Citizens, speakers, and documentation of (endangered) languages and cultures

Vatandaşlar, konuşurlar, (yok olma tehlikesi altındaki) dillerin ve kültürlerin belgelenmesi

To appear in **Journal of Endangered Languages/Tehlikedeki Diller Dergisi** Special issue on *Language Documentation in Comparative Turkic Linguistics*, edited by Éva Á. Csató and Birsel Karakoç, Uppsala University, Sweden.

Peter K. Austin
Department of Linguistics, SOAS, University of London
2023-08-30

Abstract

Over the past 30 years linguists have come to realise that there are immense threats to global linguistic diversity that could mean that around 50% of the world's 7,000 languages may no longer be spoken by the end of the 21st century because they are endangered and not being passed to future generations. One response by academic researchers has been the creation of a field of Language Documentation (or Documentary Linguistics) that has attracted a host of researchers and large amounts of grant funding, and has developed its own theorisation, recommendations for good practices, publications, and training courses. Language archives of various sorts have been established, including those with global coverage like The Language Archive (www.tla.mpi.nl) and the Endangered Languages Archive (https://www.elararchive.org/), as well as regional and local archives.

Alongside these academic developments, there have been several initiatives to collect language materials created by language speakers and others, and to catalogue and map these contributions as a form of 'citizen science'. In this paper, we will critically examine three such initiatives, Wikitongues (https://wikitongues.org/), the Language Landscape Project (www.languagelandscape.org) originated by students at SOAS, University of London, and the Endangered Languages Project (http://www.endangeredlanguages.com/) initiated by Google and now based at the University of Hawaii, Manoa, USA. Among the issues discussed are:

1. nature of the materials collected and displayed;
2. infrastructure for the projects, including web interface and data storage design, human resources, decision-making processes
3. management and vetting of user-contributed content and feedback, including possible copyright or other legal violations
4. identification of contributors and other stakeholders
5. metadata and content tagging and cataloguing
6. mechanisms for outreach and user/contributor engagement

We conclude that while citizen science and crowd-sourced data collection may appear to be attractive as research methods, there are a number of challenging issues to be overcome for them to be effective for endangered languages study.

Key words: citizen science, crowd-sourced data, endangered languages, metadata, metadocumentation

Öz

Son 30 yılda dilbilimciler küresel dil çeşitliliğine yönelik büyük tehditler olduğunun farkına vardılar. Bu, 21. yüzyılın sonuna gelindiğinde dünya üzerindeki 7.000 dilin yaklaşık %50'sinin, yok olma tehlikesi altında olup gelecek nesillere aktarılmadıkları için artık konuşulamayacağı anlamını taşımaktadır. Buna yönelik akademik araştırmacılardan gelen reaksiyonlardan biri, Dil Belgelenmesi altında, çok sayıda bilim insanının ve araştırma fonunun ilgisini çekerek kendi kuramsal temelini, iyi uygulamalar için önerilerini ve eğitim kurslarını geliştirmiş olan yeni bir alanın oluşturulması olmuştur. The Language Archive (www.tla.mpi.nl) ve Endangered Languages Archive (https://www.elararchive.org/) gibi küresel kapsamlı olanların yanı sıra bölgesel ve yerel arşivler de dahil olmak üzere çeşitli türlerde dil arşivleri kurulmuştur. Bu akademik gelişmelerin yanı sıra, konuşmacılar ve diğer insanlar tarafından oluşturulan dil materyallerini toplamak ve bu katkıları bir tür 'vatandaş bilimi' olarak kataloglayıp eşleştirmek için çeşitli girişimler olmuştur. Bu makalede, bu tür üç girişimi eleştirel bir gözle inceleyeceğiz. Wikitongues (https://wikitongues.org/), Bunlardan biri Londra Üniversitesinde, SOAS öğrencileri tarafından başlatılan Language Landscape projesi (www.languagelandscape.org), diğeri de Google tarafından başlatılan ve şu anda ABD'de Hawaii Üniversitesinde (Manoa) bulunan Endangered Languages projesidir (http://www.endangeredlanguages.com/). Makalede tartışılan konular arasında şunlar yer almaktadır:

1. toplanan ve sergilenen materyallerin niteliği
2. web arayüzü ile veri depolama tasarımı, insan kaynakları ve karar alma süreçleri dahil olmak üzere projelerin altyapısı
3. olası telif hakkı veya diğer yasal ihlaller de dahil olmak üzere, kullanıcı katkılı içerik ve geri bildirimlerin yönetimi ve güvenlik incelemesi
4. katkıda bulunanların ve diğer paydaşların belirlenmesi
5. üstveri ve içerik etiketleme ve kataloglama
6. toplumsal yardım ve kullanıcı/katılımcı yükümlülüğüne yönelik mekanizmalar

Makalede, vatandaş bilimi ve kitle kaynaklı veri toplama, araştırma yöntemleri olarak cazip görünse de, bunların yok olma tehlikesi altındaki dillerin araştırılmasında etkili olabilmeleri için aşılması gereken bir takım zorlayıcı sorunun olduğu sonucuna varıyoruz.

Anahtar kelimeler: Vatandaş bilimi, kitle kaynaklı veri, yok olma tehlikesi altındaki diller, üstdata, üst belgeleme

1. **Introduction**[1]

From the early 1990s linguistics researchers have highlighted the fact that a large number of the world's roughly 7,000 languages are under pressure from more dominant languages that are economically, socially, politically, and ideologically more powerful, leading to them becoming endangered and not being passed to children (Robins & Uhlenbeck 1991, Hale et al. 1992, Robins & Uhlenbeck 1991, Hale et al. 1992, Grenoble & Whaley 1998, Nettle & Romaine 2000, Crystal 2000, Unesco 2003). The distribution of speaker numbers across the world is highly unbalanced, with the largest 10 languages, each having more than 100 million speakers (Mandarin, Spanish, English, Bengali, Hindi, Portuguese, Russian, Arabic, Japanese), together accounting for 2.6 billion speakers (40% of global population). As Crystal (2000: 19) notes, just 4% of all the existing languages are spoken by 96% of the world's population, i.e. only 4% of the world's population speaks 96% of the languages, meaning that there are many languages that are very small (50% have fewer than 10,000 speakers, 25% have fewer than 1,000). In addition, there have been radical reductions in speaker numbers in the past 70 years across all regions of the world, together with increasing age profiles of current speakers, mostly as a result of language shift to more prestigious and socio-economically powerful regional, state, or national languages. Krauss (1992) argued that "the coming century will see either the death or the doom of 90% of mankind's languages"; others offer less extreme estimates, but a commonly quoted figure (see Austin & Sallabank 2011) is loss of 50% of linguistic diversity by the end of the 21st century (i.e. approximately 3,500 of the existing languages).

This academic concern for threatened languages has co-occurred with growing awareness of linguistic and cultural rights of minorities, ethical and equality considerations (the global history of genocide, fights for land rights etc.), and the exercise of political power by discriminated individuals and groups. Indeed, the United Nations has declared 2022-2032 the International Decade of Indigenous Languages.[2] In addition, the past 20 years has seen the rise of social media platforms (Facebook, Twitter, Instagram, WhatsApp, WeChat), blogs, and websites that have created spaces for minority languages previously excluded from mainstream media. Changes in communications, media and information technologies, the availability of smart mobile devices, and powerful apps has placed capabilities to record, edit, distribute, and consume multimedia in the hands of virtually everyone, including speakers of endangered languages, and non-tech-savvy academics and laypeople.

2. **Responses**

One scholarly response to this situation has been increased interest in describing, documenting, and archiving endangered languages (Austin 2010, 2016, Woodbury 2011, Seifart et al. 2018), aimed at creating illustrative corpora of language performances

[1] The initial draft of this paper was presented as a talk at the *Citizen Science Lab: Sampling Language and Culture* Workshop held at the Lorenz Centre, University of Leiden, 3rd-6th April 2018. I am grateful to Nicoline van der Sijs, Roberta D'Alessandro, Hans Bennis, and Steven Krauwer for inviting me to participate in this stimulating workshop. I also thank Anna Belew, Lyle Campbell, Ebany Dohle, Samantha Goodchild, Karolina Grzech, and Charlotte Hemmings for information and discussions about the case studies discussed below. None of them is responsible for any errors of fact or interpretation. Thanks to Birsel Karakoç for the Turkish translations.

[2] http://www.un.org/development/desa/indigenouspeoples/indigenous-languages.html, accessed 2023-07-25

evidencing language in use. This has been accompanied by major boosts in research funding by agencies such as Arcadia's Endangered Languages Documentation Programme (more than $70million since 2002),[3] Volkswagen Foundation's DoBeS project ($60million, 2001-2016),[4] and the NSF-NEH joint programme Documenting Endangered Languages (more than $80million since 2003).[5] There has also been the establishment of digital language archives of several types: global (ELAR,[6] TLA[7]), regional (e.g. AILLA,[8] ANLA,[9] CLA,[10] Paradisec,[11] Pangloss[12]), or local and community-based (e.g. DAUM,[13] Anindilyakwa[14]).

The research area of language documentation has recently emerged as a response to global language endangerment. This has been defined by Gippert, Himmelmann & Mosel (2006: v) as "concerned with the methods, tools, and theoretical underpinnings for compiling a representative and lasting multipurpose record of a natural language or one of its varieties". The outcome of research in this paradigm is often taken to be an annotated and translated corpus of archived representative materials on use of a language or a variety (with an accompanying grammatical sketch and cataloguing metadata). It is seen as separate from description (which treats language as a system, expressed in the form of grammars, dictionaries and text collections – see Austin & Grenoble 2007). For further discussion of the theory and practice of language documentation see Austin (2021, section 2), and for examples of the outcomes of such projects see the archives mentioned above.

Speaker communities and individuals have responded to the challenges of language endangerment in several ways, outside of their involvement in documentation projects of the type described above:

- language revitalisation initiatives to increase domains of use and/or numbers of speakers, often through education or grass-roots activities (master-apprentice, language nests, language camps, immersion schooling, mother-tongue multilingual education). This has often resulted in development of educational materials, but much of them only exist as "grey literature"[15] with limited distribution;
- engagement in language exchange through social media, especially on Twitter, Facebook, Instagram, and through multimedia messaging apps like WhatsApp or WeChat. This has involved hundreds of languages, but most of the material created is siloed within closed groups and inside the platforms, not being accessible to outsiders. Interesting listing of some of this material on Twitter can be found in Kevin

[3] https://www.edlp.net, accessed 2023-07-25
[4] https://dobes.mpi.nl/projects/, accessed 2023-07-25
[5] https://www.nsf.gov/pubs/2022/nsf22615/nsf22615.htm, accessed 2023-07-25
[6] http://www.elararchive.org, accessed 2023-07-25
[7] https://archive.mpi.nl/tla/, accessed 2023-07-25
[8] https://ailla.utexas.org/, accessed 2023-07-25
[9] https://www.uaf.edu/anla/, accessed 2023-07-25
[10] https://cla.berkeley.edu/, accessed 2023-07-25
[11] https://www.paradisec.org.au/, accessed 2023-07-25
[12] https://pangloss.cnrs.fr/?lang=en, accessed 2023-07-25
[13] www.sofi.se/servlet/GetDoc?meta_id=1196, accessed 2023-07-25
[14] https://www.anindilyakwa.org.au/language-resources/, accessed 2023-07-25
[15] See https://en.wikipedia.org/wiki/Grey_literature, accessed 2023-07-26

Scannel's *Indigenous Tweets*,[16] and *Indigenous Blogs*.[17]

In addition, there have been some interesting recent examples of what could be called 'citizen science' relating to endangered and lesser known languages which we discuss in the following sections.

3. **Citizen science**

The term 'citizen science' was introduced in 1989 to refer to "scientific research conducted with participation from the general public (who are sometimes referred to as amateur/non-professional scientists)".[18] There are various understandings of the amount of participation and control that non-professional scientists exhibit in citizen science projects, but their primary roles are commonly seen as data collectors, monitors, classifiers, or popularisers of research within the wider community, in collaboration with professional researchers. Citizen science projects exist across a wide range of discipline areas, ranging from ecology to astronomy to climate change and other areas – indeed, the EU Citizen Science website[19] lists 268 projects engaging with schools, community organisations, and individual volunteers.

3.1 **Languages and citizen science**

Some academic researchers have realised that the language documentation paradigm as envisaged in section 2 above faces a number of difficult challenges:

- currently there are at least 3,500 endangered languages and far too few trained researchers to document and create annotated and archived corpora for even a small proportion of them;
- as Wasson et al. (2016: 641) argue: "most language archives are not meeting the needs of most users. Representatives from all user groups expressed frustration at the current design of most language archives"[20];
- the knowledge and skills required by language documenters are multi-faceted and involve a wide range of disciplines, requiring many years of training (see Austin 2008 for an overview of one example);
- data collection and analysis typically involves substantial periods of fieldwork, often under difficult personal, social, and political conditions, which can deter potential researchers, especially those from outside of speaker communities;
- annotation, translation, and provision of metadata is very time consuming, requiring tens or hundreds of multiples of time of the recordings collected (see

[16] See http://indigenoustweets.com/, accessed 2023-07-26. Notice, however, that this listing and *Indigenous Blogs* includes languages like Hausa, Kinyarwanda and Aymara, which each have millions of speakers and are not endangered currently.

[17] http://indigenousblogs.com/, accessed 2023-07-26

[18] https://en.wikipedia.org/wiki/Citizen_science, accessed 2023-07-26. Individual research activities involving non-professional data collectors, such as annual observation censuses by bird watchers, has taken place for much longer.

[19] https://eu-citizen.science/projects, accessed 2023-07-26

[20] See also Burke et al. (2022).

Austin 2010), while frequently not being recognised as a significant research activity (Garrett & Harris 2022);
- growing expectations of full engagement and empowerment of speech community members mean that research needs to address decolonisation (Austin 2018), and be made more accessible to non-professional researchers.

To address these issues, could [21]language research, especially involving endangered languages, be part of the dynamic growth area of citizen science? To investigate this question we explore three initiatives that focus on non-professionals creating audio-visual and textual recordings of language material from across the world that is freely and openly accessible via a website, namely Wikitongues (3.2), the Endangered Languages Project (3.3), and the Language Landscape project (3.4). For each project we focus on the nature of the materials displayed, the organisation and management of contributions, and the use and outreach of resources by researchers and the general public. This included interviewing in 2018 the managers of the last two sites concerning the following matters:

1. **content**

- how does the project monitor audio-visual content and does it flag up materials that are considered inappropriate? How is this done if the community of speakers is very small and there may be very few people who can understand a given audio/video recording? Is there a way for people to lodge a complaint about a recording, and if so who is the ultimate arbiter if a complaint is lodged?
- are text materials, including metadata, provided by users vetted, and if so how?
- what happens to resources that contain transcriptions or translations or subtitles that might be considered offensive, or at least derogatory?

2. **identity and intellectual property**

- how does the project decide on the identity of the resource submitters? Are they vetted in some way? Can they be anonymous? Has the project considered using, e.g. Facebook or Google, as a means of login authentication for the site?
- how does the project deal with material that might violate copyright?
- who vets the metadata provided for submissions? Are metadata tags based on someone watching/listening to the content? If so, is that scalable?

3. **Interfaces, use and outreach**

- who designed the interface for the site? Was it tested before the design was finalised? How is the site hosted and maintained? What is the back-end catalogue for the media, metadata, and the site contents?
- does the project track usage of submitted content, eg. views or downloads, and/or citations of materials or information from them?
- what mechanisms exist for outreach and citizen engagement? Word of mouth? Social media? School visits or public presentations?

[21] The Multilingual Manchester project (http://mlm.humanities.manchester.ac.uk/index.html, accessed 2023-08-24) that ran from 2010 to 2021 encouraged input from the general public via its LinguaSnap smartphone app (http://www.linguasnapp.manchester.ac.uk/, accessed 2023-08-25), however this was limited to images of multilingual signage, and did not cover spoken or signed language use, unlike the three projects discussed below.

The following sections look at the three selected projects to address these issues as instances of citizen science.

3.2 **Wikitongues**

Wikitongues is an American non-profit organization registered in the state of New York that was founded in 2014 by Frederico Andrade, Daniel Bögre Udell, and Lindie Botes, and currently involves 1,500 volunteers.[22] It publishes video recordings submitted by individuals and couples speaking in languages other than English (mostly monologues or conversations spoken directly to camera), and advertises that videos in 700 languages and lexicons in 200 languages are available on its site.[23] There is no publicly accessible catalogue of these materials, and they can only be searched and played/downloaded for individual languages. Very basic metadata about the uploaded files is given, along with the ISO 639-3 and Glottlog codes for the language, and links to Wikipedia and Open Language Archives entries, if available. In about 15% of cases, subtitles are provided for the recording, but for the vast bulk there is no transcription and no translation into a language of wider communication, making them effectively inaccessible to anyone who does not speak the recorded language. In addition, where languages show geographical, interpersonal, social, genre, or other variation, this is not indicated in the minimal amount of metadocumentation. What is given can sometimes be misleading – for example, searching for "Sasak" locates just one video, identified as "The Sasak Language of Indonesia: Raden speaking Sasak and Indonesian"[24] submitted by Nabil Berri (no further metadata). The particular location of the recording, other than "Indonesia", is not given, yet this is important as Sasak has massive local variation at the village level on Lombok Island such that particular ways of speaking from different locations may be mutually unintelligible (Austin 2003). The speaker is identified as "Raden" but this is a widely-used address term for adult males of the Sasak nobility, and not a personal name. Finally, most of the conversation is in Bahasa Indonesia, the national language, with only a few Sasak terms quoted. Another example is a search for "Gamilaraay" that returns one video of "Des speaking Kamilaroi"[25] which has Des Crump (no further metadata) speaking his heritage language (specific location not identified) for six seconds (a memorised self-introduction), followed by three minutes in English.

The videos on Wikitongues can probably be understood as specimens of individuals speaking something whose identity is unclear and value for documentation and preservation is limited. The greatest issue with the site is the lack of metadata and metadocumentation that could potentially make the submitted material understandable and usable by an interested audience, including researchers. It resembles more a Cabinet of Curiosities (Wunderkammer) than a scientific project.

3.3 **The Endangered Languages Project**

The Endangered Languages Project (ELP) was established in 2012 by Google.org, and then

[22] https://en.wikipedia.org/wiki/Wikitongues, accessed 2023-08-26
[23] https://wikitongues.org/languages/, accessed 2023-08-26
[24] https://www.youtube.com/watch?v=GohjqZQHDlM, accessed 2030-08-26
[25] https://www.dropbox.com/s/3c2pdi46n4cd39w/Des%20Crump%20-%20Kamilaroi.mp4?dl=0, accessed 2023-08-26

transitioned to the *Alliance for Linguistic Diversity* (comprising the First Peoples' Cultural Council,[26] and the Endangered Languages Catalogue (ELCat) at University of Hawaii Manoa[27]). It has a Governing Council and Advisory Committee, and is managed on a day-to-day basis by ELCat staff. Its goals are given as:[28]

> through this website, users can not only access the most up to date and comprehensive information on endangered languages as well as language resources being provided by partners, but also play an active role in putting their languages online by submitting information or samples in the form of text, audio or video files. In addition, users will be able to share best practices and case studies through a knowledge sharing section and through joining relevant Google Groups.

Most of the content uploaded to ELP is hosted on several associated Google products or services, including YouTube, Picasa, and Google Docs, with product policies and content guidelines devolved to each service. In addition (Belew, p.c.):

> all content submitted through other Google products or services must be in accordance with their associated terms. These include but are not limited to: a prohibition on content containing pornography, obscenity, pedophilia, bestiality or other sexually explicit material; hateful or violent content; harassing content or content that infringes another's privacy.

Discussions with the ELP management revealed that the more than 6,700 submitted resources are not reviewed for content by ELP, nor is the submission of a resource to ELP an indication that it is endorsed. Submitters are responsible for their own uploads, and are free to describe the materials and/or comment on the contents. Users of the website can flag inappropriate material, and the flagging system immediately removes it (material which is not yet "confirmed" as inappropriate will not be visible). To date, ELP has never had material flagged as containing inappropriate content in an endangered language; usually, materials are flagged for being the wrong language, or as irrelevant or spam. Content guidelines are extremely broad but do not cover, e.g. derogatory subtitles or voiceovers on videos. The ultimate arbiter of content suitability is someone with competence in the language, working in tandem with the Governing Council, and the person who flagged the video as inappropriate.

Identification of submitters is managed by Google as all ELP accounts are tied to Google accounts, and resource submissions and votes/flags also track submitters' IP addresses. Individual accounts are not vetted beyond the Google login process. Users are required to confirm that uploads do not violate copyright during the submission process; ELP lacks resources to check copyright on all the submissions. All hosting is done on other sites so users conform to those hosts' terms of service in addition to those of ELP. Copyright violations can also be flagged manually by other users. Metadata tags for uploaded files are assigned by the submitter; suggested tags for each resource category are provided, but users

[26] http://fpcc.ca/, accessed 2023-08-27

[27] http://ling.hawaii.edu/research-current/projects/elcat/, accessed 2023-08-27. See also Campbell & Belew (2018).

[28] https://www.endangeredlanguages.com/about/, accessed 2023-08-27

can also add any tags they wish. The website interface was designed by Google, with access provided in six ex-colonial European languages and simplified Chinese.

ELP does not allow file downloads, and Google analytics tracks pageviews (from launch until mid-2018 there were 767,498 pageviews of submitted video resources, about 4.7% of total site traffic) along with user locations and gender. ELP is represented on Facebook and Twitter, and gives presentations at conferences and gatherings of language workers and academics, plus Google events; there have been official ELP presentations at more than 30 events. Most introduction to linguistics courses at University of Hawaii Manoa include ELP in their curriculum, as do some other universities, together with a few US public high schools. The project is keen to do more school engagement, but has limited resources to do so.

3.4 **Language Landscape**

The Language Landscape (LL) project grew out of an initiative called *London's Language Landscape* by staff and students at SOAS, University of London, to map languages spoken in the city of London by attendees at *SOAS Endangered Languages Week* in May 2011. Subsequently, LL became a non-profit organization set up by a group of postgraduate linguistics students, and was funded by small grants and donations; unfortunately, the project website is currently broken due to software issues. The main organisers have moved on in their careers, so LL is effectively defunct. While it was operating, LL comprised an interactive online map,[29] and several educational outreach projects. The goals, theoretical underpinnings, and functionality of LL are described in Dohle, Grzech & Hemmings (2014), Dohle (2015), Ritchie, Goodchild & Dohle (2016), and Grzech & Dohle (2018).

The LL mapping model does not represent languages per se, but rather displays instances of language use at the mapped location (e.g. a conversation in Arabic in Edinburgh, recitation of a poem in Polish in Madrid). The rationale for this is that it maps language performances, variation, and multilingualism (all of particular interest to language documenters – see section 2), and is especially relevant in urban contexts, such as major international cities. LL aimed to reach a wide audience of educators, primary and secondary school students, university students, academic researchers, minority and endangered language communities, and social media users. It encouraged non-professional people to upload audio and video recordings of language use events (typically made on a mobile phone), tagged for geolocation and basic metadata (resulting in 1005 data points by early 2022, when the site stopped functioning properly). Figure 1 from Dohle (2015) shows the LL home page:

[29] This was originally at www.languagelandscape.org. A version of the site dated 16 January 2022 has been preserved on the Internet Archive, however this is not fully functional as it relied on the Google Maps API and Javascript, which were not archived. LL videos can be found at https://www.youtube.com/@languagelandscape/videos, accessed 2023-08-24.

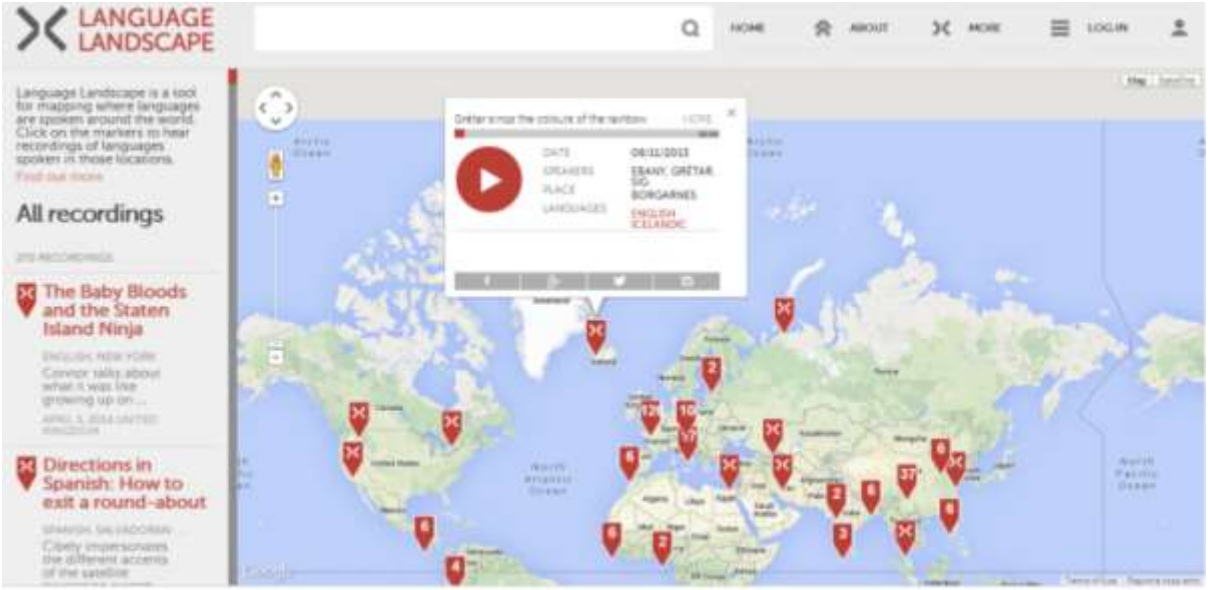

Figure 1: The Language Landscape homepage with an in-page recording.

Users could scroll around the map to locations of interest, and the website also allowed searching for instances of language events that could be filtered for language and date, as in Figure 5 from Dohle (2015).

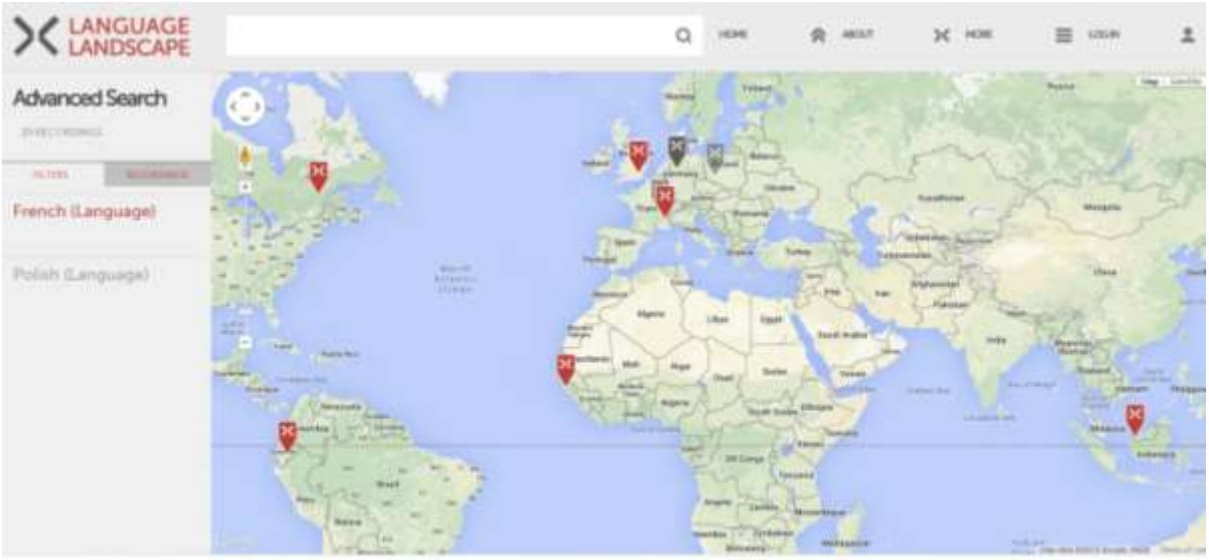

Figure 5: A custom sub-map: Recordings of either French or Polish made after 24 January 2012. The filters on the left show recordings and their corresponding search terms in different languages such as

Submitters were encouraged, but not required, to provide additional metadata about their recordings, such as details of speakers (names, ages, ethnicity, gender), topics, genre or speech style, and transcriptions and translations. For further details, see the references above.

LL organised a series of outreach activities at London schools and communities, and completed a pilot educational programme in east London, providing school students with practical training in recording techniques, and helping them to learn about issues such as multilingualism and language endangerment. The LL website was used as a starting point for discussion and activities.

In 2018 I raised the questions outlined in Section 3.1 above with the then LL organisers and received the following information. The LL website had a set of submission guidelines written in plain English which all uploaders were required to abide by. For content review, all recordings had to be approved by one of the LL administration team before they appeared on the website, and:

> currently, we are making an informed judgement as to what we publish and most of the recordings come from events run by colleagues that we are in touch with in and who can verify the content of their recordings.

Users of the LL website could flag inappropriate material by contacting LL via Twitter, Facebook, or email. Resource contributors could determine whether recording pages were editable or not (choosing "other users can edit this recording" during the upload process); if so, other uses could make changes or additions to the text materials. The LL management team discussed any contentious materials as a group. The identity of submitters was established when they created a profile in order to upload recordings; individual profiles were not vetted. Google and Facebook were not used to identify users. Individual submitters were required to confirm that their uploads did not violate copyright during the submission process; LL had no resources to check copyright on all the submissions. Copyright violations could be flagged manually by users, as was the case for ELP, discussed above (see 3.3).

All metadata tags for recordings were completed by the submitter; potential tags were pre-set and filled in on a form by users when they uploaded their files. A sub-set of metadata tags was obligatory, and tags were searchable. The website interface was designed by an IT professional (the brother of one of the student principals), funded by a Google Earth Outreach Developer grant and a grant from SOAS Alumni & Friends. The site was backed up to Amazon S3 cloud storage and maintained by a web-developer. It ran on a Python web application server and stored media files and metadata in a relational database. It also used WordPress, and Google Maps API; videos were hosted by YouTube. Activity on the site was tracked with Google Analytics.

In the areas of outreach and engagement, LL was very active at schools in London and the field sites of the affiliated post-graduate students.[30] It had a strong presence on Facebook,[31] Twitter,[32] and its blog (sadly, now defunct), attended freshers' fairs, presented at academic conferences, gave invited talks at universities, and collaborated with community centres/museums and organisations that had community outreach projects. The LL administrators noted that the website content had not been used for research purposes, but the platform was used for University-level teaching and public outreach (independently of LL), e.g. in a project at the University of Bielefeld. Internationally, LL presented its platform at the *Science is Wonder-ful* event[33] in Brussels in 2017, hosted by the European Commission as part of the European Researchers Night which was attended by 4,600 people. These examples show clearly that a citizen science-orientated project of this type can be very effective in communicating with the general public.

[30] See video at https://www.youtube.com/watch?v=iII2pcGgdmc, accessed 2023-08-27
[31] https://www.facebook.com/languagelandscape, accessed 2023-08-27
[32] https://twitter.com/langlandscape, accessed 2023-08-27
[33] See https://marie-sklodowska-curie-actions.ec.europa.eu/science-is-wonderful, accessed 2023-08-27

4. **Conclusions**

Over the past 20 years, academically-based language documentation and description research has resulted in the collection and analysis of recordings of language use events for hundreds of minority and endangered languages, much of which has been archived and made accessible to researchers and other interested parties. However, this represents only a small fraction on the world's endangered languages, so alternative approaches are need. A few citizen science-type initiatives have emerged since 2011 to enable non-professional researchers, including members of speaker communities, to document and support minority and endangered languages. Three of these provide a platform to submit publicly viewable audio and video resources to a website, where they may be accessed and searched by anyone. These are *Wikitongues,* the *Endangered Languages Project* (ELP) and *Language Landscape* (LL). Unfortunately, due to a lack of resources and career changes by its originators, LL ceased to function in 2020, but it provides a very interesting case study of collaboration between academic researchers and interested non-professionals; its structure, functions, and activities have been well described by its principals (see references), and could be instructive for other initiatives in citizen science applied to languages.

All three projects are small scale and rely heavily on volunteers for management and submission of materials. Only ELP has an institutional host (the University of Hawaii), and the precariousness of the lack of such institutional support is clear from the demise of LL. The three projects rely on good behaviour by the citizen scientists as they all require submitters to abide by publicly-available guidelines for content, copyright, and community-sanctioned behaviour. Submitted materials are not vetted, and the sites rely on a tripwire system where violations of standards are flagged by users; subsequently, corrective action is taken if necessary. All three projects have been used in university-level teaching, and both ELP and LL have been active in outreach to school and general audiences, with LL having had notable success at a major international science fair event in 2017. All of these are characteristics of impactful citizen science.

Another variant of citizen science for endangered languages has emerged in the past five years, especially in South Asia, Canada, and Australia, that looks promising as a response to the challenges outlined in 3.1 above. This involves local university academics, especially in India, Pakistan, Canada and Australia, training members of ethnic communities in situ in basic principles of language documentation and description, and encouraging them to work with members of their own speaker groups to collect audio-visual materials that could be used for knowledge preservation and to support language and cultural revival. Organisations such as *First Voices*,[34] *First Languages Australia*,[35] the *Society for Endangered and lesser known Languages* (SEL India)[36] and *Living Tongues Institute for Endangered Languages*[37] have been supporting this development, in concert with grassroots initiatives in minority communities, such as the *North Eastern Institute of Language and Culture*.[38] Another development that occurred during the Covid-19 pandemic in 2020-2021 was the development of models for 'remote fieldwork' where researchers who were unable to travel collaborated via the internet and supplied hardware with trained local researchers on data

[34] https://www.firstvoices.com/home, accessed 2023-08-30
[35] https://www.firstlanguages.org.au/, accessed 2023-08-30
[36] https://selindia.org/, accessed 2023-08-25
[37] https://livingtongues.org/, accessed 2023-08-24
[38] https://www.neilac.org.in/, accessed 2023-08-25

collection and analysis, using software tools for data and metadata management and transfer of files between the fieldwork site and the base outside. Williams, Silva, McPherson & Good (2021: 359) report on case studies in West Africa, Amazonia and Indonesia, suggesting that:

> elements of remote fieldwork should become a permanent part of linguistic fieldwork, but that such methods need to be considered in the context of decolonizing language documentation and centering the community's needs and interests

The concrete outcomes of these kinds of initiatives as a particular narrower form of citizen science will be interesting to observe in coming years.

References

Austin, Peter K. 2003. The Linguistic Ecology of Lombok. *PELBBA* 16, 165-198. http://www.peterkaustin.com/docs/Austin_2003_Sasak_PELBBA.pdf

Austin, Peter K. 2008. Training for language documentation: Experiences at the School of Oriental and African Studies. In Margaret Florey & Victoria Rau (eds.) *Documenting and Revitalising Austronesian Languages*, 25-41. Language Documentation and Conservation Special Publication No. 1. Hawaii: University of Hawaii Press

Austin, Peter K. 2010a. Current issues in language documentation. *Language Documentation and Description* 7, 12-33.

Austin, Peter K. 2010b. How long is a piece of string? Post on Endangered Languages and Cultures blog14 April 2010. https://www.paradisec.org.au/blog/2010/04/how-long-is-a-piece-of-string/

Austin, Peter K. 2016. Language documentation 20 years on. In Martin Pütz & Luna Filipović (eds.) *Endangered Languages and Languages in Danger: Issues of ecology, policy and human rights*, 147-170. Amsterdam: John Benjamins

Austin, Peter K. 2018. Colonialism in language documentation and revitalization – the times they are a changin'? Talk given at University of Malaya, 6 December 2018. Slides available at http://www.peterkaustin.com/docs/teaching/2018-12-06_UM.pdf

Austin, Peter K. 2021. Corpora and archiving in language documentation, description, and revitalisation. *Ethnorema* 17, 53-65.

Austin, Peter K. & Lenore A. Grenoble. 2007. Current trends in language documentation. *Language Documentation and Description* 4, 12-25.

Austin, Peter K. and Julia Sallabank. 2011. Introduction. In Peter K. Austin and Julia Sallabank (eds.) *Cambridge Handbook of Endangered Languages*, 1-24. Cambridge: Cambridge University Press

Austin, Peter K. & Andrew Simpson. (eds.) 2007. Endangered Languages. *Linguistische Berichte Sonderheft* 14. Hamburg: Helmut Buske Verlag.

Burke, Mary, Oksana. L. Zavalina, Shobhana L. Chelliah & Mark E. Phillips. 2022. User needs in language archives: Findings from interviews with language archive managers, depositors, and end-users. *Language Documentation & Conservation* 16, 1-24.

Campbell, Lyle & Anna Belew. 2018. Introduction: Why catalogue endangered languages? In Lyle Campbell & Anna Belew (eds.) *Cataloguing the World's Endangered Languages*, Chapter 1. London: Routledge.

Crystal, David. 2000. *Language Death*. Cambridge: Cambridge University Press.

Dohle, Ebany. 2015. Language Landscape. *Unravel Blog*, 1 June 2015. https://unravellingmag.com/dialogue/language-landscape/

Dohle, Ebany, Karolina Grzech & Charlotte Hemmings. 2014. Language Landscape: A digital platform for mapping languages. *Book 2.0*, 4(1-2), 71-89.

Garrett, Andrew & Alice Harris. 2022. Assessing scholarship in documentary linguistics. *Language* 98(3), e156-e172.

Gippert, Jost, Nikolaus P. Himmelmann & Ulrike Mosel (eds.) 2006. *Essentials of language documentation*. Berlin: Mouton de Gruyter.

Grenoble, Leonore A., & Whaley, Lindsay J. (eds.) 1998. *Endangered languages: Current issues and future prospects*. Cambridge: Cambridge University Press

Grzech, Karolina & Ebany Dohle. 2018. Language Landscape: An innovative tool for documenting and analysing linguistic landscapes. *Lingue Linguaggi* 25, 65-80.

Hale, Ken, Michael Krauss, Lucille J. Watahomigie, J., Akira Y. Yamamoto, Colette Craig, LaVerne M. Jeanne & Nora C. England. 1992. Endangered languages. *Language* 68(1):1–42.

Henke, Ryan & Andrea L. Berez-Kroeker. 2016. A Brief History of Archiving in Language Documentation, with an Annotated Bibliography. *Language Documentation and Conservation* 10, 411-457.

Nettle, Daniel, & Suzanne Romaine. 2000. *Vanishing voices: The extinction of the world's languages*. Oxford: Oxford University Press

Ritchie, Sandy, Samantha Goodchild & Ebany Dohle. 2016. Language Landscape: Supporting community-led language documentation. In Vera Ferreira & Peter Bouda (eds.). *Language Documentation and Conservation in Europe*, 121-132. Honolulu: University of Hawai'i Press.

Robins, Robert H. & Eugenius M. Uhlenbeck, (eds.) 1991. *Endangered languages*. Oxford: Berg.

Seifart, Frank, Nicholas Evans, Harald Hammarstrom & Stephen C. Levinson. 2018. Language documentation twenty-five years on. *Language* 94(4), e324-e345.

UNESCO Ad Hoc Expert Group on Endangered Languages. 2003. Language Vitality and Endangerment Document submitted to the International Expert Meeting on UNESCO Programme Safeguarding of Endangered Languages Paris, 10–12 March 2003. http://www.unesco.org/culture/ich/doc/ src/00120-EN.pdf.

Wasson, Christina, Gary Holton and Heather S. Roth. 2016. Bringing UserCentered Design to the Field of Language Archives. *Language Documentation and Conservation* 10, 641-681.

Williams, Nicholas, W. D. L. Silva, Laura McPherson & Jeff Good. 2021. COVID-19 and documentary linguistics: Some ways forward. *Language Documentation and Description* 20, 359-377.

Woodbury, Anthony C. 2003. Defining documentary linguistics. *Language Documentation and Description* 1, 35-51.

Woodbury, Anthony C. 2011. Language documentation. In Peter K. Austin & Julia Sallabank (eds.) *The Cambridge Handbook of Endangered Languages*, 159-186. Cambridge: Cambridge University Press