CHAPTER 10

Creating a Language Archive of Insular South East Asia and West New Guinea

Marian Klamer, Paul Trilsbeek, Tom Hoogervorst and Chris Haskett
Leiden University, Max Planck Institute for Psycholinguistics, KITLV/Royal Netherlands Institute of Southeast Asian and Caribbean Studies, MPI Nijmegen

ABSTRACT

The geographical region of Insular South East Asia and New Guinea is well-known as an area of mega-biodiversity. Less well-known is the extreme linguistic diversity in this area: over a quarter of the world's 6,000 languages are spoken here. As small minority languages, most of them will cease to be spoken in the coming few generations. The project described here ensures the preservation of unique records of languages and the cultures encapsulated by them in the region. The language resources were gathered by twenty linguists at, or in collaboration with, Dutch universities over the last 40 years, and were compiled and archived in collaboration with The Language Archive (TLA) at the Max Planck Institute in Nijmegen. The resulting archive constitutes a collection of multimedia materials and written documents from 48 languages in Insular South East Asia and West New Guinea. At TLA, the data was archived according to state-of-the-art standards (TLA holds the Data Seal of Approval): the component metadata infrastructure CMDI was used; all metadata categories as well as relevant units of annotation were linked to the ISO data category registry ISOcat. This guaranteed proper integration of the language resources into the CLARIN framework. Through the archive, future speaker communities and researchers will be able to extensively search the materials for answers to their own questions, even if they do not themselves know the language, and even if the language dies.

10.1 Background of the Project

The geographical region of Insular South East Asia and New Guinea is well-known as an area of mega-biodiversity. What is less well-known is that its tremendous species diversity correlates

How to cite this book chapter:
Klamer, M, Trilsbeek, P, Hoogervorst, T and Haskett, C. 2017. Creating a Language Archive of Insular South East Asia and West New Guinea. In: Odijk, J and van Hessen, A. (eds.) *CLARIN in the Low Countries*, Pp. 113–121. London: Ubiquity Press. DOI: https://doi.org/10.5334/bbi.10.

with a rich cultural and linguistic diversity across the area (Gorenflo et al., 2012). But both the biodiversity that supports the humans and all other species in the area and the traditional ethnolinguistic knowledge that helps sustain it are subject to a converging extinction crisis. We are rapidly losing the unique ways of life and the encapsulating languages of peoples in this part of the world.

Over the last 40 years, more than two dozen linguists at, or in collaboration with, Dutch universities collected multimedia materials and written documents from over 50 languages in Insular South East Asia and West New Guinea. However, as these unique records were not digitised and/or archived systematically, they were bound to get lost. The current CLARIN 'Resource Curation' project grew out of the recognition that these unique resources had to be preserved for the future. The initial goal was to archive 52 language resources along with their basic metadata, and make them accessible online. The project ran for 16 months from February 2013 to July 2014 and involved various types of tasks: (i) Collaboration with the linguists who originally collected the resources (and who are not part of this project) to systematically compile their materials and the metadata; (ii) File conversion and digitization; (iii) CMDI profile creation and metadata entry; (iv) Creating an inventory of annotations and the conventions and terminology used in them; (v) Mapping of these with terms in the ISOcat data category registry; (vi) Archiving mappings along with the data sets; (vii) Make the archive accessible online.

Materials that had not yet been digitised were digitised and archived alongside more recent digital collections in accordance with the highest standards and principles as established by The Language Archive at the Max Planck Institute in Nijmegen (Drude et al., 2012).

As a result of this project The Language Archive now contains language resources compiled by twenty linguists over the past 40 years on 48 languages spoken in Insular South East Asia and New Guinea, archived under the name of LAISEANG (Language Archive of Insular South East Asia and New Guinea), see Figure 10.1.

Through this structured digital archive, speaker communities and researchers – be they anthropologists, linguists, historians or researchers from other disciplines – will be able to extensively search the materials for answers to their own questions, even if they do not themselves know the language, and even if the language dies.

10.1.1 Research Question(s)

Research questions that may be (better) addressed using the data of this project include, but are not limited to, the following:

- *Questions relating to language structure, e.g.* What is the structure of a conversation/narrative/-paragraph/sentence/clause/word in language X?
- *Questions relating to language use, e.g.* How do bride-price negotations, religious songs, historical narratives, speeches, etc. function in language X?
- *Questions relating to the history of languages and their speakers, e.g.* Which lexical evidence supports the reconstruction of historical relations between Papuan languages of Timor-Alor-Pantar; or between the Austronesian languages of New Guinea?
- *Questions about the language of particular semantic domains, e.g.* Which numeral systems are employed in group/language X? How does language X express location and motion in space? How does language X encode pronominal reference? What formatives does language X use to express physical sensations or emotional states/experiences?
- *Questions about intonation, prosody, and tone, e.g.* What are the major intonation patterns in language X? How does prosody affect realisation of tone in language X?

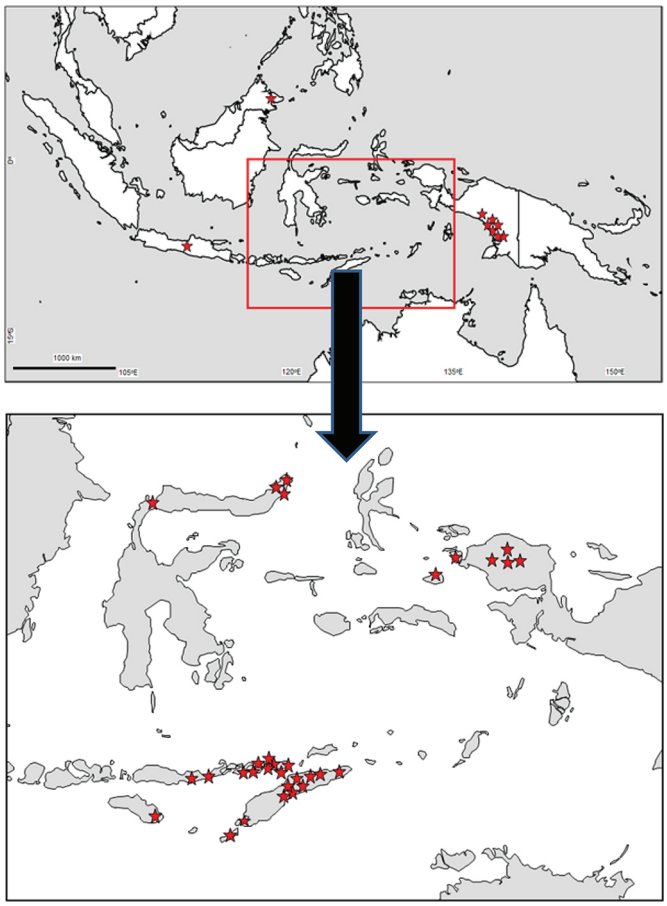

Figure 10.1: Locations of LAISEANG languages in Insular Southeast Asia and New Guinea.

- *Questions relating to comparative folklore, e.g.* What kind of omens are certain species of birds said to hold in ethnolinguistic group X?
- *Questions relating to speech registers and oral literature, e.g.* What are the poetic devices and metaphors used in ritual language? How are songs composed, in terms of tunes and/or texts?
- *Questions from speaker communities, e.g.* How do/did we produce traditional material cultural items such as houses, woven cloth or baskets? How do/did we grow and prepare certain traditional types of food? Which traditional place names are/were used in our region? What kind of mythology or ancestor stories do/did we have?

10.1.2 Research Data

Before the project started we contacted the original collectors of the linguistic data, and almost all emailed their consent for the data they collected to be included. We had to visit some collectors who currently work abroad to collect their data manually and to compile specific details on the content and format of their resources and metadata sets.

As we had anticipated, many different annotation schemes had been used by the collectors. Most of the older materials had handwritten transcriptions and annotations on paper; these paper

resources had to be converted to PDF and were archived as such. More recent language resources have often been entered and annotated as Toolbox projects (as files linking texts and lexicon), and various types of linguistic categories were employed for the annotation across the different Toolboxed resources. Where this was possible we mapped these linguistic categories to ISOcat categories; these mappings were added to the archive as separate resources.

10.1.3 Technology

Many of the resources we archived were already available in digital form. However, conversions had to take place in some cases, such that all resources conformed to standards accepted by The Language Archive as suitable for long-term preservation.

The older non-digital recordings we digitised as WAV files included (i) reel-to-reel, (ii) audio cassette tape, and (iii) video cassette tape recordings. A number of cassette tapes that had previously been digitised as MP3 by the original collector had to be digitised again as WAV. Non-digital paper materials, which constitute valuable or unique documents, were scanned into PDF format and also deposited.

The metadata descriptions of the language resources we archived were in a number of different formats: some were available in spreadsheets, some in Word documents, some on paper – and some only in the collector's mind.

For the component metadata infrastructure, we evaluated the existing CMDI profiles in the CLARIN Component Registry and decided that the IMDI CMDI profile (CMDI profile based on the IMDI metadata schema) fitted the needs of the project. In addition to the metadata categories in this profile that had already been linked to the ISO data category registry ISOcat (cf. Kemps-Snijders et al., 2009, among others), the annotation terminology of a number of languages was inventoried and linked to ISOcat as well. The archiving software at TLA automatically assigned a Handle Persistent Identifier to every archived resource and metadata record. It also made all the metadata records available for harvesting via the OAI-PMH protocol. This guaranteed the proper integration of the resources into the CLARIN framework.

10.1.4 Description

Once the resources per language and the metadata were compiled and prepared for archiving, they were put on the server of TLA. Data was structured hierarchically in trees, where each language resource would be assigned one major node in an 'Insular SE Asia & New Guinea' comprehensive corpus. Metadata were entered, and data integrated and organised with the help of the Arbil standalone tool and the LAMUS online tool. The Access Management System (AMS) was used to assign one of four levels of access to any resource or group of resources according to place in the tree structure and/or file type: 1) open, 2) controlled open (where identification and agreement with a code of conduct is necessary), 3) restricted (where individual permission by the researcher or person in charge is required) or 4) closed (to anyone except original researchers/depositors). Except for one resource, the data that were archived in this project are all of level 1: open access.

Annotations in unsupported formats such as DOC were converted to PDF, and were thus treated identically to annotations that are handwritten in (field) notebooks.

We made mappings between the annotation terms used in the data sets and the general ISOcat concept registry for the more recent resources that were already in Toolbox format. In some of these resources, the Leipzig Glossing Rules had been applied, often with an additional set of glossing conventions and labels that only apply to an individual language; the ISOcat data category registry provides a means to clarify the nature of the meaning of such particular glosses.

10.2 Results of the Project

10.2.1 Deliverables and Milestones

A deliverable is a measurable and tangible outcome of a project. They are developed by project team members in alignment with the goals of the project. Milestones are checkpoints throughout the life of the project. They identify when activities have been completed, thus implying that a notable point has been reached in the project. The Appendix contains the list of deliverables and milestones we originally aimed for, in their original order, with the relevant actors. The results are indicated with colour shading, where green deliverables and milestones have been met completely, while red deliverables have not been met. The reasons for not curating some of the resources were of variable nature: some of the resources we originally planned to curate turned out to be untraceable and/or lost, other resources were deemed by the author not yet ready to be archived, or the author decided not to archive them with LAISEANG after all. The list also includes some added deliverables: these resources were added to the archive in the course of the project, or shortly after it.

The LAISEANG archive can be accessed directly at http://hdl.handle.net/1839/00-0000-0000-0018-CB72-4@view. It can also be accessed through The Language Archive (tla.mpi.nl) by clicking on the 'Access the Archive' quick link. In the left-hand panel of the archive there is an alphabetical list of corpora, which includes LAISEANG. As mentioned earlier, except for one, all the resources in the LAISEANG archive are open to everyone.

10.2.2 Lessons Learned

Most of the linguists that contributed their resources were extremely grateful for the opportunity to get their materials digitised and archived in this way, and mentioned that their materials would otherwise never have been curated or archived. Linguistic field researchers typically wish to make their data available for everyone, rather than keep it to themselves. However, our experiences in this project suggest that it is unlikely that their data will be archived unless there is a budget to pay for the time it costs to curate and archive it. In addition, there must be a user-friendly infrastructure to help them actually *do* it, preferably online.

In some cases, we inherited quite large archives with an existing structure that was designed by the collector to archive data and metadata in a logical and transparent way (e.g. the resources on Ma'ya and Matbat, Figure 10.2). It would have been efficient to be able to automatically incorporate such existing structures into the metadata structure, for example by having an option to add resources to Arbil directly from such existing structures, using a kind of 'folder importing' script. At present, Arbil has no such option, so that the original structures unfortunately all got lost.

Another issue that we ran into was that Arbil has been developed as a local tool, originally designed for a single person to archive a single language resource. As a result, Arbil is not a tool to efficiently deal with (i) multiple language resources collected by a single person and archived by several others or (ii) single language resources collected by multiple persons and archived by yet others. Arbil does not offer an online collaborative workspace, which meant that in our project the data sets had to be treated in strict cycles and metadata had to be sent back and forth among the team members responsible for digitisation and metadata collection. It would have been more efficient to allow online collaboration on metadata in a 'Google Docs' manner, such that project members could work simultaneously on the same data sets, and such that authors could look at their own materials and add or correct information if they wished.

One result from this project that may be useful for the future is that we are now able to estimate the costs of curating and archiving a single language resource collected in the field. The size of

Figure 10.2: Dr. Bert Remijsen with his resources on Ma'ya and Matbat.

the resources for the LAISEANG archive varied enormously: from 30-minute audio recordings to 35-hour video recordings. Some resources were not yet digitised, and/or had no or little organised metadata, while others were completely digitised, with all the metadata organised in spreadsheets. As a result, the time it took to curate and archive an individual resource varied from 4 to 60 hours. With a budget of 73,000 euros and 1,800 work hours we were able to curate and archive 48 language resources. On average, a single language resource thus takes 37.5 hours and costs about 1,500 euros to curate and archive. A project which involves the collection of primary linguistic data should therefore budget *at least* this amount of time and money to prepare the data for archiving. Not included in these figures are the costs that an archive may ask for their services and long-term storage, as in our project this was offered for free by The Language Archive.

References

Drude, S., Broeder, D., Trilsbeek, P., & Wittenburg, P. 2012. The Language Archive: A new hub for language resources. In N. Calzolari (Ed.), Proceedings of LREC 2012: 8th International Conference on Language Resources and Evaluation (pp. 3264–3267). European Language Resources Association (ELRA).

Gorenflo, L.J., S. Romaine, R. A. Mittermeier, K. Walker-Painemillad. 2012. Co-occurrence of linguistic and biological diversity in biodiversity hotspots and high biodiversity wilderness areas. PNAS 109, 21: 8032–8037, doi: 10.1073/pnas.1117511109.

Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., & Wright, S. E. 2009. ISOcat: Remodeling metadata for language resources. International Journal of Metadata, Semantics and Ontologies (IJMSO), 4(4), 261–276. doi:10.1504/IJMSO.2009.029230.

Appendix: Deliverables and Milestones, with actors

FK=František Kratochvíl, MK=Marian Klamer, PT=Paul Trilsbeek, TH=Tom Hoogervorst

Deliverable (D) / Mile-stone (M)	Description	Digital media collected?	Metadata collected by:	Described in Arbil by:	Delivered by:
D	Kambera metadata + data set	yes	MK	MK	MK
D	Teiwa metadata + data set	yes	MK	MK	MK
D	Kaera metadata + data set	yes	MK	MK	MK
D	Alorese metadata + data set	yes	MK	MK	MK
D	Alor Malay metadata + data set	yes	MK	MK	MK
D	Lamaholot metadata + data set	yes	MK	MK	MK
D	Ende metadata + data set	yes	MK	MK	MK
D	Roti metadata + data set	yes	MK	MK	MK
D	Tetun metadata + data set	yes	MK	MK	MK
D	Tokodede metadata + data set	yes	MK	MK	MK
D	Lakalei metadata + data set	yes	MK	MK	MK
D	Kemak metadata + data set	yes	MK	MK	MK
D	Tetun Dili metadata + data set	yes	MK	MK	MK
D	Bunaq metadata + data set	yes	MK	MK	MK
D	Mambai metadata + data set	yes	MK	MK	MK
D	Idate metadata + data set	yes	MK	MK	MK
D	Ambai metadata + data set	yes	MK	MK	MK
D	Abui metadata + data set (by Kratochvíl)	yes	FK	FK & MK	FK
D	Sawila metadata + data set	yes	FK	FK & MK	FK
D	Subo metadata + data set	yes	FK	FK & MK	FK
D	Western Pantar metadata + data set	yes	MK	MK	MK
D	Adang metadata + data set	yes	MK	MK	MK
D	Inanwatan metadata + data set	yes	TH	TH	TH
D	Awyu-Dumut metadata + data set	yes	TH	TH	TH
D	Aghu metadata + data set	no	no	no	
D	Asmat metadata + data set	no	no	no	

(*Continued*)

D	Wambon metadata + data set	yes (Awyu Dumut)	TH	TH	TH
D	Tsaukambo metadata + data set	no	no	no	
D	Kombai metadata + data set	yes (Awyu Dumut)	TH	TH	TH
D	Citak metadata + data set	no	no	no	
D	Hatam metadata + data set	yes	TH	TH	TH
D	Sougb metadata + data set	yes	TH	TH	TH
D	Mansim metadata + data set	yes	TH	TH	TH
D	Matbat metadata + data set	yes	TH	TH	TH
D	Ma'ya metadata + data set	yes	TH	TH	TH
D	Mpur metadata + data set	no	no	no	
D	Manado Malay metadata + data set	no	no	no	
D	Bantik metadata + data set	no	no	no	
D	Mongondow metadata + data set	no	no	no	
D	Kupang Malay metadata + data set	no	no	no	
D	Javanese metadata + data set	no	no	no	
D	Begak metadata + data set	yes	TH	TH	TH
D	Blagar metadata + data set	yes	TH	TH	TH
D	Fataluku metadata + data set	no	no	no	
D	Bunaq metadata + data set	no	no	no	
D	Kamang metadata + data set	no	no	no	
D	Abui metadata + data set (by Schapper)	no	no	no	
D	Kemak metadata + data set	yes	MK	MK	MK
D	Tokodede metadata + data set	yes	MK	MK	MK
D	Woisika metadata + data set	yes	TH	TH	TH
D	Makasae metadata + data set	yes	TH	TH	TH
D	Dampelas metadata + data set	yes	TH	TH	TH
Added D	Sar metadata + data set	yes	MK	MK	MK
Added D	Klon metadata + data set	yes	MK	MK	MK
Added D	Kafoa metadata + data set	yes	MK	MK	MK
Added D	Kabola metadata + data set	yes	MK	MK	MK
Added D	Kawa metadata + data set	yes	MK	MK	MK

Added D	Hamap metadata + data set	yes	MK	MK	MK
Added D	Biak	yes	TH	TH	TH
Added D	Ambel metadata + data set	yes	TH	TH	TH
Added D	Butleh metadata + data set	yes	TH	TH	TH
Added D	Hewa	yes	TH	TH	TH
D	A synthesis of annotation issues, conventions and how they have been treated in the project, leading to a protocol/ procedure for the next six months of the project				
D	Inventory of annotations, conventions and terminology				MK
D	Mapping of resource-specific categories to ISOcat				MK & PT
D	Blagar transcribed, glossed and translated sample				MK
D	Woisika transcribed, glossed and translated sample				
D	Wambon transcribed, glossed and translated sample				
D	Kaera transcribed, glossed and translated sample				MK
D	Pre-final document describing the type of annotation issues and conventions, and the protocol/procedure used in the project to deal with these				
D	Final document describing the type of annotation issues and conventions, and the protocol/procedure used in the project to deal with these				
D	A document describing requirements and desiderata for the CLARIN infrastructure, resulting from the experiences gained through our project				MK
M	1st set of language resources online				PT
M	1st metadata harvesting test				PT
M	Inventory of annotations, conventions and terminology				MK
M	2nd set of language resources online.				PT
M	2nd metadata harvesting test				PT
M	Mapping of resource-specific categories to ISOcat				MK & PT
M	1st sample of annotated texts				MK
M	2nd sample of annotated texts				MK
M	All resources online				PT