International Journal of Applied Research 2015; 1(5): 224-229

ISSN Print: 2394-7500
ISSN Online: 2394-5869
IJAR 2015; 1(5): 224-229
www.allresearchjournal.com
Received: 28-03-2015
Accepted: 29-04-2015

Pankaj Dwivedi
Department of Computational Linguistics Mahatma Gandhi Antarrashtriya Hindi Vishwavidyalaya, Wardha, India

Correspondence:
Pankaj Dwivedi
Department of Computational Linguistics Mahatma Gandhi Antarrashtriya Hindi Vishwavidyalaya, Wardha, India

Technology, language documentation and meta-documentation

Pankaj Dwivedi

Abstract
With the revolution in the area of information technologies, it is now possible to maintain organized and long-lasting linguistic and cultural records. Software and different technical linguistic tools play a vital role in producing audio and video recordings, annotations, translation distribution and different sort of metadata of linguistic material. However, technology changes rapidly and an uncritical adoption of new tools and technologies might compromise with long-term sustainability, portability, usability and compatibility with other platforms (Bird & Simons, 2003) [4]. This paper, hence, concerns itself with the role of technology in language documentation and its possible pros and cons.

Keywords: language documentation, documentary linguistics, language endangerment, meta-documentation, meta-data, linguistic tools

1. Introduction
There are around seven thousand languages spoken across the world; and that half of these may no longer to continue to exist after a few more generations as they are not being learnt by children as first languages (Austin & Sallabank, 2011) [2]. Approximately 96% of the world's languages are spoken by less than 4% of its population, implying that 96% of the world's population speaks only 4% of the world's languages (Allwood, 2006) [1]. What can be inferred from these statements is that more than half of the world languages are on the verge of extinction. Crystal (2002) [5] claims (as cited in Allwood, 2006) [1] that the rate of language disappearance is as high as two languages each month. The people who are poor in power, number, money, and other resources usually speak these languages. There may be various sociopolitical, economic, technological and military arguments in favor of and against this situation, that is, whether linguistic diversity is useful or not, but the fact remains that the arguments don't change the situation and something must be done in order to document, if not to preserve, these languages. Therefore, in the generation when the rate of language death is at its peak, if we choose to use moribund technologies to create and preserve language data, when technologies die, unique heritage is also lost or encrypted (Bird & Simons, 2003) [4].
Till early 20th century, language documentation meant the collection of data on paper and therefore each documentation project resulted into large bodies of texts, that is, text corpora. Task such as dictation, transcription translations, elicitation and analysis were all done only on papers. Linguists and other language researchers worked painstakingly to collect and preserve these collections of texts. However, later on, technology entered into the scene and revamped the entire process all through. It provided the documentary linguists with the useful tools and resources leading to better organization and long-term preservation of the language data apart from significantly reducing the manual efforts. It became possible to store the text data in electronic form using floppy disks, CD ROMs, Cassettes, etc. Bird and Simons (2003) [4] write that the process of documenting the world's language is undergoing radical transformation with the rapid uptake of new digital technologies for capture, storage, annotation and dissemination. Nathan and Austin (2004) [11] say that language documentation embraces information and communication technologies to create digital sound and video recordings and integrate them with text and other explanatory or analytical material.
The term 'language documentation' concerns itself with principles and methods used for the recording and analysis of primary language and cultural materials, and metadata about them, in ways that are transparent and accountable, and that can be archived and disseminated for current and future generations to use (Austin, 2013) [3]. Language documentation and language

description differs from each other in that language description concerns itself mainly with the production of the dictionaries and grammars whose primary audience are linguists while language documentation primarily focuses on collection of the data, its representation and diffusion. Today, a state-of-the-art documentation is expected to include multimedia representation of a range of aspects of community life and to present the speakers' attitude towards and reflections on the language (Himmelmann, 1998) [9]. Language documentation attracts audience from a wide range of areas such as anthropology, geography, sociology, education and population studies.

A state-of-the-art language documentation actively engages members of the speech community whose language is being documented. Fieldwork makes an essential component of the language documentation. However, this fieldwork may differ in its depth and coverage depending on the goal and scope of the of the term language documentation, which usually is decided (goal and scope of the language documentation) by the linguist(s) performing the documentation. Grinevald (2003) [7] presents (as cited in Nathan & Csató, 2006) [12] a set of approaches to fieldworks, or "frameworks" that have successively evolved over time:

- fieldwork ***on*** a language: fieldwork for purely academic purposes
- fieldwork ***for*** the language community: fieldwork focuses on advocacy of the community
- fieldwork ***with*** speakers of the language community: recognizes control of the community members over the fieldwork and makes the provision of training them
- fieldwork ***by*** speakers of the language community: community partnership in linguistic fieldwork and aims to develop the products which help language maintenance and revitalization.

Today's language documentation mainly concerns itself with either "***with***" and "***by***" framework and therefore generally technological optimization and emphasis is required on certain processes and stages of documentation, not all. Good (2002) describes "one of the most important uses of metadata is to locate resource". So, the optimization and emphasis on one part does not imply towards compromise with others. One loose point might make the whole documentation project useless and therefore due cautions are required in all the stages.

2. Stages in language documentation:

Languages documentation begins with a project to work with a speech community on a language and goes through a series of stages. Austin (2006) [12] lists following five main stages in the process of language documentation. These processes make essential part of any language documentation project.

- Recording: audio, video or text
- Capture: moving analogue material to the digital domain
- Analysis: transcription, translation, annotation and metadata notation
- Archiving: creating archival objects, assigning access and usage rights
- Mobilization: publication and distribution of material in various format

2.1 Recording

A good documentation corpus, apart from texts, includes audio and/or video materials, ideally recorded in authentic settings and under good conditions (Austin, 2006) [12]. Valuable suggestions for making good audio and video recordings can be found o in textbooks such as Ladefoged, 2003 [10] and websites dedicated to language documentation, such as http://www.hrelp.org/documentation/whatisit/#8.

Though audio and video have made the documentation tasks easier, richer and more productive, their careless and uncritical adoption can make the job more complicated and even result in total loss of work.

2.1.1 Audio Recording

Audio recordings of the data can be a very effective way if the purpose of the documentation project includes working out phonetics and phonology of a language. In comparison to sound files, text files (i.e., written data) prove to be rather inadequate for the purpose. Description of sounds and sound, patterns on text files by a linguist often falls prey to linguist's own subjectivity, whereas, original audio recordings of the data can be verified in the laboratory or anywhere in the world and by anyone who is interested and therefore more suitable. Like the text files, sound files are easy to work with and a wide range of supportable software and devices are available for their recording, editing, presentation and storage. Original sounds files of a language can be of great help in determining the change in language, as spoken form a language changes sooner than the written form. Technology has already paved way for easier and longer storage of sound files.

Some good quality recorders that currently are used by linguists are ZOOM, OLYMPUS, Marantz, Fostex, Tascam, SONY, etc. It is usually recommended to use a good quality digital recorder to create sound files in WAV format and suggested resolution for audio recording for 48 KHz, 24 or 44.1 KHz, 16 bit. However, before making a selection, functionality of recorder and interportality across different platforms of recording formats must be clearly understood. Choice of the audio recorder, audio format, sampling rate, etc. depends on one's research purpose.

The following figure shows compression of a 3 minutes audio file in different audio formats:

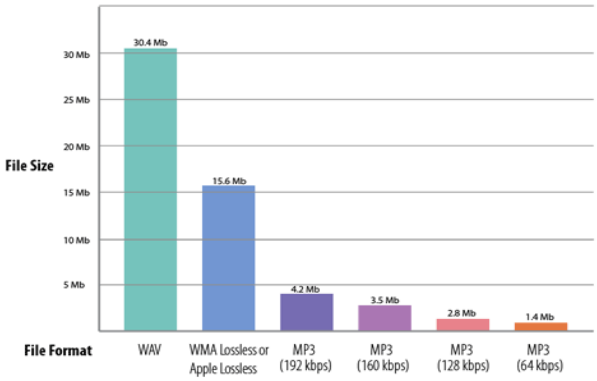

Fig 1: Comparison of a 3 minutes audio in different formats (*Source: http://www.crutchfield.com*)

Figure 1 shows while on the side file in WAV format remains least compressed, MP3 (64 KB) file is compressed to the maximum. WMA format file that stands second after WAV format in terms of compression could save merely half of what may be called recorded stuff. Figure 2 presents a general comparison among the most popular versions of audio recording formats across different parameters such as portability, propriety, cost, compression rate and sampling rates.

Table 1: Comparison of different audio formats across various parameters

Format	Encodercost	Player cost	Proprietary Implementations	Lossless Compression	Patented	Multi Channel	Sample rate
AAC	Non-free	Non-free	FAAC (encoding only),iTunes, Nero Digital Audio	No	Yes	48	8 kHz to 192 kHz
AMR-WB+	Non-free	Non-free	?	No	Yes	No	8, 11.025, 16, 22.05, 32, 44.1, 48 kHz
FLAC	Free	Free		Yes	No	8	1 Hz to 655350 Hz
MP3	Non-free	Non-free	FhG, l3enc, MP3enc, (old implementations:Xing TOMPG, SCMPX)	No	Yes	No	8, 11.025, 12, 16, 22.05, 24, 32, 44.1, 48 kHz
Vorbis (Ogg)	Free	Free		No	No	255	1 Hz to 200 kHz
WAV	Free	Free		Yes	No	256	1 Hz to 16.777216 MHz
WMA	Free for Windows OS only	Free for Windows OS only	Windows Media Player,Windows Media Encoder	Yes	Yes	6	8, 11.025, 16, 22.05, 32, 44.1, 48, 88.2, 96 kHz

2.1.2 Video Recording

Video can capture the multimodal nature of language use (with context) and allows for a far richer documentary record than that of audio. Since video can capture linguistic, paralinguistic and contextual information, it is often easier to perform the tasks such as transcription, translation and annotation. Video recordings also make the scope for interdisciplinary research such as gestural analysis and experimental pragmatics. Video recordings make a good option in the process of product development. One of the most interesting features of video recordings is that it is often particular interest to the endangered language community and can be produced by community members themselves without much assistance from researchers (Nathan and Austin, 2011) [2].

Table 2A: Comparison of various Digital Video Codecs

Codec	License	Patented compression formats	Compression method
libtheora (Theora)	BSD-style	Patented, but freely licensed	lossy
x264	GNU GPL	MPEG-4 AVC/H.264	lossy/lossless
Xvid	GNU GPL	MPEG-4 ASP	lossy
FFmpeg (libavcodec)	GNU LGPL	MPEG-1, MPEG-2, MPEG-4 ASP, H.261, H.263, VC-3, WMV7, WMV8, MJPEG, MS-MPEG-4v3, DV, Sorenson codec etc.	lossy/lossless
DivX	Proprietary	MPEG-4 ASP, H.264	lossy
Windows Media Encode	Proprietary	WMV, VC-1, (in early versions MPEG-4 Part 2 and not MPEG-4 compliant MPEG-4v3, MPEG-4v2)	lossy

Era of DV tapes has already passed for video recordings. Most of the linguists working on documentation now a days use high resolution cameras with built in or removable flash memory cards. These cameras usually provide good recording in variety of formats and codec. While choosing a camera two points, i.e., '**Digital Container Format'** and **'Video Codec**' must be kept in account, because these two things may hugely affect inter portability across different platforms and long-term preservation of a video, besides the quality. Webopedia.com refers the Container Format as "a type of file format that contains various types of data compressed by standardized codecs". Digital Container Format refers to a metafile format that makes it possible for the different data elements and metadata coexist in a single computer file, while codec video is a device or software that enables direct compression or decompression of a digital video. Webopedia.com further says that the "…container formats are essentially wrappers in that they don't specify what codec the container format uses, but rather it defines how the video, audio and other data is stored within the container."

Below is a list of the most popular Digital Video Codec **(Table 2A)** and the Container Formats **(Table 2B)** with their technical specifications.

Table 2B: Comparison of different Digital Container Formats

Format	License	Video formatssupported	Audio formatssupported
AVI	?	Almost anything through VFW	Almost anything through ACM; Vorbisis problematic
FLV (Flash Video)	Patent encumbered	Sorenson, VP6, Screen Video,H.264/MPEG-4 AVC	MP3, Nellymoser,ADPCM, Linear PCM,AAC, Speex
Matroska(.mkv, .mk3d, .mka, .mks)	Freely licensed	Virtually anything	Virtually anything
MP4	Patent encumbered	MPEG-2 Part 2, MPEG-4 ASP, H.264/MPEG-4 AVC, H.263, VC-1, Dirac, others	MPEG-2/4 (HE)-AAC, MPEG-1/2 Layers I, II, III (MP3), AC-3, Apple Lossless, ALS, SLS, Vorbis(with private objectTypeIndication), others
Ogg	Patent-free	Theora, Dirac,OggUVS, MNGand others, and almost anything else via VFW	Vorbis, FLAC, Speex, Opus,OggPCM, and almost anything else via ACM
WebM	perpetual license grant to patents	VP8	Vorbis

Apart from choosing 'right container and codec, one must choose a good quality external microphone. A good choice is an HD camera recording in MPEG-4/H.264/AVCHD to its built-in flash memory (Nathan and Austin, 2011) [2]. Some brands that make better quality camera are **Nikon, Canon** and **SONY**. However, like audio recorder, here too, before making a selection, functionality of a video recorder and interportality of recording formats must be understood.

3. Text Creation

Field notes play a very important role in the process of language documentation. Whereas taking field notes used to be tricky and cumbersome, the task has become far easier, but far trickier too – explained later, with the computers and laptop becoming handy and to the financial reach of the people in general. While earlier linguists used to be worried for the pages of the field notebook being lost or torn, or becoming illegible due to moisture, now one might lose the whole data because hard disk or software crash. ***There is many a slip between the cup and lip!*** Whatever the case is, there are several advantages of the digital text over traditional notes: digital text is compact, easy to store, access and index and can express hyper textual relationships and large number of tools are available to process digital text data (Austin, 2006). Though most of the text editors and word processors are user-

friendly and remain up-to-date, using without understanding them can lead to several problems however. For example, one must recheck the spellchecker and new word addition function properly in word-processing programs. Choosing right type of text editor helps in proper **Encoding** of the characters/fonts and file format. Some text editor might not be portable and compatible to the other platforms and software respectively.

- **Character Encoding:** Unicode, Windows/ANSI, Big5, Latin 5, etc.
- **Data Encoding:** XML, SGML, MSWord, etc.
- **File Encoding**: plain-text, PDF, MSWord, etc.

3.1.1 Character Encoding

Character encoding is the process of representing information in computers. Computer encodes written language by means of binary numbers. A character set encoding is a system of representing the grapheme or any text element of a written language. It contains at least these two components: a set of characters and some system for representing these graphical units for the purpose processing. There are many a character set standards in practice now a days. Some of them are Unicode, Windows/ANSI, Big5, Latin 5, etc. However, Unicode is of the most popular Character Encoding Systems. Unicode 7.0 (released in on July 16, 2014) in its present form consists of a total of 2,834 characters and almost every script found on earth (120). Apart from characters *Numbers, General Diacritics, General Punctuation, General Symbols, Mathematical Symbols (Western and Arabic), Musical Symbols (Western, Byzantine, and Ancient Greek), Technical Symbols, Emoji, Dingbats, Arrows, Blocks, Box Drawing Forms, and Geometric Shapes, Game Symbols, Miscellaneous Symbols, Presentation Forms, Braille Patterns and Kangxi Radicals.* More information can be found on http://www.unicode.org.

3.1.2 Markup/Data encoding

Markup or data encoding refers to the way of annotating a document so that it is syntactically distinguishable from the text. Good Markup platform helps us write unconstrained text as well as to have full information retrieval from the text. It makes the process of text processing far easier. Like logical languages markup, languages indirectly instruct computer with the help of predefined tags. Markup languages provide the documents with Descriptive Structural, Presentational and referential properties. Some much appreciated markup languages are XML, SGML, GML, HTML and MSWord.

3.1.3 File Encoding

Text files are encoded differently depending on the platform and the locale. For a common user, working with one type of script/ characters, default text file encoding fits well. However, a linguist works with a wide range of characters at the same time on the same file and this work/text/file is likely to be used by other linguists for further analysis in future. Hence, saving file in default encoding setting may come up with unforeseen results like unreadable fonts, loss of information, etc. It is advisable, therefore, for linguists to use (and finally save) the editors that are Unicode supported and are in line with OCR. Proper encoding of the files makes the work portable and sustainable, and reaches out to the others. Below is given the comparison of various text editors for their features that are usually sought by linguists.

Table 3: Comparison of various text editors

	Features	Text Editors						
		gedit	GNU Emacs	Kate	Notepad	Notepad++	TextEdit	Vim
	Cost	Free	Free	Free	Windows	Free	Mac	Free
	Open source	Yes	Yes	Yes	No	Yes	Yes (BSD)	Yes
	OS support	W,M,L,U	W,M,L,U	W (partial),M,L,U	W		M	W,M,L,U
Features	Spell Checking	Yes	Yes (plugin)	Yes	No	Yes (plugin)	Yes	Yes
	Large file support	No	Yes (64-bit OS)	No	No	No	?	Yes
	Regex compatible	Yes(plugin)	Yes	Yes	No	Yes	No	Yes
	Multiple undo/redo	Yes	Yes	Yes	No	Yes	Yes	Yes
	Protocol support	FTP, HTTP, SSH	FTP, HTTP, SSH	FTP, HTTP, SSH	No	FTP (plugin)	No	FTP, HTTP, SSH (plugin)
Language support	Encoding conversion	Yes	Yes	Yes	No	Yes	Yes	Yes
	UI-langauge	82	1	50+18	11	43	18	25
	Character Encodings	ASCII, ISO-8859, UTF-8, UTF-16	ASCII, ISO-8859, UTF-8, UTF-16	ASCII, ISO-8859, UTF-8, UTF-16	ASCII, ISO-8859, UTF-8, UTF-16	ASCII, ISO-8859, UTF-8, UTF-16	ASCII, ISO-8859, UTF-8, UTF-	ASCII, ISO-8859, UTF-8, UTF-16
	RTL-Bidi	Yes-Yes	Yes-Yes	Yes-Yes	Yes-Yes	Yes-No	Yes-Yes	Yes-No
Programming Features	Compiler Integration	Yes	Yes	Yes (plugin)	No	Yes	No	Yes
	Syntax Highlighting	Yes	Yes	Yes	No	Yes	No	Yes
	Auto completion	Yes (plugin)	Yes	Yes	No	Yes	No	Yes
	Auto indentation	Yes	Yes	Yes	No	Yes	No	Yes

4. Language Documentation and Social Media Technology

In the present setting importance of social media cannot be ignored. Social networking websites such as Facebook, Orkut, Blogs, Twitter and MySpace have become a highly popular place of interaction/ communication and sharing updated information. Popularity of these platforms can be estimated from the fact that Facebook alone boasts of 1.15 billion active users on March 2013 (source: http://investor.fb.com). This networking website can become good platforms for revitalization of endangered languages as they can connect people from different geographic areas to connect and communicate with one another. These platforms can be used

for paasing your endangered linguistic skills to younger generations. Scannell (2012) [16] says "Social media are particularly well-suited to minority language revitalization. When speaker populations are geographically scattered, or located in remote areas, or when there are large diaspora populations, social media can allow people to connect with one another in way that was impossible two to five years ago". Several studies have investigated the effectiveness of social networks in maintaining language varieties in indigenous communities (Sallabank, 2010) [14]. Hildebrandt (2004) [8] (as cited in Sallabank, 2010) [14] have found that even a small community of speakers might promote the retention of conservative linguistic features as long as access to the native language in that community is regular and unrestricted.

One very good example of language revitalization through social media is ***Indigenous Tweets*** project headed by Prof. Kevin Scannell, which searches everyone using Twitter in an indigenous language around the world. As of November 2012, Project covered around 139 minority European languages and around 46 thousand users (Scannell, 2012) [16]. Another project (by Prof. Kevin Scannell) '**Facebook in your language**' a part/extension of '***Indigenous Tweets'*** aims at encouraging indigenous language groups to use their languages in social media and there they produce translations of Facebook's interface (the menus, navigation, etc.) into as many languages as possible (Scannell, 2012) [16].

One truth of the social media platforms is that they are all subject to commercial gains and not for promoting endangered language or cultures and therefore dominating languages such as Chinese, English, Spanish, French and Hindi are used for communication. In past at times software giant's like Facebook, Google, Yahoo, Microsoft had come to provide translations services in as many as 175 languages, but had to step back due to little or no financial gains. Users on social networking platforms often hugely 'code-mix' and 'code-switch' the languages, which are even affecting the widely spoken languages. Therefore, it may be the variety that is spoken by the few hundred speakers can take so different shape in a short period of time so that it cannot be a match for the original variety.

5. Developing open source tools and free software in endangered language

With the spread of education and technological comfort among the common masses, one way of saving the endangered or lesser known languages can be development of open source software and tools in the those languages – both as a user interface and technical documentation. In past there have been these kinds of efforts with propriety software too. However, most of them met with unfortunate results for various socio-political or financial reasons. For example, 1) Efforts of translating Apple Macintosh operating system into Irish in 1990s had to be stopped for lack of due financial backing. 2) Under the pressure of the Chinese government, Microsoft Corporation dropped the planning of translating Windows into Dzongkha, the national language of Bhutan. It had happened despite of the fact that Govt. of Bhutan made investment as high as half a million dollar for the purpose (cf. Scannell, 2008) [15].

Free software and tools provide an advantage of freedom from commercial venders and therefore in the access of people at large. Software and tools in the endangered language also promote what may be referred as a sense of "community ownership". Freely available software in endangered language can also help bridge the gap between community members and rest of the world. Following is a list of the popular software and tools that are license free and have been extensively used by linguists and other people engaged in the language documentation process.

Table 4: Popular tools used in language documentation

Tools and Software	Names
Audio Annotation	PRAAT, ELAN, Annotation Pro, AVATecH, WaveSurfer
Vedio annotation	Anvil, ELAN, VATIC, MobilVOX, Annotation Pro, VideoANT, Oxdb, AAV, AVATecH etc.
Lexicon Annotation	FLEX, Shoebox, Toolbox
Discousre Annotation	Transcriber, EXMAraLDa
Transcription	IPA keyboard, G2P Converters, Transcriber
Translation	XLIFF editor
Referenceing	Endnote, Reference Manager, LateX

6. Conclusion

Language documentation, especially for the weak ones, is the dire necessity at the present time. When technology has become too deep rooted into every field of inquiry, the area of language documentation is no exception to it. Technology exerts both negative and positive influences. A careless adoption of technology may lead to unforeseen and regrettable consequences. While its thoughtful choice can make the process of language documentation far easier and sustainable, an inconsiderate adaptation or usage can put every inch of work into serious trouble on the other hand. Uses of standard open source software, tools, archives, etc. are highly recommended in the language documentation process. Creating online repositories and social media platform can be an effective way of reversing the language shift and help the process of language maintenance and revitalization.

7. References

1. Allwood, Jens. Language survival kits. Trends in linguistics studies and monographs 2006; 175:279.
2. Austin, Peter K, Julia Sallabank, eds. The Cambridge handbook of endangered languages. Cambridge University Press, 2011.
3. Austin, Peter K. Language documentation and meta-documentation. In Keeping Languages Alive: Documentation, Pedagogy and Revitalization ed. Sarah Ogilvie and Mari Jones Cambridge: Cambridge University Press, 2013, 3-15.
4. Bird, Steven, Gary Simons. Seven dimensions of portability for language documentation and description. Language 2003, 557-582.
5. Crystal, David. Language death. Cambridge University Press, 2002.
6. Good, Jeff. A gentle introduction to metadata, 2010.
7. Grinevald, Colette. Speakers and documentation of endangered languages. Language documentation and description 2003; 1:52-72.
8. Hildebrandt, Kristine A. Manage tone: Scenarios of retention and loss in two communities. PhD dissertation. University of Michigan, 2004
9. Himmelmann, Nikolaus P. Documentary and descriptive linguistics, 1998, 161-196.
10. Ladefoged, Peter. Phonetic data analysis: An introduction to fieldwork and instrumental techniques. Wiley-Blackwell, 2003.

11. Nathan, David, and Peter K. Austin. Reconceiving metadata: language documentation through thick and thin. Language documentation and description 2004; 2:179-187.
12. Nathan, David, Éva Á. Csató. "Multimedia: a community-oriented information and communication technology. Trends in linguistics studies and monographs 2006; 175:257.
13. Nathan, David. What is language documentation? 2014. Web. 4 Nov, 2014 <http://www.hrelp.org/documentation/whatisit/>.
14. Sallabank, Julia. The role of social networks in endangered language maintenance and revitalization: The case of Guernesiais in the Channel Islands. Anthropological Linguistics 2010; 52(2):184-205.
15. Scannell, Kevin P. Free software for indigenous Languages. Native Language Network, 2008. Web. 1 March, 2015. <http://borel.slu.edu/pub/ili.pdf/>.
16. Scannell, Kevin. Translating Facebook into Endangered Languages. Proceedings of the 16th Foundation for Endangered Languages Conference 2012.
17. Scannell, Kevin P. Indigenous Tweets, visible voices, and technology, panel discussion at SXSW '13, Austin, Texas, with Kara Andrade, Maite Goñi, and Peter Rohloff, 9 March, 2013. Web. March 3, 2015. <http://borel.slu.edu/pub/SXSW.pdf/>.l