elpublishing

Language Documentation and Description

ISSN 1740-6234

This article appears in: *Language Documentation and Description, vol 12: Special Issue on Language Documentation and Archiving*. Editors: David Nathan & Peter K. Austin

Increasing the future usage of endangered language archives

PAUL TRILSBEEK, ALEXANDER KÖNIG

Cite this article: Paul Trilsbeek, Alexander König (2014). Increasing the future usage of endangered language archives. In David Nathan & Peter K. Austin (eds) *Language Documentation and Description, vol 12: Special Issue on Language Documentation and Archiving*. London: SOAS. pp. 151-163

Link to this article: http://www.elpublishing.org/PID/142

This electronic version first published: July 2014

Increasing the future usage of endangered language archives

Paul Trilsbeek and Alexander König

Max Planck Institute for Psycholinguistics, Nijmegen

1. Endangered languages archives and their users

For many field linguists, digital archiving has become an integral part of their research on endangered languages (indeed Himmelmann 2006 identifies archiving as an essential characteristic of language documentation, as he defines it). It has become clear to researchers and their funding agencies that the only way to preserve the valuable material they collect for current and future generations is to store it in specialized archives accompanied by sufficiently rich metadata descriptions.

Several specialized digital archives that preserve materials on endangered languages have been established in the past decade and some already existing traditional language archives have moved into the digital era. Some, such as the Alaska Native Language Archive[1], or the California Language Archive[2], serve a specific geographical region, whereas others such as the DOBES[3] and the ELAR[4] archives are more globally oriented. According to Austin (2011), the first type of archive is most often used by members of the speaker communities, whereas the more globally-oriented archives are mainly used by their depositors and other scholars. However, both types of archives could benefit from being used by more diverse groups of people. In particular, now that some of the major language documentation funding initiatives are coming to an end[5], the question arises how maximum advantage can be gained from the archiving infrastructures that have been created, for example by

[1] www.uaf.edu/anla

[2] cla.berkeley.edu

[3] www.mpi.nl/DOBES

[4] elar.soas.ac.uk

[5] The DOBES programme had its last grant round in 2012 and ELDP SOAS has just one further grant round in 2014 according to its current funding.

Paul Trilsbeek and Alexander König 2013. Increasing the future usage of endangered language archives. In David Nathan and Peter K. Austin (eds.) *Language Documentation and Description*, *Vol 12*, 151-163. London: SOAS.

encouraging a wider range of people to engage in documenting languages and to deposit their materials into archives, as well as by drawing more users to the various archives.

2. Community involvement

An important part of the mission of many endangered languages archives is to acquire additional materials that fall within their collection policy. With the decline of funding opportunities for the documentation of endangered languages within academia, archives may want to broaden their scope and look for a wider range of possible documenters and depositors. Engaging language community members themselves in the documentation of their languages, for example, would be a way to collect more materials. Some communities are not strangers to using the Internet and services such as YouTube or Vimeo, and have shown how willingly people take the opportunity to make video recordings and share them with the world if the process is easy enough. One could think about setting up a YouTube-like portal where community members could easily upload short videos of people speaking in their languages. The endangeredlanguage.com website that was launched in June 2012 offers something along these lines. Making such resources part of an archive poses a few challenges though because archives typically strive for particular quality levels for recordings and their accompanying metadata. Moreover, they rely to a large extent on the scholarly capabilities of their depositors, and make related assumptions about the methodologies and ethical procedures that were involved in producing the recordings.

Not much is known about any of these issues for resources that are contributed by unknown depositors. The quality of recordings will reflect the limitations of the recording equipment used (mobile phone, still camera, video camcorder etc.) and the skills of the person operating them. Providing online training and advice may help to improve quality to some extent, but large variation in the quality of recordings is to be expected. Collecting high quality metadata is already a challenge for many current language documenters who are in principle already trained and obliged to provide them. Writing metadata is often seen as a boring task, but uploaded material without any metadata would be of little use. Simply providing a few free-text keywords, as is typically done on YouTube and similar sites, is probably not sufficient either. YouTube is an incredibly rich source in which one can find lots of interesting materials in a large variety of languages, but the discoverability of these materials depends on many factors such as how many other videos the depositor has uploaded, how popular the depositors' channel is, and how many views a video has

had previously, apart from how good the metadata (title, keywords, descriptions, annotations) are. In an archive that is to be used for research purposes, users would like to have more certainty about finding materials when entering the correct search terms. However, one would have to think carefully about developing a very limited set of metadata fields that are obligatory for every upload, such as language name, location, recording date, etc., and about the values that can be entered in these fields. For example, insisting on strictly controlled vocabularies for a language name field would not work since many languages are known by different names, although offering previously entered values as auto-complete suggestions may reduce the proliferation of language names. Nevertheless, one still does not know which different names actually refer to the same language. Input from expert linguists would be needed to provide mappings between language names such that results can be found regardless of which name is entered. Technological advances can help to improve the quality of metadata, for example face recognition could be used to give suggestions to the depositor regarding people who appear in video recordings or on photographs. Speaker detection algorithms could be used to do the same for audio recordings.

The fact that little is known about the methodology and ethics involved in collecting material displayed on such sites as YouTube must be made clear to users of that material. This could be achieved either by offering such materials in a distinct part of the archive's catalogue, or by clearly marking them as 'external' contributions. Within a research setting, ethical practices such as obtaining informed consent from people being recorded is common practice (du Toit 1980, Dwyer 2006). The level of awareness of such issues is unknown, however, when third party contributors deposit recordings of people other than themselves. In the DOBES archive, for example, such resources would probably in many cases breach the DOBES code of conduct (Wittenburg 2005), a set of ethical guidelines that every DOBES language documenter (and archive depositor) must adhere to. These breaches would have to be made clear to the users of the archive.

As mentioned earlier, one example of an attempt to involve community members as well as scholars and other interested people in contributing endangered languages materials to an easily accessible portal is the endangeredlanguages.com[6] website, which was launched in June 2012. This site was initially a collaboration between the University of Hawaii at

[6] www.endangeredlanguages.com

Manoa, Eastern Michigan University and Google.org, the philanthropic arm of Google Inc., as part of the ELCat (Endangered Languages Catalogue) project funded by the National Science Foundation. The ELCat project aims at compiling the most comprehensive catalogue of information about endangered languages. Google.org built the initial version of the endangeredlanguages.com website, which serves both as a front-end to the ELCat catalogue, as well as a platform where anyone with an interest in endangered languages can gather and share information and materials. The ELCat project is still ongoing, however Google.org has meanwhile handed over the site to a governance committee[7], which is led by the First People's Cultural Council in British Columbia, Canada. Further technical developments are now being carried out by ILIT at Eastern Michigan University. The site enables anyone to upload audio-visual material as well as documents about endangered languages using YouTube and Google Docs. In that sense it comes close to the scenario described above, however the site does not serve as a long-term preservation archive, nor does it currently connect to one. Upon uploading new materials to the site, the user is asked to provide metadata that is based on the OLAC[8] metadata standard. The only obligatory fields are "Language" (automatically filled in if material is added from a specific language page), "Title" and "Description". Optional fields that are provided are "Tags", "Theme", "Recorded by", "Location" and "Date". The remainder of the OLAC set ("Contributor", "Coverage", "Creator", "Publisher", "Relation", "Rights", "Source" and "Subject") is initially hidden from the depositor but can be shown upon clicking an "Additional Information" button. "Contributed by" is a field that is filled in automatically from the depositor's profile on the site and "Author" is automatically taken from the YouTube or Google Docs account that contains the material (often this is the same person as the depositor but it can be someone else in the case of existing material that someone else had previously uploaded to YouTube, for example). A year after the site was launched, about 530 of the 3,175 featured languages have had some material added to them. In total around 3,500 items have been uploaded, of which about 1,650 were audio files, 1,150 video files, 600 documents and 100 images. Inspection of a significant part of these contributions shows that the metadata fields that are predominantly used besides the obligatory ones are "Location" and "Date", if the contribution is made by

[7] The first author of this article is a member of the endangeredlanguages.com governance committee.

[8] www.language-archives.org/OLAC/metadata.html

an individual. Contributions made by an organization tend to have more fields filled out, often including some of the initially hidden fields. While most of the time it is not clear whether an individual contributor to the site is part of an endangered language community or not, there are a number of active contributors who are members of a language community. This suggests that indeed archives of endangered languages would be able to engage members of endangered languages communities in the documentation of their own languages if the interface is intuitive enough. The metadata filled in by most individuals on the endangeredlangauges.com site is minimal though, and while it is enough to find a non-specific sample of a certain language in use, specific queries e.g. for certain speech genres, certain topics of conversation (like fishing, cooking) or certain communication contexts (monologue, dialogue) yield fewer results than one would expect. Experimentation with varying the number of required fields would be needed to decide on what is acceptable.

During the last year we have also seen a number of different approaches to involving endangered language communities in the documentation of their languages using smartphone apps. One example is the *Ma! Iwaidja* app developed by Bruce Birch in collaboration with a software developer and a graphic designer.[9] This app can be used to distribute as well as to create multimedia dictionaries and phrase books. It was built for the Iwaidja community on Croker Island and comes with an Iwaidja dictionary and sound recordings, however the framework can in principle be used for any language. An upcoming version of the app should make it possible to upload to a central server new entries that users have added to the app on their own smartphones so that they could be shared with other users or deposited in an archive. Another example of an app that is built to engage communities in the documentation of their own languages is the *Aikuma* app developed by Steven Bird and Florian Hanke.[10] This app is designed to allow language community members to record stories along with some metadata, upload them to a central repository and share them with others, thereby creating a kind of social network. The app also enables users to record translations of the stories. First People's Cultural Council in British Columbia, Canada, has developed a whole series of dictionary and phrasebook apps[11] for different

[9] www.iwaidja.org/site/ma-iwaidja-phone-app/

[10] lp20.org/aikuma/

[11] www.fpcc.ca/language/FirstVoices/FirstVoices-Mobile.aspx

communities that also allow users to customize the app content by making new recordings or adding their own pictures. Mobile phone usage in many areas of the world where endangered languages are spoken has increased dramatically in recent years and while in most of these areas the percentage of smartphones on which one could use these apps is not yet very high, this is likely to change before too long. The audio and video quality that can be achieved with the latest generations of mobile (smart)phones is also approaching the quality that could only be achieved with dedicated hardware some years ago, and while currently dedicated hardware is still to be recommended for those who can afford it, the rapid decline of the number of living languages combined with the rapid increase in smartphone usage only makes it an obvious choice to use these phones in order to document as many endangered languages as possible.

3. Integrating with large-scale infrastructures

Currently there are a few large-scale European 'e-infrastructure' projects such as the CLARIN[12] and DARIAH[13] initiatives whose goal is to develop integrated digital research environments that allow researchers to combine resources and tools from various sources in a seamless way. The challenges here are the large variations in formats and standards that are used for data and metadata and the relatively poor interoperability between them. In addition, users typically need to create an account with every archive they want to use, and each archive has different conditions and mechanisms for accessing data. The e-infrastructure projects try to streamline all of this so that, for example, users can search for data across multiple archives and retrieve all the materials relevant to their query with a single account. Software tools will also be made more interoperable.

Within the CLARIN initiative, a new modular metadata model called Component MetaData Infrastructure[14] (CMDI) has been developed. CMDI allows different users or groups of users to customize a metadata schema to their own particular needs. It will be easy in this framework to create schemas that match the major current standards in the linguistics field such as IMDI[15], OLAC[16] and TEI[17], but other schemas can also be created

[12] www.clarin.eu

[13] www.dariah.eu

[14] www.clarin.eu/cmdi

[15] www.mpi.nl/imdi

if these formats do not fulfil a project's needs. It is mandatory to link each field to a concept definition in a central data category registry called ISOcat, and it is this which makes the increased flexibility possible.[18] These definitional links make it possible to search across a range of metadata schemas.

Demonstration versions of search tools have been developed to make use of the CMDI framework. There are some highly specialized search frameworks such as the CMDRSB[19] metadata browser, which is targeted at expert users and is able to fully exploit the potentially complex structures of CMDI metadata with very fine-grained queries. On the other hand, some services are very straightforward to use but do not offer the full detail of every single CMDI schema. Two examples are the CLARIN Virtual Language Observatory (VLO[20]), which aims to aggregate all existing CMDI metadata in one central place, and the NaLiDa Faceted Browser[21] which displays data collected by the German project 'Nachhaltigkeit Linguistischer Daten' (Sustainability of Linguistic Data). The approach here is to determine the fields that users are most likely to search within and make these available as 'facets' in a 'faceted browser'. Faceted browsing (also known as faceted search) is a technique that is commonly used on the World Wide Web, for example in e-commerce sites; it allows users to quickly narrow down a query to the items they are interested in by selecting values for 'facets'. After selecting a value for one facet, a user sees how many results are available for each of the remaining facets. For language archives, a user can quickly filter the available resources by language or by country (see Figure 1). The CLARIN Virtual Language Observatory currently offers the facets 'Collection', 'Continent', 'Country', 'Organization', 'Data Provider', 'Language', 'Genre', 'Subject' and 'Resource Type'.

---

[16] www.language-archives.org/OLAC/metadata.html

[17] www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html

[18] www.isocat.org

[19] clarin.aac.ac.at/MDService2

[20] catalog.clarin.eu/vlo/

[21] www.sfs.uni-tuebingen.de/nalida/katalog/app/nalida/_design/nalida/index.html

Figure 1: Selection of two facets from the Virtual Language Observatory faceted browser

Other examples of faceted browsing are the ELAR archive catalogue[22], which uses it as the main way to navigate deposits in the archive, and the latest OLAC browser[23], which shows metadata from a large number of archives.

Making use of services such as the VLO or OLAC metadata aggregators can draw more users to participating archives. Users do not have to familiarize themselves with the specific search interfaces of each archive but can perform searches across all participating archives within one interface and once they have found a resource that they are interested in they can follow a link that leads them to the actual location. The OLAC metadata aggregation service reports metrics[24] on how many monthly record views and click-throughs each participating archive receives. These numbers vary between 0 to 1399 record views and between 0 and 275 click-throughs for the month of June 2013.

[22] elar.soas.ac.uk

[23] search.language-archives.org

[24] www.language-archives.org/archives

4. Usage by other research disciplines

Archives of endangered languages contain incredibly rich collections of data that could potentially be of interest to many other research disciplines besides linguistics. Language documentation during the last decade has been done according to the recommendations of Himmelmann (1998, 2006) and others, meaning that as much as possible of the context in which the language is spoken is made part of the documentation and samples that are as broad as possible of the actual use of the language in different genres are documented. In practice this means that much of the cultural and environmental setting of the communities that speak these languages is also documented by means of video recordings, photographs and increasingly often with geographical coordinates as well, besides the more traditionally used audio recordings and written records. It is likely that researchers from related research disciplines such as anthropology, ethnobotany, ethnomusicology, history and archaeology will be able to benefit from the richness of information in these collections. There are however also collections in which unexpected information might be found such as, for example, the Alaska Native Language Archive, which turned out to contain some very detailed descriptions of star constellations by members of various Alaska Native language communities (Holton 2012) and was therefore a valuable source of information for an astronomer who was studying this very topic.

Collections of endangered languages material can only be used to their full potential for research if both linguists as well as researchers from other disciplines are able to search the collections using criteria that are relevant for them. First this means that the data should be accompanied by rich metadata descriptions. Depositors should try to look beyond the scope of linguistic use of the material and include as much information as possible that might be relevant to other disciplines in their metadata. For example, an ethnobotanist might want to search for material about plant species by their scientific names, so including such designations in the metadata descriptions of recordings would increase the chance of them being found and used by ethnobotanists. Differences in metadata terminology across different research disciplines might be overcome by making use of a central terminology registry such as ISOcat, mentioned previously. Although rich metadata descriptions can enable users to better find recordings of interest to them, the recordings themselves are not very useful to those who do not understand the language if they are not accompanied by transcriptions and translations, if one wants to do more than simply listen to the sounds and look at the videos. The creation of these transcriptions and translations is very time consuming however, and it is virtually impossible for current field linguists to fully transcribe, translate and linguistically annotate all the recordings that they make.

Peter Wittenburg, former head of The Language Archive, conducted a survey some years ago among DOBES grantees to investigate how much time was needed for transcription, translation and annotation and discovered that on average it takes 35 hours to transcribe one hour of recorded material. Translation into a major language takes on average another 25 hours and detailed linguistic analysis can take more than 100 hours per hour of recorded material (Sloetjes et al. 2011). Austin (2010) reports similar numbers for transcription and translation of conversations (cf. much lower estimates in Simons 2008, though he does not include detailed annotation).

Crowdsourcing might be one way to bridge the gap between the large amount of collected material and the small portion of it that is currently transcribed and translated. Some work in this direction has already been done on handwritten material, for example, in the *Transcribe Bleek and Lloyd* project by Ngoni Munyaradzi at the University of Cape Town.[25] Transcribing and translating audio or video material is more complicated since it requires the 'crowd' to actually understand what is being said. Using spoken translations recorded for example with the *Aikuma* or *Ma! Iwaidja* smartphone apps could be an intermediate step between the spoken endangered language and a written translation. One problem with crowdsourcing is that contributors need to be motivated to provide good results. Commercial crowdsourcing platforms give participants a small monetary reward for each performed task. Because this still does not ensure that participants perform the task to the best of their ability, transcription and translation tasks are generally given to a number of different participants and then cross-checked for consistency. A different approach to keeping people motivated in crowdsourcing tasks might be to present the task in the form of a game. Chamberlain et al. present and evaluate a number of different *Games With A Purpose* (GWAP) that are used for the creation of language resources (Chamberlain et al. 2013). GWAP could be successfully applied to different tasks such as transcription, translation and annotation, however the creation of attractive games does require a good understanding of game concepts.

Making use of state-of-the-art audio and video analysis algorithms can also speed up creating time-aligned transcriptions of audio and video. The AVATecH project[26], which was a collaboration between two Fraunhofer institutes and The Language Archive at the Max Planck Institute for

[25] boinc.cs.uct.ac.za/transcribe_bushman/

[26] tla.mpi.nl/projects_info/avatech

Psycholinguistics, tried to investigate how these algorithms could be applied to language data, including language data that was recorded under less than ideal circumstances in the field (Lenkiewicz et al. 2011). Relatively simple algorithms can, for example, detect where speech is present in an audio signal and where not. This already gives a first rough segmentation of the signal so that the annotator does not have to do this by hand. More advanced algorithms such as speaker diarization or speaker detection algorithms can indicate where in a given recording a given individual is speaking, making it easier for the researcher to find these places. Video analysis algorithms that could be used include, for example, scene detection algorithms to segment a given recording into individual scenes, or camera motion detection algorithms to indicate where, for example, zooming is used and thus where something interesting might have happened. The ELAN annotation tool[27] is able to make use of a number of these 'detectors' that have been developed within the AVATecH project.

Statistical methods that are traditionally used in the corpus linguistics domain on languages for which large written corpora are available can also be tried and adapted for languages that only have relatively small corpora, which is the case for all endangered languages. If these methods could be successfully applied, for example, to perform an automatic morphosyntactic analysis of texts in an endangered language, researchers would be saved a lot of annotation work, meaning that a larger proportion of the recorded material could be transcribed, translated and linguistically analysed. Kirschenbaum et al. (2012) have applied unsupervised machine learning techniques with reasonable success on a small corpus of the Kilivila language spoken on the Trobriand Islands in Papua New Guinea.

5. Conclusions

The endangeredlanguages.com site shows that non-linguist users are willing to contribute valuable material in and about endangered languages if the process of uploading it and giving it simple metadata descriptions is easy enough. Specialized archives therefore should be able to apply similar methods in order to increase the usage of their archiving infrastructure and to preserve more material about languages that are in a vulnerable situation. Care should be taken, however, that at least minimum requirements are met for metadata and technical quality of the resources. Archive users should be able

[27] tla.mpi.nl/tools/tla-tools/elan/

to clearly identify sources of recordings and therefore be able to assess their value for their own use.

The use of smartphone apps among members of language communities as a means of actively involving them in the documentation of their own languages looks like a very promising way to collect more language material. At the same time apps can serve as educational tools to help revitalise languages.

Archives of endangered languages should follow current e-infrastructure developments closely to make sure that their systems are compatible with upcoming infrastructures. This will make their resources accessible to a wider scientific community. Other developments, such as adaptations of state-of-the-art analytical methods previously used only on large corpora (of non-endangered languages) to now work with smaller corpora of endangered languages may help to bridge the large gap that exists between unenriched recorded material and recorded material that has been transcribed and analysed.

Archives of endangered languages contain rich collections of material that could be of interest to a wide range of research disciplines, however it might be difficult for these other disciplines to discover this wealth of information if only minimal metadata descriptions are provided and large portions of the recordings contained in these archives are not transcribed and translated. Providing rich metadata descriptions that recognise that the data could be useful beyond linguistic research and using central terminology registries such as ISOcat would be first steps towards making the data more widely usable for a greater range of purposes. Additionally, we could try to get a larger percentage of the recordings transcribed and translated by using crowdsourcing techniques, by applying state-of-the-art audio-visual analysis algorithms and by applying corpus linguistics statistical analysis methods that have been adapted to work with smaller corpora.

References

Austin, Peter K. 2010. How long is a piece of string? www.paradisec.org.au/blog/2010/04/how-long-is-a-piece-of-string/ [accessed 2013-08-23]

Austin, Peter K. 2011. Who uses digital language archives? www.paradisec.org.au/blog/2011/04/who-uses-digital-language-archives/ [accessed 2013-08-23]

Chamberlain, Jon, Karën Fort, Udo Kruschwitz, Mathieu Lafourcade and Massimo Poesio. 2013. Using games to create language resources: Successes and limitations of the approach. In Iryna Gurevych and Jungi Kim (eds.) *The People's Web Meets NLP: Theory and Applications of Natural Language Processing*, 3-44. Berlin: Springer.

du Toit, Brian M. 1980. Ethics, informed consent, and fieldwork. *Journal of Anthropological Research*, 36(3), 274–286.

Dwyer, Arienne. 2006. Ethics and practicalities of cooperative fieldwork and analysis. In Jost Gippert, Nikolaus P. Himmelmann, Ulrike Mosel (eds.) *Essentials of language documentation,* 31-66. Berlin: De Gruyter.

Himmelmann, Nikolaus. 1998. Documentary and descriptive linguistics. *Linguistics*, 36, 161-195.

Himmelmann, Nikolaus. 2006. Language documentation: what is it and what is it good for? In Jost Gippert, Nikolaus P. Himmelmann, Ulrike Mosel (eds.) *Essentials of language documentation,* 1-30. Berlin: De Gruyter

Holton, Gary. 2012. Language archives: They're not just for linguists any more. In Frank Seifart, Geoffrey Haig, Nikolaus Himmelmann, Dagmar Jung, Anna Margetts, and Paul Trilsbeek (eds.) *Potentials of language documentation: Methods, analyses, and utilization*, 105-110. Manoa: University of Hawai'i Press, Language Documentation and Conservation Special Publication.

Kirschenbaum, Amit, Peter Wittenburg and Gerhard Heyee. 2012. Unsupervised morphological analysis of small corpora: First experiments with Kilivila. In Frank Seifart, Geoffrey Haig, Nikolaus Himmelmann, Dagmar Jung, Anna Margetts, and Paul Trilsbeek (eds.) *Potentials of language documentation: Methods, analyses, and utilization*, 105-110. Manoa: University of Hawai'i Press, Language Documentation and Conservation Special Publication.

Lenkiewicz, Przemyslaw, Peter Wittenburg, Oliver Schreer, Stefano Masneri, Daniel Schneider, and Sebastian Tschöpel. 2011. Application of audio and video processing methods for language research. *Supporting Digital Humanities 2011 Proceedings.* Copenhagen, Denmark, November 17-18, 2011.

Simons, Gary. 2008. Documentary linguistics and a new kind of corpus. Paper at 5th Natural Language Research Symposium, De La Salle Univerity, Manilla, 25 November 2008.

Sloetjes, Han, Peter Wittenburg and Aarthy Somasundaram. 2011. ELAN – aspects of interoperability and functionality. *Proceedings of Interspeech 2011*, 3249-3252. Florence, Italy, August 2011.

Wittenburg, Peter. 2005. DOBES code of conduct. dobes.mpi.nl/ethical_legal_aspects/DOBES-coc-v2.pdf [accessed 2013-08-28]