. http://www.censusindia.gov.in

Eberhard, D. M. & G. F. Simons & C. D. Fennig (eds.). 2019. Ethnologue: languages of the world. Twenty-second edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Eberhard, D. M. & G. F. Simons, and C. D. Fennig (eds.). 2020. Ethnologue: languages of the world. Twenty-third edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Moseley, C. (ed.). 2010. Atlas of the world's languages in danger, 3rd Edn. Paris, UNESCO Publishing. Online version: http://www.unesco.org/culture/en/endangeredlanguages/atlas.

"Page View, Page 6." A Dictionary of the English language: a digital edition of the 1755 classic by Samuel Johnson. Edited by B. Besalke. Last modified: December 6, 2012. https://johnsonsdictionaryonline.com/page-view/&i=6

Gouws, H. R. 02 Oct 2017, Dictionaries and access from: The Routledge handbook of Lexicography Routledge Accessed on: 03 Aug 2020 https://www.routledgehandbooks.com/doi/10.4324/9781315104942.ch3