

**Practicalities of language data collection and management  
in and around Indonesia**

Klamer, Marian, Owen Edwards, Hanna Fricke, Zoe Gialitaki, Francesca Moro, Axel Palmér,  
George Saad, Yunus Sulistyono, Eline Visser, Jiang Wu

*Abstract:* Researchers use different approaches when collecting and managing primary language materials through fieldwork. Yet it is important that this work is done in a transparent way, so that they can be used by other researchers, who may have other aims, as well as for the speaker community who may want to use or take note of the collected materials. In this paper we use our research experience in language data collection in and around Indonesia in fieldwork projects of three kinds: descriptive fieldwork, linguistic surveys, and projects investigating language contact. Our aim is to provide an introductory and practical guide for students and professionals who are embarking on fieldwork in or around Indonesia. Describing practical methods of language data collection, processing and management our aim is to provide a guide for any research that involves the collection of language materials, including linguistic research, oral history or literature, and ethnography.

Keywords: fieldwork, language data, survey, language contact, oral history, oral literature, ethnography

## 1. Introduction

Researchers use different approaches when collecting and managing primary language materials through fieldwork. Different projects with their own unique set of questions call for different methods. However, it is important that the data be collected and processed in a transparent way, so that they will also be useful for other researchers, who may have other aims, as well as for the speaker community who may want to use or take note of the collected materials.

In this paper we use our research experience in language data collection in and around Indonesia in fieldwork projects of three kinds: descriptive fieldwork, linguistic surveys, and projects investigating language contact. Our aim is to provide an introductory and practical guide for students and professionals who are embarking on fieldwork in Indonesia, or a neighboring Southeast Asian country. We describe practical methods of language data collection, processing and management and our aim is to stay general enough to be useful for any research that involves the collection of language materials, not limited to linguistic research but also including research on oral history, oral literature, or ethnographic research.

By providing information on the practical and methodological basics of language fieldwork, along with appendices that cover a range of practical topics, we hope to answer some of the basic questions that beginning language fieldworkers may have. We focus on discussing practical low-effort realities that are effective, and include common mistakes; we do not present ideal theories, sophisticated methods and top notch technologies. In this sense, the present article complements textbooks such as (Bower 2008; Meakins, Green, and Turpin 2018). Unlike these sources, the current paper does not focus on describing or documenting a single language from a holistic point of view (cf. Meakins et al. 2018: 8). Instead, that kind of classic linguistic fieldwork is discussed here as only one of three different types of field research, alongside language surveys and language contact studies. Each of these types of fieldwork has its own aims. The aims of surveys and contact studies are different from the description or documentation of a single language. Neither does this paper a handbook for linguistic fieldwork (Chelliah and de Reuse 2011; Thieberger 2011). Here, we do not touch on all aspects of fieldwork; only on methods of language data collection and management. By thus limiting our aims and scope, we hope to provide a source that is easy to read, free Open Access, and practical to use in the field.

This paper is unique in its strong geographical focus on Indonesia, Malaysia and Timor-Leste. Most textbooks and handbook articles on language description and

documentation published over the last decades deal with fieldwork on endangered languages in the US, Australia, or South America, while fieldwork situations in other parts of the world, including Island SE Asia, are underrepresented.<sup>1</sup> At the same time, Island SE Asia, along with Papua, is known to have the highest percentage of living undocumented languages in the world (Hammarström and Nordhoff 2012, 26), and the number of local Indonesian, Malaysian, and Philippine researchers involved in language fieldwork is growing.<sup>2</sup> There is thus an increasing need for publications clearly explaining protocols, practicalities, and challenges of collecting language data in the context of this particular region.

The paper is organized as follows. In section 2, we describe the methods and materials that most of the linguistic fieldwork projects discussed in this paper have in common. This is followed by a description of three different types of fieldwork: descriptive fieldwork on one language (section 3), surveys of dialects and languages in different locations (section 4), and investigations of language contact in one or more locations (section 5). All the projects took place in small, rural communities in Southeast Asia; most of them in eastern Indonesia, one in Malaysia. The linguists who did the research were all foreign to the communities where their research took place. In section 6 we discuss how language data collections can be archived in online repositories. The organization of this paper is intended to make it easy for the reader to read only the general section, and/or to focus on those subsection(s) that describe a project that is most similar to theirs. This means that there is occasional overlap between the sections.

It is hoped that the materials collected in the projects discussed in this paper will be useful for other researchers who are interested in knowing more about the language, culture or history of the communities that were visited, or who want to use our materials for cross-linguistic comparison. All the data is accessible online, and downloadable from the archive that is discussed in Appendix 1.<sup>3</sup> Appendix 2 provides some brief practical “cookbooks” for fieldwork focused on Island SE Asia, addressing questions such as: What goes in a fieldwork research plan? How to find consultants and compensate them for their work? How to obtain

---

<sup>1</sup> For example, of the 44 languages mentioned in Meakins et al. (2018), only one is a spoken language from Island SE Asia.

<sup>2</sup> Examples include local linguists associated with the Language & Culture Unit, Kupang (UBB), and the researchers from Indonesia and the Philippines involved in *The Oceanic and South East Asian Navigators (OCSEAN)* project, funded by the European Commission under the Horizon 2020 Marie Skłodowska-Curie Research and Innovation Staff Exchange program (MSCA-RISE-2019, [Project Number 873207](https://www.wordsandbones.uni-tuebingen.de/ocsean/?staff-dept=member)), see <http://www.wordsandbones.uni-tuebingen.de/ocsean/?staff-dept=member>.

<sup>3</sup> The appendices to this paper are available online at the Zenodo Open Repository ....

informed consent, and what would a useful informed consent form look like? What are the steps involved in a recording session with a video camera? Which kinds of metadata are collected? How to transfer data between ELAN and FLEx?

Throughout the paper we refer to native speakers who collaborate with the foreign researcher in a linguistic fieldwork project as '(native speaker) consultant' or 'research participant'; these terms are used as synonyms.

## **2. The collection and processing of fieldwork data**

This section describes the methods and materials that most of the linguistic fieldwork projects discussed in this paper have in common. It is organized following the sequence of fieldwork. It first discusses the research materials and visual stimuli used in the data collection (section 2.1), followed by recording equipment and set-up (section 2.2), methods of data collection (section 2.3), software tools used (section 2.4), data processing (section 2.5), transcription (section 2.6), and annotation (section 2.7), compensating consultants (section 2.8), ending with a discussion of some common challenges and pitfalls (section 2.9).

### **2.1. *Written materials and visual stimuli***

In the projects described in this paper we used a variety of written materials and visual stimuli including: word list, family chart, sociolinguistic questionnaire, cultural questionnaire, pre-recorded videos, and pictures books. Each is briefly described and discussed in turn.

In descriptive projects and cross-linguistic surveys, collecting word lists is one initial method of data collection. In our projects we used the 600-item LexiRumah list (Kaiping and Klamer 2018; Kaiping, Edwards, and Klamer 2019). The LexiRumah list contains basic vocabulary, region-specific vocabulary, and highly borrowable words in English and standard Indonesian.<sup>4</sup> The basic vocabulary of the list comprises the 200-item Swadesh list<sup>5</sup> combined with words that are specific to the region and cultures in and around Indonesia, such as 'betelnut', 'rice grains', 'bride price', 'mosquito'. Because we also wanted to study lexical borrowing, we also included words in LexiRumah that are known to be highly borrowable (Haspelmath and Tadmor 2009; Robinson 2015), giving preference to concepts that are commonly used and are culturally relevant (e.g., 'church', 'mosque', 'to pray'). To

---

<sup>4</sup> The LexiRumah word list is downloadable in a variety of formats from <https://lexirumah.modeling.eu/lexirumah/>, by selecting a single language and downloading the list of that language as e.g. Excel or csv file.

<sup>5</sup> For more information on Swadesh lists, see [https://en.wikipedia.org/wiki/Swadesh\\_list](https://en.wikipedia.org/wiki/Swadesh_list).

collect basic kinship terminology, we used kinship diagrams ('family charts') of two generations: one diagram with ego's generation (+0) and one generation above ego, and another diagram with ego's generation (+0) and one generation below ego. Examples of kinship diagrams are easily found online.

Sociolinguistic information on the speakers that are recorded was collected with a questionnaire asking for their personal details (name, gender, date of birth, place where they grew up, highest education, current place of living, current occupation), their own language background, use and attitude, as well as that of their family members. (An example of this questionnaire can be found in the appendix of Saad (2020)). To investigate the cultural diversity in the region we used a questionnaire of cultural traditions and practices. The list contained ~100 questions addressing the following domains: (1) the linguistic situation of the community, (2) subsistence, (3) kinship, (4) inheritance, (5) marriage, (6) settlement patterns, (7) dwelling (or house), (8) political system, (9) naming practices, (10) registers, (11) rituals and myths, (12) material culture, and (13) traditional adornment. The cultural survey was done through interviews in Indonesian with selected speakers in the community, such as displayed in Figure 1.

Figure 1: Speakers of the Kaera community being interviewed for the cultural survey  
(Abangiwang, Pantar island, May 2016)



To elicit linguistic utterances without using an intermediate language, visual stimuli were used. The *Surrey Stimuli* (Fedden, Brown, and Corbett 2010)<sup>6</sup> are a set of 40 short video clips showing brief actions (e.g. a man pulling another man, a man bumping into a tree), events (e.g. a coconut falling from a tree) and states (e.g. a bent person on all fours with a rock on his back). The set of clips was originally designed to elicit pronominal reference markers in languages of eastern Indonesia and to depict events that differ in being active vs. stative, as well as involving one or two participants who are animate vs. inanimate and volitional vs. non-volitional.

Another set of stimuli used was the *Event and Position List* (Moro and Fricke 2020). This list contains a selection of 38 video clips and pictures developed by the Language and Cognition Department of the Max Planck Institute for Psycholinguistics (see <http://fieldmanuals.mpi.nl/>) and 8 additional video clips shot by F. Moro and H. Fricke to elicit 'give' events (e.g. a girl giving flowers to another girl). The 'give' video clips were designed to study cross-linguistic variation in the expression of three-participant events.<sup>7</sup>

Narratives were elicited using the *Frog Story* (Mayer 1969), the Totem Field Storyboards *Chore Girl* and *The Woodchopper* (<http://totemfieldstoryboards.org/stories/>), the Chicken thief story (Rodriguez 2010), and the video of the *Pear story* developed by Wallace Chafe in 1975 (<http://pearstories.org/>).

Pictures from the *Questionnaire on Information Structure* (Skopeteas et al. 2006) were used to elicit constructions involving a semantic undergoer participant. Topological relation pictures from MPI were used to collect data on the usage of locative markers.<sup>8</sup>

## 2.2. Recording equipment and set-up

Most of the recordings in the projects discussed below were made using a recording set-up with a video camera and an audio recorder. The camera was set on a tripod some distance away so as to capture all the speakers present. It was connected to an external microphone with a 2-3 meter long cable. The microphone was set on a table or a chair close to the speaker(s). The cable can be taped down to prevent tripping hazards. The audio recorder was placed near the microphone to function as a backup device. For some tasks, such as

---

<sup>6</sup> Downloadable from <http://www.smg.surrey.ac.uk/projects/lor-pantar/pronominal-marking-video-stimuli/>. For a description of the clips in the Surrey list see Fedden and Brown (2017).

<sup>7</sup> The clips are available at

<https://www.youtube.com/playlist?list=PL0RRGmSasZc812xw6PyK4G3jVWZR3ayB2> and can be downloaded from <https://vici.marianklamer.org/media.html>.

<sup>8</sup> [http://fieldmanuals.mpi.nl/download/1992\\_Topological\\_relations\\_picture\\_series.pdf](http://fieldmanuals.mpi.nl/download/1992_Topological_relations_picture_series.pdf)

describing video clips, the speaker was asked to wear headphones to enable the participant to hear the sound of the clips, and to avoid being influenced by bystanders.

Because most video recorders with inbuilt microphones have poor sound quality, it is useful to use a video camera that has a 'line in' (also known as 'sound in', 'audio in' or 'mic in') audio jack to which an external microphone can be connected. Most cheaper cameras do not have this, so a researcher on a tight budget can also use the sound recordings of the audio recorder, which will often have inbuilt microphones of better quality.

Video recording was chosen because it captures the visual dimension of the language, such as gestures and facial expressions of speakers, (lip) pointing as well as the physical setting of the recording. Video recording is especially useful in conversations where people may get up, walk around, or point to certain referents; and it allows us to see whom a speaker is addressing. Also, native speakers who help with transcribing recordings generally find it easier and more engaging to transcribe video than audio, which may speed up the transcription process. Making video-based archives is currently considered best practice for linguistic documentation and description and most funding agents would require linguistic data to be video-recorded.

Besides being used as a back-up device, an audio recorder was used in situations where video was deemed unfit or impractical. They can be used in situations where a minimal set-up time, a less intrusive way of recording, or saving battery power is required.

Batteries of audio and video recorders typically die without giving an audio signal, so battery level has to be constantly visually monitored and the batteries replaced as soon as the battery level is low. Some audio recorders only save the file when the recording is stopped, so that the entire file is lost if the battery dies during the recording session. It is good practice to always carry some spare batteries to a recording session, and replacing the old ones if a particularly long recording session is anticipated. Certain audio recorders (e.g. Zoom H4) can take a long time to start the recording; the bigger the SD card the slower the start. SD cards of 4GB hold about six hours of audio, for video recording sessions of similar length SD cards of 16GB or 32GB are more suitable.

At the beginning of each recording session, it is good practice to check the sound of the recording by putting on a headphone and doing a few test sentences. This ensures that sound is actually being recorded (which may not be happening because either the microphone or the video camera is not switched on, or the cable connecting the microphone and the camera is not connected properly), and it allows the linguist to check the sound settings.

The researcher starts the recording by providing information on the language being recorded, the place, the date, the name of the participant(s), and the researcher's own name. This is to make sure that even when the filename of the recording is lost or mixed up, it is still clear what the recording is about.

### **2.3. Methods of data collection**

#### *2.3.1. Eliciting word lists and other lexical material*

Eliciting word lists is fraught with problems. This is especially the case in surveys when there is limited time available, and the researcher collecting the data does not (yet) speak the target language so that a third language must be used as an intermediate language. The risk of collecting bad or noisy data further increases when only one speaker is consulted, and the risk can become very high when this speaker has lived away from the place where the variety is originally spoken and has not used it for extensive periods. For these reasons, we apply the following best practices where possible.

Eliciting word lists (i) takes place in the location where the variety is actually spoken, (ii) involves a small group of three to six native speakers who feel confident about their language and speak the same variety with each other on a daily basis, (iii) involves native speakers who have sufficient time for compiling a word list which they consider to be representative of the forms used in their local language variety, (iv) involves a linguist who has in-depth knowledge of at least one and preferably two languages that are spoken in the region. Such a background enables the linguist to interpret the responses to the word lists more quickly and detect possible misunderstandings and other 'noise' in the responses given. Furthermore, (v) the word list used for elicitation should not provide only a single word in a gloss language, but give a clear definition of the meaning to be elicited, and (vi) specify criteria of which word(s) should be included if there is more than one word that can be used to express the intended meaning, either because they are synonyms, or because each of them is more specific than the (generic) meaning requested.

For descriptive and survey works, the following materials and protocol were used. Elicitation of the LexiRumah list (see section 2.1) was in Indonesian, the national language of education, media and government, and spoken by virtually everyone in Indonesia as a second language. For surveys in East Timor, we also used Indonesian, as this was the language of education before the independence of Timor-Leste in 2002 and is still used widely by adults to communicate with Indonesians in western Timor and beyond. A "notes"



column in the list provides extra information about Indonesian prompts that often raise questions and/or need some extra clarification.

Before the first compilation stage, the linguist went through the word list and the notes to familiarize him/herself with what was going to be asked. Then several speakers of the local variety were invited. The speakers had to be willing and able to translate the Indonesian words into their own language, and have sufficient time for the task. The linguist and the speakers worked through the list together as illustrated in Figure 2.

Figure 2: Compiling a survey word list with speakers of the Adonara Lamaholot community (Lewat, Adonara island, May 2015)



When the speakers had reached a consensus about which word was the best translational equivalent of the Indonesian prompt, the linguist repeated this word until his/her pronunciation of it was accepted by the speakers, and wrote it down in International Phonetic Alphabet (IPA). In our survey region in the Lesser Sunda Islands, we worked with fluent native speakers who lived in a stable social context with few distractions. In such situations the elicitation of a list of 600 words (using Indonesian prompts) took at least half a day, but could also last one or two days. The speed of collection might differ depending on the region where the elicitation takes place, the fluency of the speakers, and/or their cultural context.

After the first compilation stage was finished, the linguist filled in a new (blank) list with the local words, now using the Indonesian orthography (not IPA), if at all possible. This

was done so that a local speaker would be able to read the word that had been written down (reading IPA is hard for untrained speakers). A second appointment was made to audio/video record the word list.

On this second appointment, the copy of the word list with the words in Indonesian script was given to one native speaker who was willing to be recorded. An informed consent form was filled in. The linguist (or an assistant) kept the list with the IPA transcriptions. On the recording, the linguist (or assistant) read out the Indonesian word once, and the speaker repeated this word twice in his local language. The speaker had the written word list in front of him/her as a reminder, but the linguist/assistant sat next to him to assist when he felt uncomfortable reading from that list. The linguist/assistant checked if the response that was uttered was identical to the word written on the list. If there is a difference, the speaker was invited to comment on the difference. Usually, differences are due to a transcription error by the linguist, or an erratic choice of words by the recorded speaker (e.g. because of the pressure felt by being recorded). Speakers who are reasonably comfortable reading their language will often note small errors in the way the linguist captured the words in written form (e.g. an [n] should be [ŋ], a final glottal stop should be an unreleased [k], etc.). These transcription errors are corrected during the recording. In this way, the recording session not only provides a recording of the word list, but also a double-checked transcribed word list.

There are situations where the vernacular language of investigation is very closely related to the national or 'standard' language; for example vernacular varieties of Malay or Indonesian that are spoken alongside standard Malay or Indonesian. In such cases, words in both varieties have very similar shapes and meanings and the challenge is to capture the small differences. Typically, the standard language has an orthography and is prestigious, while the local vernacular is unwritten and has low prestige. In such contexts, speakers are less likely to correct errors in transcriptions or pronunciations if their corrections would lead to an increased difference between the standard orthography of the word and the orthography of the vernacular word. If the standard language is also the intermediate language used when communicating with the researcher, they are even less inclined to point out where the vernacular diverges from it. Moreover, speakers often code-switch and borrow from a standard language and are less consciously aware of the boundaries with the vernacular. This poses challenges for obtaining 'clean' data on the vernacular.

One way of dealing with these issues which was found to be particularly effective in Malaysia is to work with a group of consultants that consists of both older and younger speakers, or with parents and their grown-up children. Older speakers usually have a larger

vocabulary in the local vernacular, or they know words that were used formerly. However, they may have trouble translating words into their own vernacular, especially when the prompts are in the standard language. Elder speakers may then respond by giving a synonym in the standard language rather than a translational equivalent in the vernacular. Younger speakers are usually more aware of the differences and boundaries between their vernacular and the standard language because of formal education conducted in standard variety, and because their vernacular vocabulary is smaller than that of older speakers. As a result, they can assist in eliciting the vernacular words from the older speakers.

Elicitation of kinship terminology has its own challenges. Because kinship terms differ per culture and depend on the relative position of the speaker ('ego') in the family, they cannot be elicited by simply going through a word list with prompts in an intermediate or national language. Using a visual representation of 'family diagrams' with a position for 'ego' is more appropriate. The diagram is filled in collaboration with an adult speaker of the community, and usually a number of bystanders are also present to assist the speaker. There are two types of kinship terms: terms of *address* (e.g. 'mum') and terms of *reference* (e.g. 'mother');. In our work, we wanted to collect the terms of reference. In and around Indonesia, personal names are not typically used as terms of address, and terms of address are sometimes but not always identical to terms of reference. As a result, many speakers struggle to not mix up the two while they do the task. For this reason, we first asked them to fill in the personal names of actual people in their own family: their mother, father, aunts, uncles, grandparents, daughters, sons, and so on; this was considered a pleasant and easy task. Then the question was asked: "So how do you refer to these people when you speak about them to someone else? For example, how do you refer to *Nina* (pointing to e.g. daughter Nina in the diagram) when you say to me something like: "Tonight I am going to tell you about the time when Nina was born, Nina, who is my daughter". What would be your way to say 'my daughter'?" After the answer was provided, it was checked whether this was indeed a term of reference and not a term of address by asking: "And how would you call Nina when she is in next room?" If a different term was provided, this was likely to be the term of address. To make sure that this was indeed the case, and to make the speaker aware of the different use of both types of terms, they would be discussed in more detail, asking for further situations where one would use one term or the other. A similar discussion would be held for the following items in the chart, until the speaker was confident in keeping the notions apart. Collecting kinship terms of reference for three generations may take up to two hours.

### *2.3.2. Eliciting information on cultural traditions and practices*

Interviews on cultural practices are held with at least two elders who have experience and knowledge of the culture and traditions used in their community. In our surveys, they typically take place with two or more middle-aged or elderly speakers, usually men, one of whom would be the lead speaker, with a variable number of by-standers present, who would occasionally add their own contributions or corrections. In practice, it may not be possible to find two elders available during the days of a survey visit, in which case young adults can be interviewed. A cultural features interview usually takes a full day -- five to six hours interspersed with a short break about every hour, and a long break for a meal. Alternatively, it can be held in sessions spread over two or three days. Needless to say, a survey interview held by outsiders with a few elder speakers of a community will only provide knowledge that is already commonly known across the community, and the information will not be very specialised or deep. It is only the first step in charting cultural similarities and differences between communities.

### *2.3.3. Using visual stimuli to elicit language utterances*

Most of the visual stimuli used in our project (see section 2.1) were shown to the participants on a laptop operated by the researcher, with the instruction “describe what you see in the picture/video”. To familiarize the participant with the task, the researcher showed two video clips from the Surrey list and two video clips from the Event and Position list, and gave an example of how to describe the video clips in Indonesian. With the Pear Story video, the video was played muted, and the consultant was asked to narrate what was going on while watching the video. Another way to use the Pear Story video is to ask the consultant to retell the story after watching it as many times as he or she wishes. For the Frog Story, a printed copy of the book was used, and speakers would hold the booklet, look at the pictures and flip through the pages while they were being recorded. For some speakers it was necessary to stress that the tasks were meant to record how they normally spoke in everyday contexts, not about ‘good’ or ‘bad’ ways of saying things. In some cases it was necessary to give an example in Indonesian of how to tell the Frog Story.

## **2.4. Software tools used**

The free software applications that are mentioned in several of the projects discussed below are: (i) *ELAN* (‘EUDICO Linguistic Annotator’) (<https://tla.mpi.nl/tools/tla-tools/elan/>), a tool to transcribe recorded materials, (ii) *Toolbox* (<https://software.sil.org/toolbox/>), and (iii) *FLEx* (FieldWorks Language Explorer) (<https://software.sil.org/fieldworks/>). Both FLEx

and Toolbox are tools to build a corpus of parsed and interlinearized texts which are linked to a separate word list that can be turned into a glossary or dictionary of the language. One advantage of Toolbox over FLEx is that the information is stored as easily accessible, transferable and stable .txt files. FLEx has information stored as .xml files. On the other hand, the FLEx interface is more intuitive than Toolbox and, unlike Toolbox, FLEx is still maintained by SIL with an active user community.<sup>9</sup> Finally, in all projects we used MS Office *Excel* or OpenOffice spreadsheet applications to compile data and metadata in a way that allows it to be searched and sorted easily. As Excel or OpenOffice are not long-term stable formats, for archiving purposes the spreadsheets should also be saved as tab/comma separated (.tsv, .csv) text files.

## **2.5. Data processing**

### *2.5.1. Renaming raw data*

After the recording has been made, the first step is to copy the recording from the flash disk of the recording device onto a laptop. This is best done on the same day that the recording was made.

As the file names given by cameras or audio recorders are usually meaningless codes, copying the files also involves renaming them. The exact naming conventions should fit the purposes and content of the particular project. In general we advise to at least include in the file name the following information: (i) an abbreviation of the language name or its ISO code, (ii) the date of the recording, (iii) the location name where the recording was made, (iv) an abbreviation of the person who made the recording, and (v) an indication of its content. This naming convention has the advantage that the files can be sorted according to language and in the order in which they were recorded, and basic metadata is available at a glance without having to consult separate metadata files.

Files names should find a balance between transparency and not being overly long. Do not include spaces into your file names but rather use an underscore or hyphen to separate types of information as spaces in file names can cause problems for certain software. Full stops are only used preceding the file type extension. Video devices often automatically cut long recordings into smaller chunks of 10-15 minutes. In such cases, a number can be added to the files relating to one recording session, such as “\_1of4”, “\_2of4”, and so on.

---

<sup>9</sup> <https://groups.google.com/forum/#!forum/flex-list>

An example of a file name used in the Central Lembata project (see section 3) is “LHHF\_2016\_04\_04\_Conversation1.mp4”. This is a video recording in *mp4* format, of the Lamaholot (LH) language, made by Hanna Fricke (HF) on the 4<sup>th</sup> of April 2016, which was the first conversation recorded during the project. In case a researcher is working on different dialects of a single language that has only one ISO code, it is useful to add additional codes to identify the sub-varieties on the recordings. For instance, in the case of Amarasi (ISO code *aaz*) the Ro'is variety of Amarasi was coded as *aaz-RO*. If there is no dedicated name for the particular variety, one can use the name of the village where the recording was made to distinguish the varieties.

The part of the file name containing information on the content of the file may use a genre name, such as “Conversation”, “FrogStory”, “Legend”, “Prayer”, etc. An additional more specific content-related name helps to immediately recognize a particular recording and remember its content. For instance, instead of naming a recording “Conversation1”, it may be more useful to name it after the context in which it occurred, such as “Conversation\_Breakfast”. Depending on how useful sorting on genres is for the specific research, genre names, such as “conversation”, can also be omitted altogether in the file name. In any event, the genre of every recording will also be noted in the metadata sheet that accompanies it. If the research involves recordings of different participants doing the same task (as in the language contact studies, see section 5), each speaker must be identifiable from the file name by e.g. including their first name or a pseudonym. In case age and gender are important speaker variables, these may also be included in the file name for easy reference. An example of such a file name convention is ABGS2015\_06\_30S1\_SS\_1\_16yr\_M, where AB is the code used for Abui, GS stands for the researcher George Saad, the date of the recording was 30<sup>th</sup> of June 2015, the genre was SS, the *Surrey Stimuli* (see section 2.1) and the recording was of a participant identified uniquely as number 1, a 16 year old male. If there are speakers with identical names, they may be differentiated by an additional number (e.g. Nurmala1 and Nurmala2). If the recording contains responses to a survey (a word list or a list of cultural questions), the file name may include a reference to the survey type and the particular words or question number(s) that are being discussed on the recording.

The extension (.wav, .mp4) is generated by the computer according to the type of file and folder options can be set such that extensions become visible as part of file names. Visible extensions are very useful when the same file name is given to all the files related to a single recording as the files would be sorted together and only be differentiated by the file extension.

If two devices (e.g. a video and audio recorder) are used to record the same event, both files should receive the same name. If they also have the same extension (e.g. both produce .wav files) this means the filename of one of them has to be adapted slightly. If both files recorded exactly the same time span, one of them will be used as the main recording and the other as the backup, with “\_backup” added to the file name. In case of separate video recordings made of the same event, e.g. from different perspectives, the files could be distinguished by additions about their angle, such as “\_wide”, “\_closeup”, and so on.

### *2.5.2. Metadata*

The metadata of the recording is collected right before or after the recording and may be filled in a metadata spreadsheet (see Appendix 2.5). The transferring of the metadata to a spreadsheet is best done at the same time as when the recordings are transferred from the device to the laptop and backed up. The metadata is placed in the same folder together with the recordings (and the transcriptions of the recordings that will be made later), and all are backed up together.

### *2.5.3. Consent forms*

Consent may be asked and given before the recording is made, or on the recordings themselves, by filming the speakers reading and signing the consent form (see Appendix 2.3 for an example of such a form). In some cases when spontaneous recordings were made, consent to use the recording was sought afterwards. Any signed consent forms should be photographed or scanned as soon as possible after the recording, renamed according to the same system that was used when renaming the recordings, filed together with these recordings, and backed up. Recordings which contain the speakers’ consent also have this indicated in their file name.

### *2.5.4. Folder structure*

Many devices automatically split a long recording session into several different files. Such files may need to be merged into one before being processed further. All original, but renamed, recordings may be stored in a separate folder for raw data. The original files remain in this folder and are not directly used for further processing. A backup of this folder should be made to an external hard drive that is stored separately from the laptop, or to a folder in a cloud. Ideally, more than one backup is made and stored in separate physical locations.

A second folder with ‘working data’ keeps the recordings together with all the files that relate to them. It includes the annotations of the recordings, such as ELAN

transcriptions, FLEx texts and a FLEx dictionary, text documents and spreadsheets. For each video recording transcribed with ELAN, the working folder would typically include the following contents: (i) one video file (.mp4 or .mov), (ii) one audio file (.wav), and a transcription file produced by ELAN (.eaf). Later, a FLEX or Toolbox text file, PDFs, pictures, or other documents with notes relating to this recording could be added to the folder. If all files belonging to one recording have the same file name only distinguished by their file type extension, they will appear one after another when they are sorted by their file name. The lower level folder structure will be determined by the nature of the project. For example, in a project recording first and second languages speakers (see section 5.1) of a language, the recordings are put into two subfolders: L1 SPEAKER and L2 SPEAKER, which, in turn, contain subfolders for each speaker.

If the metadata of the recordings are compiled in a spreadsheet summarizing the metadata of all the recordings, it should be stored under the main folder, together with a list of all the data collected in the project.

In sum, a sensible folder structure keeps original recordings separate from working data, keeps together all the files relating to same recording session, and has a structure that reflects categories relevant for the particular research involved.

## ***2.6. Data transcription***

While recordings are the fundamental basis for any grammatical description, they are of limited use for those who do not speak the language without a transcription. Ideally, all recordings should be transcribed.

Software that can be used for transcription includes Transcriber, SayMore, and ELAN. While ELAN is quite popular among linguists, it requires relatively sophisticated technical skills to be set up adequately, and a fairly powerful laptop to run without crashing regularly. It can be more straightforward to use Audacity or VLC media player to replay the recordings and write the transcription in a Word document (which is regularly saved as a PDF). One strong advantage of using common software tools like these is that in places such as Indonesia many native speakers who have secondary schooling can use a laptop with this kind of software immediately, while they would need extensive training in using ELAN.

Depending on the recording device used, there may be several steps in file processing before one can start transcribing with ELAN. A video file has to be in an ELAN readable format, such as .mp4 or .mov. If the device produces files in the right format they can be copied as such from the raw data folder into the working data folder. If the files produced by



the device have a different format, they first have to be converted to .mp4 or .mov. This can be done using free conversion software.

The sound .wav file to be used with ELAN is either extracted from the original video file, or it is recorded with a separate (audio) device. The advantage of extracting the .wav file from the video file is that both the video and audio file are perfectly time-aligned and can be used in ELAN without any further steps needed. Wav files are extracted from video using *FFmpeg* (<https://www.ffmpeg.org/>). Having a separate .wav file is important for transcription with ELAN, as it enables the visualization of sound waves, allowing for easy segmentation. If it is not possible to make good quality .wav recordings using a video recorder and a high-quality external microphone, the .wav file produced by the audio device can also be used. In this case, the audio file and video file of the recording stem from different devices, so they first need to be time-aligned in ELAN (which has an inbuilt tool for this purpose) before the researcher can start segmenting and transcribing the recording. Next, the ELAN file is created, using both the .mp4 file and the .wav file, as well as a FLEx-ELAN template file designed for the appropriate number of speakers.

In the initial stages of a project, the only way to make a transcription will be with the assistance of a native speaker of the language(s) spoken in the recording. Writing down what is being said on a recording is a task that is usually unfamiliar for non-linguists, and individuals differ in how much they like it, or how good they are at it. Moreover, for a language with no established and well known orthography, native speakers will need training in how to write their language. Especially when the local language is phonologically or morphologically more complex than the national language/lingua franca, or has a different set of phonemes, native speakers are likely to feel uneasy or hesitant to write their language. They may also produce improvised transcriptions based on the national language's orthographical system, and such transcriptions are likely to be incomplete, inconsistent or erroneous.

For transcription, it is therefore best that a linguist and native speaker work together, both wearing headphones, listening to the same recording. To attach two headphones to a laptop, a splitter cable is needed. The recording is played back, and the native speaker repeats each utterance, which is then written down. Initially it would probably be the linguist who writes, with the speaker checking and correcting where necessary. As soon as the speaker feels comfortable writing, the roles can be reversed, with the linguist checking. Only when both speaker and linguist feel comfortable about the accuracy of each other's transcriptions can the task be left to just one of them.

When a speaker has had sufficient practice, it is possible to let them do the transcription first, after which the linguist listens through the recordings, adds time markings to the transcription (if no software for time-alignment is used) and flags things that do not match the recording. These parts are then double checked with the speaker. Often the non-matching parts are corrections of utterances that the speaker who did the transcription deemed 'not right' for some reason or other, or they concern repetitions and false starts that the speaker did not transcribe. Both the aligning and the diverging transcription are noted down, as both provide valuable information on the language. Letting speakers do the transcription not only saves a lot of time for the researcher, it also instills a level of trust and responsibility in speakers to master a given software and do a task that is important in the documentation of their own language. A linguist can only do a transcription alone after being exposed to the language enough to understand almost all (85-95%) of it; the remaining 5-15% that is still unclear can then be checked with a native speaker.

### ***2.7. Data annotation***

Most transcriptions will be translated, analysed and grammatically annotated (glossed). Together, all the transcribed and annotated texts will form the corpus on which the grammatical research will be largely based. It is important that all the data is part of one corpus, so that a single search can cover all materials. Either Toolbox or FLEx can be used to build such an annotated corpus.

A first pass translation of the recording can often be made during the original transcription. At initial stages the native speaker can usually supply a summary translation of parts of the text in a language shared by the linguist and native speaker. While these translations may not be completely accurate, it is important to save a copy before editing them. At later stages as the linguist gains proficiency in the target language, it may be that they only have to ask the meaning of particular unfamiliar words or phrases.

For one of the projects (the Amarasi project, see section 3) most of the translations were made by the local speaker who also made his own recordings. He translated his own recording roughly word for word into the local variety of Malay (Kupang Malay [mkn]). At initial stages the linguist relied on these translations to understand the Amarasi text. It was only at later stages when the researcher had gained proficiency in Amarasi that translations into English were made. In this case the English translations were based on the original Amarasi, not the Kupang Malay, though the original Kupang Malay translations were preserved.

In many cases, the corpus of texts is supplemented by sentences that were collected during working sessions with speakers. Typically, not all such sentences are recorded. To keep them separate from the recorded texts, they can be given an ID that indicates that they were elicited. One way of keeping natural and elicited data separate is to create a separate FLEx or Toolbox text for elicited sentences. A researcher who prefers taking handwritten notes when working with a speaker should type the elicited sentences into a FLEx/Toolbox text as soon as possible after the session, with a reference to the original page of the notebook in which it was written. It is important to keep the original notebook as there might be corrections, additions, or notes that are not carried over into the FLEx/Toolbox text because the researcher thought them to be irrelevant at first but which turn out to be insightful later.

### *2.8. Compensating consultants*

In most of the projects discussed below, speakers that were recorded were compensated for their time. In the Indonesian projects, local language experts and consultants were reimbursed 100,000 IDR (€6 in 2015-2018 which is roughly the equivalent of a teacher's daily salary) for a full day of transcription. For a session recording a speaker doing tasks which would last one to two hours, a speaker was compensated 50,000 IDR (about half a teacher's daily salary). Giving such financial compensation also allowed the researcher to find a good number of speakers in a relatively short time. A spontaneous offer to tell a short narrative or a joke, or to sing a song to be recorded was deemed a gift and was not paid. More information on how to compensate consultants in Indonesia is given in Appendix 2.2.

### *2.9. Common challenges and pitfalls*

In our field sites it can be a challenge to find a relatively quiet place to make recordings. Recording in a yard is likely to include noises from bystanders, children, chicken, dogs, motorbikes and vehicles passing by. Recording inside a house is often too hot and too dark, and may be culturally inappropriate. A recording location outside which has shade, and is removed from the (main) road is usually the best. In the evening or night time there is often insufficient light to make good recording and there may be noise from a power generator or animals. Often, it is just inevitable that recordings will have background noise.

Another reminder for data recordings is to make sure all cables are well-connected (see Appendix 2.4) as a loose microphone cable can result in a video recording without sound. Also, most recorders signal that the battery is dying with only a blinking light which will go unnoticed when a researcher is concentrating on the people being recorded and the

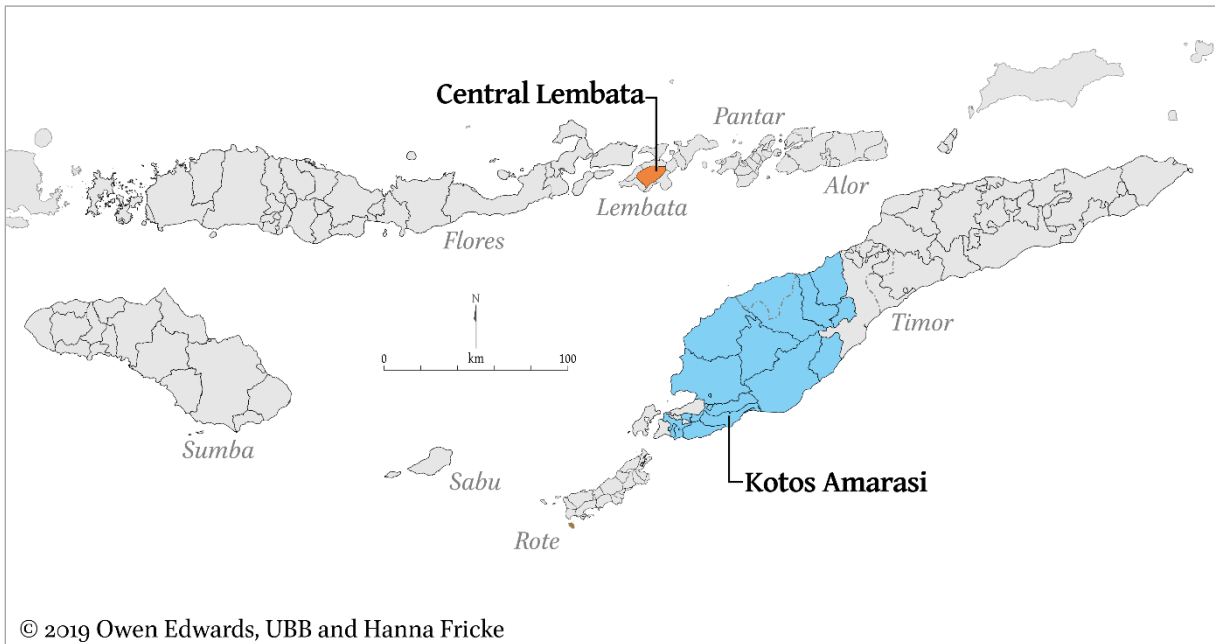
camera is a few meters away. All of the projects discussed below have lost parts of recordings this way, which were saved by the backup device.

One challenge which deserves special mention is transcription. Un-transcribed recordings without translations are generally of little use. Most beginning fieldworkers are not fully aware that transcription and annotation of recordings involves an immense amount of manual work. A common pitfall is then to make many recordings while in the field, without allowing sufficient time (or energy) for transcription and annotation. When the research focuses on a single language, a realistic goal is to transcribe and annotate 40-60 minutes of recordings per week. The time spent on processing the recording will decrease as the researcher becomes more familiar with the language, but even for a proficient speaker of a language, an hour of recording will always take a multiple of that to transcribe and annotate: depending on the phonetic and morpho-phonological complexities of the language and the experience/familiarity of the researcher, one hour of recording can take ten to forty hours to transcribe. While the manual work involved in transcribing recordings is laborious and intensive, it is not a waste of time. Being intensively exposed to the language is a good way to make oneself familiar with it relatively quickly.

### **3. Descriptive fieldwork on one language**

In this section, we describe projects involving descriptive fieldwork on one language. Such fieldwork typically aims to write a grammar of (parts of) a language which has not yet been (fully) described. For descriptive linguistic projects which aim to describe a single (variety of a) language, the study of variation according to dialects, social group, or age is not the primary aim. In the two case studies discussed here, the field research was part of a PhD project. One fieldwork project took place in Lembata island, by Hanna Fricke, collecting data on the yet unstudied Lamaholot variety of Central Lembata (ISO 639-3 lvu, Glottocode cent2336). The other fieldwork project was carried out by Owen Edwards, and took place in West Timor, collecting data on the Amarasi language (ISO 639-3 aaz, Glottocode koto1251), see Figure 3. The theses produced as the result of these projects are Fricke (2019b) and Edwards (2016, 2020) respectively.

Figure 3: The locations of Central Lembata and Amarasi in eastern Indonesia



The period of fieldwork in Lembata was carried out for a total of about nine months between 2015 and 2018. Out of these nine months, two months were spent working with native speakers of Central Lembata who lived in Yogyakarta, a city on the island of Java. 38 native speaker consultants were involved. They had different roles, such as responding to elicitation tasks, telling stories while being recorded, helping with translating recordings or participating in spontaneous conversations that were recorded. The fieldwork in West Timor totaled seven months between 2013 and 2016, and was almost exclusively conducted in Nekmese' village. A total of 60 native speaker consultants had some involvement, most as speakers in a recorded narrative or conversation. Two consultants carried out the bulk of the transcription.

### ***3.1. Data types collected***

For the Central Lembata Corpus, the following types of data were collected: (i) video and audio recordings, (ii) lexical data, and (iii) hand-written notes of elicited sentences. The recordings included word lists, descriptions of visual stimuli, free narratives, conversations and cultural practices. Visual stimuli were used to elicit brief descriptions of events, activities and states that would yield results comparable across different speakers and languages. The Surrey Stimuli, the Event and Position List, the Frog Story and the Totem Field Storyboards were used for this purpose. Free narratives were collected on topics such as the village's origin, local rituals, local traditions, everyday life activities, and so on. Conversations were

recorded with the permission of the speakers by placing a recorder in a small (3–4 people) group conversing with each other. Either the researcher or a local speaker who operated the audio recorder were part of this group. Consent was sought and given either before or after the recording was made. Cultural practices such as weaving were staged, and a local speaker gave explanations in the target language about the practice itself, and the instruments used in it.

A sociolinguistic and a cultural questionnaire were conducted in Indonesian and recorded. The aim of these recordings was to collect structured information on the sociolinguistic background of the speakers that had been recorded in the other tasks, and on the socio-cultural practices of the speaker group as a whole. These recordings were not used for the building of a corpus in the target language, so they were not transcribed. The answers to the questions were transferred to spread sheets.

Lexical data was extracted from the transcribed recordings and hand-written notebooks to build a dictionary in FLEx. Later on, this lexical database was exported as a publishable dictionary using the software LexiquePro by SIL and published as Fricke (2019a).

Notes of sentences and words were taken while talking to speakers of Central Lembata and asking for vocabulary and sentences by pointing to objects or events, or by using Indonesian as an intermediate language. Such data were initially hand-written in notebooks and then the notes were added to the FLEx database, to become part of the corpus (as a text containing elicited sentences) and the dictionary.

For the Amarasi project the same types of data were collected. However, unlike the Central Lembata project only a small number of video recordings were made. Video recordings were only made when this was initiated by the linguist's main collaborator who had his own video camera. Another difference between the Lembata and Amarasi project was that in the Amarasi project, lexical data was compiled and stored in Toolbox rather than in FLEx.

### ***3.2. Data recording, processing, and annotation***

For the Central Lembata project, the recording set-up and procedure was similar to the one described in section 2.2 above. For recordings of narratives or responses to video clips, usually only the researcher and a local speaker were present. While other people joined and listened, they did not speak during the recording. For recordings of conversations and cultural practices, the circumstances were less controlled. Often several people were present and all of them would speak on the recording.

For the Amarasi project, two audio devices with in-built microphones were used for recording: one for the researcher and one for the local collaborator. One particular characteristic of the Amarasi data set was that much of this data collection was initiated and carried out by the local collaborator, a native speaker of Amarasi, Heronimus Bani. This speaker carried the audio device around with him and made many spontaneous recordings, including his own speech (e.g. giving instructions on how to vote in an upcoming election), conversations he had with others,<sup>10</sup> or he asked other native speakers to tell a story. As a result, the data that he collected is about as natural as linguistic data can get. However, in several cases the recording quality was quite low.

The data processing in the two projects went along the lines discussed in section 2.5 above. For Central Lembata, the files were organized first in a raw data and a working data folder, as in Figure 4, and the working data folder was organized as in Figure 5.

Figure 4: File organization in the Central Lembata project

















Name	Date modified	Type	Size
Consent forms	4-10-2018 14:53	File folder	
RAW DATA	25-4-2018 10:31	File folder	
WORKING DATA	4-10-2018 14:52	File folder	
LHHF_Metadata.xlsx	27-9-2018 18:29	Microsoft Excel W...	41 KB

In the working data folder, many files relate to one recording. In Figure 5, the files of two records (“Monologue2” and “Interview3”) are shown. There is the main ELAN (.eaf) file, two further working files created by ELAN (.eaf.011 and .pfsx), the video file (.mp4) along with a converted version of this file which was needed for archiving, the audio file (.wav), a FLEX text (.flextext) which is exported from ELAN and can be imported into FLEx, and text file (.txt) which was produced by Toolbox at an earlier stage of the project.

---

<sup>10</sup> In some cases of recorded conversations, not all participants were aware that they were being recorded. Oral consent for the use of this data was then sought after the completion of the recording. Most speakers found it funny that they had been recorded and gladly consented.

Figure 5: Example of files in the working data folder

	LHHF_2015_08_14_Monologue2.eaf	22-9-2016 16:29	ELAN EAF Docum...	102 KB
	LHHF_2015_08_14_Monologue2.eaf.001	26-9-2016 12:46	001 File	102 KB
	LHHF_2015_08_14_Monologue2.flextext	23-10-2015 12:32	FLEXTEXT File	97 KB
	LHHF_2015_08_14_Monologue2.mp4	14-8-2015 5:02	MP4 Video	1.710.259 KB
	LHHF_2015_08_14_Monologue2.mp4.converted.mp4	22-11-2018 22:49	MP4 Video	1.041.862 KB
	LHHF_2015_08_14_Monologue2.pfsx	26-9-2016 12:46	PFSX File	8 KB
	LHHF_2015_08_14_Monologue2.wav	14-8-2015 12:25	Wave Sound	258.843 KB
	LHHF_2015_08_14_Monologue2_TB.txt	21-8-2015 6:53	TXT File	33 KB
	LHHF_2015_08_24_Interview3.eaf	15-9-2016 15:18	ELAN EAF Docum...	88 KB
	LHHF_2015_08_24_Interview3.eaf.001	12-12-2015 1:43	001 File	88 KB
	LHHF_2015_08_24_Interview3.flextext	23-10-2015 12:53	FLEXTEXT File	60 KB
	LHHF_2015_08_24_Interview3.mp4	24-8-2015 11:18	MP4 Video	730.340 KB
	LHHF_2015_08_24_Interview3.mp4.converted.mp4	22-11-2018 23:08	MP4 Video	964.562 KB
	LHHF_2015_08_24_Interview3.pfsx	15-9-2016 15:18	PFSX File	9 KB
	LHHF_2015_08_24_Interview3.wav	24-8-2015 12:20	Wave Sound	110.525 KB
	LHHF_2015_08_24_Interview3_TB.txt	2-9-2015 3:33	TXT File	18 KB

#### 4. Surveys of dialects or languages

Unlike the descriptive fieldwork on one language discussed in the previous section, surveys of dialects and languages typically involve fieldwork in different locations. Dialect surveys aim to investigate the internal diversity of a language, as well as the boundaries of that diversity. Language surveys typically collect lexical data to establish how languages are affiliated to each other as well as to established family groupings. In addition, surveys can be used to collect other comparative materials on the community, such as their oral history, or their material and immaterial culture. By their nature, language and dialect surveys collect relatively ‘shallow’ data, unlike descriptive linguistic or ethnographic fieldwork that focusses on one language or community.

##### 4.1 Dialect surveys

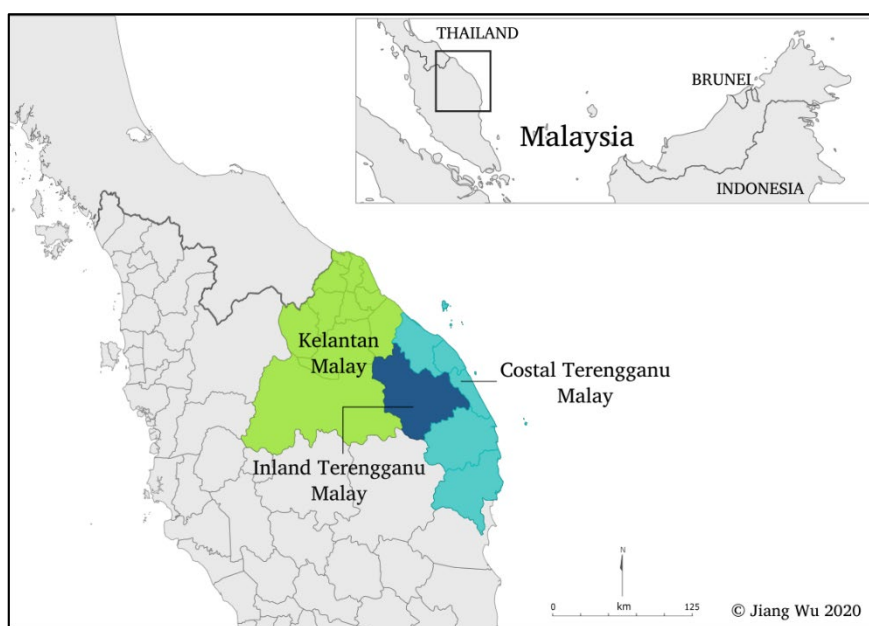
Here we discuss two case studies of dialect surveys: one on Malayic varieties spoken in Malaysia, and another on varieties of the Alorese language spoken in eastern Indonesia. Both studies are parts of ongoing PhD projects, running from 2017 to 2021/2022, and data collection and analysis are still in progress.

In the first study, carried out by Jiang Wu, the survey covers three dialects/dialect groups in the states of Kelantan and Terengganu in northeast Malay Peninsula: Kelantan Malay, Inland Terengganu Malay and Coastal Terengganu Malay (see Figure 6), all of which are members of the Malayic subgroup within the Austronesian language family. The vernacular Malayic varieties spoken in this area have been conventionally considered ‘dialects’ of Standard Malay, but initial work suggests that they are as different from



Standard Malay as other languages in the Malayic subgroup. The project investigates the relatedness of these varieties to each other and to other known Malay varieties. By applying the comparative method, their historical development can be reconstructed and their genealogical position within the Malayic subgroup can be determined.

Figure 6: Map with the locations of the Malayic varieties studied



The first fieldtrip for the Malay project was conducted in August to October in 2018, in the locations listed in Table 1. More fieldtrips are planned for 2020 onwards. The duration of fieldwork in each village varied. Because all three of the dialects also exhibit intra-dialectal differences, more than one village was visited in each location, and comparative lexical data was collected when possible.

Table 1: Malayic varieties studied, with their locations and the duration of the first fieldwork

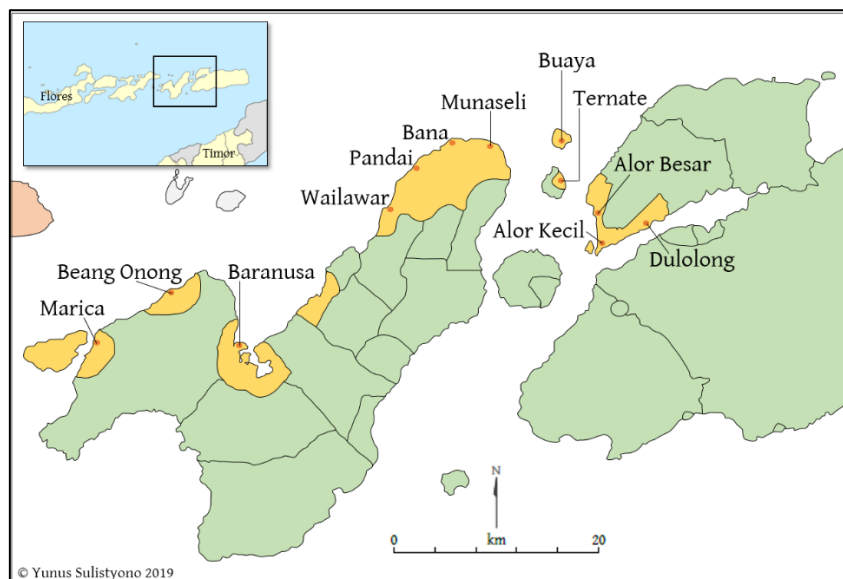
<i>Variety</i>	<i>Primary location</i>	<i>Duration</i>
Kelantan Malay	Kampung Kusial Bharu, Tanah Merah, Kelantan	3 weeks, 2018
Inland Terengganu Malay	Kampung Dusun, Ulu Terengganu, Terengganu	1 month, 2018
Coastal Terengganu Malay	Kampung Gong Sentul, Kuala Nerus, Terengganu	2,5 weeks, 2018

The first fieldtrip involved working with five Malay consultants: two speakers in Kelantan, one in Inland Terengganu and two in Coastal Terengganu.

The second study discussed here, carried out by Yunus Sulistyono, investigates varieties of Alorese (ISO 639-3 aol, Glottocode alor1247), an Austronesian language spoken on the coast of the islands of Alor and Pantar in eastern Indonesia. The aim of the study for which the survey took place is to investigate the history of the Alorese people on the basis of their language as well as oral and written historical sources. Linguistically, the relations between the Alorese dialects are investigated by reconstructing their common ancestor and describing the historical changes in the phonology and morphology of these dialects, as well as by studying patterns of lexical borrowing.

Initial fieldwork for the Alorese project took place from May–July 2018. During this fieldwork, the researcher went to a dozen villages located on the northern coast line of Alor and Pantar, where settlements of Alorese speakers are concentrated (see Figure 7). The total time spent on collecting this dataset was 12 weeks. A second fieldwork trip is scheduled for 2020 to collect additional grammatical data.

Figure 7: Map with Alorese villages surveyed



#### 4.1.1. Data types collected

Good historical comparative reconstruction is grounded in a solid understanding of the synchronic phonology and morphology of the languages compared. Since the varieties studied in the two projects discussed here have not yet been well-documented, the research

also aims to provide a basic description of the synchronic phonology and morphology (and ideally also the morpho-syntax) of the varieties investigated. This description would serve as the basis for further historical comparison. The data types collected therefore also show overlap with those collected in descriptive work on a single language (section 3).

The linguistic data collected during the first phase of the fieldwork in both projects include word lists, narratives, elicited materials, spontaneous conversations, and discussions. The collection of free-style story telling was unsuccessful in the Malay project, see below.

The word lists collected in the Malay research combine the Swadesh list with previous word lists used for research in Malayic varieties, and additional concepts added by the researcher. For the elicitation the project used the materials described in section 2.1. The same applied to the Alorese dialect project, while in this project, additional historical material was collected through oral histories told by local speakers. The historical materials were used to reconstruct elements of the socio-cultural history of Alorese speakers, such as information on the order in which different locations were settled and who the local rulers were in the past.

The historical data was collected through focused interviews, following a questionnaire with 22 questions on the history of the community developed by the researcher. In addition, if a community had written documents on their history (e.g. written historical accounts produced by local authors), these documents were photographed or scanned by the researcher. Additional information was collected using a cultural questionnaire (see section 2.1), sociolinguistic information on the recorded speakers, and village census data for all the villages that were visited. Informed consent was sought and recorded for all speakers that were recorded.

#### *4.1.2. Data recording, processing and annotation*

The Malayic varieties were recorded with a video recorder and an audio recorder, both used at the same time whenever possible, see Figure 8. This did not apply to spontaneous conversations, for which only audio recordings were made.

Figure 8: Recording set up used in the Kelantan Malay survey



Procedures and protocols to collect word lists, as explained in 2.3.1., were also followed in this research. Free-style story telling was unsuccessful in the setting of this project where the languages of investigation and the national language are very similar. Especially when the researcher was present, consultants felt as if they were being interviewed in a formal setting, and therefore switched to Standard Malay very easily (see section 2.9). Using visual stimuli such as pictures, picture books or video clips increased the chance that the speakers actually told the narrative in their vernacular, and would switch to Standard Malay less.

The influence of Standard Malay was also strong in the recording of spontaneous conversations where the researcher was present: speakers would switch to Standard Malay to talk to the researcher. To solve this, the audio recorder would be left in front of speakers while they continued talking (with their consent), and the researcher withdrew from the scene.

The organisation of data followed that described in section 2.5, with data for each variety stored in sub-folders.

The transcription of survey word lists involves several steps. The first transcription is always phonetic, but after a phonological analysis of the variety has been carried out, a phonemic transcription can be added, or the phonetic transcription has to be revised. Relating to the Malay project, it was important to treat each variety (dialect) as unique and separate, and not transcribe it through the lens of Standard Malay. A large percentage of words in the vernacular word lists are cognate with Malay words and certain sound correspondences can easily be spotted. The potential danger here is that the transcription of

word lists (and the further phonological analysis) of the vernacular varieties will be easily influenced by the researcher's knowledge of the standard language. It is important to be faithful to the recorded data, and not rely on preconceptions.

#### *4.1.3. Challenges and mistakes*

Some challenges and potential pitfalls are particularly pertinent to surveys of dialects or language varieties. This type of fieldwork can be seen as a combination of descriptive work on one language and lexical surveys of many languages. Given that the project has time limits, the challenge is to find a balance between collecting sufficient data to do a basic grammatical analysis and not to get drawn into doing a full grammatical description, which would double the research load.

For the Malay project, the biggest challenge, as mentioned above, is the close relatedness between the vernacular target languages and the intermediate standard language, which often influences consultants' answers and judgements. This requires special consideration in data collection and extra care during data transcription and data analysis. One way this was handled was to ask the consultants to listen to their own speech, and let them point out which parts might have been influenced by the standard variety.

The most optimal data for the Malay project are data from spontaneous conversations, as these appear to show the least influence from the intermediate standard language. But conversations may be difficult to obtain as explained above, and they are difficult to transcribe. It is advisable to first use elicitation and narrative data to familiarize oneself with the language, and to do a preliminary analysis. Further analysis should rely on naturalistic conversations as much as possible.

#### *4.2. Language surveys*

In the survey reported here, Marian Klamer surveyed eleven Austronesian and Papuan languages spoken on the Lesser Sunda islands listed in Table 2 with their locations and the times of the fieldwork. Besides collecting lexical data for cross-linguistic (historical) comparison, the aim was also to chart some of the cultural diversity of the region. Given that there was limited time available for the survey, the researcher visited the individual language communities for a relatively short stay, typically one week or less. There were three trips, one to East Flores and the adjacent islands of Adonara and Lembata, one to East Timor, and one to Pantar and Pura, see Figure 9. Doing the survey took approximately 4–5 days per language community, while travelling between the various locations took up to 2 days.

Figure 9: Map of fieldwork survey locations

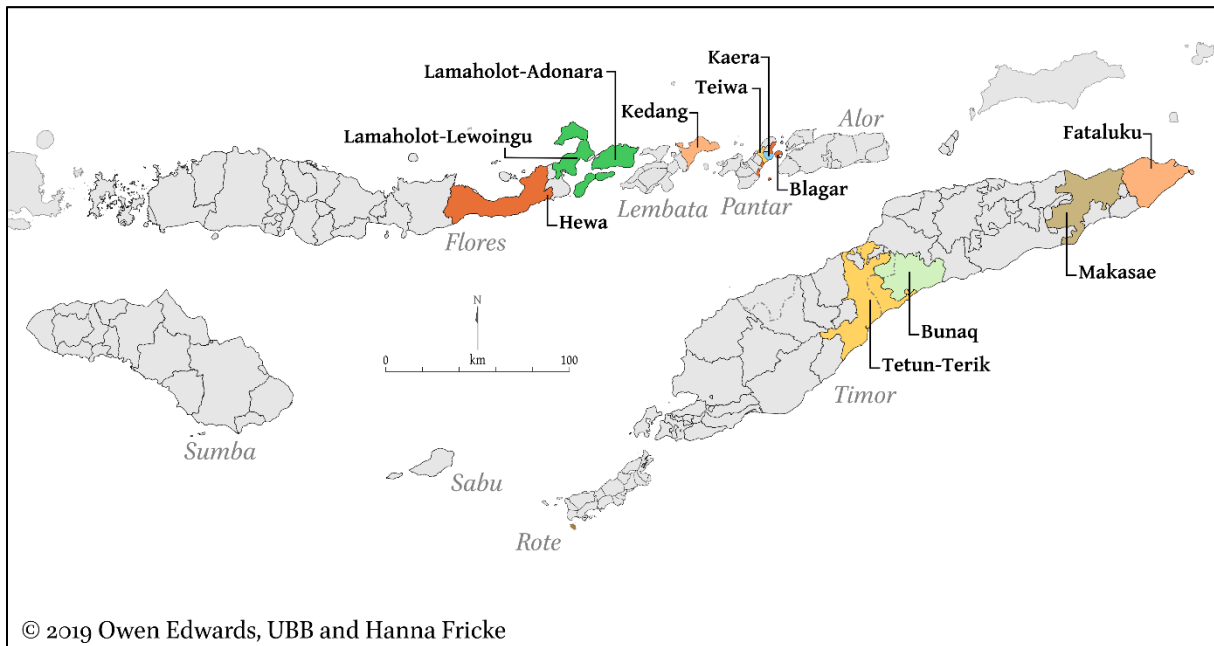


Table 2: Languages surveyed, with their locations and durations of the fieldwork

<i>Language</i>	<i>Island location</i>	<i>Duration</i>
Hewa	East Flores	several days, May 2015
Lamaholot-Lewoingu	East Flores	several days, May 2015
Lamaholot-Adonara	Adonara	several days, May 2015
Kedang	Lembata	several days, May 2015
Tetun-Terik	East Timor	several days, January 2016
Bunaq	Central/East Timor	several days, January 2016
Fataluku	East Timor	several days, January 2016
Makasae	East Timor	several days, January 2016
Teiwa	Pantar	several days, May 2016
Kaera	Pantar	several days, May 2016
Blagar	Pura	several days, May 2016

#### 4.2.1. Data types collected

The data types collected for each language in the survey included the LexiRumah word list, kinship charts for three generations, a questionnaire on the cultural traditions and practices

of the community, and sociolinguistic questionnaire for each of the speakers who participated in the recordings (section 2.1).

#### *4.2.2. Data processing and annotation*

All the data collected used the methods outlined in section 2.3, and was organized according to the language surveyed, following the steps outlined in section 2.4. Each file name includes the usual reference to language code, linguist, date, and the content of the recording. The recordings of the interviews about the cultural practices, which were held in Indonesian, were not literally transcribed, but summaries in English of the answers given to the survey questions were entered in spreadsheets, to enable comparison across the languages. For each summary, the time code of where the answer was given in the recording was also included to make it easy to go back to the original full answer if necessary. The transcription of the word list was also done directly into spread sheets.

It was decided to skip the step of transcribing the interviews and word lists in ELAN for two reasons. First, the files produced by the video camera used in this project were in AVCHD format, and in order to be used in ELAN they had to be reformatted to MP4 files. An average of 8-10 hours of video recordings was collected per community. On the laptop which was brought in the field, to reformat this amount of video recordings took an excessive amount of time and battery supply. Using the laptop for this task would mean that other tasks that had to be done with help from members of the community, such as transcription of the word list and summarizing the cultural interviews, could not be done in the limited time available. Second, all the recordings already followed a previously established 'script': a list of words and a list of questions, so that these lists can be used to locate a particular word or answer in a particular recording. For this reason, the 12-minute clips produced by the video recorder were kept as such, and each of these was given a file name that also included numbers referring to the number of the words in the word list on that particular clip, or the number of the cultural question(s) discussed on the clip.

After each trip, a folder was created combining all the data and transcriptions of that trip, as shown in Figure 10. This overall folder had subfolders according to the tasks: a folder with responses to the cultural features questionnaire, a folder with the word list responses, a folder with pictures, one with the metadata and one with the sociolinguistics questionnaire of the speakers, and bundles of scanned consent forms, organized per language community, as shown in Figure 11. At the end of the project, all recordings were reformatted before they were archived.

Figure 10: Main folders per field trip in the survey project

Fieldwork 2015 MK (Hewa Lmh_Lewoingu Lmh_Adonara Kedang)	29-11-2019 10:28	File folder
Fieldwork 2016 January MK (Bunaq Tetun_Terik Makasae Fataluku)	7-1-2019 13:55	File folder
Fieldwork 2016 May MK (Blagar Kaera Teiwa)	7-1-2019 13:54	File folder
Fieldwork 2018 May MK (Sar)	25-4-2019 10:29	File folder

Figure 11: Subfolders within each main folder in the survey project

Name	Date modified	Type	Size
Fieldwork 2016 May cultural features responses	2-5-2019 11:06	File folder	
Fieldwork 2016 May pictures	7-12-2018 12:59	File folder	
Fieldwork 2016 May recordings	21-3-2019 16:36	File folder	
Fieldwork 2016 May speaker metadata and sociolinguistics	22-5-2019 14:05	File folder	
Fieldwork 2016 May word lists responses	4-11-2019 11:26	File folder	
Informed consent forms Fieldwork May 2016 Alor Pantar Blagar_Pura.pdf	7-1-2019 13:53	Adobe Acrobat D...	361 KB
Informed consent forms Fieldwork May 2016 Alor Pantar Kaera.pdf	7-1-2019 13:49	Adobe Acrobat D...	199 KB
Informed consent forms Fieldwork May 2016 Alor Pantar Teiwa.pdf	7-1-2019 13:51	Adobe Acrobat D...	240 KB
Metadata about the recordings.txt	7-12-2018 12:19	Text Document	1 KB

As mentioned, transcriptions of the word lists were made during the trip. For these transcriptions, the linguist listened back to the recording and transcribed what was there, while consulting the list that was written up in IPA during the very first session when the word list was compiled. This third and final check ensured that the transcription and annotation of the words in the list would reflect both what had been discussed and what was recorded. Transcription of the word lists was all in broad IPA.<sup>11</sup>

The most optimal situation is such that the linguist who transcribes the word list is the same person that was also present at the compilation stage and at the time the recording was made, because this person has written notes of the earlier sessions, including the corrections suggested by the speakers before or during the recording. If the transcriber of a recording is a different person than the original collector or recorder, this may cause confusion that influences the transcription, and should be indicated as such in the metadata of that recording.

<sup>11</sup> By “broad” IPA transcription we mean all basic IPA vowel and consonant symbols, and indication of (primary) stress, nasality and segment length. Our transcriptions are phonemic for the languages for which the phoneme inventory was already known (e.g. Teiwa, Kaera, Fataluku); otherwise they are in broad IPA, but phonetic.



#### 4.2.3. Challenges and mistakes

The following are some issues that arose in the language survey project. Being aware of these issues means they can be addressed in the elicitation process. If they are not addressed, it will be harder to find cognate forms necessary for historical reconstruction, such that the varieties under consideration may appear less related than they actually are.

First, speakers may translate the Indonesian word differently in their own language, because the Indonesian prompt word is polysemous. For example, Indonesian *susu* refers to ‘milk’ or ‘breast’; and the word *sempit* ‘narrow’ in Indonesian can mean ‘narrow’ (road), ‘crowded’ (house), or ‘tight’ (clothes). In elicitation it must thus be specified what the target is. There may also be words in Indonesian with a meaning that is too generic to be translatable. For example, the Indonesian general preposition *di* ‘at, in’ often does not have an equally generic counterpart in the target language, so speakers will provide a semantically more specific adposition, or, in cases in which the language does not have adpositions, an expression that contains a locational verb (‘be at’, ‘sit’) or noun (‘inside’, ‘top’) will be given. Third, the target language may have more than one translational equivalent for the Indonesian prompt, e.g. *pukul* ‘to hit’ may render different lexemes for e.g. ‘hit (a drum)’ and ‘hit (a dog)’. Finally, not all Indonesian words have a translational equivalent in the target language. For example, causal conjunctions such as Indonesian *karena* ‘because’ are not directly translatable into Alor-Pantar languages because causal relations between clauses are not expressed with conjunctions in these languages. It may also be the case that languages lack a word for a particular concept, e.g. *murah* ‘cheap’ may be translated with various expressions such as ‘low price’, ‘short price’, ‘light price’ or ‘price goes down’. Issues of translational non-equivalence and polysemy will always cause word lists to have unclear or incomparable data. We tried to minimize the amount of such “noise” by applying best practices and a uniform protocol described in section 2.3.1 above.

A challenge of a different nature particularly concerns surveys like the one discussed here, which covered many different languages and locations, collecting data during different fieldwork trips, across different years. Without being aware of it, the researcher started to use a different folder organization and different file naming conventions for the first 2015 collections and those collected in January and May 2016. As a result, a lot of files had to be renamed and reorganized before they could be archived systematically at the end of the project.

And finally, for a survey that has only limited time available with the speech community, it is useful to bring two laptops: one to reformat video recordings and one to do

other work on. However, if the survey involves travelling to locations on foot or on the back of a motorbike, as it did in our case, then the amount of equipment that a single person can carry in a backpack alongside their personal luggage, gifts for the community, and food or water supplies, is very limited and adding a second laptop may not be feasible.

## **5. Investigating the effects of language contact in bilinguals**

Language contact studies are concerned with studying the effect of one language on another language. One type of contact study involves the investigation of the language variety spoken by adult second language (L2) speakers. The general aim of such studies is to find out what kind of changes have taken place in the language of these L2 speakers under influence of their first language (L1). An example of such a study is described in section 6.1. Another type of language contact study investigates the changes in a minority language under influence of a dominant (e.g. national) language. By studying language variation in speakers belonging to different age groups, the aim is to find out whether the variation could be induced by contact with the dominant language. An example of such a study is discussed in section 6.2.

### ***5.1. Changes in second language (L2) under influence of first language (L1)***

Bilingual speakers with a first (L1) language and a second (L2) language can show changes in their second language. These changes can be directly contact-induced when they stem from the influence of the speaker's L1 on their L2. However, they can also be indirect, when they are found in grammatical areas (e.g. inflectional morphology) that are vulnerable in all L2 grammars, regardless of the nature of the speaker's L1.

If a language community has had (and still has) a large number of L2 speakers, it is likely that the language of the whole community has undergone contact-induced changes in phonology, lexicon, morphology, and syntax. This is the case with Alorese, an Austronesian language spoken on the islands of Alor and Pantar, which has been in contact with neighboring Papuan languages for about 600 years (Klamer 2011). There is evidence that Alorese was learned as an L2 by many Papuan speakers (Klamer 2012; To appear; Moro 2018; 2019; Moro and Fricke 2020). Studying the on-going changes in the L2 of speakers today allows us to make inferences about the changes that have happened in the past, and help us reconstruct the history of Alorese and the reasons it has the structures it has today (e.g., having an impoverished morphology, genealogically unexpected grammatical patterns, etc.).

To detect on-going changes, both a quantitative and qualitative analysis is necessary. Ideally three types of samples are collected: a sample of Alorese L1 speakers, a sample of Alorese L2 speakers (e.g. Adang-Alorese bilinguals), and a sample of L1 speakers of language X, with no knowledge of Alorese, where language X represents the L1 of the Alorese L2 speakers (e.g. Adang). This L1 differs per speaker, as Alorese L2 speakers come from different linguistic backgrounds.

The Alorese L1 sample serves as a baseline to detect divergence between the speech of L1 and L2 speakers. The sample of L1 speakers of language X is necessary to demonstrate the direction of change if one wants to argue that a given change in the language of the L2 speakers stems from their L1. The Alorese L1 and the non-Alorese L1 samples can be smaller than the sample of Alorese L2 speakers, as L1 speakers are generally expected to be more homogeneous than L2 speakers.

During a fieldwork trip, from May to August 2016, the researcher Francesca Moro collected data from 13 Alorese L1 speakers, and 24 Alorese L2 speakers on the islands of Alor and Pantar. The locations of data collection are indicated in Figure 12, and the number of speakers for each sample is given in Table 3.

Figure 12. Map of fieldwork locations of the Alorese project

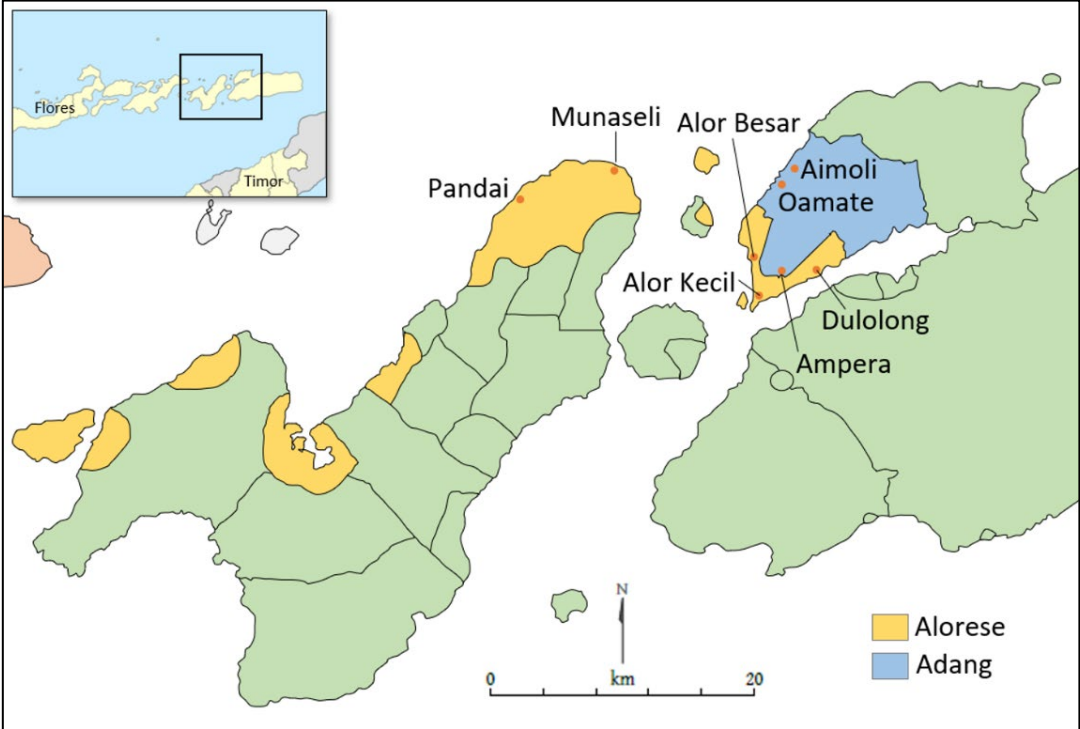


Table 3: Number of speakers per sample in the Alorese project

Island	Villages	Date of fieldwork	Language	L1 speakers	L2 speakers
--------	----------	-------------------	----------	-------------	-------------

<b>Alor</b>	Alor Besar, Alor Kecil, Dulolong	May-June 2016	Alorese	6	12
<b>Alor</b>	Oamate, Aimoli, Ampera	May-June 2016	Adang	7	-
<b>Pantar</b>	Pandai, Munaseli	July-August 2016	Alorese	7	12
			Total	13 (Alorese) 7 (Adang)	24

The difference between the islands of Alor and Pantar is that on Alor, Alorese is in contact with only one language, Adang; therefore, the Alorese L2 speakers recorded on Alor all have Adang as their L1. For this reason, a sample of seven Adang L1 speakers was collected as well. On Pantar, Alorese is in contact with different languages, so the background of the L2 speakers is more heterogeneous. In terms of their L1s, the breakdown of the 12 Alorese L2 speakers recorded on Pantar is the following: Kroku (five speakers), Blagar (three speakers), Teiwa (one speaker), Sar (one speaker), Kaera (one speaker), Klamu (one speaker). Because of time limitations, it was not possible to record a sample of L1 speakers of each of the six different first languages.

For historical and culture-specific reasons, almost all speakers recorded are women (there are only two men, one in the Alorese L1 sample and one in the Alorese L2 sample). This choice was motivated by the fact that in the patrilocal Alorese society, it is usually the woman who moves into the husband's village. Hence, today, as well as in the past, the majority of L2 speakers living in Alorese villages are women who married an Alorese man.

#### 5.1.1. Data types collected

The types of data collected are video recordings of speakers performing four production tasks and a sociolinguistic interview. The production tasks were: (i) a free narrative in which speakers were asked to tell a fairytale or a personal experience, (ii) the Frog Story, (iii) the Surrey Stimuli, and (iv) the Event and Position list. The free narrative and the description of the Frog Story elicit (semi-)natural speech, and were used for data mining research. The two elicitation lists constrain the speaker to tell what she sees in the video clips, and they target specific grammatical constructions that are expected to be vulnerable in language contact (e.g., *give*-constructions, see Moro and Fricke 2020). Together, these types of data provide a varied sample that ensures ecological validity and comparability across speakers.

The researcher recruited participants thanks to the assistance of some Alorese and Adang community members to whom she had explained the aim and the methodology of

her research. First, they thought of a number of L1 and L2 speakers in the village or neighboring villages that matched the requested profile, then they acted as ambassadors, by visiting the house of these speakers and, when possible, fixing an appointment for the recording on the next day(s).

#### 5.1.2. Data recording, processing and annotation

After the equipment was set up, the speaker was asked to read and sign a consent form written in Indonesian (see section 2.5.3 and [Appendix 2.3](#)) An illustration of the recording set up is Figure 13.

Figure 13: Video recording of one Adang speaker in the village Oamate, Alor, June 2016



All participants started by re-narrating the Frog Story while leafing through the book. Once the Frog Story was recorded, the speaker was asked to tell a free narrative (a traditional story or a personal experience). Then, the Surrey Stimuli list was recorded, followed by the Event and Position list. Finally, the participant was asked a number of sociolinguistic questions concerning her life and language history. The sociolinguistic interview was carried out in Indonesian to facilitate the subsequent extraction of information by the researcher.

The tasks were ordered in this way to follow a cline from more demanding to less demanding, as the Frog Story turned out to be quite a difficult task for many speakers (Klamer and Moro To appear), while the video clip description gave very little problems. The sociolinguistic interview was the easiest task, and therefore was done last.

The speakers were recorded in familiar environments, in or by their own homes, or they were invited to the local house where the researcher was staying, the village church, or a local community health clinic. They were recorded individually but the presence of (many) onlookers was inevitable, see Figure 13.

The data were processed following the procedure outlined in section 2.4. Because the aim of this research was to discover differences between L1 and L2 speakers, this determined the primary division in the folder structure. The .MTS files were stored in a raw data folder which also contains the backup files created by the audio recorder, while the .mp4 and .wav files were stored in a working data folder. Within the raw data and working data folders, the recordings were put into two subfolders: “L1 speaker” and “L2 speaker”, these, in turn, contain subfolders for each speaker, see Figure 14.

Figure 14: Folder organization for the Alorese project



The video recordings were transcribed and translated into Indonesian during the fieldwork with the help of native speakers who were paid for their work, following the procedure outlined in section 2.7. A speaker of Alorese and a speaker of Adang did independent transcriptions of their respective languages, which were later checked by the researcher. Investing time and energy to train native speakers to do transcriptions and translations can be a very rewarding experience for both.

The Alorese transcriptions are more reliable than the Adang transcriptions. While the researcher could carefully check the Alorese transcriptions as she was becoming more and more familiar with the language, the same cannot be said for the Adang files. It is a challenge to work with two or more languages from different families because the languages are lexically and grammatically very different, and it is hard to familiarize oneself with both in a fieldwork of just four months.

The transcribed files were imported into five different FLEx corpora: one corpus for Alorese L1 speakers on Alor, one corpus for Alorese L2 speakers on Alor, one corpus for Adang L1 speakers on Alor, one corpus for Alorese L1 speakers on Pantar, one corpus for Alorese L2 speakers on Pantar.

### *5.1.3. Challenges and mistakes*

One of the main challenges of the Alorese second language project was to collect enough quantitative data in a limited amount of time. Finding L2 speakers with the right criteria (e.g. Papuan language as L1, learned Alorese in adulthood, relatively fluent in Alorese) was sometimes difficult, and it had to be explained many times that the researcher was looking for speakers with a Papuan language as L1 not just any foreign language (sometimes the researcher was asked to interview Alorese L2 speakers coming from Flores or Timor). It also happened that in the quest for L2 speakers, the researcher made an agreement to record a young lady who was supposed to meet the criteria. Only when starting the recording, it became clear that she could barely speak Alorese and she performed all the tasks in Indonesian. Unfortunately, when native speakers made appointments for the researcher, especially in other villages, it was not possible to verify the proficiency of the speaker beforehand.

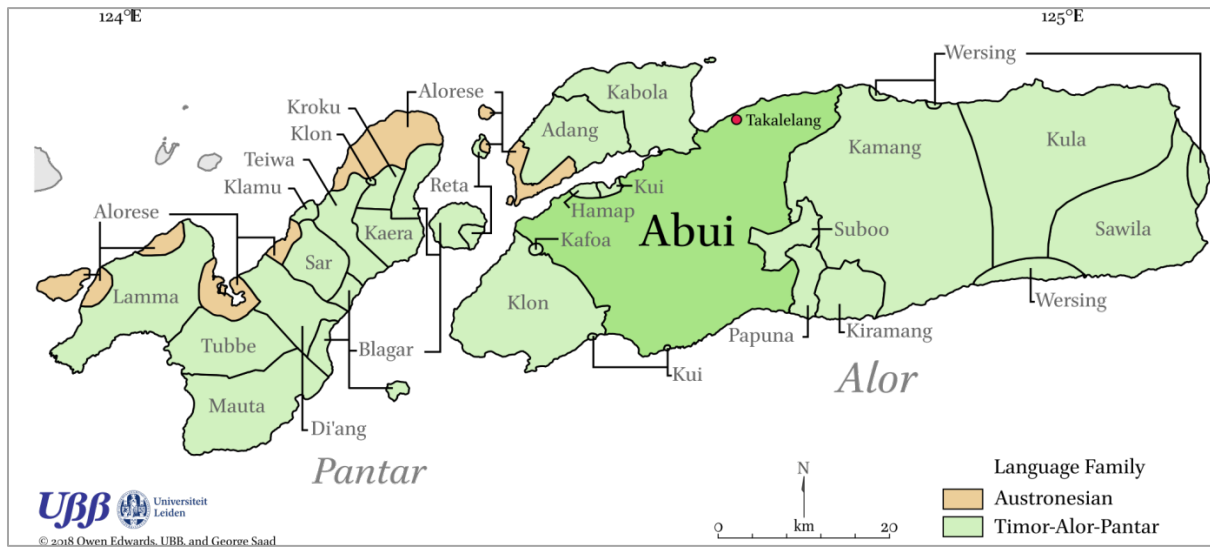
Another issue that emerged was that the Alorese language community would feel offended if the researcher had started recording L2 speakers before recording L1 speakers, and if only women were recorded without recording any men. Even though the researcher had explained the purpose of the research to the head of the village and to prominent figures in the village, they all felt confused by the fact that I only wanted to record the Alorese spoken by non-native speaker L2 women. So, out of respect, some sessions with Alorese men were recorded as well. This had the advantage that people began to familiarize with what the researcher was doing and, especially women, were less hesitant to perform the tasks when they were recruited in the following days.

### *5.2. Changes in a minority language under influence of a dominant language*

There are also contact studies investigating the influence of a majority (e.g. national) language on a minority language. In the study discussed in the present section, Abui is the indigenous minority language, while Alor Malay functions as the majority language. Abui (ISO 639-3 abz; Glottocode abui1241) is a Papuan (non-Austronesian) language spoken on the island of Alor in eastern Indonesia, see Figure 15. Alor Malay is a regional variety of Malay spoken as a lingua franca on Alor and Pantar. Speakers of Abui often consider Alor Malay and Indonesian (the national language of Indonesia, which is lexically very similar to Alor Malay) as a single language. The research was conducted in Takalelang, on the north coast of Alor Island in eastern Indonesia, as part of a PhD project by George Saad. Data collection involved three fieldtrips: ~2,5 months in 2015, ~2 months in 2016, and ~2 months in 2017. The research results are reported in Saad (2020).



Figure 15: Abui in the Alor Archipelago



In the Abui community, younger speakers are more dominant in Alor Malay, while older speakers are more dominant in Abui. It was observed that the Abui spoken today by youngsters and by older speakers showed interesting grammatical variation. To see whether any of this variation might have been induced by contact with Alor Malay, linguistic data from younger speakers was compared with data from older speakers. For the study, Abui data was collected from over 60 speakers, divided across four age-groups: (pre-)adolescents aged 9–16 years, young adults aged 17–25 years, adults aged 26–34 years, and elders aged 40+ years. As life-stages may differ across cultures, ethnographic interviews were conducted to determine which culturally relevant life-stages could be used to establish age-groups based on ‘emic’ notions. In addition, sociolinguistic data on all the participants was collected, to discover: (i) how the language use and exposure of the Abui speakers might have changed over time, and (ii) if there were other possible variables besides contact that might explain the observed linguistic variation. For example, exposure to, and use of the minority language could be affected by the fact that pre-adolescents have different access to certain speech registers or language practices than older speakers. Some of the observed variation could also be due to factors such as children’s residence with grandparents, or being the child of a school teacher.

### 5.2.1. Data types collected

Two types of data were collected: (i) linguistic data, and (ii) ethnographic and sociolinguistic interview data. The linguistic data consisted of experimental data and conversational data. The experiments involved a production task using the Surrey Stimuli video clips set (see



section 2.1), and a comprehension ('forced choice') task. The variation found in the results of the production task was used to identify four linguistic variables that could potentially differ across age-groups, and to study how significant they were. (For example, one variable was the use of a dedicated reflexive possessor affix in possessed object NPs, versus using a different, more general possessive affix in such contexts (Saad, Klamer, and Moro 2019; Saad 2020). As a second step, it was investigated whether the production of these variable features differed from their comprehension, by using a forced choice task with sixty participants (many of whom had also done the production task). The forced choice task consisted of 30 video clips with 60 accompanying sentences (two for each clip). The task required the participants to watch a video clip, then listen to two expressions describing the event on the clip (spoken by a native speaker of Abui), after which they had to decide which of the two descriptions fit the video best.

The conversational data in this project was collected as a dataset that is more 'ecologically valid' or 'natural' than the materials elicited with video clips, the latter of which could serve as a background against which to evaluate the experimental data. It included both spontaneous and directed conversations. Spontaneous conversations entailed the recording of speakers who were already engaged in conversations, while directed conversations consisted of the researcher recruiting specific individuals and asking them if they would like to sit down together and converse freely on everyday topics. The directed conversations were held to collect materials from younger speakers in particular, as they were often seen running around as opposed to sitting in one place conversing.

The ethnographic and sociolinguistic interview data were conducted to understand the speech community in more detail. Specifically, ethnographic interviews discussed community-wide matters with elders in the community, such as distinct Abui notions of age-groups and the language they use, and the history of schooling in the community and how it might have affected the shift to Alor Malay observed in youngsters. The sociolinguistic interview data provided information on the participants of the experiments, such as their age, the languages they speak, and with whom and where they speak these, the language(s) of their parents, their residential history, education, and so on (see the appendix in Saad 2020).

#### *5.2.2. Data recording, processing and annotation*

The recording set-up was as described in section 2.2. An illustration of a recording set-up is shown in Figure 16. The set-up of the forced choice task is shown in Figure 17.

Figure 16: Ethnographic interview with several speakers



Figure 17: Collection of Forced choice task



Once the data was collected, the same protocol as described for other projects in previous sections was followed. Interview data and linguistic data involve different processing methods. The sociolinguistic interviews were recorded, and answers were instantly entered into a master spreadsheet with the questions. The recordings of these interviews were not transcribed. As for the ethnographic interviews, many parts of them were transcribed. This was done using ELAN after having returned from the field. Since the

interviews had been conducted in Alor Malay, the researcher could transcribe them without help.

The linguistic data were transcribed and translated with the help of several Abui speakers that were already skilled at using ELAN, because they had been involved in linguistic research before. The researcher could hire their services so that a large number of recordings could be processed in relatively short time. The researcher always made sure to check the transcriptions. A full day of transcription would usually amount to 15–30 minutes of transcribed and translated text; the variation in time depended on the genre of the recording.

Having multiple people transcribe recordings also allowed for the metalinguistic judgment of variation. In the first phase of the project, this was crucial for establishing the linguistic variables to focus on for the research. Speakers were always instructed to transcribe the speech as closely as possible to what was being uttered. They were specifically asked to avoid any prescriptive judgments, and be true to any variation and deviations from the norm they would observe. They were, however, also afforded room to engage with these deviations, as they were told to mark down those sentences that appeared to deviate from the norm, and could suggest “corrections” in a separate note tier in ELAN. This made it possible to both collect data on variation, as well as *judgments* on that variation. It also provided a battery of examples from which to select a linguistic variable for further investigation.

### 5.2.3. *Challenges and mistakes*

Four challenges were encountered. First, it would have been ideal to have conversational data from each of the participants from whom experimental data was collected, so as to have a complete portfolio for every speaker, including: (i) a sociolinguistic questionnaire, (ii) a Surrey Stimuli task, (iii) a forced choice task, and (iv) a conversation. This would have allowed a comparison of experimental and conversational data. In reality it turned out to be impossible to have all the speakers that had taken part in the Surrey Stimuli experiment sit down in smaller groups and converse. While this was done for a few speakers, this data was not sufficient and as a result only the data from the Surrey Stimuli and the forced choice task could be used to study variation across the age-groups. In addition, while the researcher did collect many conversations, finding the time to transcribe them all proved difficult. The Surrey Stimuli, on the other hand, were very easy to collect and transcribe.

Second, many of the recordings made during the first field trip had missing data in the Surrey Stimuli task. This was due to two reasons. In the first year, it was still unclear which variables the research would focus on. Therefore, if there was a specific video prompt with e.g. a man falling over, and the speaker would provide a description not including a description of the act of falling over, the researcher would ask them to repeat and try to incorporate the act of falling in their description. This was also aggravated by the fact that at that stage, the researcher's proficiency in Abui was not yet at a level where he could understand whether the video clip was being described accurately. In later fieldwork trips, the researcher ensured that speakers were focusing on the right elements of the clips, although this still meant that not all data points were present.

Third, data was collected from speakers across a time-span of three years. This complicated the age group division used for the study. For instance, if data from a speaker born in 2000 was collected in 2015, they would have been 15 years old and thus included in the group of (pre-)adolescents; however, if the first recording of their classmate, also born in 2000, was made in 2017, the classmate would have been 17 years old and thus be placed in the group of young adults. What complicated matters further was that the decision to do a forced choice task was made after the Surrey Stimuli data had already been collected in 2015 and 2016. As the forced choice task was collected in 2017, there were a few instances where a speaker would be in one age-group for the production task, and in another age group for the comprehension task. For these cases, the age of the speaker when they were originally recorded for the Surrey Stimuli was taken as their age for the forced choice task. Thus a 15 year old (pre-)adolescent recorded in 2015 doing the Surrey Stimuli, was counted as a (pre-)adolescent for both the production and the forced choice task, even though he was 17 years old when doing the forced choice task.

Fourth, the sociolinguistic questionnaire was edited several times, so that it became difficult to collect comparable data for all speakers involved. In 2015, many of the questions were answered; however, in 2016 and 2017, many new, more relevant questions were added and some of the old (now deemed irrelevant) questions were left unanswered. This means that there are a number of holes in the sociolinguistic data collected across the three field periods.

## 6. Archiving the materials with the language archive PARADISEC

PARADISEC has guides explaining how to deposit material available online at <https://www.paradisec.org.au/deposit/> . Here we provide a brief summary of the four main tasks that need to be done for archiving with PARADISEC.

### 6.1 Creating a collection

The first task to be carried out is to get in contact with PARADISEC to establish a collection where your files will be stored. Consult the guide *Getting started with PARADISEC* at <https://www.paradisec.org.au/deposit/> for how to do this and contact the archive at [admin@paradisec.org.au](mailto:admin@paradisec.org.au). Your collection will be created with an identifier (CollectionID), typically the initials of the collector, or another ID determined by collector(s) in consultation with the archive manager.

Once your collection is established, you will need to archive your data. A single recording session in PARADISEC is referred to as an *item*. Each item can be associated with multiple files, e.g. a sound recording, video recording, transcription, photos etc. These are the four tasks you need to do to archive items in your collection.

1. Filling out the metadata
2. File re-naming
3. Filling out the deposit form
4. Sending the data to the manager

### 6.2 Metadata

The first task is filling out the metadata. If you have followed the instructions in section 2.5.2 above, and Appendix 2.5 of this guide, you will already have much of this metadata, but it needs to be converted into the format accepted by the archive. PARADISEC provides a *basic metadata spreadsheet* (downloadable from <http://www.paradisec.org.au/deposit/>) which contains the basic metadata which can be associated with an item.<sup>12</sup> Each row in this spreadsheet belongs to a specific item, and each column contains different kinds of metadata. The column headings explain what kind of information can go in each column (not all cells need to be filled in). Some can only receive certain words from a list (given in the headings) and some are “free text” meaning there are no restrictions. You can provide as much or as little information in the metadata as you want. We recommend giving lots of information.

---

<sup>12</sup> More metadata not taken from the spreadsheet itself (such as geographical location) can be added once the collection is set up. If there is metadata that needs to be added beyond that provided by the basic metadata spreadsheet this can be directly added to the collection and item levels.



One example of an item (recording session) in the Amarasi collection is *aaz20130905\_01* ([https://catalog.paradisec.org.au/collections/OE1/items/aaz20130905\\_01](https://catalog.paradisec.org.au/collections/OE1/items/aaz20130905_01)). (See below for item naming advice) There are seven different files associated with this item, which are explained in the item description.

- video recording (in three separate formats: .mp4, .mxf, and .webm)
- audio recording (in two separate formats: .mp3 and .wav)
- transcription
- transcription with morpheme-by-morpheme gloss and free translation

The name of the item does not need to be descriptive as there are separate metadata fields for *item title* and *item description* where such information can go, as well as many other metadata filed which provide information on the item.

### 6.3 File re-naming

Once you have filled out the metadata, you need to re-name your files according to PARADISEC's conventions (<https://www.paradisec.org.au/deposit/file-naming/>). The main purpose of the name of the file is to differentiate it from other files. The metadata is stored separately, as discussed above. Each filename in PARADISEC has three parts separated by hyphens: CollectionID-ItemID-ContentFile. One example of a file in the item given above is: *OE1-aaz20130905\_01-transcription.pdf*.

The first part *OE1* is the name of the collection. This is predetermined by how your collection is named. In our example *aaz20130905\_01* is the name of the item. This needs to differentiate items (recording sessions) from one another. Thus, we recommend choosing as your item name some feature which varies and differentiates all your recording sessions. If they are all of different speakers, you could use speaker initials. If they are all of different varieties/languages (as in a survey) you could use the name of the varieties. In the case of the Amarasi collection, the main variable was the date of recording, *aaz20130905\_01* is the first (*\_01*) recording made on the fifth of may 2013 (*20130905*). The language ISO 693-3 code is also redundantly in the file name (*aaz*).

The final part of a file name is the *ContentFile*. In the Amarasi this was used to differentiate the different files associated with each item: *OE1-aaz20130905\_01-transcription* is a transcription, *OE1-aaz20130905\_01-recording.wav* is the original recording, *OE1-aaz20130905\_01-video.mp4* is a video recording, and so on. This part of the file name can also be used to mark multiple parts of a long recording session.

#### 6.4 Sending off your archive

Once you have renamed all your files, you need to download and fill out the deposit form (available <https://www.paradisec.org.au/deposit/>) and send your collection to the archive manager. Consult the guide *Getting started with PARADISEC* at <https://www.paradisec.org.au/deposit/> and get in contact with the archivist ([admin@paradisec.org.au](mailto:admin@paradisec.org.au)). They will arrange a way to have your data transferred. They will also check that items are named appropriately and that metadata has been filled out appropriately.

### 7. Summary and conclusions

In this paper we have discussed a number of best practices and tips for collecting and managing data that apply to all the projects presented here. These are potentially useful to future projects to be carried out by students and colleagues in or outside Island South East Asia. Here we present a summary of these.

- Make recordings with a video and audio recorder, using the audio recorder as the backup.
- Record the consent procedure if possible.
- Make back-ups of the recordings at the end of every recording day.
- Use adequate file naming conventions.
- Fill in metadata immediately after each recording session.
- Organize data so that raw data is separated from working data.
- Use a folder structure that reflects the research questions of the project.
- Make regular backups of all data and keep these in different physical locations.
- Reserve sufficient time (and energy) to transcribe, annotate and translate all the recordings that contain language data.
- Extract and fill in sociolinguistic data of speaker(s) immediately after recording, while the memory of the person interviewed is still fresh and details are still remembered.
- Invest time in training a native speaker to do transcriptions.
- If possible, bring two laptops to the field, reserving one for a native speaker transcriber, or for reformatting files.

While many methodological aspects are similar across different projects, there are also differences in how specific projects carry out their data collection and management. For

example, file names always should contain the date of the recording and an abbreviation for the language, but which other elements are essential depends on the project. Such information could include genres, abbreviations for the content of the recording, abbreviations of speakers' names or it could include the age and gender of speakers, or information on whether they are recorded speaking their first language or second language. In addition, the folder structure has to be in line with the project aims. If the file names contain enough information and no essential subcategorization of recordings is necessary, all recordings may be stored in one folder (see Central Lembata project in section 3). However, for a survey project, such as the Alorese dialect project (section 4.1) or the survey of different languages (section 4.2), subfolders for villages, locations, or languages are a useful way to organize the data. In a project that looks at different groups of speakers according to socio-linguistic variables (such as the projects on language contact (section 5), these variables may be used to organize the data in subfolders.

As members of the same team at Leiden University, we have decided to write this article, bringing together our fieldwork experiences, with the hope that others (as well as ourselves) can learn from each other's practices and mistakes. To all future language data collectors, we wish them the best of luck with their fieldwork.

**Appendix:** The appendices to this paper are part of the version in the Zenodo Open Repository.



## Appendix 1: Data sets discussed in the paper

For those who would like to see the data sets collected through the fieldwork discussed in the present paper, or who wish to continue to work with our data, we include information on where they have been archived in online Open Access repositories. The Amarasi data (section 3) is archived with PARADISEC and can be found at <https://catalog.paradisec.org.au/collections/OE1>). All the other data sets are archived with The Language Archive (TLA; tla.mpi.nl) as the Language Collection “Eastern Indonesia and East Timor”, (Klamer et al., n.d.) The persistent identifier of the collection is <https://hdl.handle.net/1839/06afa50e-ae9-4adb-a6a7-d7496a8a47fc>. The collection contains data on 25 language varieties. Each recording in the archive has a field “Detailed Metadata” as part of its archive entry. Data in The Language Archive can be searched by search filters (by Language, by Contributor, by Genre etc.). Many users would not use the search filters, but instead want to browse through the collection, so we present the structure of the archive to enable easy browsing.

The first tier of our collection “Eastern Indonesia and Timor Leste” contains four sub-collections based on the *geographical region* where the data was collected: (i) the Alor Archipelago, (ii) Flores and the Solor Archipelago, (iii) the Maluku Archipelago, and (iv) the Timor Archipelago. Within each of these sub-collections, there is a second-tier with sub-collections organized by *language*. For example, in “Eastern Indonesia and Timor Leste” we find the folder of the geographical region “Alor Archipelago”, and within that folder we find a number of sub-collections that are organized by language including: Abui, Adang, Alor Malay, and so on, see Figure 18. Under each language node, the researcher has determined how their data would be best organized in the archive. For instance, consider the geographical region “Flores and Solor Archipelago” folder in Figure 19, which has sub-collections themed on individual languages: Atadei Painara, Central Lembata, Ende, etc.. Each of these language nodes comprise bundles of files that are different in size, type and content. For instance, the Atadei Painara Collection has only a few recordings (a word list and a prayer), while the Central Lembata corpus contains dozens of files of different types.

Figure 18: Structure of the Alor Archipelago sub-collection

The screenshot displays the Language Archive interface for the Alor Archipelago sub-collection. On the left, there is a search bar and a filters sidebar. The filters include:

- Access Level:** Restricted (300) and Open (193).
- Contributor:** Francesca Moro (276), George Saad (189), Marian Klamer (15), Amos Sir (3), and Jeroen Willemsen (2).
- Language:** Abui (190), Alorese (169), Indonesian (74), Adang (54), Alor Malay (22), Alor (3), Sar (3), and Kaera (2).
- Country:** Indonesia (498).
- Genre:** Elicitation Stimuli (175), Narrative (108), Sociolinguistic Questionnaire (86), Forced Choice (46), Conversation (36), Interview (12), Wordlist (5), Transcriptions (4), Word list (4), and EAF transcripts (3).

On the right, the main content area shows the "Alor Archipelago" sub-collection with a list of language-specific folders:

- Abui corpus (Corpus collected by George Saad from 2015 to 2017, Part of Marian Klamer VICI p...)
- Adang Collection
- Alor Malay Collection
- Alorese Collection
- Blagar Dadibira Collection
- Blagar Manatang Collection
- Kaera Abangiwang Collection
- Reta Hula Collection

Figure 19: Structure of the Flores and Solor Archipelago sub-collection

The screenshot displays the 'Browse Archive' interface for the 'Flores and Solor Archipelago' sub-collection. The top navigation bar includes 'Browse Archive' and 'Browse by'. A search bar is located at the top left. The left sidebar contains filters for 'Access Level' (Restricted, 104), 'Contributor' (Hanna Fricke, Alexander Elias, Marian Klamer, Yunus Sulistyono), 'Language' (Central Lembata, Indonesian, Lio, Hewa, Eastern Atadei (Lerek), Ende), 'Country' (Indonesia), and 'Genre' (Narrative, Elicitation task, Conversation, Word list, Elicitation Stimuli, Prayer, Sociolinguistic Questionnaire, Interview, Elicitation, Metadata). The main content area shows the title 'Flores and Solor Archipelago' and a list of sub-collections, each with a folder icon and a description: 'Atadei Painara Collection', 'Central Lembata Corpus' (with a detailed description), 'Ende Collection', 'Hewa Collection', 'Kedang Collection', 'Lamaholot Adonara Collection', and 'Lamaholot Lewoingu Collection'.

Having the archive organized by geographical region and language not only allows easy browsing but, equally important, it also allows any new data collected in the region (e.g. a new language, or additional data about a particular language) to be added easily to the existing archive structure. While at present, most of the sub-collections for individual languages contain data collected by a single researcher, our structure allows adding data collected by other researchers. The structure down to the geographical regions is rigid, but within each language sub-collection, the structure reflects the type of data and wishes of the individual researchers.

## Appendix 2: Fieldwork Cookbooks

### 2.1 *What goes in a fieldwork research plan?*

- a description of the goals of the fieldwork
- a description of the type of data that will be collected
- how the data will be collected (methodology)
- the questionnaire or elicitation materials to be used
- the number of speakers that must/will be recorded
- the location(s) of the fieldwork
- the equipment that is necessary
- a time line, specified by week
- a budget including costs for travel, equipment, living, accommodation, medication, consultant payment, transportation, communication
- contacts in the field or a close-by location, if available

### 2.2 *How to find consultants and compensate them? Considerations from Indonesia*

In Indonesia, it is often not a problem to find consultants to help with a short survey. To find people to record narratives or dialogues, you can ask for people who know a traditional story, a fable or a myth, or to tell something about the history of the village. For procedural monologues you can ask someone to explain how to grow rice, how a wedding is organized, how a house is built, how a local dish is cooked, etc. The hardest part is to find someone to help you transcribe the recordings. For this, you need someone to work with, who is reliable and has time.

We have often found consultants through the head master of a local secondary school. School students (16–18 years of age, almost finishing school, or just having left school waiting for a job) usually have enough time to work on transcriptions frequently. It is also possible to ask the family that is hosting you, or their neighbors, or the village head, or a religious leader. Other adults, such as school teachers or the village head can also be great consultants, but they are often called away suddenly for urgent matters, and/or have other duties in the afternoon and weekends, so they are much less available than young adults.

It is better to start working with 2 or 3 different consultants and not focus only on a single person, to avoid becoming too dependent on that person, and running the risk of getting biased or collecting idiolectal information on the language.

If you are not transcribing, but want to work on the lexicon or on grammatical judgment, it is also good to work with 2–3 people at the same time. You could invite e.g. the village head and ask him/her to bring 2–3 others to work with you for one or more mornings or afternoons per week.

In Indonesia it is very acceptable to ask people to work with you in a polite and straightforward way. In general people are very willing to help, but will not offer to help spontaneously, e.g. because they are shy, or because they think they are not good enough. Thus they need to be asked directly.

Regarding compensation for consultants that you work with in the field, an Indonesian PhD student who works with local consultants himself suggested the following. Do not discuss any payment; because to pay for the help that people give one another is seen as not done or even offensive. However, people do appreciate receiving money in compensation for their time. It should be considered a gift, not a salary, and it should not be negotiated. A suggested amount for a day's work would be the equivalent of a teacher's daily salary (in our project this was about 100,000 IDR for a full day of work). Notes that are smaller than 50,000 IDR are considered so small that it can be offensive as a gift. In that case it is better not to give anything at all, and compensate with a small gift in kind. If someone comes regularly, e.g. 3–4 times a week, it is good to give money after each

meeting, and not in advance, to keep the person motivated to come again the next time. Give the money when the session is finished and the person is about to leave, putting a note in his/her hand in a kind of off-hand manner (like giving a tip to a porter), while saying "thanks very much, see you next time". It is advised to keep track of how much money has been given to whom and for what, to avoid awkward situations where you pay one person more than another for the same type or amount of work. People will be disappointed if they are not treated equally. It is not done to discuss payment of individuals when others are around, let alone in public.

**2.3 How to obtain informed consent, and what does an informed consent form look like?**

The researchers asked for consent before the recording equipment is set up. If the speakers agree, they are shown the consent form, and asked to read and sign it. This form can either be read out by the researcher, or read by the speakers themselves. The reading and signing of the form may also be video-recorded.

A sample consent form as used in the projects described above is the following. (The version used in the field was in Indonesian or Malay.)

*My name is [researcher's name]. I am from [name of university] and I would like to learn more about your language, how you use it and why. I want to learn because I want to understand better how people speak, think, and live in places where many languages are spoken.*

*I would like to record what you say and keep that record so that other people may also learn from you that way. I will ask you about your language, how you use it, when, and why and also about how you say certain words and sentences and how to describe things properly. If there are things you would want to record (stories, jokes, about your life), we can also record those. What I record will be kept in my university and in an archive that can be found through the internet.*

---

My name is: \_\_\_\_\_  
I was born in \_\_\_\_\_ (place, year)  
I speak the following languages: \_\_\_\_\_

*I feel good about talking about my language with you and I know and understand that:*

1. *what I say can be recorded and other people may listen or watch it;*
2. *If I want to I can say to you: "Do not show to other people what I told you.";*
3. *If I want to I can say to you: "Don't use my name.";*
4. *If I want to I can say to you: "Delete my recording."*
5. *I can tell you: "Change what I told you.";*
6. *I can stop teaching you about my language any time;*
7. *I can ask you if I do not understand what you are doing;*
8. *I can ask you to give me back a copy of what I said to you;*
9. *I can ask other people at your university or school to tell me about what you are doing.*

*What I tell you about my language is to help you talk or write about it, that's all.*

Signature: \_\_\_\_\_ Date: \_\_\_\_\_  
(Parent for children)

#### **2.4 What are the steps involved in a recording session with a video camera?**

The following is an example of a to-do list for a recording session with video camera. As it was written for one particular type of video camera and microphone, certain details may not apply to other devices.

##### *Before the recording*

- 1) Keep all the equipment together in a dedicated bag/backpack that is packed in a systematic way
- 2) Check that battery of the video camera, audio recorder and microphone are full
- 3) Check that spare battery/batteries is/are full
- 4) Check that there is an empty flash card in the camera [tip: change flash card after each recording session]
- 5) Put the video camera on the tripod (with the regulating lever toward you)
- 6) Put the microphone on the tripod and then on the table
- 7) Connect the microphone to the video camera with a cable (insert the cable into the red hole in the video camera)
- 8) Tape down the cable to a surface
- 9) Put the recorder on the tripod and then on the table, next to the microphone (the recorder functions as a backup)
- 10) Turn on the video camera and check battery
- 11) Turn on the recorder and check battery
- 12) Turn on the microphone and check battery
- 13) Connect the headphones to the video camera (insert cable in the green audio jack)
- 14) Check that all cables are well-connected, in particular the cable connecting the external microphone and the video recorder
- 15) Do a sound check for the video camera and for the audio recorder
- 16) Check the recording level of both devices (check which level is good: with a Zoom recorder 100 is a good level). You may also broaden the angle of the microphones if that is possible: 90° is good for one person, 120° is good for about 4 people sitting in a semicircle
- 17) Check the settings of the video camera: adjust Exposure (to regulate bright/dark), Spot-Focus (to adjust the brightness) or Low-Light (if the room is too dark). The image quality should be 50p – this can only produce MTS files
- 18) Start recording with the recorder (if you have done a sound check, the red light of REC will be blinking. If you press the REC button for a few seconds, it starts recording)
- 19) Start recording with the video camera
- 20) Don't start the session immediately, let it run for a few seconds
- 21) Say: the name of the language, the date, the place, your name, name of the speaker(s).
- 22) Give the speaker(s) the consent form and let him/her sign it, after reading it out loud if possible

##### *After the recording*

- 23) Switch off the recorder
- 24) Switch off the microphone
- 25) Switch off the video camera
- 26) Pack up the equipment in a systematic order, so as not to leave elements behind
- 27) Create raw data and working data folders in your laptop
- 28) Extract the flash-card from the video camera and put it in the reader, or connect the video camera to your laptop with a cable

- 29) The MTS files are in the folder PRIVATE/AVCHD/STREAM, mp4 files are in the folder ROOT (these locations may differ, depending on the device used)
- 30) Copy the raw data on your laptop (in the raw data folder) and change the name of the file (e.g. languagecodeyourinitials\_yyyy\_mm\_dd\_nameofrecording), e.g. If the video camera splits the session into several files (see section 2 above), names would be of the type AOLFM\_2015\_12\_08\_frogstory(1of5), AOLFM\_2015\_12\_08\_frogstory(2of5), etc.
- 31) Extract the flash-card from the recorder and put it in the reader
- 32) Copy the raw data from the recorder (in the raw data folder) and change the name in (for example) AOLFM\_2015\_12\_08\_frogstory(1of5)\_Backup (add Backup so that you don't mix up the files in the future, when you will have two .wav files)
- 33) When the flash cards are full, you can format them directly in the video camera or in the recorder
- 34) Return the flash card to the video camera
- 35) Convert the files produced by the recorder (e.g. MTS) to MP4, as MP4 is what you can use for ELAN
- 36) Use Free (MTS) Converter Software to merge the files that were previously split (1of5, 2of5, 3of 5 etc), so that you get back one single file of your recording session.
- 37) Convert MTS to MP4 (be aware that it takes a while to convert files) and save it into the working data folder
- 38) Use FFmpeg to convert MP4 to WAV filesOpen a new ELAN file and select both the MP4 and the WAV files, so that you have the video and the spectrogram

### ***2.5 Which kind of metadata are collected and filed?***

The metadata of a recording would constitute information such as the following:

- (i) the filename
- (ii) the date of recording
- (iii) the name of the language on the recording
- (iv) any alternative names for the language
- (v) the location where the recording was made (in geographical coordinates)
- (vi) the location where the language community lives (in geographical coordinates)
- (vii) the (village) name of the location where the recording was made
- (viii) the names of village(s) where the language community lives
- (ix) the topic of the recording
- (x) the length of the recording
- (xi) who made the recording (the name of the researcher and/or the assistant)
- (xii) personal details about speaker(s) who are on the recording, such as: name, gender, date of birth, place where they grew up, highest education, current place of living, current occupation, language of father, language of mother, language of spouse, language spoken with children
- (xiii) a short description of the content or topic of the recording

At later stages, additional information could be added to the metadata sheet, such as:

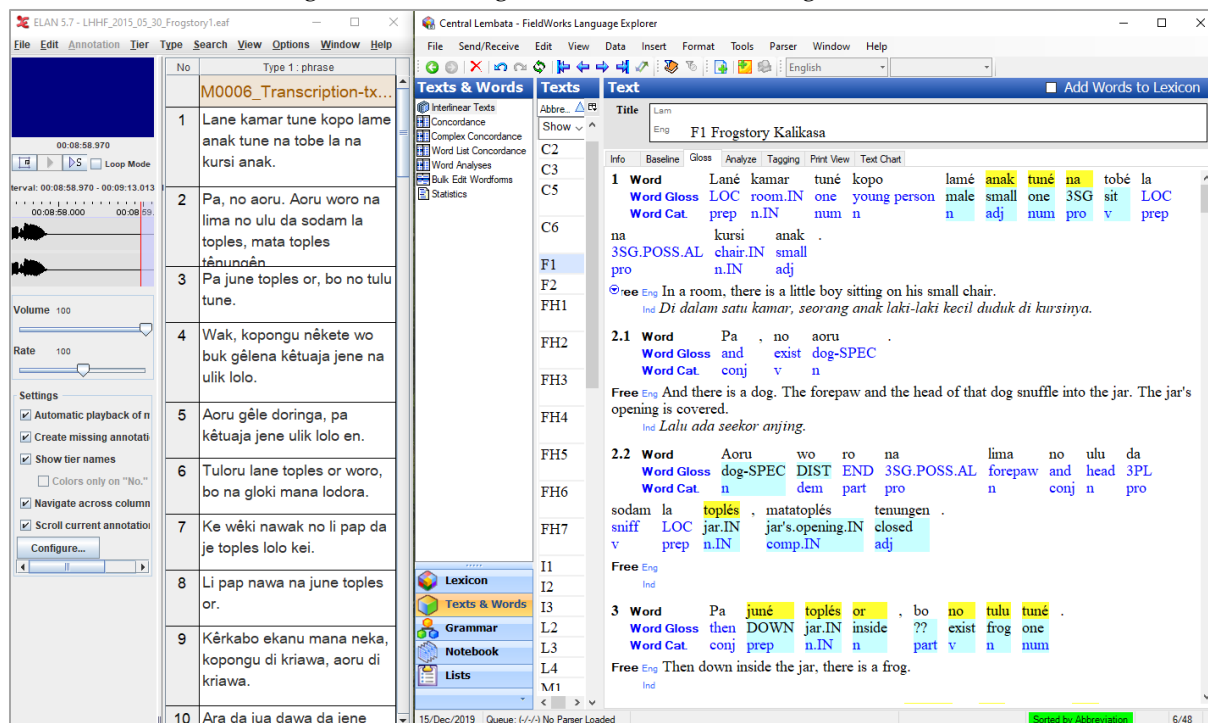
- (xiv) who transcribed the recording
- (xv) who translated it into the national language/lingua franca
- (xvi) who glossed it
- (xvii) who translated it into English
- (xviii) any earlier names of the file (in case it has been renamed)

## 2.6 How to transfer data between ELAN and FLEx?

ELAN (Wittenburg et al. 2006) is freeware which can be used to annotate audio and video files. It has been proven to be a very useful tool for linguistic transcription of recordings. FLEx is freeware developed by SIL, available at <https://software.sil.org/fieldworks>. The program allows linguists to build an annotated text corpus that is connected to a lexical database that can be built and expanded while glossing. FLEx allows for consistent glossing and builds a well-searchable database with help of regular expressions. In most projects described above both programs were used. As ELAN and FLEx are not inherently connected, the researcher needs to follow a specific workflow when using both of them with the same data. We recommend the workflow originally described by Tim Gaved and Sophie Salfner, available online at <https://www.soas.ac.uk/elar/helpsheets/file122785.pdf>. Following this workflow, a transcription from ELAN can be exported and opened in a FLEx corpus where it can be glossed and analysed further. It is equally possible to re-import the glossed text back into ELAN and reconnect it to the audio and video file. Because files that are re-imported into ELAN cannot be exported to FLEx a second time, the step of re-importing the glossed text to ELAN is best done at the very end of the project, for example before archiving the ELAN files with its video and audio files. Note that whenever the researcher wants to create an ELAN file that is exportable to FLEx, the ELAN files need to be set up in the way described in the workflow mentioned above from the start. Applying the necessary changes to an ELAN file that had been set up in a different way is possible but can be difficult and time-consuming.

When working on the glossing and analysis in FLEx, the researcher might feel the disadvantage of not having the audio file integrated into the FLEx corpus. A way around this is the possibility of working with the audio/video file in ELAN and the FLEx corpus opened at the same time, see Figure 20. The annotation numbers can be used to navigate to the right segment in both programs.

Figure 20: Working with ELAN and FLEx together



An issue arises when the researcher wants to change or correct transcriptions after the text has been exported to FLEx. The transcription, and possibly also the free translation, now exist twice, once in the ELAN file and once in the FLEx corpus. Here, we describe two ways to deal with this issue.

One solution is to only make changes in the transcription and the translation in the FLEEx corpus. This means that, from the moment the text is in FLEEx, the ELAN file is only used for listening back to the audio but no annotations are made or changed in the ELAN file. When re-importing the text back into ELAN, the corrected transcriptions and translations are imported and matched with the audio.

Another solution is the following workflow, suitable for projects that a) already started annotating in ELAN and/or b) where the original transcription is frequently edited:

- Field trip 1: record new data, transcribe data with consultant in ELAN
- Intermediate period 1: check transcriptions from Field trip 1, note down questions
- Field trip 2: record new data, transcribe data with consultant in ELAN, update transcriptions from Field trip 1 with help of questions
- Intermediate period 2: check transcriptions from Field trip 2, note down questions
- Field trip 3: record new data, transcribe data with consultant in ELAN, update transcriptions from Field trip 2 with help of questions
- Intermediate period 3: check transcriptions from Field trip 3, note down questions
- [etc. ad infinitum]
- Last field trip: update transcriptions from previous field trip with help of questions.

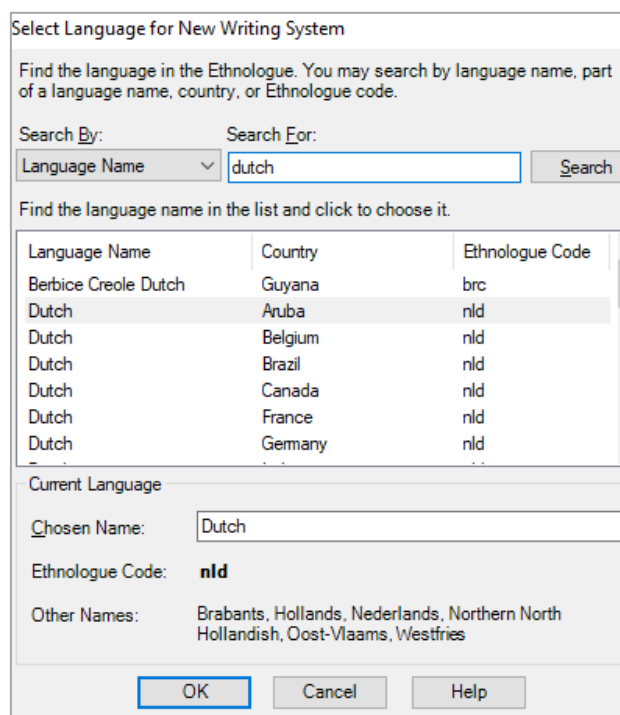
In this case, it does not make much sense to move data to FLEEx until after the end of Field trip 2, when the first somewhat “final” transcriptions are ready. This working method means that the whole corpus is never in one place, so the advantage of using FLEEx is somewhat lost. However, after Field trip 3, all recordings from Field trip 1 and 2 can be migrated, so that a big portion of the corpus is in FLEEx. After the last field trip, the last transcriptions can be moved and the researcher can make full use of the FLEEx database.

In addition to transcriptions and translations, audio and video files can be annotated with notes in ELAN on a separate tier. Notes are then connected to the utterances they refer to (e.g. to indicate when someone out of the frame is being talked to). Ideally they are linked to the exact utterance they refer to. That way one only needs one screen while working with a consultant, and it also allows the researcher to make notes quickly when working on their own.

In the following, a way to export these notes together with the transcriptions from ELAN and import them into a FLEEx corpus is described. To make these notes appear in the right spot in the FLEEx corpus later, the note tier in ELAN has to be set up as a so-called Translation tier with tier type note (just like the tiers for the translation to English and the national language, such as Indonesian). The language code used for the note tiers should be a code of a language that is not used in the project, for example the code “nld” for Dutch. Language codes can be found in FLEEx by opening FLEEx > Format > Set up writing systems > Add Analysis Writing Systems > type in the language of your choice > Search, see Figure 21. The three-letter code is shown next to the language name.



Figure 21: Finding Language Codes in FLEx



The result is that both in the lexicon (where it is not needed, Figure 22) and in the text corpus (Figure 23) an extra analysis language is added which can be used for notes. When the texts are re-imported in ELAN, the notes will be in a separate tier.

Figure 22: A lexical entry in FLEx

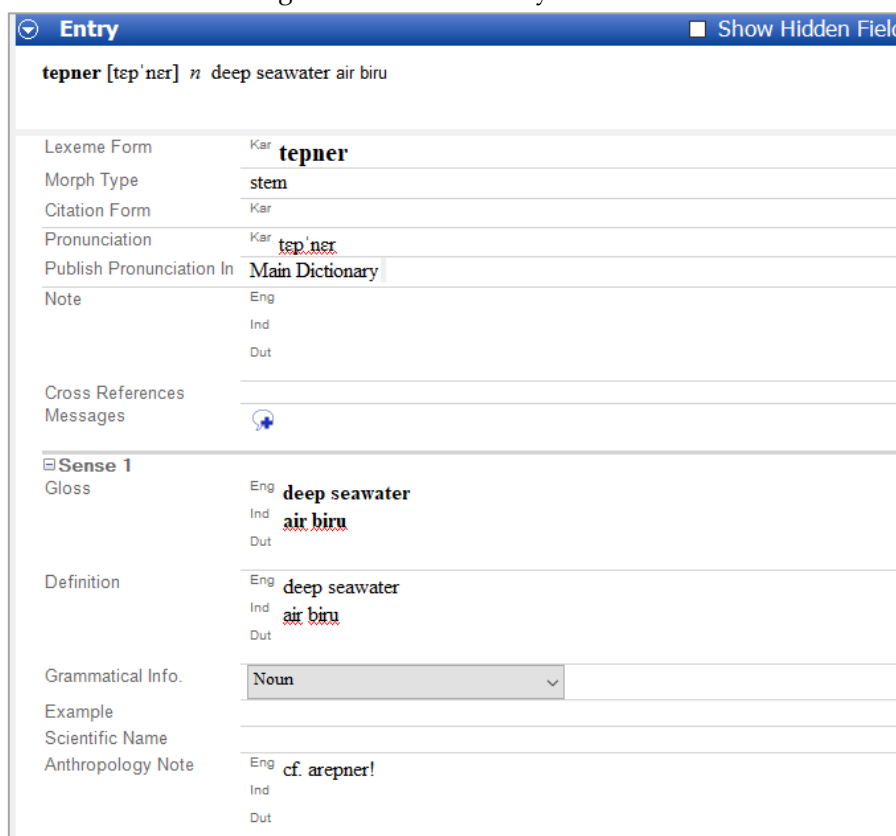


Fig 23: An annotated fragment in FLEEx with a grammatical note

<b>95 Word</b>	wowa	nonakonggoa		to
<b>Morphemes</b>	wowa	nona	-konggo	-a
<b>Lex. Entries</b>	wowa	***	-konggo <sub>1</sub>	-a
<b>Lex. Gloss</b>	aunt	***	an.loc	foc
<b>Lex. Gram. Info.</b>	n	***	n:(case)	Inflects any category
<b>Word Gloss</b>	aunt	at Nona's		
<b>Word Cat.</b>	n	***		
<b>Free Eng</b>	aunt Nona's family			
<b>Dut</b>	nona-onggo is ungrammatical			
<b>Ind</b>	di wowa Nona dong to			

### 1. Importing words from Toolbox to FLEEx

If you have already started a project in Toolbox, but want to migrate it to FLEEx, you must import the Toolbox project into FLEEx as follows:

1. Export the Toolbox file in Toolbox. File > Export... > Standard Format > OK. Make sure you saved your file in Standard Format. You may need to add the extension .sf to the file name manually.
2. Create a new FLEEx project in FLEEx. File > New Fieldworks Project...
3. Import the Toolbox file in FLEEx. File > Import... > Standard Format Lexicon > follow the steps in the pop-up window.

If there are any mismatches between fields used in Toolbox and in FLEEx, FLEEx will tell you so, and will import residue to a dedicated field called Import Residue. If Toolbox fields are imported to the wrong field in FLEEx, this can be easily dealt with as follows. For example, one researcher had used the field \ge (Gloss English) in Toolbox for definitions rather than glosses. Toolbox \ge is imported to the field Gloss in FLEEx. One then wants to copy the Gloss fields to the Definition fields. Operations like these can be done with a function called Bulk Edit in FLEEx. In this case, the steps were as follows:

1. In Lexicon view, click Bulk Edit Entries in the menu top-left.
2. In the tabs at the bottom, click Bulk Copy.
3. Source Field: Glosses. Target Field: Definition. Apply.

When importing a word list from Toolbox to FLEEx, make sure you know where your personal notes and notes you wish to publish end up. The notes fields in Toolbox (\nt, \na, etc.) exist in FLEEx but are not very accessible, whereas another field called Notes is. After the transition, be consistent in your use of the different notes fields. Bulk Edit (described above) may be used to move all notes to the same field. Alternatively, notes in FLEEx can be kept in the message field (📧 Messages 📧), which is useful for collaborating researchers and for tracking changes.

### 2. Importing elicited data from Word to FLEEx

You may find that you have loose fieldnotes and elicited sentences (typically copied from a paper notebook) in Word, Excel, or other non-durable formats. It is a good idea to import these to FLEEx because a) that way they will be saved in a durable, archivable format and b) you can annotate them and they will be added to your corpus of naturalistic data, so that you can easily search your entire corpus. It is recommended that you tag your notes in a similar way as your naturalistic data, e.g. with a unique identifier. You can tag your field notes by date, by topic, by questionnaire, by notebook page or whatever suits your data. FLEEx also offers the possibility of adding other metadata, such as the source of a questionnaire, comments, participants and locations. Importing sentences from Word (or a similar format) into FLEEx is done as follows:

1. Copy your sentences in the Word file. Make sure they are unnumbered.
2. In the Texts & Words environment, click Insert > New Text.
3. In the top menu, where it says Normal, select Numbered Text.
4. In the Baseline tab, paste the sentences that you copied into Word.
5. Add a unique identifier and other metadata in the Info tab.
6. Annotate your sentences in the Analyse tab.

## References

- Edwards, Owen. 2016. *Metathesis and Unmetathesis: parallelism and complementarity in Amarasi, Timor*. PhD Thesis: The Australian National University.
- Edwards, Owen. 2020. *Metathesis and Unmetathesis in Amarasi*. Berlin: Language Science Press. <https://langsci-press.org/catalog/book/228>
- Bowern, Claire. 2008. *Linguistic Fieldwork: A Practical Guide*. New York etc: Palgrave Macmillan.
- Chelliah, Shobhana L., and Willem J. de Reuse. 2011. *Handbook of Descriptive Linguistic Fieldwork*. Dordrecht etc.: Springer.
- Fedden, Sebastian, and Dunstan Brown. 2017. "Participant Marking: Corpus Study and Video Elicitation." In *The Alor-Pantar Languages: History and Typology (2nd Ed.)*, edited by Marian Klamer, 403–46. Berlin: Language Science Press.
- Fedden, Sebastian, Dunstan Brown, and Greville Corbett. 2010. "Conditions on Pronominal Marking: A Set of 42 Video Stimuli for Field Elicitation. University of Surrey." 2010. <http://dx.doi.org/10.15126/SMG.25/1>.
- Fricke, Hanna. 2019. *Kamus Tiga Bahasa: Atadei Demon – Indonesia – Inggris*. UBB Language & Culture Series, A-11. Kupang: Unit Bahasa dan Budaya (UBB), GMTI.
- Hammarström, Harald, and Sebastian Nordhoff. 2012. "The Languages of Melanesia: Quantifying the Level of Coverage." In *Melanesian Languages on the Edge of Asia: Challenges for the 21st Century*, 13–34. Language Documentation & Conservation 5. Honolulu: University of Hawai'i Press. <http://scholarspace.manoa.hawaii.edu/handle/10125/4559>.
- Haspelmath, Martin, and Uri Tadmor, eds. 2009. *The World Loanword Database (WOLD)*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://wold.clld.org/>.
- Klamer, Marian. To appear. "From Lamaholot to Alorese: Morphological Loss in Adult Language Contact. In David Gil and Antoinette Schapper (Eds), Amsterdam: Benjamins." In *Austronesian Undressed: How and Why Languages Become Isolating.*, edited by David Gil and Antoinette Schapper. Amsterdam: Benjamins.
- — —. 2011. *A Short Grammar of Alorese (Austronesian)*. Munich: Lincom Europe.
- — —. 2012. "Papuan-Austronesian Language Contact: Alorese from an Areal Perspective." In *Melanesian Languages on the Edge of Asia: Challenges for the 21st Century*, edited by Nicholas Evans and Marian Klamer, 72–108. Language Documentation & Conservation Special Publication 5. Honolulu: University of Hawai'i Press. <http://scholarspace.manoa.hawaii.edu/handle/10125/4561>.
- Klamer, Marian, Hanna Fricke, Francesca Moro, George Saad, and Eline Visser. n.d. "Language Collection 'Eastern Indonesia and Timor Leste.'" The Language Archive. <https://hdl.handle.net/1839/06afa50e-ae9-4adb-a6a7-d7496a8a47fc>.
- Klamer, Marian, and Francesca Moro. To appear. "What Is 'Natural' Speech? Comparing Free Narratives and Frog Stories in Indonesia." *Language Documentation & Conservation*.
- Mayer, Mercer. 1969. *Frog, Where Are You?* New York: Dial.
- Meakins, Felicity, Jennifer Green, and Myfany Turpin. 2018. *Understanding Linguistic Fieldwork*. London and New York: Routledge.

- Moro, Francesca. 2018. "The Plural Word Hire in Alorese: Contact-Induced Change from Neighboring Alor-Pantar Languages." *Oceanic Linguistics* 57 (1): 22.
- — —. 2019. "Loss of Morphology in Alorese (Austronesian): Simplification in Adult Language Contact." *Journal of Language Contact* 12 (2): 378–403. <https://doi.org/10.1163/19552629-01202005>.
- Moro, Francesca, and Hanna Fricke. 2020. "Contact-Induced Change in Alorese Give-Constructions." *Oceanic Linguistics*.
- Robinson, Laura C. 2015. "The Alor-Pantar (Papuan) Languages and Austronesian Contact in East Nusantara." In *Language Contact and Austronesian Historical Linguistics*, edited by Malcolm D. Ross. Canberra: Pacific Linguistics.
- Saad, George. 2020. *Variation and Change in Abui: The Impact of Alor Malay on an Indigenous Language of Indonesia*. PhD thesis Leiden University. Utrecht: LOT Publications. <https://www.lotpublications.nl>.
- Saad, George, Marian Klamer, and Francesca R Moro. 2019. "Identifying Agents of Change: Simplification of Possessive Marking in Abui-Malay Bilinguals. 4(1), 57." *Glossa: A Journal of General Linguistics* 4 (1): 57. <https://doi.org/DOI:10.5334/gjgl.846>.
- Thieberger, Nicholas, ed. 2011. *The Oxford Handbook of Linguistic Fieldwork*. Oxford: Oxford University Press.
- Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. "ELAN: A Professional Framework for Multimodality Research." *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*, 1556–59.