

## Making an online dictionary for Central Australian sign languages

<p><b>Margaret Carew</b></p> <p>Batchelor Institute of Indigenous Tertiary Education</p> <p>margaret.carew@batchelor.edu.au</p>	<p><b>Jennifer Green</b></p> <p>The University of Melbourne</p> <p>jag@unimelb.edu.au</p>
-------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------

### Introduction

In Central Australia, sign languages are used alongside speech, gesture and other semiotic systems such as sand drawing. These sign languages have been described as ‘alternate’, as they are not generally the primary mode of communication in these communities but rather used instead of speech in particular cultural circumstances (Green & Wilkins, 2014; Kendon, 1988 [2013]). In this paper we discuss a sign language documentation and online resource development project for Indigenous sign languages from Central Australia. In particular we track our workflow, from sign recording sessions through to the publication of selected video clips of sign in an online sign language dictionary ([www.iltyemiltyem.com](http://www.iltyemiltyem.com)). This project represents the first comprehensive attempt to document sign language knowledge in the Central Australian region since Kendon’s research in the 1980s, and his in-depth analysis of the sign languages found in some Central Australian communities provides a foundation for the current research (Kendon, 1988 [2013]).

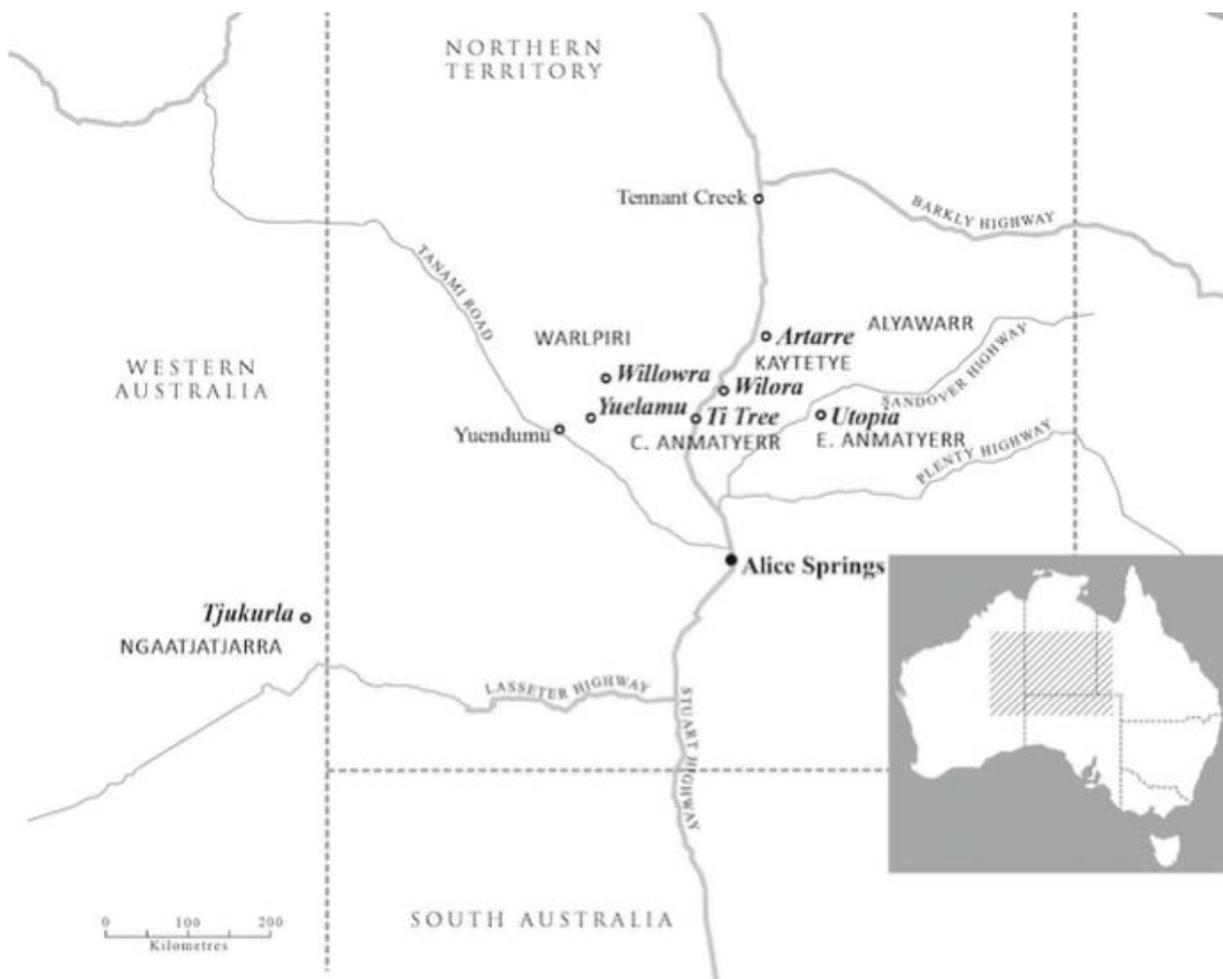
The various requirements of media publishing in environments typical of remote Central Australia has led to a divergent workflow that has two distinct paths – one that is primarily a film editing workflow and the other more suited to presenting media and metadata online. Working in a multilingual and multimodal environment also presents particular challenges for the design of sign language corpora and resources. Attention to workflow design enables multiple outcomes from language documentation – materials presented through a range of media, resources relevant to academic and community audiences, curated archival data sets and a refined corpus that enables further research.

The web-based dictionary project is titled *Iltyem-iltyem*, an Anmatyerr term meaning ‘signaling with hands, using handsigns’. The key purpose of *Iltyem-iltyem* is to support the maintenance, teaching and learning of sign languages in Central Australia by providing an accessible and contemporary online media resource. We aimed to create a media product using tools that are open-source, widely used, aesthetically pleasing and ‘pertinent’ (Nathan, 2006) to a community audience. A related aim was to design the dictionary so it could be used by sign and gesture researchers. The online repository of signs provides a refined and growing corpus that is searchable on several parameters. Corpus and workflow design providing for multiple uses of data is consistent with best practice in language documentation work. As stated by Thieberger (2011, p. 463), “we now take it for granted that all documentation should include a media corpus, that various data sources can be made to work together, and that outcomes of linguistic work be created in an archival form with derived forms for presentation”. Although archiving and future access to preservation copies of media and metadata is a central concern, in this paper we focus on our solutions for deriving web-site posts from complex sign data sets. We have found a linear workflow cannot account for the multiple uses of the sign video data. Rather, a divergent workflow with two distinct paths meets the multiple needs identified in our workflow design.

The project was designed as a language documentation and publishing partnership between signers and speakers of Indigenous languages, linguists and multi-media designers. The working practices of the project team were established in collaboration with local leaders and structured around mentoring relationships within a skilled and culturally diverse group. This involved the development and testing of a prototype website, raising awareness of issues involved in internet publishing via a project blog<sup>1</sup> and ongoing review and consultation over the use and archiving of recorded material.

A group of Anmatyerr and Warlpiri speaker/signers from Ti Tree, 200 km north of Alice Springs, and Ngaatjatjarra speaker/signer Elizabeth Marrkilyi Ellis from Tjukurla in the Western Desert region participated in the early stages of the project. The project has also recorded sign at Wilora (Stirling), Artarre (Neutral Junction), Yuelamu (Mt Allan) and Utopia, and the website is currently being expanded with contributions from these communities. Many of the participants from these communities have long-standing experience working on education, training and language documentation projects over many years (Green, 2003, 2010). Figure 1 shows the locations of communities involved in *Iltyem-iltyem* and the languages spoken there.

**Figure 1:** Map of Central Australia, showing main languages (in capitals) and names of communities (in bold italics) involved in the *Iltyem-iltyem* project



1. The blog appears on the front page of the *Iltyem-iltyem* website (<http://iltyemiltyem.com>). For a discussion of this as a form of digital outreach in language documentation, see Gawne (2015)

The website was designed to add to the considerable suite of language learning resources already developed for Indigenous communities in Central Australia by articulating with the semantic domains and graphics used in the IAD Press Picture Dictionary Series (see Green, 2003). The site was launched in Alice Springs in September 2013 and it contains close to 400 clips available for public view by registered users, who can browse and search across a range of categories.

Janie Pwerrerl Long,<sup>2</sup> a member of the Ti Tree team, has asserted the importance of sign documentation work in terms of both language maintenance and the need for suitable resources for teaching and learning sign:

*Anwern mpwarem iltyem-iltyemel arelh mapel akwer maparl akalty-irretyek. Website-warn anwern arrernem. If inang website altywer-ilem, ina can arerl iltyem-iltyem nthakenh apek. Anwerneh akalty-anthek iltyem-iltyemek angerr-pat mapel – anwernek imperl-alhek. Anengkerrant alkenty ina rrkwek angerr-pat mapel ant hand-em over-ilerlapetyart, passing on anwernek. Lyet anwern want-em-irrem akwerek pass-em on-irretyek. We want to website-warn arrernerl anwern-kenh angkety so they can iltyem-iltyem yanhek akalty-irrerl.*

*All of us women are working on the sign project so that the children can learn. We are putting the signs on a website. If they open the site then they'll be able to see how signs are done. The elders taught us sign language – they handed it down to us. They held that knowledge from the Dreaming and they handed it over and passed it on to us. Now we want to pass it on to our children. We want to put our language on the web so that the children can learn sign language (Janie Pwerrerl Long, Hanson River, 29 June 2011).*

### Considerations in website design

There are a number of technical and cultural considerations related to the appearance of the website and its functionality as a community resource. A key consideration is that the website should suit the corpus and reflect the multimodality of the recordings. It is also essential to validate regional linguistic identities and to respect cultural norms and protocols. Another important consideration is the curation and archiving of recorded material and its associated metadata. In addition, the design of the website and project workflow needed to reflect the constraints imposed by limited internet download speeds in remote communities in Central Australia. We discuss each of these considerations briefly below, before returning to outline the workflow, from sign recording to the posting of selected sign clips on the website.

Some features of our sign corpus reflect the complex communication ecologies of the desert regions, and they have theoretical implications as well as presenting practical challenges. Multimodal utterances – composites of sign and speech – are typical of many communicative contexts in these language communities (Green & Wilkins, 2014; Green, Woods, & Foley, 2011; Kendon, 1988 [2013]). To date, the focus has been on recording sign knowledge from hearing signers,<sup>3</sup> and consequently the majority of the recordings are sign/speech composites which

2. Janie Long Pwerrerl is the daughter of Lucky Long Peltharr, one of the Anmatyerr women at Ti Tree with whom Kendon worked on sign documentation in the 1980s.

3. We are yet to conduct studies of the ways that deaf individuals in the region acquire and use sign, although there is some anecdotal evidence that some traditional sign is used by Indigenous deaf with each other and with hearing members of their communities (see Adone & Maypilama, 2012, 2014; Bauer, 2012, 2014; Cooke & Adone, 1994; Maypilama & Adone, 2013; O'Reilly, 2006; Power, 2013).

include a range of spoken languages: Anmatyerr, Kaytetye, Alyawarr, Warlpiri and Ngaatjatjarra. The online resource is designed to present both speech and sign in selections of audio-visual recordings of signers from across this range of language groups. Elicitation of sign was conducted in the spoken languages of the communities. In the annotation and analysis of sign language recordings, we have thus identified the basic communicative unit as a ‘sign utterance’ comprising one or more signs, and with or without accompanying speech.

The second issue relating to website design centers around challenges inherent in making a set of community resources that have coherence in terms of local language identities and affiliations. From the perspective of the signers regional variations in sign ‘identity’ are based on a complex set of factors, but predominantly on the variety of speech (if used) of the signer and hence on their language and cultural identity. So, for example, an Alyawarr or a Kaytetye person may employ identical signs in most domains, yet speak distinct languages and belong to particular geographical areas within Central Australia. We identify instances of language-specific sign production – for example Alyawarr sign, Kaytetye sign, or Warlpiri sign – even though the sign systems used in some regions of Central Australia are essentially the same, apart from minor lexical differences (Kendon, 1988 [2013]).<sup>4</sup> As many Indigenous people in these communities are multilingual, it is also common for a particular signer to use sign in communicative contexts where one or another of several spoken languages predominate. A sign/speech composite thus may consist of a sign that is shared across the Central desert region, but coupled with regional and community specific varieties of speech – either one of the Arandic languages or one of the neighbouring languages such as Warlpiri. Both of these issues have important consequences for the implementation of a practical system for identifying particular signs - their ‘Sign IDs’, and for the ways sign lexemes are represented on the website. We will return to the problem of Sign IDs below.

We designed a project workflow to select and extract sign utterances from longer recordings, and to label individual signs within sign utterances, thus making the online resource searchable for these individual signs. Given the predominance of both speech and sign in sign utterances, we needed to display information about both in sign utterance clips. These issues have consequences for the treatment of the media and metadata throughout the annotation and online content building workflow.

Internet download speed was a technical constraint for the website design. Speed testing of an early prototype of the website at Ti Tree community in early 2012 helped to establish compression parameters for online video.<sup>5</sup> We found that extremely slow internet speeds at Ti Tree (and in Alice Springs) resulted in a significant lag in playing video, even clips of only several seconds duration. Vimeo and similar video hosting services are designed for faster Internet speeds than are currently available in Central Australia. Instead of external hosting, the project now hosts its own video files.

### Workflow design

A key challenge for the online project was to design a workflow which identifies suitable sign utterances from longer video recordings and then extracts them and their associated metadata

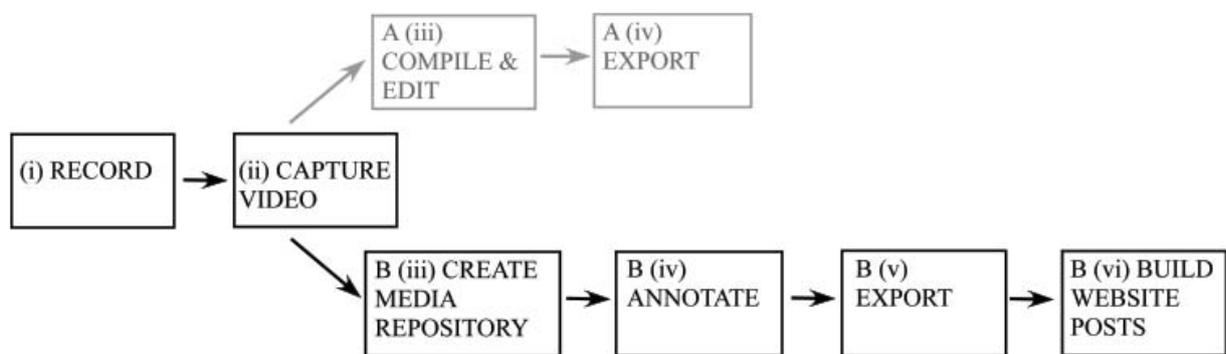
4. There are differences in “sign language development and use” between sign systems used by Anmatyerr, Kaytetye, Alyawarr, and Warlpiri people in the Central region, and those found in the Western Desert (an area fanning out from the tri-state border of Western Australia, South Australia and the Northern Territory) (Kendon, 1988 [2013], p. 53-54). The *ltyem-iltyem* website includes some examples of Ngaatjatjarra sign, but further research is needed to explore these differences.

5. Internet speed testing was conducted via <http://www.speedtest.net> in 2012. Testing of the Ti Tree School Internet showed the speed was slower than 85% of the rest of Australia. At that time, the speed over a Telstra Next-G mobile device used at Ti Tree was slower than 65% of the rest of Australia.

for presentation on the website. There are two main approaches to undertaking this. The first is to select relevant sections in a media file by start and end time codes and then call them up as ‘snippets’ (e.g. by HTML 5), and stream them from a host server which holds a repository of archival files, which remain intact. The alternative is to create ‘clips’ – a suite of secondary files that are then presented as independent items. Clips are small and more manageable for media streaming, especially over slow internet connections. This was the approach we chose for this project, even though creating clips can potentially lead to data management problems by replicating media and creating secondary files.<sup>6</sup> Although perhaps not regarded as best practice in language documentation (Thieberger, 2011), we found it to be appropriate in this context given the limits on internet speed (cf. Bowden & Hajek, 2006). Managing the large number of secondary files requires a workflow design that maintains provenance between archived media and the clips created for online publication. Our management strategy combined the potentials of the annotation software ELAN<sup>7</sup> (Crasborn & Sloetjes, 2008) with the media production tools Ffmpeg and Wordpress.

We have identified a number of stages in processing the sign language media and metadata. Figure 2 shows how the project workflow takes divergent paths after the video capture stage (ii). These paths have different outputs: path A is an editing workflow, in which media files are compiled and edited using proprietary media compilation software to make films for various purposes. Path B is a workflow for creating sets of preservation files and for selecting media to present online on the *Iltyem-iltyem* website. In the following discussion we focus on path B, which comprises the following stages: (i) Record, (ii) Capture video, (iii) Create media repository, (iv) Annotate, (v) Export and (vi) Build website posts.

**Figure 2:** The *Iltyem-iltyem* project workflow



We illustrate our workflow by tracking one sign utterance from elicitation and recording, through to annotation, and finally to presentation on the website. This sign utterance can be viewed online at: <http://iltyemiltyem.com/sign/anmatyerr/kwaty-2/>.<sup>8</sup> Henceforth we refer to this clip as the KWATY/WATER example.

6. For a discussion of issues about the presentation of audio and video snippets or clips in academic publications see <http://listserv.linguistlist.org/cgi-bin/wa?A2=ind1303&L=RESOURCE-NETWORK-LINGUISTIC-DIVERSITY&F=&S=&P=3335>.

7. ELAN is downloadable from <http://tla.mpi.nl/tools/tla-tools/elan/>.

8. To access this example, viewers are required to register to enter the dictionary component of the *Iltyem-iltyem* website.

*(i) Record*

The recording sessions follow a methodology outlined previously (Green et al., 2011), which attempts to adapt some of the conventions used in primary sign language recording to remote conditions. This has resulted in sign recordings that are consistent in terms of the lighting and positioning of the signers: for example signers are generally filmed seated in front of a blue background with the figure of the signer prominent in the frame.

The sessions involved structured elicitation in local spoken languages, based around the semantic categories used in the Picture Dictionary series, and alongside more spontaneous explorations of particular signs. Green initially led elicitation, but as the team's familiarity and confidence grew, the sessions became more collaborative. Typically this involved senior women sitting alongside Green and offering both spoken and signed prompts and coaching for those being filmed. This provided a learning opportunity for those who approached the task with low confidence – possibly due to lack of knowledge of sign but also unfamiliarity with the role of a language consultant. The sessions were greatly enriched by the multilingual capabilities of senior members of the project team.

We turn now to the featured sign utterance, the KWATY/WATER example. Figure 3 is a still taken from a video of a recording session in which this particular sign featured.

**Figure 3:** KWATY/WATER video still



Source: <http://iltyemiltyem.com/sign/anmatyerr/kwaty-2/>

The KWATY/WATER post on the website comes from a video recording made on 30 June 2011 at Ti Tree community. It is one of four different sessions filmed on that day. Each involved a team of Anmatyerr and Warlpiri women who took turns to demonstrate the signs on camera. This particular recording contains 38 sign utterances and a total of 32 unique lexical signs.<sup>9</sup> The signs span a number of semantic categories: kinship, person types, food, motion and feelings. Like most sign utterances in the corpus, they include speech as well as sign and gesture. The recording also contains elicitation prompts and requests for clarification from the signer. This diverse profile is typical of the recordings found throughout the corpus.

9. Session name: SIGN-20110630-0

*(ii) Capture video*

Video recordings are captured from the camera using Final Cut Pro X software (FCPX). This creates folders of high-resolution video files, which we refer to as ‘primary video files’. As FCPX captures files from the camera, it compiles them into folders called ‘Events’. We then re-name the media in the Events folders according to the protocols established for the project. These names identify files by date and session, and remain constant for all associated files derived from the primary media – for example the .wav and .mp4 files used in annotation, and the ELAN annotation files themselves. Establishing consistent file naming practices from the start is essential for efficient management of complex sets of media (Johnson, 2004, p.148-49).

FCPX can be used to compile and edit short films from primary files, following Path A illustrated in Figure 2. In this project, we frequently edit sign material into short films and publish them to DVD or place them onto USB devices, to ensure community members receive copies of highlights from their recording sessions soon after they occur. These are also used as consultation tools to contextualise discussions about online publishing and other uses of recorded sign language material. For some communities, where there is little or no internet access, DVD or USB based copies of sign recordings may be the only feasible way to provide copies of sign recordings. Primary files can be retained in the FCPX Events folder and in the capture format if future video editing is anticipated, but they are not in a suitable preservation format for archiving.

*(iii) Create media repository*

The project’s media repository is large, consisting of multiple files from approximately 100 recording sessions conducted between 2011 and 2015. Creating the media repository involves preparing media by transcoding it from the primary files in the FCPX events folder to appropriate formats for archiving, annotation and online delivery. This creates two other video file types – ‘preservation’ files and ‘access’ files. The distinctions between these types of files are based on the optimal formats for different uses of video files. The preservation set are unedited files in a suitable high-resolution format. The access files are smaller and suitable for both annotation and online delivery of media. A summary of video file types and their uses are presented in Table 1.

**Table 1:** Summary of video file types and formats used in the *Iltyem-iltyem* project.

<b>Workflow</b>	<b>Capture and editing</b>	<b>Annotation and online delivery</b>	
<b>Storage location</b>	FCPX Events folder	Project media repository	
<b>File type</b>	Primary	Preservation	Access
<b>Purpose</b>	Video editing	Long term archival deposit	Annotation in ELAN
<b>Outputs to</b>	Compiled and edited films	Unedited video files	Export as clips for website
<b>Format</b>	High quality Compression: Apple ProRes Wrapper: Quicktime (.mo)	Highest quality MPEG-2 or MPEG-4	Lower quality Compression: H.264 Wrapper: mp4 <sup>10</sup>

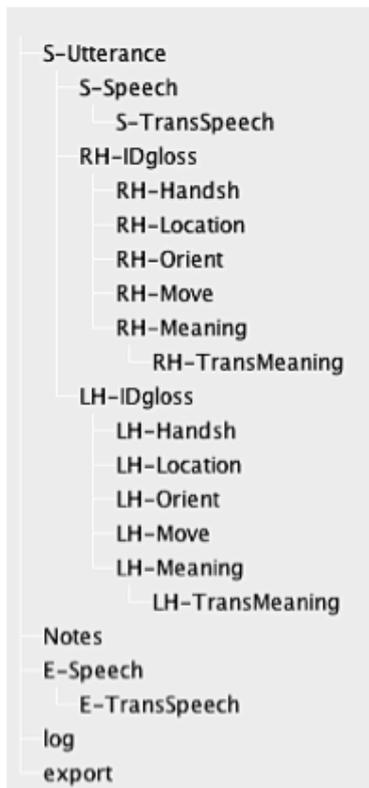
10. H.264/mp4 access files are scaled to the following parameters: frame size: 640x360 pixels; audio: 44.1 kHz; frame rate: 25fps; data rate: limited to 8000 Kbps.

*(iv) Annotation*

We now describe how project files are annotated and how clips are derived from these annotations. The software ELAN is used, and clips destined for the *Iltyem-iltyem* website are exported, along with their metadata, directly from ELAN. We use a multi-tiered ELAN template first developed by Green et al (Green et al., 2011) and informed by guidelines developed for the annotation of corpora of primary sign languages such as Auslan (see Johnston, 2014). Annotations of varying degrees of granularity – chunking sign utterances, transcribing co-signing speech, and identifying individual signs and the distinguishing features of signs, such as handshape, hand position and orientation, and hand movement – are time-aligned with their video media.<sup>11</sup>

Figure 4 shows the ELAN tier hierarchy of the *Iltyem-iltyem* project’s annotation template. The top level of the tier hierarchy, called S-Utterance, marks a composite unit of sign and speech. The S-Utterance tier is the parent of a set of analysis tiers, such as RH-IDgloss, LH-IDgloss (where sign identification labels for signs made with the right and left hand are noted) and S-Speech which is a transcript of the co-signing speech if it occurs. As discussed by Johnston (2010:125) the optimal number of annotation tiers for sign language corpora is “yet to be determined” and is very much a matter of “trial and error”.<sup>12</sup>

**Figure 4:** ELAN annotation template used for the *Iltyem-iltyem* project



11. At this stage we have not annotated any non-manual features of the signs in the corpus. In general facial expression, eye-gaze and posture-shift have no formal position in the system of signs, and only play a role at the discourse level (Kendon [1988] 2013, p. 113). However, it may be the case that Indigenous signers who have had exposure to other sign languages such as Auslan use mouthings and other facial expressions alongside manual signs.

12. Johnston (2014, p. 9) notes of the 60 or so tiers used in the standard ELAN template for the Auslan corpus that “most tiers have yet to have any annotations entered in them for the vast majority of video files”.

Annotation is undertaken in a series of passes (Johnston, 2010, p. 116), initially segmenting the video footage into sign utterances, then proceeding to assign sign ID glosses, transcribe associated speech and develop free translations. Over time, continuing work on annotation and transcription of the corpus will enable searchability and recognition of patterns through the examination of large data sets. This will increase its value as a research tool to further explore various aspects of Central Australian sign languages. Although the ELAN template for this project allows for more detailed annotation of sign forms, in the initial stages the objective was to prepare as many sign examples as possible for export to the website.

There are a number of challenges inherent in consistently and uniquely applying sign identification labeling. Ideally, a reference lexical database such as the Auslan Signbank is needed to do this effectively. Johnston (2010, p.23) suggests that the creation of a corpus without such a lexical database is “unlikely to succeed”. In the *llyem-iltyem* project, sign identification labels are developed heuristically throughout the annotation and analysis process. For the on-line dictionary, signs are labeled with a ‘Sign ID’ comprising a gloss (in the spoken language of the signer) and an English equivalent. This iterative process inevitably leads to revisions to the schema as the searchable corpus grows, and as new signs and variations to well known ones are identified.

The prevalence of sign polysemy poses particular problems in assigning unique Sign IDs. For example, in Central Australian Indigenous languages a range of kin terms is lexically differentiated in speech, yet there are less kin signs than spoken kin terms. An example is the sign for Anmatyerr spoken kin terms: *angey* ‘father, father’s brother’, *awenh* ‘father’s sister’, and *aler* ‘man’s child, a person’s brother’s child’. Each of these kin terms is signed the same way: a horizontally extended index finger taps the chin. A related situation exists for many flora and fauna terms where a single sign refers to more general categories or taxonomic groupings. In such instances, co-signing speech often serves to disambiguate what is being referred to in sign. In these circumstances, the question arises as to whether to identify the unique sign form with one Sign ID label and then describe the multiple meanings that occur in context, or alternatively to identify them individually as sign form/referent complexes.<sup>13</sup>

For the purposes of the *llyem-iltyem* website, Sign IDs comprise two parts – the first based on a speech equivalent from the spoken language of the community, and the second an English gloss that approximates the meaning of the sign. Table 2 illustrates this with Sign IDs for ‘water’ taken from four different speech communities. Four Sign IDs correspond to three distinct sign forms. The Anmatyerr Sign ID for the sample clip is KWATY/WATER, as *kwaty* is the Anmatyerr speech term for water. The Warlpiri Sign ID equivalent for KWATY/WATER is NGAPA/WATER, *ngapa* being Warlpiri for water. Warlpiri and Anmatyerr speakers sign WATER in the same way. In the Kaytetye language, *arntwe* ‘water’ leads to a Sign ID ARNTWE/WATER, and the sign is formed by using a flat hand articulating with the upper chest. In Ngaatjatjarra, the corresponding Sign ID is KAPI/WATER, and the sign is formed with a flat hand that touches the contralateral then the ipsilateral upper chest.<sup>14</sup>

13. For a discussion of the determination of sign homonyms in Auslan see Johnston (2010, p. 124). A single sign form will be given a separate sign ID gloss if meanings are “completely distinct and unrelated”.

14. Anmatyerr signers may employ both the rotated fist and the flat hand forms to sign KWATY/WATER. This variation may be understood as dialectal or language-specific variation. Alternatively there are two signs for water that may encode semantic distinctions that are not immediately apparent.

**Table 2:** Sign IDs and sign forms for WATER in Anmatyerr, Warlpiri, Kaytetye, and Ngaatjatjarra

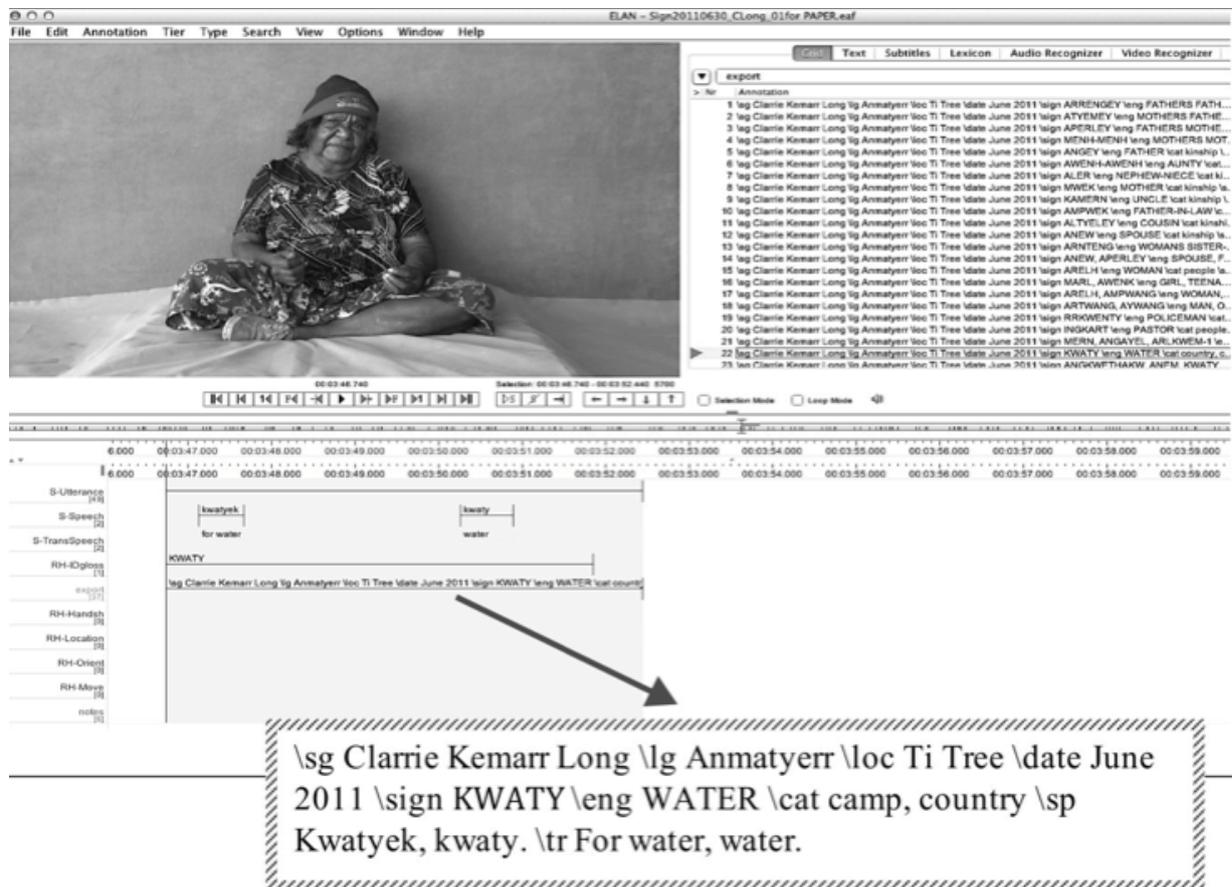
Language	Handshape	Description	Sign ID
Anmatyerr		Fist with thumb upwards is rotated rapidly several times	KWATY/WATER
Warlpiri			NGAPA/WATER
Kaytetye		Flat hand articulates with upper chest	ARNTWE/WATER
Ngaatjatjarra		Flat hand touches contralateral then ipsilateral upper chest	KAPI/WATER

Although sign forms could be given a unique and abstract value,<sup>15</sup> it is important to present sign glosses on the website in a way that is not overloaded with technical language and linguistic or abstract information. This enables intuitive searching of the website using English words and Indigenous language words. As our annotation of sign data is an on-going process, analysis of sign polysemies and preparation of clips for the website proceeds hand in hand.

For sign utterances selected for use on the website, a separate ELAN tier called 'export' is employed (see Figure 4). Metadata for these segments are entered into this tier, following a structured template that includes a range of metadata values. The metadata in the export tier for the KWATY/WATER example are shown in Figure 5.

15. For example we have considered cross-referencing signs in our corpus to Kendon's numerical sign identifiers.

**Figure 5:** Annotation window in ELAN for project file <20110630\_CLong\_01> showing the export script for the KWATY/WATER clip



(v) *Export clips and metadata*

Once the selections have been made in the export tier, and structured metadata entered for each selection, both clips and metadata are ready for export from ELAN. This is a two-stage process. An export script<sup>16</sup> commands the transcoding program Ffmpeg to create a clip of each selection.<sup>17</sup> Ffmpeg assigns a file name to the clip, which comprises the original project file name appended with the beginning and end time code values of the position of the selection within that file. To illustrate, the file name of the derived KWATY/WATER clip assigned by the ELAN export script command is <Sign20110630\_CLong\_01\_226740\_232440.mp4>.

The metadata and the time codes for selected clips are exported from the export tier via the ‘export as tab delimited text’ command in ELAN. This provides an array of structured metadata annotations alongside the time codes for each segment selected from the source media file. Figure 6 shows part of a tab delimited text file exported from ELAN and opened in a spreadsheet. Each row contains the metadata for a selected clip and the KWATY/WATER example is highlighted. Note the time codes in the first two columns correspond to those in the file name of the example clip.

16. The export script is located in the clip-media.txt file in the ELAN application folder. The required script is as follows: `<ffmpeg -y -i $in_file -ss $begin(sec.ms) -t $duration(sec.ms) -vcodec h264 -acodec aac $out_file>`. To enact this script requires installation of Ffmpeg, see: <http://www.ffmpeg.org/>.

17. For details about batch exporting of clips see the ELAN manual, section 4.9.2.8, at <http://www.mpi.nl/corpus/html/elan/ch04s09s02.html>.

**Figure 6:** Part of a tab delimited text file exported from ELAN with KWATY/WATER example highlighted

Time	Label	Text
207600	214000	\sg Clarrie Kemarr Long \lg Anmatyerr \loc Ti Tree \date June 2011 \sign INGKART \eng PASTOR \cat people \sp Inkart, Inkart. \tr This is the sign for a pastor or priest.
216120	223360	\sg Clarrie Kemarr Long \lg Anmatyerr \loc Ti Tree \date June 2011 \sign MERN, ANGAYEL, ARKWEM-1 \eng FOOD, HUNGRY, EAT \cat food, feelings, actions \sp Mern, angayel arkwetek mern. \tr Food, when s
226740	232440	\sg Clarrie Kemarr Long \lg Anmatyerr \loc Ti Tree \date June 2011 \sign KWATY \eng WATER \cat country, camp \sp
235400	241200	\sg Clarrie Kemarr Long \lg Anmatyerr \loc Ti Tree \date June 2011 \sign ANGAYEL, ANEM \eng HUNGRY, SIT \cat feelings, positions \sp Angayakw, angayakw anem. \tr Hungry, sitting down feeling hungry.
241960	247400	\sg Clarrie Kemarr Long \lg Anmatyerr \loc Ti Tree \date June 2011 \sign ANGKETHAKW, ANEM, KWATY, KWENY \eng THIRSTY, SIT, WATER, WITHOUT \cat feelings, positions, camp \sp Angkethakw anem, kw

(vi) *Build website posts*

We used Wordpress to build the *llyem-iltyem* website. Wordpress is an open source content management service, with a built-in capacity to publish online content with titles, text, embedded video, images and audio. The content can be categorized, searched and commented upon. People can register accounts with a Wordpress site, and editorial functions can be restricted to registered members. The *llyem-iltyem* project has extended these features for better site usability and manipulation of sign data, using a combination of publicly available and custom plugins.<sup>18</sup> Figure 7 shows a screen capture of the KWATY/WATER example clip and metadata, as a published post on the *llyem-iltyem* website.

**Figure 7:** The KWATY/WATER post

Language  
 ↓  
**Anmatyerr - Camp, Country**

Semantic categories  
 ↓  
**Anmatyerr - Camp, Country**



**KWATY** ← Sign ID      Signer →  
**WATER**

**Kwatyek, kwaty** ← Co-sign speech  
**For water, water**

This post was generated by the Wordpress engine, following three steps. First the clips were uploaded to the host server via Wordpress. Next, the corresponding metadata text files (as shown in Figure 6) were imported into Wordpress. The final step involved linking time code information in the metadata to time codes in clip file names, and building posts that combine

18. Plugins are packets of code that extend the core behavior of existing software.

metadata and clips. The metadata is displayed on screen with the video clip, identifying the signer, the Sign IDs (Indigenous language and English), co-sign speech and speech translation. The signer's spoken language identity and the semantic categories the sign clip belongs to are displayed above the clip. The website is searchable across all these parameters – for example, viewers are able to search for all clips contributed by any given signer or recorded in a particular community, all the clips containing particular local language and English terms, and all clips belonging to different semantic categories. The search function includes filters, which allow searches to be constrained.

The *Iltyem-iltyem* website enables various levels of access to content and material in both the front and back ends of the website. At its basic level it is open for subscriber access, which allows non-editable access to open material. There are also editor, researcher and administrator roles. This allows others who are not part of the project team to review, comment on and edit posts.

## Conclusions

The *Iltyem-iltyem* website is unique in several ways. Although there are online dictionaries for some primary sign languages, *Iltyem-iltyem* is the first searchable online dictionary of an Australian Indigenous sign language. To our knowledge it is also the first of an 'alternate' sign language to include embedded video and to represent multiple 'alternate' sign languages in a single online repository accessible to both community and academic audiences.<sup>19</sup>

Consultation about access to sign language material continues with communities. Although the searchable sign dictionary focuses on sign utterances filmed in elicitation contexts, we are also expanding the capacity of the website to host short films showing sign language in use, and demonstrate the multimodal nature of communicative practices in Central Australia. This includes recordings of sand stories, where sign is used along with drawing, speech and song (Green, 2014). We are planning for a project review in 2016 and at this time all consultants or their families will be able to add and/or withdraw material, and reflect on the consequences of being 'online'. Post 2016, the website will be archived, but we also aim for it to continue as a live media product.<sup>20</sup>

Sign language annotation is time consuming and requires specialist knowledge (Green & Wilkins, 2014). Building a corpus of time-aligned annotations linked to media provides access to a refined data set in which representative clips are labelled and searchable across a range of parameters. Extending this to an online repository of curated sign language material thus represents increased value for money for the research investment. The online repository enables researchers to work collaboratively on research questions, such as the degree to which sign languages in Central Australia differ across language groups and what similarities and differences exist between these and Indigenous sign languages found further afield.

19. Examples of on-line sign language dictionaries include the Auslan Signbank Dictionary (<http://www.auslan.org.au>), the British Sign Language dictionary (<http://www.british-sign.co.uk/british-sign-language/dictionary/>), and the New Zealand Sign Language dictionary (<http://nzsl.vuw.ac.nz>) – see also McKee & McKee (2014). The Summer Institute of Linguistics (SIL) is currently developing and testing a program called SooSL (See Our Own Sign Language), designed to support the creation of video-based dictionaries for sign languages of the world (<http://www.sil.org/about/news/new-technology-supports-language-development-signed-languages>).

20. The <http://iltyemiltem.com> website will be archived through the National Library of Australia's Pandora project: <http://pandora.nla.gov.au/subject/12>. Sign recordings from the project and associated derived files are archived at the Endangered Languages Archive at SOAS, London (<http://www.elar-archive.org/>) and at AIATSIS (<http://www.aiatsis.gov.au/collections/ava.html>).

The workflow we have outlined in this paper is designed for multiple uses of sign language data: edited films published on DVD or USB for a community audience and for use as consultation tools, selected clips presented online, and a research corpus of annotated sign language recordings. The export functions of ELAN are used to create the clips that are uploaded to the website, and provenance is maintained between the original media and the use of selected parts of it on the website through media file names and their metadata. In particular, the project aims for best practice in language documentation by ensuring the metadata on the website reflects the curatorial structure of the archival deposit. Planning for long-term preservation of language recordings and metadata is central to best practice in language documentation and a key part of the *Iltyem-iltyem* project. However, capacity constraints and community needs are also fundamental. In this paper we have demonstrated how various factors influence our project workflow. These include the multilingualism of the communities we work with, the multimodal nature of sign language recordings and the need to create a diverse range of outputs tailored for a range of audiences.

### Acknowledgements

The *Iltyem-iltyem* project is a partnership between community language teams in Central Australia, Batchelor Institute's Centre for Australian Languages and Linguistics (CALL), and Jennifer Green, a Postdoctoral Fellow at The University of Melbourne and a member of RUIL (the Research Unit for Indigenous Language). Research that forms the basis of this paper was funded by an ELDP (Endangered Languages Documentation Programme) Postdoctoral Award (IPF0173), and an ARC (Australian Research Council) Postdoctoral Fellowship (DP110102767). We thank the many signers who have participated in the project, including Elizabeth Marrkilyi Ellis, Clarrie Kemarr Long, Eileen Pwerrerl Campbell, Molly Napurrurla Presley, April Pengart Campbell, Helen Nakamarra Ross, Marie Nakamarra Long and June Nakamarra Long. The website development project was funded for 2013-2016 by the Australian Government's Indigenous Languages Support program. We are most grateful to Ben Foley for his technical expertise in developing the website and for his assistance in formulating the workflow for the *Iltyem-iltyem* project. We also thank the two anonymous reviewers for their constructive comments on a previous version of this paper. An early version of this paper was presented by the first author at the Australian Linguistics Society Annual Conference in 2013 (Carew, 2013).

### References

- Adone, D. & Maypilama, E. (2012). Yolngu Sign Language; A sociolinguistic profile. In U. Zeshan & C. de Vos (Eds.), *Sign languages in village communities. Anthropological and linguistic insights* (pp. 401-403). Berlin: De Gruyter Mouton.
- Adone, D. & Maypilama, E. (2014). *A Grammar Sketch of Yolŋu Sign Language*. Munich: Lincom Europe.
- Bauer, A. (2012). *The use of signing space in a shared language of Australia*. PhD Thesis, Universität zu Köln.
- Bauer, A. (2014). *The Use of Signing Space in a Shared Sign Language of Australia*. Berlin, Boston: De Gruyter Mouton.
- Bowden, J. & Hajek, J. (2006). When best practice isn't necessarily the best thing to do: dealing with capacity limits in a developing country. In L. Barwick & N. Thieberger (Eds.), *Sustainable data from digital fieldwork* (pp. 45-56). Sydney: Sydney University Press.
- Carew, M. (2013). *Dimensions of a non-linear workflow for Iltyem-iltyem - a Central Australian sign language documentation project*. Paper presented at the Australian Linguistics Society Annual Conference 2013, University of Melbourne, Parkville.

- Cooke, M. & Adone, D. (1994). Yolngu Signing – gestures or language? *Centre for Australian Languages and Linguistics (Batchelor College) Working papers*, October 1994, 1-15.
- Crasborn, O. & Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In O. Crasborn, E. Efthimiou, T. Hanke, E.D. Thoutenhoofd & I. Zwisterlood (Eds.), *3rd Workshop on the Representation and Processing of Sign Languages*. Paper presented at the Sixth International Conference on Language Resources and Evaluation, Marrakech, 28-30 May 2008 (pp. 39-43). Paris: ELRA.
- Gawne, L. (2015). *Language documentation and division: Bridging the digital divide. Digital Studies / Le champ numérique*. Retrieved from [http://www.digitalstudies.org/ojs/index.php/digital\\_studies/article/view/290/355](http://www.digitalstudies.org/ojs/index.php/digital_studies/article/view/290/355).
- Green, J. (2003). *The Central Anmatyerr Picture Dictionary*. Alice Springs: IAD Press.
- Green, J. (2010). *The Central and Eastern Anmatyerr to English Dictionary*. Alice Springs: IAD Press.
- Green, J. (2014). *Drawn from the ground: Sound, sign and inscription in Central Australian Sand Stories*. Cambridge: Cambridge University Press.
- Green, J. & Wilkins, D.P. (2014). With or without speech: Arandic sign language from Central Australia. *Australian Journal of Linguistics*, 34(2), 234-261.
- Green, J., Woods, G., & Foley, B. (2011). Looking at language: appropriate design for sign resources in remote Australian Indigenous communities. In N. Thieberger, L. Barwick, R. Billington & J. Vaughan (Eds.), *Sustainable data from digital research* (pp. 66-89). Parkville: University of Melbourne.
- Johnson, H. (2004). Language documentation and description, or how to build a better corpus. In P. Austin (Ed.), *Language documentation and description* (Vol. 2, pp. 140-153). London: SOAS.
- Johnston, T. (2010). From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, 15(1), 104-129.
- Johnston, T. (2014). *Auslan Corpus Annotation Guidelines*. Retrieved from Auslan Signbank: [http://media.auslan.org.au/attachments/JohnstonAuslanCorpusAnnotationGuidelines\\_14June2014.pdf](http://media.auslan.org.au/attachments/JohnstonAuslanCorpusAnnotationGuidelines_14June2014.pdf).
- Kendon, A. (1988 [2013]). *Sign languages of Aboriginal Australia. Cultural, semiotic and communicative perspectives*. Cambridge: Cambridge University Press.
- Maypilama, E. & Adone, D. (2013). Yolŋu Sign Language: An Undocumented Language of Arnhem Land. Learning Communities: *International Journal of Learning in Social Contexts*, 13, 37-44.
- McKee, R. & McKee, D. (2014). Making an online dictionary of New Zealand Sign Language. *Lexikos*, 23(1), 500-531.
- Nathan, D. (2006). Proficient, Permanent, or Pertinent: Aiming for Sustainability. In L. Barwick & N. Thieberger (Eds.), *Sustainable data from Digital Sources: from creation to archive and back* (pp. 57-68). Sydney: Sydney University Press.
- O'Reilly, S. (2006). *Indigenous Sign Languages and Culture: the interpreting and access needs of deaf people who are Aboriginal and/or Torres Strait Islander in far North Queensland*. Kewarra Beach: Disability Services Queensland.
- Power, D. (2013). Australian Aboriginal Deaf People and Aboriginal Sign Language. *Sign Language Studies*, 13(2), 264-277.
- 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. doi: 10.1093/ijl/ecr027

