The Nisvai Corpus of Oral Narrative Practices from Malekula (Vanuatu) and its Associated Language Resources

Jocelyn Aznar, Nuria Gala

The Nisvai Corpus of Oral Narrative Practices from Malekula (Vanuatu) and its Associated Language Resources

Jocelyn Aznar[1], Núria Gala[2]
[1]Aix Marseille Université, EHESS, CREDO
[2]Aix Marseille Université, Laboratoire Parole et Langage, LPL CNRS UMR 7309
contact@jocelynaznar.eu, nuria.gala@univ-amu.fr

Abstract

In this paper, we present a corpus of oral narratives from the Nisvai linguistic community and four associated language resources. Nisvai is an oral language spoken by 200 native speakers in the South-East of Malekula, an island of Vanuatu, Oceania. This language had never been the focus of a research before the one leading to this article. The corpus we present is made of 32 annotated narratives segmented into intonation units. The audio records were transcribed using the written conventions specifically developed for the language and translated into French. Four associated language resources have been generated by organizing the annotations into written documents: two of them are available online and two in paper format. The online resources allow the users to listen to the audio recordings while reading the annotations. They were built to share the results of our fieldwork and to communicate on the Nisvai narrative practices with the researchers as well as with a more general audience. The bilingual paper resources, a booklet of narratives and a Nisvai-French French-Nisvai lexicon, were designed for the Nisvai community by taking into account their future uses (i.e. primary school).

Keywords: Nisvai, Vanuatu, language resources, oral language, narrative practices, annotated corpora

1. Introduction

The Nisvai language is spoken by a community of around 200 native speakers in the South-East of the Malekula island, in Vanuatu. Prior to the work that we present in this paper, the Nisvai had not been subject of a research study. A lexical survey of 19 languages of the South of Malekula was conducted (Charpentier, 1984) and resulted in an linguistic Atlas of 3900 lexical items in each of those languages, plus its translation in French, English and Bislama (the Vanuatu vehicular language). During his linguistic survey, Charpentier (1984) counted only 20 native speakers of Nisvai, but he acknowledged that the language was understood and could be spoken by around 720 people in the area. This observation reveals the local status of the Nisvai: apart from Charpentier's atlas, very little information on this linguistic community is available, except some comments by Lynch et al. (2001) on the number of speakers, or Guiart (2011), who wrote some notes during a fieldwork in 1950 about the number of inhabitants and the name of their villages, or Charpentier, in Simeoni's geographical atlas (Siméoni, 2009), who commented on the increase of the Nisvai population (he compared with the data he collected in 1984).
The fieldwork leading to the production of the resources for the Nisvai language presented in this paper was undertaken from 2012 to 2015 with the support of the local community: one of the local elders, the local kinder-garden and primary school director, wished for the creation of pedagogical resources to help teach the Nisvai and French[1]. A collaboration with the local community was established in order to study the oral narrative practices and to create several language resources.

The paper is organized as follows: in section 2, we present related work regarding the problem of building resources for under-resourced languages. In section 3 we provide a description of the linguistic area of Malekula and then focus on the Nisvai language and community. Section 4 describes the methodology for building the Nisvai corpus and the different layers of its annotation, while sections 5 and 6 describe the resources. We finally discuss the distinction we made between the initial corpus of narratives and the language resources we built during the whole documentation process.

2. Related work: building linguistic resources for under-resourced languages

Since the call for documenting endangered languages (Hale et al., 1992), the network of researchers carrying out language documentation has been growing. Helped by the recent advances of language documentation software and techniques, many hundreds of language documentation projects have been conducted to produce sustainably archived audio and video collections. Thanks to projects like online archives as Paradisec (Thieberger and Barwick, 2012) or PANGLOSS (Michailovsky et al., 2014) and the Open Language Archives Community (OLAC) standards which help develop consistent interfaces between these infrastructures (Simons and Bird, 2003), the resulting language resources can be rendered available. Scientific assessment on endangered languages has also been conducted by field linguists and supported by institutions such as UNESCO [2]. Work on under-resourced and endangered languages is promoted for multiple reasons, among which we can emphasize those two:

[1]In the Vanuatu, both French and English can be used within schools as a result of previous colonization by both France and United-Kingdom. There are currently discussions within the Vanuatu government about having Bislama, the national vehicular language, as an educational language as well.

[2]See the survey framework promoted by the UNESCO (2008) to assess the endangered status of a language.

- it provides unique linguistic data that will not be accessible anymore if the language disappears and no documentation is available on it (Hale et al., 1992),
- it questions the paradigms devised with data coming from mainstreams languages (Stanford and Preston, 2009, p.1),

2.1. From language documentation to linguistic resources construction

Building resources like corpora, lexicons or dictionaries, is one of the inherent tasks associated with language documentation (see Cablitz et al. (2007), Austin and Sallabank (2011), Thieberger (2011)). Thieberger insists on the relevance of multimodality: developing resources which combine various types of data (audio, video, text) may offer multiple scientific interests, such as enabling traceability from primary data to the different analyses and to ensure sustainability, reusability and portability (Bird and Simons, 2003). Nathan (2006) associates multimodality to the notion of 'mobilization': the possibility for a resource to be used by different users (local speakers, researchers). We clearly place ourselves in this trend.

Using Natural Language Processing (NLP) techniques for language documentation is a current research issue (Aznar, 2019). Recent initiatives from the NLP community have been launched to produce massive documentation and resources for oral under-resourced languages. Such initiatives lay on a normalization of the process of collecting language data and on the use on NLP techniques to annotate and analyze linguistic items and structures: the BOLD-PNG project (Bird et al., 2013) for Papua New-Guinea or BULB (Adda et al., 2016) for three different language : Basaa, spoken in Cameroon, Myene in Gabon and Embosi in Congo-Brazzaville, are two recent examples. Both projects are inspired by the Basic Oral Language Documentation protocol (BOLD) proposed by Reiman (2010). The protocol encourages the recording of linguistic events, a portion of which should be respoken by a different speaker and translated into a more vehicular language.

Many NLP techniques require the language to be written[3]. Oral language communities that do not have their own writing system are often familiar via the school or the religious practices (Lüpke, 2011) with an already existing writing system, language documentation projects will then rely on the writing system associated with the community's religion, such as the Latin, Arabic or Korean alphabets. Language documentation projects can then rely on a writing system which has already gone through the process of informatisation[4].

2.2. Challenges and pitfalls

The lack of written conventions or orthography is one of the first challenges a documentation project has to deal with. Developing a normative way of using the alphabet is often the solution adopted by a language documentation project[5]. If the written conventions are designed to be used by the language community, taking into account the electronic devices, and their limitations, that will be used by the local community to write the language is a requirement, so that they can actually use the script on their devices[6]. In Vanuatu, only Guérin (2008, p. 58) and Aznar (2019) have taken explicitly this constraint into account.

When documenting a language is considered as producing as many documents as possible, transcription and translation can be regarded as tedious tasks slowing down the documentation process[7]. In their article, Seifart et al. (2018) encourage research on under-resourced –and often endangered– languages by taking advantage of technological helps for automating time-consuming aspects of documentation work. To contribute to the solving of that issue, NLP language documentation projects offer algorithms that provide transcriptions based on the native speakers' or linguistics' transcriptions or translations.

3. The Nisvai language and its community

Vanuatu presents the highest linguistic density in the world (François et al., 2015): 138 vernacular languages for about 260,000 inhabitants. Many of those languages are not documented yet: no records exist, nor oral nor written. Charpentier (1984) described the phonological systems of 19 languages of the South Malekula as a base to undertake his lexical survey. His work layed down the foundations for other linguistic studies in the area.

The Nisvai language is an Oceanic language, according to Lynch (2016), part of the Southeaster Malekula sub-group, an embranchment of Easter Malekula group. As it is common in this area, there is a distinction between alienable and inalienable objects, alienable objects are further subcategorised into eatable objects or not. The language does not possess many inflections, they only occurred on inalienable nouns to mark the person owning the object and on transitive verbs to refer to an object already known to the speakers and , by the process of an assibilation, to mark the verb as transitive.

According to our census during a fieldwork in 2014, the Nisvai community consists of around 200 people. The community is distributed in five villages: Renivier, Blaksand, Asuk Malekula, Levetbao and Bwenahai. The study on the Nisvai narrative practices, locally called *nabol* (Aznar, 2019), is the first one in linguistics, and social sciences in general, which focus on the Nisvai language community. It provides a description of the linguistic means used by the Nisvai speakers to organize their narrations and shows how the oral performances are differentiated depending on the age group of the speakers or the enunciative situation.

[3] If the oral language community uses its own written system which has not be adapted to digital devices, there exists some reflection about how to implement an existing writing system (Berment, 2004).

[4] Informatisation could be defined, using Berment (2004, p.18), as giving the user the means to use digital devices with the writing system associated with their language.

[5] There are other possibilities, like accepting graphical variations as proposed by Sebba (2009 01) and Clifton (2016).

[6] Many communities around the world do not use tactile devices and have access only to devices that limit the number of characters available.

[7] A language documentation project can focus on the transcription or the translation process, automatising that process would be a non-sense.

4. Methodology to create the Nisvai corpus

The process of annotation, done manually, was twofold: first during the fieldwork and second back at the office. Our prime concern was to transcribe and to translate with a local speaker the narrative events that we recorded. Secondly, back from Vanuatu, we described semantically the lexicon of the transcribed data and then analysed how the speaker organised the narratives into a plan, that is a text with an introduction, peripetia and a conclusion. Back at the office, the transcription and its French translation were reworked according to the state of the linguistic analysis and following the set of transcription conventions devised for showing the structure of the language and developing the language resources.

The annotation process was implemented using ELAN (Brugman and Russel, 2004). Seven layers (or tiers) were defined during the annotation process (see Figure 1). The unit of segmentation is the intonation unit as proposed by Chafe (1994), a prosodic unit that allows the segmentation of a recorded text without having a phonological or morphological analysis. Each intonation unit has a unique identifier within the corpus to contribute to the "accountability" (Bird and Simons, 2003) of the data.

The annotation structure can be schematized as follows:

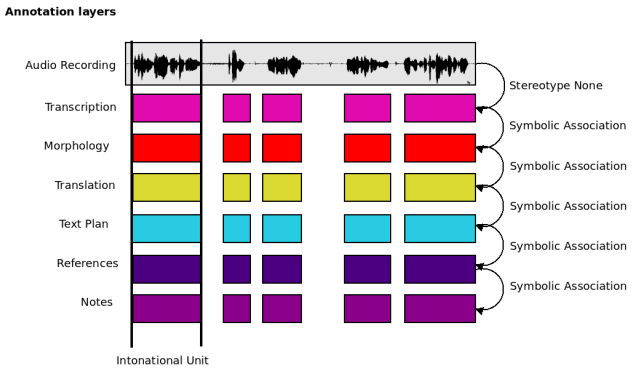


Figure 1: Scheme of the annotation structure

Elan requires the user to define a relationship between the annotation tiers. *Stereotype None* is used to allow the annotator to segmentate the audio recordings into time-align annotations. *Symbolic Association* indicates that the annotation layer matches with the other layer of annotations is refers to.

The following subsections describe the different tiers of annotation. The first layer, the audio recording of the performed oral text was done with the help of a portable audio recorder and a dynamic vocal microphone which produced lossless files, according to the quality standards of a language documentation project (Bird and Simons, 2003). The recordings were made mostly inside the kitchen house of the people narrating, or outside when the weather condition allowed for it. These recordings are considered as annotations as they result from a particular point a view, or a point of listening over the event.

4.1. Transcription

As transcription is the result of theoretical choices (Ochs, 1979) and has to be designed according to the purpose of the research (Du Bois, 1991), the conventions of transcription thus have to be designed according to the purpose of the research and its applications. In the case of the study of Nisvai narrative practices, our purpose was to design a transcription that would fulfill the wishes of the local community for a written system similar to those they know already, that is English, French and Bislama. At the same time, these transcription conventions had to be compatible with the aim of creating linguistic resources for the local school. Finally, the transcriptions were also the basis to help us describe the textual organisation of the Nisvai narratives and its variations. These choices have consequences on the way Nisvai speakers will read and write their texts, especially if the written conventions are used as fixed orthography conventions and not simply as a proposition.

The transcription, based on the phonological analysis of the Nisvai language and produced during the fieldwork, tried to stay as closed as possible to the original oral text, transcribing the speakers' disfluencies as well.

For the transcription of Nisvai texts, we retained graphical choices made by the Nisvai speakers observed during the fieldwork while doing transcription sessions. One of these choices was to use a space for what was morphologically analyzed as clitics. Using a space is not obvious, other possibilities would have been not representing them or using a dash or another character. Using another character would have been more precise in terms of linguistic information represented but would have rendered the Nisvai texts less similar to other written practices.

4.2. Morphological layer

The annotations at the morphological level allow a fine description of the Nisvai lexical units: a French equivalent is associated to each Nisvai lexical unit, and a linguistic concept is provided for each Nisvai grammatical unit.

This layer works as a pivot between the Nisvai transcription and the French translation. Two links are created, the first between the Nisvai transcription and the Morphology layer, the second between the morphological annotations and the French translation. The link between the Nisvai transcription is analytical and follows the convention adopted by interlinearisation annotations (Martin et al., 2015): the order of the annotated unit in the first line, separated by a space, is respected in the annotation layer in the second line. The second link, between the morphological layer and the French transcription highlights the semantic order: the French terms used to translate the Nisvai intonation units are turned into lexical units in the morphological layer. So it is possible to semantically link the two layers when the layers are read. This only concerns the lexical units as the Nisvai grammatical morphems are annotated with linguistic concepts (they are not systematically present in the French translation).

The following example, referenced **T1.2013.154**, is taken from the corpus (the English translation was added for the article and is not part of the corpus):

(1) Ara=silvar ga=hav ili dry kal:
3PL=discuss 3SG=end ASP.R COO.VB say
"Ils finissent de discuter et disent :"
"They finish discussing and they say:"

It shows three layers of annotation: first, the transcription, then the morphological annotation and finally the French transcription. In that example, *silvar* is associated with "discuss", *hav* with end and *kal* with "say". During the analysis of this intonation unit, three lexical units are created.

4.3. Translation From Nisvai to French

As the field of traductology made obvious, translation is a social practice which differs depending on the context and the purposes of the translator. Berman (1985) proposes that the translators explicit the targeted audience to render possible criticisms and comments on translated text. In our context, the first audience of the French translation are the teachers and students of the local school.
The first choice retained to translate the Nisvai narratives into French was to choose a textual genre adapted to convey the social role of the practice. We thus translated the Nisvai narratives with verbal tenses associated with the textual genre 'discussion' in French because, from the Nisvai perspective, a narrative is a kind of 'discussion'.

4.4. The Text Plan

The "text plan" as defined by Adam (2002) refers to each intonation unit associated to a part of the text in the narrative plan: introduction, dialogue introduction, initial situation, events (peripetia), final situation, conclusion, conclusive dialogue, song. This layer of annotation is used to segment the texts into paragraphs, each time there is a change in the text part associated with an intonation unit, a paragraph break is written.
To annotate the text plan, two different kinds of information were used: the language processes used by the Nisvai speaker to produce their narration and the understanding of the narrative events after hearing them. The correspondence between these two information

4.5. The Notes

An intonation unit can also have a note which contains the comments and information the Nisvai speaker helping the annotation process gave during the annotation session. These information given by the speaker helped understand some lexical items, clarified the action in which the characters were engaged or added socio-cultural information required to understand the context or the social implications of what was happening in the story.

4.6. The References

Having a unique identifier for each composant of the corpus contribute to the accountability of the corpus (Bird and Simons, 2003). The unique identifier of the Nisvai narrative corpus is composed of three bits of informations. Here is an example of identifier : **T1.2011.4**. The first bit, **T1**, refers to the text within the corpus. The second bit, **2011**, refers to the fieldwork during which the recording was made, and the last one, **4**, is a number corresponding to the counting of intonation units within the text.

5. The Resulting Corpus of Nisvai Narratives

5.1. Quantitative Analysis

The corpus is made of 32 narratives which have been produced by 19 different speakers. The distribution of the speakers is as follows (see Table 1):

Age group	Sex	Count
Adult	Male	4
	Female	0
Children	Female	4
	Male	3
Elders	Female	2
	Male	6
Total		19

Table 1: Distribution of recorded speakers in the corpus.

The corpus is made of 3,135 intonation units whose length can be described according to the sex and the age group, as showed in Table 2:

Age group	Sex	Count	Mean
Adult	Male	928	2454.9
Children	Female	289	1858.7
	Male	154	1904.4
Elders	Female	1072	2478.2
	Male	693	2496.9

Table 2: Distribution of intonation units.

The shortest mean length of an intonation unit of a female child is 1858.7 milliseconds (1.9 seconds) while the longest mean can be found among the male elders : 2496.9 milliseconds.

The corpus contains 31,552 tokens, of which 11,346 are lexical morphemes, and 20,206 are grammatical morphemes. Among the lexical morphemes, there are 1,250 unique lexical items, mostly common names and verbs (1,173) but also proper names (77). These entries include phonological variations. The total length, in terms of minutes, is 174 minutes. On average, a story lasts about 7 minutes. The shortest is 45 seconds, and the longest is 13'32".
An analysis of the length of the texts according to the age group shows that there is a tendency for children to produce short stories while adults and elders produce longer texts (a greater corpus would be necessary to confirm this tendency). Moreover, there is also a very strong relationship between the number of intonation units in a text and its length, the longer a text, the more the intonation units.

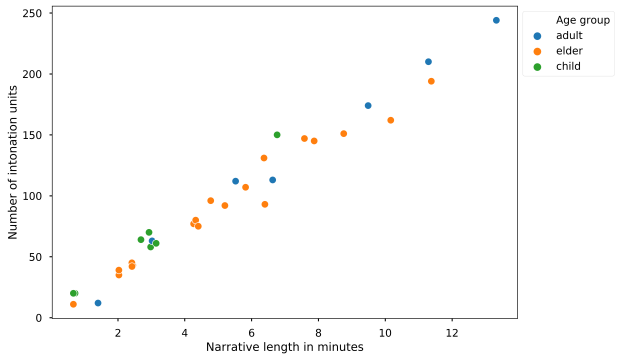


Figure 2: Texts' length according to the age group of the speaker.

5.2. General Outline about the Narrative Content

The contents of the oral narratives are varied. Among the corpus, seven stories tell the events about the "Five Finger Brothers": *Bat*: "thumb", *Keskisvas*: "index finger", *Subiarvlu*: middle finger, *Vingotngot*: "ring finger" and *Vierarar*: "little finger". They often have to deal with the ogress *Livenbumbrao* who tries to devour them. Other stories are focused on the relationship between people : a child and his father or her mother, the encounter between a young man and a young woman, a man breaking a taboo told by his wife. Finally, some stories relate events dealing with animals, like rats, turtles, fishes and birds.

6. Resources Associated to the Nisvai Narrative Corpus

Based on the annotated corpus of Nisvai narratives, four resources were developed[8]. The files with the raw data and its annotations were compiled into a database, thus yielding different linguistic multimodal information views. Each view is a linguistic resource, where the data is displayed either via a Web interface or as a pdf printable document. In this section, we first describe the two Web interfaces (the annotation audio-viewer and the text audio-viewer), and the two printable resources (the Nisvai-French French-Nisvai lexicon and the Nisvai-French story booklet).

6.1. Two Web interfaces to see and hear the Nisvai narratives

The two Web interfaces are designed to provide the scientific communities an access to the corpus and its annotations[9]. Both interfaces provide an interactive access to the annotations. Users can thereby read and listen to the annotations synchronously. The differences are situated in the way the annotations are displayed. The annotation audio-viewer shows the annotations per annotation or group of annotations while the text audio-viewer presents the whole text and its structure.

6.1.1. The annotation audio-viewer

This interface proposes to enter a form to query the annotations present in the database. Different options are possible to specify what kind of annotations the query will be made or the span of the result displayed. The user can look either for an morphological annotation by selecting the *Annotations* mode, or a lexical unit, by selecting *Lexique* or query both the transcription and the translation layers at the same time.

The results are displayed as interlinear examples (Martin et al., 2015) (see Figure 3): the first line corresponds to the Nisvai transcription, the second line to the morpho-lexical description and the third one to the French translation.

6.1.2. The Audio-visual interface

This interface allows the user to select a text from the corpus and to display the whole narrative. Its purpose is to offer a view of the text that shows how the Nisvai narrations are organised. To The interface relies on punctuation marks, dots, paragraphs and line breaks, which are reflecting Nisvai processes used by the speakers to linguistically mark the text organisation.

When the user selects a text, the Nisvai transcription is displayed on the left side of the page, while its French translation appears on the right. An audio player enables to listen to the performance. To better follow what the Nisvai speaker is saying, the intonation unit currently being pronounced is highlighted in blue both in the Nisvai transcription and in the French translation (see Figure 4).

6.2. Two paper resources for the local Nisvai school

To build the paper resources we kept in mind the requirements of the local school (the targeted audience): to help the teacher to find French translations for the Nisvai narratives. The lexicon is a list of Nisvai words with their equivalents in French and an example of use. The narrative bilingual booklet was designed to be a reference book for the Nisvai school teachers. Having the oral narrative represented as written texts adds significant value to the local narrative practices and shows how the language can be transcribed. The lexicon and the booklet were designed to be used together. The reference system used both in the lexicon and the Nisvai-French booklet enables the readers to find the text from which the examples are taken.

6.2.1. The Nisvai-French French-Nisvai lexicon

The bilingual lexicon is a book organized in two parts and which contains 1,027 verbs and common nouns. The first part of the lexicon displays the Nisvai transcriptions and then shows the French equivalent. The second part of the lexicon displays the same entries but starts with the French translation followed by the Nisvai transcription.

Each entry is a lexicon unit, either Nisvai or French, with its equivalent in the intonation unit where it appeared in the corpus of narratives. The entry is followed by up to three examples extracted form the corpus. Every example is an

[8]The resources are available at `https://jocelynaznar.eu`.

[9]Most of the members of the Nisvai community neither have a computer nor an Internet access. At present, only one member of the community owns a computer and has access to the Internet from time to time. Unfortunately, he lives in another island, Tanna.

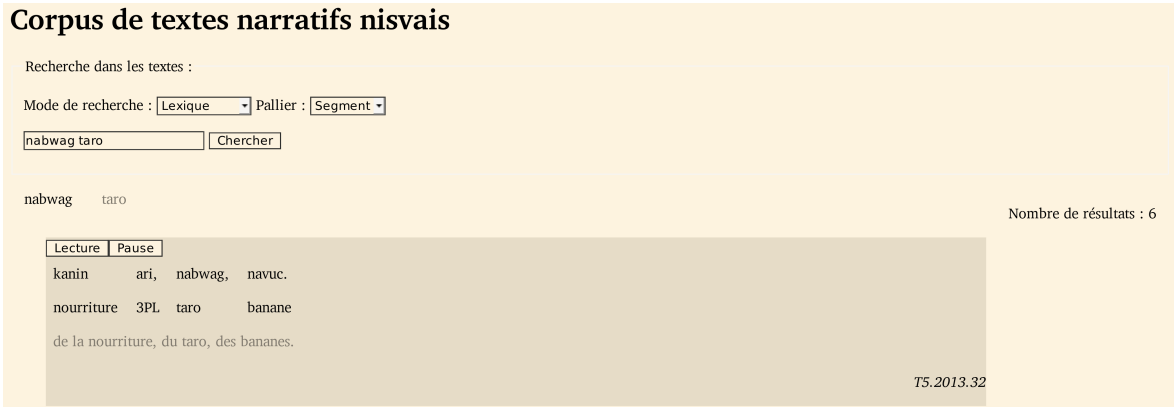


Figure 3: Interface for seeing and listening to the interlinear annotations.

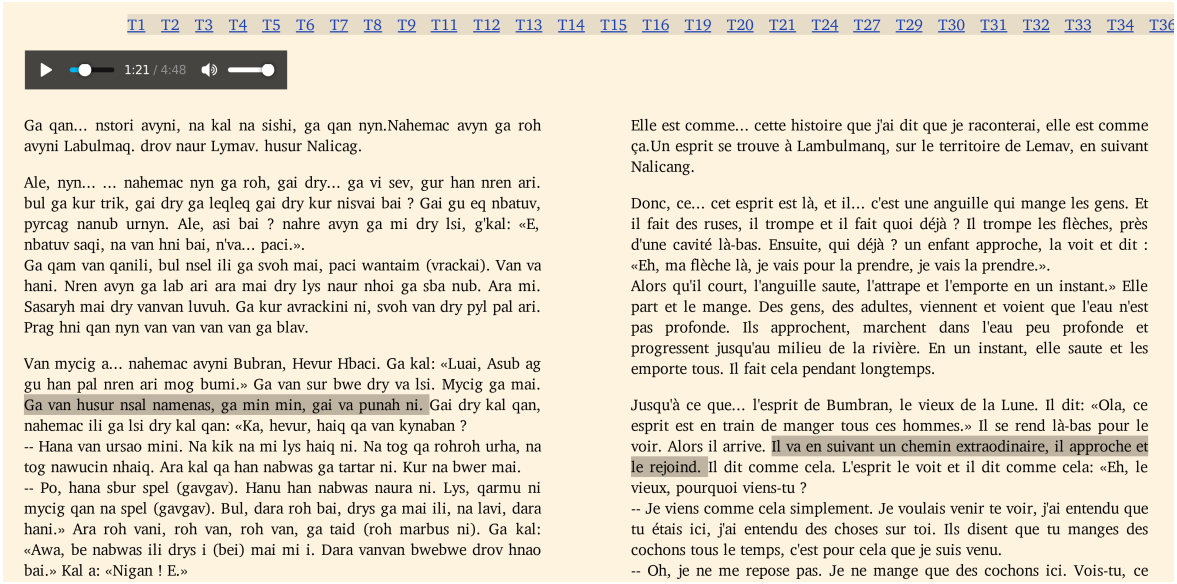


Figure 4: Interface for seeing and listening the recordings as a text.

intonation unit which contains the Nisvai term and its translation in French. If a Nisvai term is translated by another French term, or if a French term is used to translate different Nisvai terms, they are presented as different entries in the lexicon (see Figure 5).

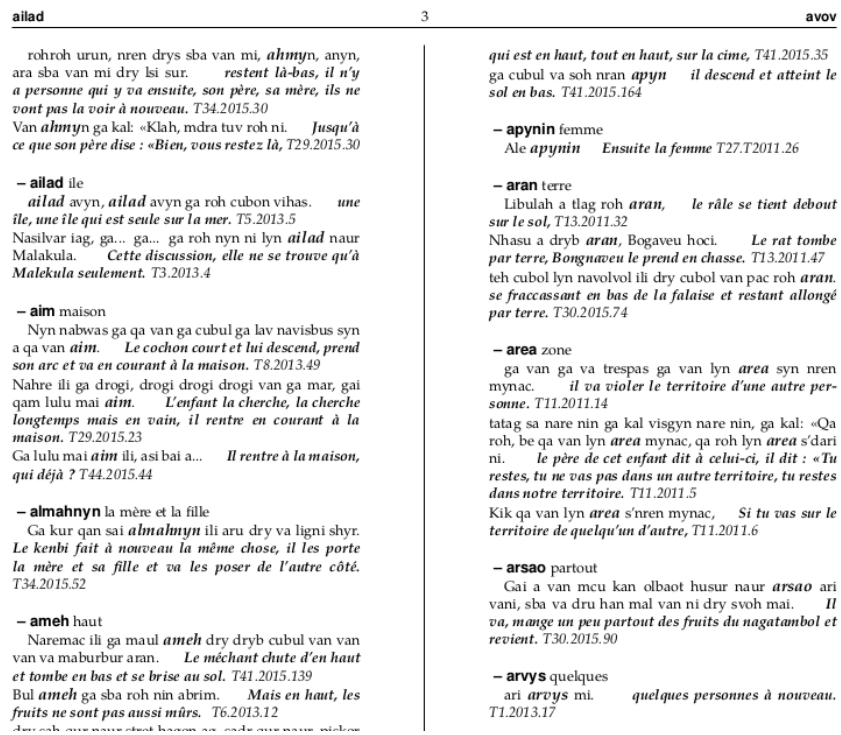
6.2.2. The Nisvai-French Story Booklet

The purpose of the Nisvai-French story booklet is to provide reading material for the teachers at the local nursery and primary school. The booklet comes along with an audio reader that contains the recordings from which the transcriptions are derived (see Figure 6). It has been designed to help the local teachers, who are not native Nisvai speakers, so that they can have references in both languages. This resource can also be useful to the researchers studying narrative practices in the Malekula (or south est Oceania) area. They will find the source data together with the analysis of the narratives.

7. Discussion

The term ‘corpus’ is employed in this article as a set of primary working data for the researcher. The notion of ‘language resource’ has been used to design a more specific outcome designed for different targeted audiences. For us, a resource inscribes itself into a discourse genre known by the targeted audience (a lexicon, a booklet, a parallel corpus) while a primary corpus collected from native speakers and annotated by a researcher matches a set of requirements according to a theoretical framework depending on the research study (in our case, the oral narratives studied from a linguistic –enunciative pragmatics– point of view).

If, in practice, a corpus can be considered as a resource (i.e. it is going to be used by a group of targeted people), as far as our corpus is concerned that possibility is limited to those having the technical skills for doing so (computer linguists and computer scientists). Indeed, our primary corpus is a set of XML files and audio recordings, annotated with ELAN which is not accessible to a lay reader.



ailad 3 avov

rohroh urun, nren drys sba van mi, ***ahmyn***, anyn, ara sba van mi dry lsi sur. ***restent là-bas, il n'y a personne qui y va ensuite, son père, sa mère, ils ne vont pas la voir à nouveau.*** *T34.2015.30*
Van ***ahmyn*** ga kal: «Klah, mdra tuv roh ni. ***Jusqu'à ce que son père dise : «Bien, vous restez là,*** *T29.2015.30*

– ailad ile
ailad avyn, ***ailad*** avyn ga roh cubon vihas. ***une île, une île qui est seule sur la mer.*** *T5.2013.5*
Nasilvar iag, ga... ga... ga roh nyn ni lyn ***ailad*** naur Malakula. ***Cette discussion, elle ne se trouve qu'à Malekula seulement.*** *T3.2013.4*

– aim maison
Nyn nabwas ga qa van ga cubul ga lav navisbus syn a qa van ***aim***. ***Le cochon court et lui descend, prend son arc et va en courant à la maison.*** *T8.2013.49*
Nahre ili ga drogi, drogi drogi drogi van ga mar, gai qam lulu mai ***aim***. ***L'enfant la cherche, la cherche longtemps mais en vain, il rentre en courant à la maison.*** *T29.2015.23*
Ga lulu mai ***aim*** ili, asi bai a... ***Il rentre à la maison, qui déjà ?*** *T44.2015.44*

– almahnyn la mère et la fille
Ga kur qan sai ***almahnyn*** ili aru dry va ligni shyr. ***Le kenbi fait à nouveau la même chose, il les porte la mère et sa fille et va les poser de l'autre côté.*** *T34.2015.52*

– ameh haut
Naremac ili ga maul ***ameh*** dry dryb cubul van van van va maburbur aran. ***Le méchant chute d'en haut et tombe en bas et se brise au sol.*** *T41.2015.139*
Bul ***ameh*** ga sba roh nin abrim. ***Mais en haut, les fruits ne sont pas aussi mûrs.*** *T6.2013.12*

qui est en haut, tout en haut, sur la cime, *T41.2015.35*
ga cubul va soh nran ***apyn*** ***il descend et atteint le sol en bas.*** *T41.2015.164*

– apynin femme
Ale ***apynin*** ***Ensuite la femme*** *T27.T2011.26*

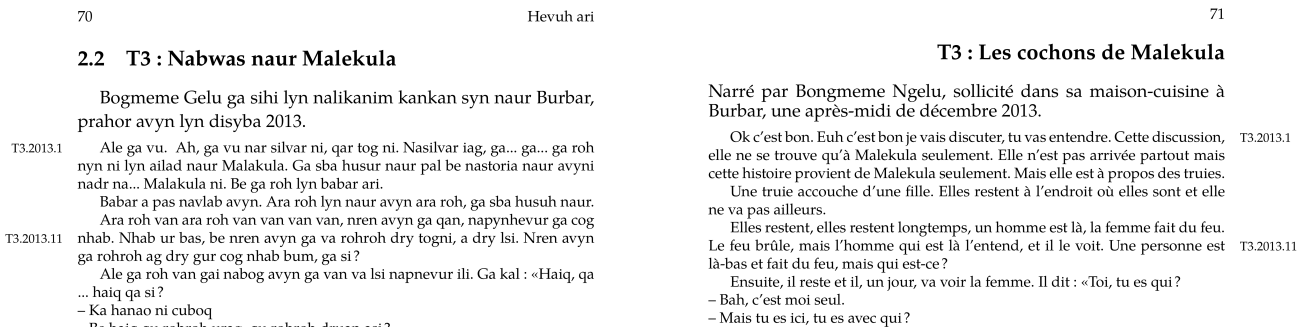
– aran terre
Libulah a tlag roh ***aran***, ***le râle se tient debout sur le sol,*** *T13.2011.32*
Nhasu a dryb ***aran***, Bogaveu hoci. ***Le rat tombe par terre, Bongnaveu le prend en chasse.*** *T13.2011.47*
teh cubol lyn navolvol ili dry cubol van pac roh ***aran***. ***se fraccassant en bas de la falaise et restant allongé par terre.*** *T30.2015.74*

– area zone
ga van ga va trespas ga van lyn ***area*** syn nren mynac. ***il va violer le territoire d'une autre personne.*** *T11.2011.14*
tatag sa nare nin ga kal visgyn nare nin, ga kal: «Qa roh, be qa van lyn ***area*** mynac, qa roh lyn ***area*** s'dari ni. ***le père de cet enfant dit à celui-ci, il dit : «Tu restes, tu ne vas pas dans un autre territoire, tu restes dans notre territoire.*** *T11.2011.5*
Kik qa van lyn ***area*** s'nren mynac, ***Si tu vas sur le territoire de quelqu'un d'autre,*** *T11.2011.6*

– arsao partout
Gai a van mcu kan olbaot husur naur ***arsao*** ari vani, sba va dru han mal van ni dry svoh mai. ***Il va, mange un peu partout des fruits du nagatambol et revient.*** *T30.2015.90*

– arvys quelques
ari ***arvys*** mi. ***quelques personnes à nouveau.*** *T1.2013.17*

Figure 5: Bilingual lexicon Nisvai-French.



70 Hevuh ari

2.2 T3 : Nabwas naur Malekula

Bogmeme Gelu ga sihi lyn nalikanim kankan syn naur Burbar, prahor avyn lyn disyba 2013.

T3.2013.1 Ale ga vu. Ah, ga vu nar silvar ni, qar tog ni. Nasilvar iag, ga... ga... ga roh nyn ni lyn ailad naur Malakula. Ga sba husur naur pal be nastoria naur avyni nadr na... Malakula ni. Be ga roh lyn babar ari.

Babar a pas navlab avyn. Ara roh lyn naur avyn ara roh, ga sba husuh naur.

T3.2013.11 Ara roh van ara roh van van van, nren avyn ga qan, napynhevur ga cog nhab. Nhab ur bas, be nren avyn ga va rohroh dry togni, a dry lsi. Nren avyn ga rohroh ag dry gur cog nhab bum, ga si ?

Ale ga roh van gai nabog avyn ga van va lsi napnevur ili. Ga kal : «Haiq, qa ... haiq qa si ?
– Ka hanao ni cuboq

71

T3 : Les cochons de Malekula

Narré par Bongmeme Ngelu, sollicité dans sa maison-cuisine à Burbar, une après-midi de décembre 2013.

Ok c'est bon. Euh c'est bon je vais discuter, tu vas entendre. Cette discussion, T3.2013.1
elle ne se trouve qu'à Malekula seulement. Elle n'est pas arrivée partout mais cette histoire provient de Malekula seulement. Mais elle est à propos des truies.

Une truie accouche d'une fille. Elles restent à l'endroit où elles sont et elle ne va pas ailleurs.

Elles restent, elles restent longtemps, un homme est là, la femme fait du feu. Le feu brûle, mais l'homme qui est là l'entend, et il le voit. Une personne est T3.2013.11
là-bas et fait du feu, mais qui est-ce ?

Ensuite, il reste et il, un jour, va voir la femme. Il dit : «Toi, tu es qui ?
– Bah, c'est moi seul.
– Mais tu es ici, tu es avec qui ?

Figure 6: Booklet of parallel corpus of narratives (Nisvai-French).

8. Conclusions and Future Work

In this paper, we have presented a corpus of annotated Nisvai narratives and four associated language resources designed for two different audiences: the Nisvai community and its school, and the scientific community interested in linguistics and ethnology of that geographical area. The corpus is the first set of written documents available for the Nisvai language. It provides different kind of annotations (e.g. intonation units, morphemes, enunciative blocks, translations into French). The different resources derived from this corpus allow a link between the primary data, the audio records and the linguistic analysis.

The Nisvai community have already received a first version of the paper resources, with an audio player to play the records. A future fieldwork will allow a discussion with the teachers in order to collect their feedback to improve the resources (e.g. foresee new formats better adapted to the specific uses at the school). Moreover, we are already working on improving the computer treatments for producing the different language resources (from a set of scripts in various computer languages to a single framework to handle the whole documentation process). One main reason motivates this change: it seems to us essential to simplify the documentation process to make it easier for sharing our data and practices with other researchers working on this area.

Acknowledgements

This research would not have been possible without the financial support of the *Centre de Recherche et de Documentation sur l'Océanie* (CREDO) at Aix Marseille Univ. (France) and the *École des Hautes Etudes en Sciences Sociales* (EHESS), particularly with the personal directions given by Prof. Laurent Dousset, who made the fieldwork possible by introducing Jocelyn Aznar to the Gelu family. We also wish to thank the Gelu family who welcomed Joce-

lyn Aznar for many months, and the members of the Nisvai community for having patiently worked with him. We are finally thankful to the Vanuatu Kultural Senta for allowing this research, and to three anonymous reviewers for their suggestions and comments.

References

Adam, J.-M. (2002). Plan de texte. In Patrick Charaudeau et al., editors, *Dictionnaire de l'analyse du discours*. Seuil.

Adda, G., Stüker, S., Adda-Decker, M., Ambouroue, O., Besacier, L., Blachon, D., Bonneau-Maynard, H., Godard, P., Hamlaoui, F., Idiatov, D., Kouarata, G.-N., Lamel, L., Makasso, E.-M., Rialland, A., de Velde, M. V., Yvon, F., and Zerbian, S. (2016). Breaking the unwritten language barrier: The bulb project. *Procedia Computer Science*, 81:8 – 14. SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09-12 May 2016 Yogyakarta, Indonesia.

Austin, P. K. and Sallabank, J. (2011). Why worry about language endangerment? In *The Handbook of Endangered Languages*, pages 6–11.

Aznar, J. (2019). *Narrer une nabol: la production des textes nisvais en fonction de l'âge et de la situation d'énonciation, Malekula , Vanuatu*. Phd thesis, Ecole de Hautes Etudes en Sciences Sociales (EHESS), Marseille, France.

Berman, A. (1985). *La traduction et la lettre ou l'Auberge du lointain*.

Berment, V. (2004). *Méthodes pour informatiser les langues et les groupes de langues "peu dotées"*. Phd thesis, Université Joseph Fourier, Grenoble, France.

Bird, S. and Simons, G. F. (2003). Seven dimensions of portability for language documentation and description. 79(3):pp. 557–582.

Bird, S., Chiang, D., Frowein, F., Berez, A. L., Eby, M., Hanke, F., Shelby, R., Vaswani, A., and Wan, A. (2013). The international workshop on language preservation: An experiment in text collection and language technology.

Brugman, H. and Russel, A. (2004). Annotating multimedia/multi-modal resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pages 2065–2068.

Cablitz, G., Ringersma, J., and Kemps-Snijders, M. (2007). Visualizing endangered indigenous languages of french polynesia with LEXUS. pages 409–414. IEEE.

Chafe, W. (1994). *Discours, consciousness, and Time*. The University of Chicago Press.

Charpentier, J.-M. (1984). *Atlas linguistique du Sud-Malakula (Vanuatu).* Coll. Langues et cultures du Pacifique, 2. SELAF.

Clifton, J. M. (2016). Orthography as social practice: Lessons from papua new guinea. 34(1).

Du Bois, J. W. (1991). Transcription design principles for spoken discourse research. 1(1):71–106.

François, A., Lacrampe, S., Franjieh, M., Schnell, S., Australian National University, and Asia-Pacific Linguistics. (2015). *The languages of Vanuatu: unity and diversity*. Studies in the Languages of Island Melanesia (SLIM). Asia-Pacific Linguistics.

Guiart, J. (2011). *Malekula : L'explosion Culturelle au Vanuatu*. Le Rocher à la Voile, 1e édition expérimentale restreinte edition.

Guérin, V. (2008). Writing an endangered language.

Hale, K., Krauss, M., Watahomigie, L. J., Yamamoto, A. Y., Craig, C., Jeanne, L. M., and England, N. C. (1992). Endangered languages. 68(1):1.

Lynch, J., Crowley, T., Pacific, A. N. U. R. S. o., and Studies, A. (2001). *Languages of Vanuatu: A New Survey and Bibliography*. Pacific Linguistics. Pacific Linguistics, Research School of Pacific and Asian Studies, The Australian National University.

Lynch, J. (2016). Malakula internal subgrouping: Phonological evidence. 55(2):399–431.

Lüpke, F. (2011). Orthography development. In *The Handbook of Endangered Languages*, The Cambridge Handbook, pages 312–353. Cambridge University Press.

Martin, H., Balthasar, B., and Bernard, C. (2015). Leipzig glossing rules : Conventions for interlinear morpheme-by-morpheme glosses.

Michailovsky, B., Mazaudon, M., Michaud, A., Guillaume, S., François, A., and Adamou, E. (2014). Documenting and researching endangered languages: the pangloss collection. 8:119–135.

Nathan, D. (2006). Thick interfaces: Mobilizing language documentation with multimedia. In *Essentials of language documentation*, Trends in linguistics. Studies and monographs, page 363 à 380.

Ochs, E. (1979). Chapter 3 transcription as theory. In *Developmental Pragmatics*. Academic Press.

Reiman, D. W. (2010). Basic oral language documentation. *Language Documentation & Conservation*, 4:254–268.

Sebba, M. (2009-01). Sociolinguistic approaches to writing systems research. *Writing Systems Research*, 1(1):35–49.

Simons, G. and Bird, S. (2003). The open language archives community: An infrastructure for distributed archiving of language resources. *Literary and Linguistic Computing*, 18(2):117–128.

Siméoni, P. (2009). *Atlas du Vanouatou (Vanuatu)*. Géoconsulte.

James N. Stanford et al., editors. (2009). *Variation in indigenous minority languages*. Number 25 in Impact studies in language and society. Benjamins. OCLC: 551806532.

Thieberger, N. and Barwick, L. (2012). Keeping records of language diversity: The pacific and regional archive for digital sources in endangered cultures (PARADISEC). *Language Documentation & Conservation*, 5.

Thieberger, N. (2011). Building a lexical database with multiple outputs: Examples from legacy data and from multimodal fieldwork. *International Journal of Lexicography*, 24(4):463–472.

UNESCO. (2008). Vitalité et disparitions des langues.