Documenting Simpa: Advances in language documentation[1]

Nana Ama Agyeman

Lecturer
Language Centre
University of Ghana, Legon, Ghana
E-mail: agyeman@ug.edu.gh

Abstract

Documentary linguistics,[2] also known as language documentation, a relatively new branch of Linguistics, advocates for the fundamental need to collect records of language use and practices in various forms from diverse genres for multiple purposes. Such purposes include language description, language development, language maintenance, and language revitalisation. Such a record of a language serves to feed not only linguistic research but also research in other disciplines, such as anthropology, history, and ethnography. Language documentation is recognized as an ultimate response to language endangerment. This paper explores language documentation with specific reference to Simpa, an under-described, minority language of Ghana. The paper reviews theories, approaches, methods, and tools of language documentation to highlight how they were employed and attuned to take care of the Simpa context. Thus, the discussion dilates on specific field methods and tools adapted for obtaining a balanced set of data from three complementary event types, viz., natural communicative events, staged communicative events, and elicitations, to build a language documentation corpus. Data processing, data annotation, and data management practices applied in building the corpus, as well as dissemination of the research outcomes are also addressed. Furthermore, fieldwork ethics used in the study are discussed. Finally, for consideration in future research, the paper reflects on some challenges that were encountered in documenting Simpa.

Keywords: language documentation, language description, language endangerment, Simpa, Efutu, Ghana

---

[1] This paper is partly based on chapters from Agyeman (2016), a PhD thesis by the author.
[2] Abbreviation: ELAR=Endangered Languages Archive, ELDP=Endangered Languages Documentation Programme, HRELP=Hans Rausing Endangered Languages Project, L1=first language/mother tongue, L2=second language, OCE=observed communicative event, OLB= observable linguistic behaviour, SCE=staged communicative event. UNSD=United Nations Statistics Division

Introduction

This study discusses a documentation of Simpa, a South Guan language spoken in Winneba, a coastal town in the Central Region of Ghana. The paper attempts to define language documentation and its associated elements to relate them to the Simpa documentation experiences. The discussion begins with a definition and a detailed characterisation of language documentation. This is followed by a sociolinguistic depiction of Simpa. After that, a detailed description of the documentation of Simpa in terms of its contents, methods, tools, and outcomes is presented. Then, the paper outlines apparent constraints that were encountered. Finally, the paper concludes the discussion with some recommendations.

Language Documentation

Definition

Language documentation, also referred to as documentary linguistics, is a fairly new sub-discipline of linguistics. Discussions in language documentation dates back to the 1990s. Major research publications in the area include Himmelmann (1998), the Language Documentation and Description volumes which started in 2003, Gippert, Himmelmann, and Mosel (2006), Grenoble and Furbee (2010), and Boerger et al. (2018), among other publications. The field of language documentation could be said to be still evolving. Austin (2003, p.6) asserts that "language documentation as a research area and activity is not yet understood", especially "as distinct from collecting data for linguistic description". A similar view from Furbee (2010) suggests that it will be premature to consider language documentation as having a theory yet, because current definitions emphasize conducts and contents, and urge exemplary ethical and best practices instead of theory-building. Nevertheless, there is a sound delineation of what language documentation entails. For instance, language documentation is described as "a lasting, multipurpose record of a language" (Himmelmann, 2006, p.1). Furthermore, language documentation is distinguished from language description which is conceived as "a system of abstract elements, constructions and rules that constitute the invariant underlying structure of the utterances observable in a speech community" (Himmelmann, 1998, p. 166). Rather, language documentation primarily involves "the making and keeping of records of the world's languages and their patterns of use (Woodbury, 2003, p. 35); it "deals with creating multipurpose records of languages through audio and video recording of speakers and signers and with annotation, translation, preservation, and distribution of the resulting materials" (Austin, 2012).

Language documentation thus, is concerned with the creation and maintenance of a comprehensive record of a language and language use in all conceivable genres and contexts (Himmelmann, 1998, 2006; Austin, 2003; Woodbury, 2003). Such a record may include all varieties of a language, depending on the specific goals of the documentation. Ideally, all facets of a language should be included in the record, nevertheless, the goals of a research may require a documentation of specific aspects of a language. Essentially, language documentation takes on ethnographic approach in the sense that language is observed through the lens of culture (Austin, 2010a; Hill, 2006; Franchetto, 2006; and Harrison. 2005). Furthermore, language documentation is fundamentally multi-disciplinary in nature (see again Austin, 2010a; Hill, 2006; Franchetto, 2006; and Harrison, 2005; see also Widlok, 2005; and Barwick, 2005). Above all, language documentation adheres to ethical standards (Austin, 2010a; Thieberger, 2007; and Dwyer, 2006).

Language documentation is of immense value. Most significantly, it is a response to language endangerment (Hale, 1998; Crystal, 2003; Foley, 2004; Matras, 2005; Himmelmann, 2006). Many of the world's languages, especially minority ones, face endangerment, a situation where a language ceases to be in use because it is replaced by another language or all of its speakers have died. Language death could result from a natural disaster, such as an earthquake, or unnatural causes, such as war and migration, or as a result of repression from a dominant language (Kraus, 1992; Hale, 1998; UNESCO, 2003; Crystal, 2000; Dalby, 2003).

Reasons for language documentation

Although language endangerment is not a new phenomenon, concerns about it have been raised increasingly in recent times, especially in the last three decades. For instance, Hale (1998) projected that out of the estimated 6000-plus languages in the world, 3000 will perish and further 2400 will be near extinction in a couple of decades. A similar prediction states that at least a third of the world's languages will disappear in the 21st century (SIL, 2018). There is therefore a call to document languages, before any misfortune befalls them, because once a language becomes extinct due to any factor, such a language can no longer be documented. Languages, endangered or otherwise, should therefore be documented for posterity and for other good causes.

But why at all is it necessary to document an endangered language; why document a language if it is going to die anyway? The answer to these questions is not farfetched. Indeed, linguistic diversity is crucial to human existence, similar to the way in which diversity in species is critical to the ecosystem. Arguably, the knowledge and belief systems of a people are encoded in their language. Their

beliefs and knowledge about governance, commerce, subsistence, socialization, healing, recreation, and indeed about every other aspect of life, are rooted in their language. Language is said to be the bedrock of a culture's world view and values, a key to individual and community identity (SIL, 2018). A similar view on the uniqueness and importance of individual languages is echoed in Hale (1998). Thus, when an undocumented language becomes extinct, the implication is that all such knowledge of the speakers die with it with no possibility of recovering such information. But when a language receives documentation, even if that language becomes extinct, the record from the documentation could be used for its revitalisation if so desired (see Raymond, 2007; Asfaha, Kurvers, & Kroon, 2007; Nathan & Csato, 2007; and Nathan & Fang, 2009). Besides revitalisation, the record of a language documentation can be used for language description, language planning, and language support and development activities. Furthermore, data from documentation corpora could be used in diverse linguistic enquiries, for instance, to test the validity or otherwise of theories. Beyond the linguistic arena, the record of a language can feed research in other disciplines, including the fields of anthropology, history, ethnography, and music, among other fields.

Forms of data

As previously stated, the main traditional methods of capturing primary data in language documentation are audio and video recording (Woodbury, 2003, p. 36; Austin, 2006, p. 89; Himmelmann, 2006, p. 1; Nathan, 2010b; 2009, 2006, Austin & Grenoble, 2007; Thiebberger & Musgrave, 2007, p. 29; Ashmore, 2008). Besides audio and video recording, still photos taken with a digital camera and field notes taken with pencil or pen on paper also form an important part of the primary data of any language documentation (Himmelmann, 2006, p. 1). Audio and video recorders, and digital cameras therefore, are indispensable equipment in language documentation (Woodbury, 2003). The quality of the primary data has to be of a high standard to the extent that the use of particular types of audio and video recorders are recommended, making the choice of equipment a crucial responsibility in language documentation (see Ashmore, 2008 and Nathan, 2010b). Some important recommendations in capturing primary data regard the use of external microphones rather than reliance on a recorder's internal microphone, as well as the use of a set of closed headphones for both audio and video recording. Overall, the format of the recordings should be of the right standard in order to achieve long preservation, ease of retrieval, and easy accessibility.

Data from the records of a language may be classified into event types. A classification by Himmelmann (2006, pp. 7-11) provides two formats, namely,

observable linguistic behaviour (OLB) and metalinguistic knowledge. The OLB data comprises samples of natural occurrences of "all kinds of communicative activities in a speech community" (Himmelmann, 2006, p. 7), including narratives, conversations, arguments, fights, gossips, sermons, songs, dispute settlement, political debates, selling and buying, and teaching and learning, among other discourse types. The metalinguistic knowledge format on the other hand records "the tacit knowledge speakers have about their language" mainly through elicitation (Himmelmann, 2006, p. 8). Metalinguistic knowledge data basically represents the native speakers' "interpretations and systemizations of linguistic units and events" (Himmelmann, 2006, p. 8). Data on information regarding grammaticality and grammatical paradigms, phonological and morphological paradigms, varieties (for example, young speakers version versus older speakers'), usage and appropriateness, as well as folk taxonomies, all fall under metalinguistic knowledge. The two types of data, viz., data from OLB and that from metalinguistic knowledge are complementary: while the former represents evidence for how speakers actually communicate naturally in their everyday settings (Himmelmann, 2006), the latter provides an avenue for recovering certain aspects of the language that are not accessible from the analysis of OLB (Himmelmann, 2006.

Another classification of primary data from language documentation recognizes three formats, namely observed communicative events (OCE), elicitations, and staged communicative events (SCE) (Lüpke, 2009). The main variation in the three types of primary data is attributed largely to their methods of collection, especially with regard to the level of influence of the researcher (Lüpke, 2009, p. 60). In collecting OCE data, the researcher has no influence whatsoever on the events being captured, except for their presence. The OCE data therefore consists of naturally occurring communicative events, including planned and spontaneous monologues and interactive discourse (Lüpke, 2009). The OCE data is comparable to Himmelmann's (2006) OLB data in terms of its naturalness and least influence of the researcher. Data from the elicitation method is "heavily influenced linguistically by and only created for the sake of the researcher" (Lüpke, 2009, p. 60). The elicitation method employs carefully designed linguistic tools, including questionnaires and other types of stimuli to collect wordlists, paradigms, and acceptability judgements (Lüpke, 2009). The elicitation data is comparable to Himmelmann's (2006) metalinguistic knowledge data in terms of contents and the heavy influence from the researcher. Lastly, data from the SCE derive from a blend of elicitation method with OCE method, with some influence from the researcher but not as much as in elicitations. The SCE method, as the label suggests, is staged

through the use of non-linguistic prompts such as pictures, videos, and games to obtain responses (Lüpke, 2009).

Data annotation and metadata

The primary data of a language documentation usually comes with annotations (Schultze-Berndt, 2006; McGill, 2009); raw data without any form of annotation may be meaningless and difficult to work with (Berge, 2010, p.55). The form of annotation to accompany the data is determined by the research goals, and in some cases, the archiver[3]. Common forms of annotation in language documentation include transcriptions and translations. Annotations could also be done with the help of various types of linguistic software (see McGill, 2009 and Schultz-Berndt, 2006). Metadata is another obligatory constituent of a language documentation (see Bergqvist, 2007; and Nathan & Austin, 2004). Metadata gives detailed description and explanation as well as access restrictions of the primary data and the annotations. Thus, without a metadata, the usefulness of a primary data is hugely reduced. Archiving is another obligation in language documentation (see Nathan, 2010a, 2008; Trilsbeek and Wittenburg, 2006; Johnson, 2004). In order to ensure proper preservation and accessibility, the language documentation corpus which comprises all the primary data with its annotations and metadata is archived. Players in language documentation may include several parties, each of whom assumes specific roles, responsibilities and expectations, based on the research goals. Austin (2003, p. 8) identifies at least six parties, namely, funding agency, archiver, researcher or research team, speaker community, user and general public.

Sociolinguistic background of Simpa

Simpa is spoken in Winneba, a coastal town in the central region of Ghana. It is one of Ghana's close-to-fifty[4] indigenous languages[5], and a minority one. It is officially called Efutu[6] (also spelt Effutu), such that all documents, formal and informal, government or otherwise, use the name 'Efutu'. Together with Awutu and Senya, Efutu is classified as a dialect of Awutu which is further classified as a South-Guan language belonging to the Kwa branch of the Niger-Congo family of languages (Hall, 1983; Eberhard, Simons & Fennig, 2019). Absolute number of speakers of Simpa (and indeed for each of the two sister-dialects) is not known, however, a total number of speakers of all three dialects is estimated at 129,000, based on a 2013 UNSD survey (Eberhard, Simons & Fennig, 2019).

---

[3] Further discussion on archive appears in section 4.5.
[4] Ethnologue estimated the number of indigenous languages of Ghana at 68 and 73 in 2013 and 2016, respectively. However, the figure is drastically reduced to 48 in the 2019 update. It appears that in the recent publication of Ethnologue, some forms are considered as dialects and therefore are not counted separately.
[5] 'Language' is used loosely here, although Simpa is considered as a dialect, as indicated later in this paragraph.
[6] ISO 639-3: afu; 5^0 12' 0" North, 1^0 19' 0" West.

Although speakers call their language as well as their township, clan, and people 'Simpa', outsiders call it Efutu, as indicated earlier. When asked for the reason for the difference in language name, speakers explained that the term 'Efutu' is an Akan[7] expression with the meaning 'mixed up', and that non-Simpa people call the language so because they perceive the language as containing vocabulary from other languages. For this reason, the researcher decided to use the name Simpa, (alongside Efutu) as an attempt to bring attention to how the speakers call their language and also as a form of respect for the speakers.

Simpa could be said to be somewhat threatened. In Winneba, the dominant language is Fante, a dialect of Akan, which is spoken in the inland parts of the town, while Simpa is spoken in the suburbs along the coast by predominantly fisher folks. (Welmers, 1974, p. 11; Agyeman, 2016, p. 31). Speakers of Simpa are bilingual in Simpa and Fante which they speak as a second language (L2). In the Simpa speech community and in Winneba in general, English, the official language of Ghana, and Fante are the languages used in primary and secondary schools both as media of instruction and as examinable subjects.[8] Simpa however is prohibited in schools to the extent that its use, whether inside or outside the classroom, attracts punishment (see Agyeman, 2013, p. 268; Ansah & Agyeman, 2015, p. 97; Agyeman, 2016, p. 29). Simpa-Speaking children are therefore obliged to suppress their mother tongue in favour of the curricular languages in school.

Not only is Simpa threatened but it is also under-described. Studies in the language include Welmers (1973), Forson and Gingiss (1977), Dolphyne and Kropp-Dakubu (1988), Boafo et al. (2002), Abaka (2006), Obeng (2008), Agyeman (2011, 2013a, 2013b, 2015, 2016), Ansah and Agyeman (2015), and Agyeman and Akrofi-Ansah (2013, 2016), with Agyeman (2016) representing the most detailed study of the language so far. The fact that Simpa is vulnerable in terms of language vitality has been pointed out in Welmers (1973), Abaka (2006), Agyeman (2013a, 2013b, 2016), and, Agyeman and Akrofi-Ansah (2013). Indicators for measuring language vitality include absolute number of speakers, population of speakers within a total population, intergenerational transfer, language policy and attitudes, availability of literacy materials, domains of use, and response to new domains and media, among other criteria (Kraus, 2000; UNESCO, 2003). In consideration of these indicators, Simpa fails to meet the standards in many regards. For example, the absolute number of Simpa speakers cannot be that substantial, considering the fact that the estimated figure of 129,000 includes speakers of all three dialects of Awutu. Moreover, the

---

[7] Akan (ISO 639-3: aka), also a Kwa language, is a linguistic neighbour of Simpa, and a dominant language in Winneba and indeed, in Ghana.

[8] This is as a result of Ghana's language of education policy which prescribes the use of English, the official language, in combination with one indigenous language for basic education. For this reason, 11 dominant languages from various regions have been developed for the purpose. See Agyeman (2013), Ansah (2014), and Ansah and Agyeman (2015).

population of Simpa speakers, in comparison with the total population of Winneba is marginal, as intimated above. Furthermore, although there is intergenerational transfer, children ultimately reduce their Simpa speaking rate or stop speaking it altogether in favour of the school languages, due to school language policy and attitudinal factors (again, see Agyeman, 2013a, 2013b, 2016; Ansah & Agyeman, 2015). Furthermore, there is not much response to new domains and media, for instance, Simpa songs are rare, and the only available public media experience is a one-hour-per-week radio programme by the community radio station, Radio Peace 88.9 FM. To add to the above points, there are no materials for Simpa language education and literacy. All these outlined factors work together to render Simpa vulnerable, hence the need for documentation.

Documentation of Simpa

Preamble

The Simpa documentation project was undertaken in the Linguistics Department of the School of Oriental and African Studies (SOAS), University of London, as part of a PhD study. The research was funded mainly by the Endangered Languages Documentation Programme (ELDP)[9], and partly supported by the Office of Research, Innovation and Development (ORID), University of Ghana. The fieldwork was carried out in Winneba, Ghana, the home of the Simpa speakers, and participants were mostly fish workers, including fishermen and fish sellers, all of whom were native speakers. Other participants included school teachers and heads of schools, and traditional leaders. The immediate research goal was to document Simpa and use the annotated corpus to describe aspects of the grammar of the language, as well as depositing the annotated corpus in an archive. Ultimate goals included further descriptive analyses of the language and the development of practical materials, including wordlists, primers, and a dictionary, for the speech community as well as the linguistic research community. The sections that follow discuss various aspects of the documentation of Simpa.

Language Consultants

As noted by Austin (2003), the speaker community constitutes a key player in language documentation, as language consultants are recruited from this community. In this study, the term language consultants, sometimes described as participants or respondents, refers to speakers who participated in the fieldwork. Their role included supply of data and its annotation, as well as corroboration of the annotations. They were selected randomly and recruited through negotiation. Most

---

[9] The project award ID is 'ELDP: IGS-0096'

importantly, their informed consent was sought before engaging their services, in conformity with ethical standards (see Dwyer, 2006; Austin, 2010). In this process, the researcher explained their exact role and the prospect of earning remuneration upon participation. Also, the fact that their acceptance to participate is voluntary and that they could withdraw their participation at any point was explained. Besides, the intended uses of the data and other possible uses were made clear. In this research, consent was mostly verbal and was recorded on audio or video[10]. The interactions with the consultants were conducted in Fante.

Equipment

In line with recommendations[11], equipment choice in this study was based not only on the research goals but also on the fieldwork environment. Thus, the windy and the noisy nature of the Simpa fishing beach where most of the recordings took place, as well as the frequent interruption of electricity power supply[12], all informed equipment choice. Specific equipment included Toshiba satellite laptop, Asus Eee netbook laptop, Zoom h4n audio recorder, Canon hg10 video camera, Canon Ixus 220 digital camera for still pictures, three different microphones, Rode Deadcat windshield, Rode blimp windshield, and a set of close headphones. The microphones included a Rode NTG2 Shotgun, a cardioid microphone and a lavalier microphone. These varieties of microphones were included to cater for different recording situations. The choice of the Zoom h4n[13] and the Canon hg10 was helpful because they had outlets for external microphone and headphones[14]. The blimp windshield was added especially because of the Simpa recording environment; it was used for recordings at the beach and other outside environment in order to reduce the extreme background noise in the recordings. The Asus Eee netbook was chosen especially for its long battery life, in case of electricity power cut, and also for its portability. Other equipment included a tripod for the video camera and a mini tripod for the audio recorder. Essential consumables, such as batteries and SD cards for the recorders, a card reader, and an external hard drive were not left out. A life jacket was acquired for precaution. A peli case was included for transporting and protecting the equipment.

---

[10] See Dwyer (2006, p. 39) for advice on verbal consent.
[11] See discussion in section 2.
[12] The fieldwork was done at a time when Ghana had a major electricity power crisis.
[13] See a review of the Zoom h4n by Bernard Howard at www.hrelp.org/archive/review/zoom_h4n_review.html
[14] See section 2 for discussion of recommendations on the use of external microphones.

Data: Types, tools, methods, and annotation

Issues on data have been briefly outlined. In this section, the discussion focuses on the types of data as well as their collection tools and methods in the documentation of Simpa. Primary data in this study included four main types of media materials: video, audio, still photos, and field notes. In terms of event types, the data included all three traditional types, namely, OLB or OCE, metalinguistic knowledge or elicitation, and SCE[15]. The following subsections expatiate on data.

Elicitation

Elicitation or metalinguistic knowledge data in this research relied on four main tools: (i) Ibadan wordlist of 400 basic items[16], (ii) tense, aspect, mood, and polarity (TAMP) questionnaire, (iii) pronoun questionnaire, and (iv) serial verb construction (SVC) questionnaire. These elicitation questionnaires were administered with the help of trained speaker-consultants.

The Ibadan wordlist contained 400 English vocabulary items which are common to most West African cultures such that equivalent vocabulary exists in their languages. The vocabulary items were from different domains, including body parts, household objects, food ingredients, farm tools, kinship terms, cardinal numbers, verbs, and adjectives. The main method of administration of the wordlist was by mentioning the word in English first, followed by its Akan equivalent, after which the respondent supplied the Simpa equivalent. Thus, Akan was used as an intermediary language while English and Simpa functioned as the source and target languages, respectively. The elicitation sessions were recorded on audio. A couple of respondents who could read and write[17] opted to complete the list on their own, where they used Akan writing system. In addition, the list was administered to them orally for audio recording[18].

The second elicitation tool, viz., the TAMP questionnaire, was designed based on a framework by Dahl (1985). Dahl (1985) hypothesizes the existence of a set of cross-linguistic TAMP categories to which language-specific categories can be assigned, and proposes prototypical contexts in which a given TAMP category may occur cross-linguistically. In creating the TAMP questionnaire for this research, 31 verbs were selected from the Ibadan wordlist and some elicitation data[19] to compose sentences in English which were then translated into Simpa with the help of speaker-consultants. In administering the TAMP questionnaire, respondents

[15] See section 2 for discussion of the various types of data.
[16] This is a common but an unpublished wordlist used by many language documentation researchers. Further description of the wordlist appears in the next paragraph.
[17] Both consultants had tertiary education.
[18] The oral method was useful for ascertaining tone in words
[19] This was data from the video watching and description.

were tasked to reproduce each sentence based on a specified context. Each verb was used in all the various contexts. For each sentence, the negative equivalent was also elicited. Each response was repeated at least twice for confirmation. These elicitations were recorded on audio, alongside note-taking with pencil and paper.

The third elicitation tool which collected data on pronouns was self-designed. The pronoun questionnaire included a set of carefully constructed English sentences with highlighted nouns and noun phrases. In this elicitation, respondents were tasked to initially produce a Simpa translation of the English sentence. Once the Simpa translation was rendered, the respondent was further asked to repeat the Simpa sentence in which they substituted pronouns for highlighted nouns and noun phrases. These elicitations were audio-recorded.

The fourth tool, that is the SVC questionnaire, listed SVCs in Akan[20]. In this elicitation, respondents were tasked to render Simpa equivalents of the Akan sentences. The sessions were audio-recorded.

For the various elicitation methods, metadata was recorded at the beginning and also at the end of each elicitation session. Metadata at the beginning included location and date of recording, equipment used for the recording, consultant's name, age, occupation, hometown, L1, L2, and other languages. Metadata at the end included access restrictions where the consultant was asked to indicate whether any part of the data could not be published, and whether any persons or groups may not access the data.

Observable Linguistic Behaviour (OLB)

In this study, OLB or OCE[21] included conversations, arguments, jokes, and other forms of interactions among speakers. Some of the OBL recordings involved more activities with minimal or no speaking; they included activities like fish preparation (including cleaning, washing, salting, drying, smoking, etc), net mending, boat mending, boat painting, and repairing of outboard motors. One major event capture in the OLB data was a celebration of a local festival, viz., Petu festival, which involved various activities, including durbars, processions, drumming, singing, dancing, gun firing, funfairs, and several rituals. The OLB data were recorded on video and audio; events involving less speech and more activities were recorded on video while those involving more speech, such as conversations and arguments, were recorded on audio, sometimes coupled with video. For the OLB data, all metadata were collected only after the recording.

Another type of naturally occurring data included in the corpus were recordings from a radio programme by the community radio station, Radio Peace,

---

[20] The Akan SVCs were mainly from Agyeman (2002; 2003), Kambon (2013), and Osam (1997).
[21] See section 2 for definition of OLB/OCE.

88.9 FM. The programme, which was done in the Simpa language, discussed social and political issues affecting the community. The discussion was carried out by a host and a panel of guests, and it included listeners' participation. This data was classified as natural data because of its occurrence being independent of the research.

Staged communicative events (SCE)

One of the tools used to collect SCE data was a non-verbal stimulus in the form of a picture book by Mayer (1969) with the title _Frog Where Are You?_ The aim of this exercise was to construct a story based on the pictures in the book. In this method, the consultant initially previewed all the pictures. After that, he went over the pictures again from the beginning to construct a story about the pictures. The whole process was recorded as one session in both video and audio.

Another method for collecting SCEs yielded relatively short narratives termed as 'prompted narratives'. In this method, consultants were asked to talk about a given topic, for instance, to describe the various fishing equipment and their functions, or how to process salted fish, etc. In some cases, consultants talked about their preferred topic rather than the suggested one. Data from the SCE methods were predominantly monologues.

Another SCE tool which involved video watching and description, was self-designed and it targeted complex predication, especially, verbal constructions, including serial verb constructions (SVCs)[22]. In this method, some of the videos from the natural events (OLB/OCE)[23] were selected based on their contents. The selection included videos with high content of activities, such as videos of women cleaning, cutting, washing, and arranging fish on a mesh for smoking; children playing different games at the beach; fishermen conversing, sewing or mending nets; fishermen launching or docking a boat, among such other videos. These videos were played back for consultants to watch and describe the activities. The sessions were recorded on audio and video. Finally, folk songs and stories were also recorded and they were classified under SCEs.

Each of the various SCEs was recorded in both video and audio. Metadata was collected at the end of the session. The above-described SCE methods could be differentiated from the OLB and the metalinguistic knowledge methods. Unlike OLB which occurred naturally without any prompts from the researcher, data from the SCEs were produced upon the request or prompts from the researcher. Nevertheless, there was a level of flexibility with the SCE methods in the sense

---

[22] One of the research goals was to analyse SVCs in Simpa hence, their elicitation.
[23] See section 2., above, and also section 4.4.2., below, for discussion of OLB/OCE.

that the consultant chose or manipulated the type of responses to give, unlike in metalinguistic knowledge method in which consultants were somewhat restricted to supply only certain kinds of responses.

As mentioned earlier, still photographs were included in the corpus. Images in the photographs included artefacts, fishing tools, fishing boats, fish species, consultants and other participants, and scenes from places like the fishing beach and the fish market. In addition, interviews were conducted among community members, including traditional leaders, school teachers, and church leaders. Such interviews investigated matters on topics like language use, language attitudes, language ideologies, aspects of Simpa history, and linguistic neighbours.

Data annotation

In this study, for practical reasons, not every data in the corpus received annotation. For instance, data with unclear sound quality were excluded. Moreover, due to time factor, only a selection based on the researcher's discretion was annotated. Thus, data that would serve the immediate needs of meeting the research goals were given priority. Furthermore, for technical reasons, annotated data included only audio recordings (see further details below). Data annotation mainly involved transcription and translation. Linguistic software used for data annotation included Transcriber (a tool for annotating speech signals), ELAN (a tool annotating audio and video recordings), and Toolbox (a tool for building lexical data, and for parsing and interlinearizing texts)[24].

On data from elicitation, phonetic transcription involving the use of the International Phonetic Alphabet (IPA) with tone marking, and free translation were done manually with pencil and paper by the researcher. For all other data types, i.e., OLB and SCE, transcription was initially done electronically in Transcriber by a trained native speaker from the community. In this process, the annotator listened to the Simpa recordings in Transcriber through a set of closed headphones to transcribe them with the use of Akan writing system since Simpa has no standard writing system. This was followed by a free translation of the transcribed Simpa data into English with pencil and paper by the trained native speaker, viz, the annotator. After that, a second level transcription of the Simpa data in the form of phonetic transcription was done manually with pencil and paper by the researcher, through listening from a set of closed headphones and with the help of the native speaker's transcription. This was necessary in order to get the right pronunciations of words for accurate transcription. The next level of annotation involved interlinearisation in Toolbox after importing the Transcriber file into Toolbox for interlinear glossing

---

[24] See SIL international for further description of these linguistic tools.

into English. This was made possible because the Toolbox project had earlier been entered with lexical data of Simpa with their English glosses. Thus, it was possible for Toolbox to parse the Simpa sentences into their basic constituents with their corresponding English glosses. ELAN was used minimally, mainly for metadata rather than data annotation. To ensure accuracy, all transcriptions were read out to native speaker consultants for confirmation. A summary of the various data types and subtypes in audio files and their associated Transcriber files, as well as their respective total length in minutes is presented in Table 1.

Table 1: Summary of Audio Data

Event Type		Audio Files		Transcribed Files	
No. of Files		**Total Length in Mins.**	**No. of Files**	**Total Length in Mins.**	
Elicitations	Ibadan wordlist	5	218		
	TAMP questionnaire	46	152		
	Pronoun questionnaire	12	42		
	SVC questionnaire	36	302		
Observable Linguistic Behaviour (OLB)	Natural events	16	152	2	13
	Radio programme	2	109	1	16
Stages Communicative Events (SCE)	Frog story	1	8	1	8
	Prompted narratives	40	283	14	111
	Video watching and description	11	133	2	12
	Folk songs and stories	10	27	5	19

Archiving and dissemination

Archiving and dissemination of the documentation outcomes are major responsibilities in language documentation. In this study, the corpus from the Simpa documentation, which consisted of all the primary data and their annotation[25] was deposited with the endangered languages archive (ELAR), a digital archive. As a requirement from ELAR, the only acceptable format of metadata is CMDI maker

---

[25] All transcriptions in Transcriber were migrated into ELAN, another linguistic software, before archiving. This became necessary, as CMDI maker, the only acceptable metadata format does not support Transcriber files.

metadata, a web tool for generating a scientific metadata (see ELAR website for further description of CMDI maker)[26]. Thus, all data in the entire corpus were formatted in CMDI maker metadata for digital archiving. Such a digital archive enhances accessibility of the deposit and constitutes a significant way of dissemination. Besides ELAR, the researcher is making arrangements to deposit copies of the primary data elsewhere, especially in the speech community. This will be done in consultation with speech community members to ensure that the materials are accessible to them in the most convenient format. Besides the archive deposit, data from the corpus was used to write a descriptive grammar of Simpa as a PhD thesis (see Agyeman, 2016). In addition, in consultation with the speech community, data from the corpus was used to produce a booklet on Simpa fishing beach for the community members. The booklet contains pictures of the fishing beach and fishing tools, with short sentences in Simpa describing the pictures. Other practical materials for the community include an ongoing project of a tri-lingual dictionary – a Simpa-Fante-English dictionary. Furthermore, findings from the research have been disseminated through publication in the form of journal articles and also through conference and seminar presentations (See Agyeman, 2011, 2013a, 2013b, 2015, 2016, 2017; Ansah & Agyeman, 2015). Also, with support from the College of Languages Education of the University of Education, Winneba (UEW), a couple of language and literacy workshops which advocate the use of Simpa in schools have been organised for teachers in basic schools in Winneba.

Field ethics

Linguistic research, like research in any other discipline, must observe ethical principles (see Dwyer, 2006; Austin, 2010; Rice, 2010). In documenting Simpa, the researcher observed field ethics. Before starting the fieldwork, permission was sought and obtained from leaders in the community. These leaders included the chief of Winneba, the leader of the fishermen, the founder and manager of the community radio station (Radio Peace FM), and the dean of the faculty of Ghanaian languages of UEW. In conducting the fieldwork, informed consent was sought and obtained from all consultants. Besides, all consultants were duly remunerated upon completion of their tasks. Furthermore, as a form of giving back to the community, multiliteracy workshops were organized for a group of school teachers in the community.[27] Moreover, the production of the booklet on Simpa fishing was also a form of giving back to the community.

---

[26] All Transcriber files, video files and audio files were imported into CMDI maker through ELAN software.
[27] The multiliteracy workshops constituted a way of addressing the issue of prohibition of Simpa speaking in schools. They were organized in conjunction with Dr Ari Sherris, a Fulbright visiting scholar and expert in multiliteracy matters.

Reflections

There were a few matters that are worth reflecting on. One of such matters involved an apparent language barrier in certain cases. Although most speakers speak and understand Fante, as intimated earlier in the paper, there were a few participants with very low proficiency and comprehension in Fante. This necessitated the hiring of an interpreter who helped with translation and interpretation in those cases. Another point of reflection concerns some belief systems in the community. One of such beliefs relates to the prohibition of fishing on Tuesdays. It was and is still believed that engaging in fishing on a Tuesday is a taboo, a violation of which will lead to a misfortune. Nevertheless, one could go to sea to catch crabs on Tuesdays. Another taboo regarded filming or photographing a pregnant woman. Thus, the researcher was confronted at one point for violating this taboo[28]. These and other belief systems in the community could have implications for or even pose a challenge for researchers. For instance, the researcher became more cautious in videoing and photographing women, in case they were unnoticeably pregnant. On the positive side, the researcher chose Tuesday as one of the days for visiting the fishermen for interviews, consultation, and other interactions. Another point of reflection regards software, specifically, ELAN. Although the software had very useful features, overall, it was complicated to use, in comparison with other software. It will be worthwhile updating future versions of the software to make it more user-friendly. Another point of reflection regarded the nature of the environment in relation to the equipment, including the recorders and laptops. As a result of the sandy and humid nature of the beach, the researcher was constantly concerned about the protection of the equipment.[29] It would be worthwhile to invest in, say, water resistant equipment in such an environment, if one's project budget allows it; or, at least, in a less costly alternative, such as a plastic cover to protect the equipment.

Conclusion

The field of documentary linguistics is still evolving. Language documentation, as conceived in the literature and in practice attests to be a starting point of addressing language endangerment, besides other gains. Nevertheless, language endangerment is widespread in many parts of the world, including Ghana. Out of the approximately fifty indigenous languages in Ghana, many of them are in a similar or even worse situation as Simpa in terms of vitality. Meanwhile, most indigenous languages remain undocumented. Even, the dominant ones, including the

---

[28] The researcher explained to the accusers that it was only a myth and went further to show them from her laptop pictures of herself during pregnancy and pictures of the child from that pregnancy as a form of assurance.

[29] The equipment survived the project anyway, nevertheless, maintaining its security created a constant anxiety for the researcher.

11 institutional languages have not received adequate documentation, as conceived in documentary linguistics. Notwithstanding, the importance of linguistic diversity cannot be overemphasized. Although these facts paint a gloomy picture, all hope is not lost, as it may seem. For instance, support for language documentation is available from some governments and organizations. One such organisation is ELDP which provides support in the form of training through workshops in various parts of the world[30]. ELDP also provides various categories of research grants for documentation of endangered languages across the world (see Austin, 2010b). Another way of approaching the issue is for language and linguistics departments to include the study of language documentation in their programmes to train students to take up responsibilities of documentation. Besides, community members could be mobilized to document their language. With the outcomes of the documentation, further actions could be taken to develop languages in order to reverse the trend of language endangerment.

---

[30] A number of such workshops have been organized in West Africa, with three of them taking place in Ghana in the past decade.

References

Abaka, E.. N. (2006). The Efutu vocalic phonology. A paper presented at the Legon Trondheim Linguistics Project Colloquium on the Typology, Lexicography and Development of the Languages of the Volta Basin. Accra, Linguistics Department, University of Ghana. 9th – 13th Jan. 2006.

Agyeman, N. A. (2016). *A descriptive grammar of Efutu (southern Ghana) with a focus on serial verb constructions: a language documentation study*. (Doctoral Dissertation). SOAS, University of London.

Agyeman, N. A. (2015). Agreement in Efutu (Simpa) - a language documentation study. Paper presented at the first School of Language Conference (SOLCON-1), University of Ghana, Legon, 27-29 October 2015

Agyeman, N. A. (2013). Language use in Winneba - some preliminary observations. In Youngberg, Connor & Laura Kipp (Eds). *SOAS Working Papers in Linguistics,* 16. 259-277.

Agyeman, N. A. (2013). Language Policy of Education in Ghana: Implications for minority languages. Paper presented at the Language Endangerment: Language Policy and Planning Conference. CRASSH, Cambridge. 26th July 2013

Agyeman N. A. (2011). Serial verb constructions in Efutu. Paper presented at the 2011 International Workshop on the Description and Documentation of the Ghana-Togo-Mounting Languages, Ho, Ghana

Agyeman, N. A. (2003). On object sharing and referent sharing in Akan serial verb constructions. *Studies in the Languages of the Volta Basin*, *1*, 3-8.

Agyeman, N. A. (2002). *Serial verb constructions in Akan*. (MPhil thesis). Department of Linguistics, NTNU, Trondheim.

Agyeman, N. A. & Akrofi-Ansah, M. (2013). Language use in two South-Guan language communities in Ghana: Implications for language documentation. Paper co-presented at the Institute of African Studies Seminar Series, University of Ghana

Agyeman, N. A. & Akrofi-Ansah, M. (2016). Vowel harmony systems in two South-Guang languages (Efutu and Leteh): A case of dominant language influence. Paper presented at the 9th Linguistics Association of Ghana (LAG) Conference, GILLBT, Tamale, 26-28 July 2016

Ansah, G. N. (2014). Re-examining the fluctuations in language in-education policies in post-independence Ghana. *Multilingual education*, *4* (1), 12.

Ansah, M. A. & Agyeman, N. A. (2015). Ghana language-in-education policy: the survival of two South-Guan minority dialects. *Per Linguam,* 31(1), 89-104.

Asfaha, Y. M., Kurvers, J. & Kroon, S. (2007). Literacy use and acquisition in multilingual Eritrea. In Peter K. Austin (Ed.). *Language Documentation and Description* (Vol. 4, pp. 195-206). London: SOAS.

Ashmore, L. (2008). The role of digital video in language documentation. In Peter K. Austin (Ed.). *Language Documentation and Description* (Vol. 5, pp. 77-102). London: SOAS.

Austin, P. (2012). Language Documentation. *obo* in Linguistics. DOI: 10.1093/obo/9780199772810-0075.

Austin, P. K. (2010a). Communities, ethics and rights in language documentation. In Peter K.

Austin (Ed.). *Language Documentation and Description* (Vol. 7, pp. 34-54). London: SOAS.

Austin, P. K. (2010b). applying for a language documentation research grant. In Peter K.

Austin (Ed.). *Language Documentation and Description* (Vol. 7, pp. 285-299). London: SOAS.

Austin, P. K. & Grenoble, L. A. (2007). Current trends in language documentation. In Peter K. Austin (Ed.). *Language Documentation and Description* (Vol. 4, pp. 12-25). London: SOAS.

Austin P. K. (2006). Data and language documentation. In Jost Gippert, Nikolaus P. Himmelmann and Ulrike Mosel (Eds.). *Essentials of Language Documentation* (pp. 87-112). Berlin: Mouton de Gruyter.

Austin, P. K. (2003). Introduction. In Peter K. Austin (Ed.). *Language Documentation and Description,* (Vol. 1, pp. 6-14). London: SOAS.

Barwick, L. (2005). A musicologist's wishlist: come issues practices and practicalities in musical aspects of language documentation. In Peter K. Austin (Ed.). *Language Documentation and Description,* (Vol. 3, pp. 53-62). London: SOAS.

Bergqvist, H. (2007). The role of metadata for translation and pragmatics in language documentation. In Peter K. Austin (Ed.). *Language Documentation and Description,* (Vol. 4, pp. 163-173). London: SOAS.

Boerger, Brenda H.; Moeller, Sarah Ruth; Reiman, Will; Self, Stephen (2018). *Language and culture documentation manual*

Crystal, D. (2003). Endangered languages: What should we do now? In Peter K. Austin (Ed.). *Language Documentation and Description* (Vol. 1, pp. 18-34). London: SOAS.

David, C. (2000). Language death. Cambridge: Cambridge University Press.

Dahl, O. (1985). *Tense and aspect systems*. Oxford: Blackwell.

Dalby, A. (2003). *Language in danger: The loss of linguistic diversity and the threat to our future*. Columbia University Press.

Dolphyne, F. A. & Kropp-Dakubu, M. E. (1988). The Volta-Comoe languages. In Kropp- Dakubu, M. E. (Ed.) *The languages of Ghana* (pp. 50-90). London: Kegan Paul International for the International African Institute.

Dwyer, A. M. (2006). Ethics and practicalities of cooperative fieldwork and analysis. In Jost Gippert, Nikolaus P. Himmelmann and Ulrike Mosel (Eds.). *Essential of Language Documentation* (pp. 31-66). Berlin: Mouton de Gruyter.

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (2019). Ethnologue: Languages of the world. 22nd edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Foley, W. A. (2004). Language endangerment, language documentation and capacity building: Challenges from New Guinea. In Peter K. Austin (Ed.). *Language Documentation and Description* (Vol. 2, pp. 28-38). London: SOAS.

Forson, B. & P. J. Gingiss. (1977). Efutu. In Kropp-Dakubu, M. E. (Ed.). *West African Language Data Sheets* 1. (pp. 180-185).

Furbee, N. L. (2010). Language documentation: Theory and practice. In L. Grenoble & N. L. Furbee (Eds.). *Language documentation: Practice and values* (pp. 3-24). Amsterdam: John Benjamins.

Gippert, J., Himmelmann, N. P., & Mosel, U. (Eds). (2006). *Essentials of Language Documentation*. Berlin: Mouton de Gruyter.

Grenoble, L. & Furbee, N. L. (2010) *Language documentation: Practice and values*. Amsterdam: John Benjamins.

Hall, E. (1983). *Ghanaian languages*. Accra: Asempa Publishers.

Hale, K. (1998). On endangered languages and the importance of linguistic diversity. In Lenore A. Grenoble and Lindsay J. Whaley (Eds.), *Endangered Languages; Language Loss and Community Response* (pp. 192-216). Cambridge: Cambridge University Press.

Harrison, K. D. (2005). Ethnographically informed language documentation. In Peter K. Austin (Ed.), *Language Documentation and Description* (Vol. 3, pp. 22-41). London: SOAS.

Hill, J. H. (2006). The ethnography of language and language documentation. In Jost Gippert, Nikolaus P. Himmelmann and Ulrike Mosel (Eds.), *Essential of Language Documentation* (pp. 113-128). Berlin: Mouton de Gruyter.

Himmelmann, N. P. (1998). Documentary and descriptive linguistics. *Linguistics* (Vol. 36, pp. 161- 195).

Himmelmann, N. P. (2006). Language documentation: What is it and what is it good for? In Jost Gippert, Nikolaus P. Himmelmann and Ulrike Mosel (Eds.), *Essentials of Language Documentation* (pp. 1-30). Berlin: Mouton de Gruyter.

Franchetto, B. (2006). Ethnography in language documentation. In Jost Gippert, Nikolaus P.

Himmelmann and Ulrike Mosel (Eds.), *Essentials of Language Documentation* (pp. 183-212). Berlin: Mouton de Gruyter.

Johnson, H. (2004). Language documentation and archiving, or how to build a better corpus. In Peter K. Austin (Ed.). *Language Documentation and Description* (Vol. 2, pp. 140-153). London: SOAS.

Kambon, O. (2013). *Serial verb nominalisation in Akan*. PhD Thesis. Department of Linguistics, University of Ghana.

Krauss, Michael (2000) Preliminary Suggestions for Classification and Terminology for Degrees of Language Endangerment.

Krauss, M. (1992). The world's languages in crisis. *Language*, 68 (1), 4-10. Lüpke, F. (2009). Data Collection methods for field-based language documentation. In Peter

K. Austin (Ed.), *Language documentation and description* (Vol. 6, pp. 53-100). The Hans Rausing Endangered Languages Project (HRELP). London: SOAS.

Matras, Y. (2005). Language contact, language endangerment, and the role of the 'salvation linguist'. In Peter K. Austin (Ed.), *Language Documentation and Description* (Vol. 3, pp. 225-251). London: SOAS.

McGill, S. (2009). Documenting grammatical tone using Toolbox: An evaluation of Buseman's interlinearisation technique. In Peter K. Austin (Ed.). *Language Documentation and Description* (Vol. 7, pp. 236-250). London: SOAS.

Nathan, D. (2010a). Language documentation and archiving: From disk space to My space. In Peter K. Austin (Ed.). *Language Documentation and Description* (Vol. 7, pp. 172-208). London: SOAS.

Nathan, D. (2010b). Sound and unsound practices in documentary linguistics:towards an epistemology for audio. In Peter K. Austin (Ed.). *Language Documentation and Description* (Vol. 7, pp. 262-284). London: SOAS.

Nathan, D. (2009). Audio responsibilities in endangered languages documentation and archiving. In Peter K. Austin (Ed.). *Language Documentation and Description* (Vol. 6, pp. 101-116). London: SOAS.

Nathan, D. (2008). Digital archives: Essential elements in the workflow for endangered languages documentation and revitalisation. In Peter K. Austin (Ed.). *Language Documentation and Description* (Vol. 5, pp. 103-119). London: SOAS.

Nathan, D. (2006). Thick interfaces: Mobilizing language documentation. In Jost Gippert, Nikolaus P. Himmelmann and Ulrike Mosel (Eds.). *Essential of Language Documentation* (pp. 363-379). Berlin: Mouton de Gruyter.

Nathan, D. & Austin, P. K. (2004). Reconceiving metadata: Language documentation through thick and thin. In Peter K. Austin (Ed.). *Language Documentation and Description,* (Vol. 2, pp. 179-188). London: SOAS.

Nathan, D. & Csato, E. A. (2007). Multiliteracy, past and present, in the Karaim communities. In Peter K. Austin (Ed.). *Language Documentation and Description* (Vol. 4, p. 207). London: SOAS.

Nathan, D. & Fang, M. (2009). Language documentation and pedagogy for endangered languages: a mutual revitalisation. In Peter K. Austin (Ed.). *Language Documentation and Description* (Vol. 6, pp. 32-60). London: SOAS.

Obeng, Samuel Gyasi. (2008). *Efutu Grammar*. Munchen: Lincom Europa.

Osam, E. K. (1997). Serial verbs and grammatical relations in Akan. *Typological studies in languages* (35, pp. 253-280).

Raymond, M. (2007). Literacy work in Papua Guinea: The accidental and the planned. In Peter K. Austin (Ed.). *Language Documentation and Description* (Vol. 4, pp. 174-193). London: SOAS.

Rogers, C. & Campbell, L. (2015). Endangered languages. *Oxford research encyclopaedias*. Oxford University Press.

Schultze-Berndt, E. (2006). Linguistic annotation. In Jost Gippert, Nikolaus P. Himmelmann and Ulrike Mosel (Eds.). *Essential of Language Documentation* (pp. 213-253). Berlin: Mouton de Gruyter.

SIL International. (2018). *Why care about endangered languages?* Ethnologue: Languages of the world. Online. Retrieved on 10th October 2018

Simons, G. F. & Fennig, C. D. (2018). *Ethnologue: Languages of the world*, 21st Ed. SIL International. Online version. Retrieved on 10th October 2018.

Thieberger, N. & Musgrave, S. (2007). Documentary linguistics and ethical issues. In Peter K. Austin (Ed.). *Language Documentation and Description* (Vol. 4, pp. 26-37). London: SOAS.

Trilsbeek, P. & Wittenburg, P. (2006). Archiving challenges. In Jost Gippert, Nikolaus P. Himmelmann and Ulrike Mosel (Eds.). *Essentials of Language Documentation* (pp. 311-336). Berlin: Mouton de Gruyter.

UNESCO Ad Hoc Expert Group on Endangered Languages. (2003). Document submitted to the *International Expert Meeting on UNESCO Programme Safeguarding of Endangered Languages*. Online. Retrieved: January 2013.

Welmers, W. E. (1974). *African language structures*. University of California Press.

Widlok, T. (2005). Cross-disciplinarity in the documentation work of an anthropologists and linguists. In Peter K. Austin (Ed.). *Language Documentation and Description* (Vol. 3, pp. 12-21). London: SOAS.

Woodbury, A. (2003). Defining documentary linguistics. In Peter K. Austin (Ed.). *Language Documentation and Description* (Vol. 1, pp.35-51). London: SOAS.