# Under-Researched Languages: Phonetic Results from Language Archives

D. H. Whalen
Joyce McDonough

**1942**

Language archives are becoming increasingly available, allowing new possibilities for phonetic analysis, especially for under-researched and endangered languages. A large number of the world's languages lack adequate phonetic documentation and depend on impressionistic descriptions of their sound patterns. Many of the world's most typologically unusual contrasts or phonological patterns have been reported in languages with small numbers of speakers that consequently have not received attention from academic researchers. To gain a greater understanding of the phonetic patterns present in human languages and to better inform phonetic and phonological theory, data from under-resourced and endangered languages and language communities is crucial. Archives allow the efficient use of resources by making acoustic records available, but the decisions made in the creation and curation of archives affect what analyses are possible, based on the level of detail in the analysis accompanying the sound files. In this chapter, we describe some currently existing language archives, focusing on those that contain data primarily from endangered languages. We summarize a selection of the papers that have performed phonetic analysis on data from those archives and close with some recommendations for archives that would make phonetic analysis of archived material more feasible and that would also provide greater reward for depositors of this valuable material.

## Introduction

The upsurge in documentation of languages, especially endangered ones, that has occurred in the past decades has produced a large amount of potential data for linguistic research, including phonetic research. The increasingly common inclusion of spoken material in archived audio and video recordings potentially provides the means to document the sound and gesture patterns and phonetic properties of spoken and signed language, aspects of language not recoverable from written text. The utility of these recordings depends on several factors. First of all, the data need to be archived and accessible to a broad group of researchers, beyond those who originally collected the material. Second, methodologies and research questions need to be continually developed to use data that have not been collected specifically for phonetic research. Third, the needs of the various stakeholders—academics, community members, archivists and others—must be kept in balance. In this chapter, we will address the use of existing language archives of under-resourced or small language communities (henceforth "small languages") and the implications of this data for phonetic research.

At present, hundreds of languages are falling silent (Hale et al., 1992); indeed, entire language families are at risk (Whalen & Simons, 2012). Some languages can be, and have been, revived from documentary descriptions (Amery, 1995; Hinton, 2001; Leonard, 2007; Littledoe, 1998), but from a phonetician's point of view, phonetic detail is likely lost in the process. With an acoustic archive, however, patterns of sounds that link contemporary speakers to past generations can be recovered in revitalization efforts, if methodological phonetic analyses are performed.

In typical experimental phonetics, laboratory sessions are organized to elicit specific types of data related to research topics in phonetic description and theory. Experimental phonetics emerged at the end of the 19th century (Rousselot, 1897-1908), but limitations of equipment kept the amount of data small. Nonetheless, researchers have taken experimental methods into the field to collect phonetic data (Whalen & McDonough, 2015). In recent decades, the Phonetics Laboratory at UCLA under Ladefoged and Maddieson produced an important body of knowledge on the phonetics of speech sounds in the languages throughout the world, working with academic linguists and community members, and using carefully constructed word lists and other materials in their fieldwork, to exemplify particular sound contrasts under investigation (archived at the UCLA Phonetics Lab Archive). The researchers who have contributed to this archive also developed field methods for collecting speech production data and performing speech perception experiments. Knowledge about sound systems in human language, the structure of the vowel space and consonant contrast patterns in phonemic inventories, came out of this work (Becker-Kristal, 2010; Ladefoged & Maddieson, 1996b; Lindblom, 1990; Maddieson, 1984; Schwartz, Boë, Vallée, & Abry, 1997). However, despite this work, and perhaps for historical reasons, phonetic documentation has not been fully incorporated into language documentation practices. Often, phoneticians are not included as part of more broadly aimed language documentation projects. To some extent, this is due to the small size of most documentation teams. Nonetheless, phoneticians are becoming more involved as linguists become aware of the importance and function of fine-grained phonetic detail in linguistic structures and phoneticians become more adept at analyzing the spontaneous speech that constitutes a large proportion of current corpora.

All types of speech corpora require careful transcription, translation and annotation. The trend in recent language documentation efforts has emphasized the collection of “natural language,” that is, narrative and other types of non-elicited “natural” speech. While this task is a challenge for spoken word corpora from any language, it is especially problematic for small and under-resourced languages, which are generally the focus of language documentation efforts. These projects in particular may not have sufficient resources to implement the translations and annotations necessary to produce useful corpora. The task of annotation and transcription is well known as a time-intensive task, one that requires training. Collection of spoken language corpora continues in spite of these challenges. There are opportunities for using these resources to undertake phonetic analyses that will both increase our knowledge of possible sound patterns occurring in linguistic systems, and provide an important asset to documentation efforts in language communities.

Small language communities are less studied and present difficulties that are not encountered in research on larger and better documented language groups. First, small languages can be typologically distinct from each other and from better-studied groups, presenting challenges to underlying assumptions about the proposed linguistic structures drawn exclusively from more commonly studied groups. Second because these projects often involve comparatively well-funded researchers working in communities with fewer monetary resources, this work involves tackling persistent and recalcitrant political, social and ethical issues. Third, significant issues arise in the analysis of data from communities with small numbers of speakers. While working with a small number of speakers itself in not uncommon in phonetic research, it is generally done within the context of language associated with a broad knowledge base. However, work on small languages lacks this base, underlining the value and importance of both the work and the development of field methodologies, archives and annotation practices that allow access to that data. Because speakers of small languages represent a more significant portion of the community as a whole, their requirements for the design and use of the material need greater attention than those from large languages. Data that suffer from any of the many problems that elicitation entails—poor recording conditions, inadequate design, unexpected interpretations by the speakers (Niebuhr & Michaud, 2015)—can be remedied at a later date for large languages. This may not be possible for small ones.

Linguistic theories and generalizations are based on data from a relatively small number of the world's languages, often ones spoken by large populations. In general, the languages and dialects that are rapidly disappearing are those that we know little about, and they frequently belong to small or disenfranchised communities of speakers. They may also be undergoing rapid generational change. Furthermore, a large portion of available language data is in written form, understandably a poor basis for information about the sound patterns of a speech community. Without spoken data, our knowledge of the phonetic structures found in human languages–the phonetic patterns and the relationship between phonetic structure and other linguistic phenomena–is limited by our dependence on an incomplete and typologically narrow set of data. Finally, spoken language recordings are difficult to collect and to archive, and accessibility to this data is a challenging issue. Standards of accessibility, for instance, can vary greatly depending on several intransigent factors from academic and community restrictions to the availability of professional archiving facilities.

In this chapter, we will provide an overview of the issues for phonetic science that arise in the collection, annotation, availability and analysis of archival data along with a survey of published papers reflecting such analysis. Our paper focuses on what we call a secondary use of archival data, in which the analysis postdates the archiving. This excludes cases where the data collected is by a phonetician and subsequently archived (as is the case, for instance, with most of the studies of the UCLA archive). This decision focuses attention on the need to develop practices that use language archives for phonetic analyses, as well as other linguistic research. While such archiving is essential for the continued expansion of our phonetic knowledge, we will focus on the use of archives on small languages by phoneticians, the issues encountered, and the work that has been done using these databases.

Our recommendations and assessment of the survey will illustrate our belief that analysis of existing archives is urgently needed. (Throughout this chapter, we use the term "archives" to refer by extension to the contents of archives, since archives themselves may use various designations for their contents ("collection," "records", “material,” “documents”).) This use of the term also allows distinguishing archived materials from unarchived ones. There are two reasons supporting the immediate use of existing archives versus leaving the task to future users. First, although it is true that archives are intended to last indefinitely (and many print archives have lasted decades if not centuries), digital archives are inherently fragile, requiring more maintenance than print ones and a larger investment of time and money on the part of digital than print archives. (This issue is by no means limited to language documentation; see, e.g., Cottingham, 2008). In times of budget constraints, such expenditures are subject to examination in relation to the use of the archival material. The Digital Preservation Coalition, for example, suggests that “for items selected for permanent preservation it is anticipated that review and de-accessioning will occur in rare and strictly controlled circumstances” (Digital Preservation Coalition, 2015, Retention-and-review). Language archives may or may not fall into this category, depending on the institution and initial agreements. The DPC goes on to say: “For other collection levels such as mirrored or licensed resources review criteria may include: A sustained fall of usage to below acceptable levels.” It is easy to imagine language archives having low usage values. Further, if archives begin to address the current limited accessibility of some audio archives, the added costs could come into play (Clement, 2014). Digital archives are so new a phenomenon that the issues about their fragile stability across time are only beginning to be understood and addressed. Second, many aspects of the recording situation are known to the depositor(s) that do not enter into the metadata for the archive. Being able to ask questions of the depositors while they are still active is often crucial to correct interpretation of the phonetic data.

Phonetic work on archives can sometimes be done with written text, to the extent that some phonetic/ phonological aspects (subphonemic alternations, stress patterns) may be reconstructed from certain types of orthography and texts (such as poetry) but most analyses of phonetic patterns depend on acoustic recordings and associated transcriptions and annotations. Even imperfect transcriptions should be archived; some transcription is better than none As examples, determination of stress versus pitch accent and meter in Latin has been argued to be derivable from poetry (e.g., Pulgram, 1954) as has the "darkness" of Latin /l/ (Sen, 2015). When acoustic analysis is possible, such secondary evidence does not make a significant contribution to theoretical debates.

Many current linguistic documentation practices tend to focus on the collection of spoken narratives and conversations, rather than specific elicitation techniques, such as wordlist recitation to elicit data on phoneme contrasts, or the methodologies of higher-level prosodic elicitations, for instance. As Pike (1947) pointed out, all elicitation techniques bring in their own set of biases, arguably including carefully controlled experiments. One advantage in the collection of narrative is that permits more natural speech and discourse related events. Moreover narratives also allow for the likelihood of providing documentation of higher level prosodic phenomena such as utterance level prominence and intonation patterns (Bowern, McDonough, & Kelliher, 2012; Fletcher & Evans, 2002), as well as the effects of position-in-utterance on the realization of phonemic contrasts. With this in mind, information on the phonetics of phoneme inventories, including subphonemic details and cues to contrast, can be extracted from narratives (C. T. DiCanio, Nam, Amith, Whalen, & Castillo García, 2015), and including, by extension, archived materials.  By contrast, work in many small under-resourced communities that may lack a baseline description of their sound patterns, narrowly focused research questions or, for instance, perception experiments to test hypotheses that are performed in a lab setting, will need to have methodologies adapted to the knowledge available on a language (Remjisen's work on supresegmental prosody in Dinka and Shilluk) (Remijsen, 2013; Remijsen & Ayoker, 2014). But this lab-in-the-field type work has much to contribute to general feature theories (Clements & Ridouane, 2011; Whalen & McDonough, 2015).  With careful attention to the limitations of the data, a description of the main characteristics of phonetic realization of phonemic structure and prosodic patterns are possible from current archives and narratives, with the discussion and development of methods to do this work (see also Winter, 2015).

Phonetic description is ideally based on the productions of fluent speakers. Ideally, documentation would include both men and women, and at least six of each (Ladefoged & Maddieson, 1996a). (Of course, this does not cover all aspects of language, and sacred speech, child language, etc. would all require different populations.)  Recording a variety of speakers allows the possibility of studying the range of individual variation and provides more support for the commonalities across speaker.  For small languages, having data from a large number of speakers is often not a feasible goal. In communities where only a few speakers remain, or in which the remaining speakers no longer use the language, or are very elderly, or associate it with reduced status, the question of how many speakers to record is different and more subject to local conditions. The languages nonetheless remain examples of human language that deserve careful documentation.

Some statistical methods used on large corpora are not suitable for this latter type data, given that the number of tokens will not provide sufficient power. Instead, good descriptive statistics and a transparent presentation of methods and results can provide a solid basis for all types of further language work well into the future (Hargus, 2007; McDonough, 2003). The problem then is shifted to one of comparing the results and analyses done on large versus small language corpora and datasets, but this is hardly an issue for linguistic analysis alone. Comparisons with standard experimental phonetic work are also valuable. Archiving experimental results can serve to expand the knowledge base on a language (Whalen & McDonough, 2015) and increase our understanding of phonetic phenomena.

Acoustic data, though a powerful tool, is limited to what can be inferred about the speech production process. Articulatory data provide another source of evidence and thus can also play a role in language documentation and revitalization efforts, as well as adding to our knowledge of speech and sound patterns (e.g. Miller, 2016). Obtaining data from speech production is more challenging than collecting acoustic recordings, and it is less frequently performed, especially in the field. This scarcity of production evidence persists despite the fact that it was done very early on in the history of experimental phonetics. Pliny Earl Goddard (1905, 1907) was a pioneer in taking articulatory measurement devices into the field. He showed, for example, that Hupa uses a more anterior articulation for /d/ than English, exemplified by static palatography from a bilingual speaker (1905:614). Only sporadic studies were performed for the next several decades, with more consistent efforts being spearheaded by Peter Ladefoged and Ian Maddieson, mostly using static palatography and airflow. Currently, those techniques continue to be used (Anderson, 2008; Flemming, Ladefoged, & Thomason, 2008), but they have been augmented with electroglottography (EGG) (C. T. DiCanio, 2009, 2012; Kuang, 2013; McDonough & Tucker, 2012), electromagnetic articulometry (Wieling et al., 2016), and ultrasound (Gick, Bird, & Wilson, 2005; Miller, 2007; Whalen, DiCanio, & Shaw, 2011). The Pangloss Collection hosts EGG together with audio (see: http://lacito.vjf.cnrs.fr/pangloss/corpus/list_rsc_en.php?lg=Na&aff=Na) The 40 documents with EGG recordings in that language are marked with a special icon with 'EGG' written on it: EGG. AILLA has recently received some EGG data from DiCanio's work on Itunyoso Trique (ISO 639-3 trq) (C. T. DiCanio, 2012) (see, e.g., https://ailla.utexas.org/islandora/object/ailla%3A243997). The increased ease of use of these techniques makes it likely that ever-increasing amounts of articulatory data will become available. Current phonetic archives, however, are still primarily acoustic in nature, and so the remainder of this paper will focus on studies that make use of this modality.

While phoneticians are increasingly using large language corpora as a source of data for research questions to examine contemporary issues (for instance the finding of the uniform variability across speakers in realization of a phonetic property shared by multiple speech contrasts) (Chodroff & Wilson, 2017), the number of studies that make use of archival material from small languages is still small. As we have noted, the issues that arise in using small data are not unique to phonetics or linguistic research, nor is it intractable. The greater risk is not using the archives at all (Schmidt et al., 2006). We have underlined the importance of this type archived data in research; the issues that researchers encounter in using these databases will be discussed in the next section.

## Archives

Language archives have seen a surge in submissions in recent years as the practice of language documentation has evolved to include archiving collected materials, including digital materials, and their associated metadata. The US National Science Foundation now requires an archiving plan in its grant proposal submissions, for example. Large language archives such as the Linguistic Data Consortium provide corpora and allow access of various types (primarily through a payment schedule), but in small language archives such as the Endangered Language ARchive (ELAR), the Alaska Native Language Archive (ANLA) or the Archive of the Indigenous Languages of Latin America (AILLA), the relationship between the data, the researchers and the language communities is more personal with concomitant and mutual obligations. Different types of issues arise in both depositing these materials as well as in general access to these archives, which must be responsive to cultural restrictions and constraints. But archiving also implicates access privileges. Across digital archives, access to material is quite variable, but restrictions on access are an important aspect of small language corpora. Access has been handled by access protocols, such as the tiered protocol access instituted by AILLA at the University of Texas at Austin and this continues to be an important aspect of archives.

Not everything defined as a corpus resides in a digital archive, and we discuss only archival results here. An important criterion for being classified as an archive is being in a "trusted digital repository" (Research Libraries Group, 2002) which crucially aims to provide long-term curation for digital materials. Individual corpora on computers in laboratories, then, are not archives by definition. Curation of digital material in general is one of the biggest issues research libraries are now facing, and a dialogue between professional librarians and linguists is likely to be mutually beneficial in the continuing development of and access to these archives. In language documentation and description practices, such efforts as the OLAC standards (Bird & Simons, 2001), the GOLD ontology (Scott Farrar & Lewis, 2007; Scott Farrar, Lewis, & Langendoen, 2002), and META-SHARE OWL (McCrae et al., 2015) also work to standardize formats and metadata, but with variable results. Given the highly individual nature of the experience of field work, it is unlikely that all but the broadest and most generous standards will be successfully adopted, such as the definition of metadata and enabling (both with funding and training) fieldworkers to archive field materials in trusted repositories. Flexible standards, then, can still result in greater transparency than a complete lack of standards.

In this spirit, we list some of the current prominent archives below. Most of these archives are members of the Digital Endangered Languages and Musics

Archives Network (DELAMAN; http://www.delaman.org/members/), an international network of archives of data on linguistic and cultural diversity, in particular on small languages and cultures under pressure:

- ANLA, Alaskan Native Language Archive (www.uaf.edu/anla), at the University of Alaska, Fairbanks, focusing on Alaska Native languages.
- AILLA, Archive of Indigenous Languages of Latin America (http://www.ailla.utexas.org), at the University of Texas at Austin, covering Latin America.
- APS Digitial Library, American Philosophical Society (https://diglib.amphilsoc.org/audio/access), primarily focused on Native American languages of North America.
- ATM, Archives of Traditional Music (http://www.indiana.edu/~libarchm), University of Indiana, Bloomington, containing various language recordings.
- CLA, California Language Archive (http://cla.berkeley.edu), at the University of California, Berkeley, with material primarily from California languages but with other collections as well.
- DoBeS, Documentation of Endangered Languages at the Max Plank in Nijmegen (http://dobes.mpi.nl/), funded by Volkswagen Foundation; now part of The Language Archive.
- ELAR, the Endangered Languages Archive (https://elar.soas.ac.uk/), part of the library at SOAS University of London, primarily archiving results of

research funded by the Endangered Languages Documentation Programme (ELDP: http://www.eldp.net/).

- LDC, Linguistic Data Consortium (https://www.ldc.upenn.edu), a repository and distribution point for universities, companies and government research laboratories which creates, collects and distributes speech and text databases, lexicons, and other language resources.
- Pangloss (formerly Lacito Archive) (http://lacito.vjf.cnrs.fr/pangloss/), itself part of the broader repository Cocoon (https://cocoon.huma-num.fr/exist/crdo/) containing varying amounts of data for over 130 languages.
- Paradisec, Pacific and Regional Archives for Digital Sources in Endangered Languages (http://www.paradisec.org.au/), a consortium of University of Sydney, University of Melbourne and Australian National University.
- TLA, The Language Archive (https://corpus1.mpi.nl), containing material from MPI Nijmegen and other sources (including DoBeS).
- UCLA Phonetics Lab Archive (http://archive.phonetics.ucla.edu/), an archive of the recordings of languages from around the world primarily collected by the Ladefoged-Maddieson team and their students, maintained by the Department of Linguistics.

It should be noted that the Illustration of the IPA series in the Journal of the IPA typically have recordings associated with them. Those from 2001 to 3 years before the present are publicly available (https://www.cambridge.org/core/journals/journal-of-the-international-phonetic-association/illustrations-of-the-ipa-free-content). (Note that to access the audio files, you have to click on the title of the article and then click on the "Supplementary Material" tab near the top of the page.) Most if not all of the examples in the text have an accompanying sound file, and all Illustrations have a connected text. This is often the story "The North Wind and the Sun," but some use an alternate text, such as the frog story. These are valuable examples, but are too limited for full phonetic analysis.

Nearly all of these archives contain spoken language corpora, though the data formats, metadata and access limits vary. Since digital data, as opposed to print, is subject to obsolescence and thus quite fragile data migration and updating must be implemented as an ongoing enterprise (e.g., Beagrie, 2008). But beyond these issues common to all digital databases, there are specific issues that arise in language archives.

### *Annotation*

One of the most challenging issues that arises in spoken language corpora is the technical one of annotating speech. There are currently no satisfactorily accurate automated transcription systems that work to the level required by phoneticians and other linguists. The ideal of automating a usable transcription has been promoted since the beginning of the computer age, but it is not yet a reality. Strides have been made for major languages, including usable first pass text–to-speech alignment algorithms. Advances are likely to help with work on small languages, as exemplified by Adams et al. (2017). However, for the present, accurate transcription and annotation requires trained human annotators, a time-consuming enterprise and an expensive commodity.

An acoustic signal that has a corresponding transcription and annotation, therefore, is necessary for phonetic analysis, but the level at which the two are aligned determines the types of analyses that can be performed. If a corpus is large enough to allow for substantial phonetic analysis, it will be too large to analyze by hand. The acoustic signal needs to be aligned to phonetic elements, and thus the detail of the phonetic annotation needs to match the scale of the proposed analysis. Locating the segments is a minimum. Identification of phonetic components (such as dividing stops into a closure portion and aspiration after release) is useful. Allowing overlapping annotations, such as those needed for tones, can also be useful. Automatic procedures can help here, as already demonstrated (C. T. DiCanio et al., 2013; Kempton & Moore, 2014), and the continuing progress in the field indicates that the tools will continue to improve.

*Archive access and interpretation of transcription*

Several issues arise in using archives based on data from small language communities for phonetic research. Often the expertise needed and difficulty of addressing these issues in the development of database archives may be vastly underestimated and can thus reduce their usefulness as crucial resources. The two main issues at this writing are first, protocol and access to the data, and second, the practices used in the translation, transcription and annotation of the speech. First, the databases almost by necessity vary widely, primarily due to the differences in the types of documentation undertaken and data collection practices that accompany the documentation.  In the archives, the type of data collected, the practices used in transcription and annotation, the research goals of the project and funder requirements on archiving and access of the data all determine the shape of the resulting database.

The terms *transcription* and *translation* refer to the act of providing a representation of digital speech recordings in some orthographic system (including the IPA), and the translation of the speech into a commonly shared language such as English, Russian or French. The transcription itself may or may not be in the IPA, and is subject to the orthographic conventions of IPA or the language community, if they exist. In general, this aspect of the process must be done in close collaboration with a native fluent speaker of the language who can provide the translation and help to correct the produced transcriptions. The speech can sometimes differ quite radically from a native speaker's more proper understanding and transcription of the spoken word, especially in 'natural language' contexts such as conversations, lectures or storytelling, which can contain pauses, idiolectal pronunciations and expressions that are glossed over by the transcriber. Developing viable working relationships with language consultants and collaborators from small language communities is an important skill for fieldworkers; the role of these community consultants and advocates in developing these resources cannot be overestimated (Dwyer, 2006). This collaborative aspect of small language archive development is critical to small language archives, and is not as significant an issue in larger language archives  because of the greater resources available, though it may well arise in the transcription of non-standard dialects.

The term *annotation* refers to the addition of further analysis levels.  Of particular use for phonetics is the alignment of this orthographic transcription to the physical speech signal. In contemporary work, the annotation to be aligned generally uses the IPA for a reasonably accurate phonemic representation. Several types of software exist that serve as platforms for doing these annotations by hand; currently the best developed and most used are ELAN and Praat, the latter especially developed for phonetic research. In both these programs, the annotations are deliverable as text files with time codes easily imported into other programs. The annotation of the speech signal generally proceeds in two steps depending on the type of annotation being done; as previously mentioned, first is the transcription of the speech and its translation, this is followed by the alignment of the orthographic transcription to the speech signal.  This alignment allows the annotation of phrases and words, and, especially integral to phonetic research, ultimately the annotation of segmental and suprasegmental elements.  There are, of course, other important types of annotation of speech, such as the annotation of syntactic and semantic units, morphological and part-of-speech parsing, phrasal groups and sentence types.  These latter types require less fine grained attention to alignment than do phonetic investigations, but all require alignment of text to the speech signal, whether gross or fine.

The practice of aligning a symbol with its specific realization as a consonant or vowel sound in the acoustic signal involves some arbitrary decisions, as boundaries are made to be non-overlapping even though the segments themselves influence overlapping portions (e.g., Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). For example, aligning the text symbol associated with the phoneme /a/ to the beginning and end of the vocalic segment of the phone [a] is standard practice, but it ignores the shared influence of adjacent consonants during the formant transitions. Nonetheless, such an alignment allows a good estimate of the major influence on the signal and can further carry the alignment of a pitch peak and a tone. Some boundaries are more difficult than others, even acknowledging the partially arbitrary nature of the decision. In a series of sonorants, for instance, the boundaries between segments are not clear and an annotator's skills are required in the recognition, and reporting, of the patterns in the speech that aid useful segmentation; delimiting fundamental frequency (F0) is an even more elusive task, F0 being subject to multiple influences across large time spans.

Annotation, therefore, is an acquired skill informed by training and practice to recognize and understand the patterns in the acoustic speech signal. The annotation of speech careful enough for linguistic research requires training in phonetics because these patterns differ from language to language and among dialects and speakers; they are critical to natural language structures and are at the crux of the phonetics / phonology interface. The transcriptional elements chosen reify the phoneme/allophone patterns (Errington, 2007: 23), and they thus evidence little recognition of the ubiquitous speech variability that confounds speech recognition.

Conversely, annotation and segmentation by hand is famously labor intensive and time-consuming. Automatic transcription and alignment algorithms exist and can reduce some of the work but even the best automatic aligners are not accurate enough for some kinds of phonetic research, and require annotation by trained human annotators (but cf. Adda et al., 2016). Finally, even archives that are developed for other types of analyses not directly related to phonetic research (such as syntactic or morphological analysis) will need careful phonetic transcriptions. Speech is the primary vehicle of communication, and the intersection of prosody and syntax is an area much in need of research.

Another issue that arises in using these archives for phonetic research is the archive's approach to protocol and access. This is partly due to the wide range of services these databases provide, relating to their original purposes, funder requirements and their growth patterns. While there is an increasing trend toward allowing access, practice still varies widely. Restriction of access seems to derive from the practice of print archives, where a single copy of a text would exist at a specific location. Although electronic formats can be duplicated without loss, many archives restrict them as much as print archives. Those that do allow copying have imposed graded access for those objects that have restrictions (cf AILLA's statement: http://www.ailla.utexas.org/site/howto_use.html#access), but there are still restrictions on some collections without such restraints, primarily an agreement not to use the data for commercial purposes. A major funder, the US National Science Foundation (NSF) specifically requires data management plans for grantees "to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data created or gathered in the course of work under NSF grants" (https://www.nsf.gov/sbe/SBE_DataMgmtPlanPolicy.pdf). Although this is quite explicit, it contradicts another aspect of NSF policy: In North America especially, this policy can be in direct conflict with tribal requests to control access to their language data, which the NSF recognizes as intellectual property. Some datasets, such those in LDC, are created as intellectual property, and some of those are only available for an access fee, while others only require payment of a media fee. Currently, these conflicting imperatives are settled on a case-by-case basis, resulting in a continuing inconsistency in access even in new acquisitions and new archives.

The Endangered Language Documentation Programme (ELDP) at the SOAS University of London has a firmer policy of archiving and access, requiring annual archiving and open access for registered users of ELAR to the majority of data collected during the duration of the grant and this basic principle must be agreed upon at the time of the funding. Any access restrictions must be justified in the meta data and ELDP and ELAR support flexible solutions in case of later encountered access issues. Compliance failures with regard to depositing, namely not depositing the documentary materials collected with ELDP funding at all might result in barring future funding to both the grantee and the grantee's host institution.

Funding for language documentation by The German Volkswagen Foundation Documentation of Endangered Language (DoBeS) project, another important source of past funding for work on small language communities, required that the data collected during funding follow a set of standardized metadata categories and be digitally archived according to open standards. The DoBeS Archive, presently part of The Language Archive (TLA) at Max Planck Institute for Psycholinguistics in Nijmegen, continues to implement upgrades to metadata and data formats. Additionally, DoBeS has provided funding for research projects based on use of the data in the archive, though very little of it has been used in phonetic research.

Archives not directly tied to funding usually have explicit access policies as well. Housed at the University of Texas at Austin and supported by the University of Texas Libraries, AILLA focuses specifically on indigenous languages south of the Rio Grande and contains copies of digitalized indigenous language recordings, field notes, photographs, and other research materials. AILLA was developed to make language resources openly accessible with a focus on the speakers of the languages represented in the archive. AILLA has a 4-level access protocol system, designed to protect intellectual property and indigenous community rights.

The best known archive for phonetic research remains the UCLA Phonetics Lab Archive, the archive developed by Ladefoged and Maddieson It contains much of their material, collected on field trips funded by the NSF, as well as many recordings made by UCLA students and other researchers. It is presently maintained by the UCLA Linguistics Department. Its stated aim is to provide source materials for phonetic and phonological research, to illustrate phonetic structures found among human languages. This material may include phonetic transcriptions, significant because they were done by phoneticians. The archive's contents are all open access, under a Creative Commons license, for non-commercial use. Much of the phonetic work in existence done on endangered languages and small language communities is based on data found in this database. Many of these papers appeared in the *UCLA Working Papers in Phonetics*, and summaries appear in Ladefoged and Maddison's (1996b) *Sounds of the World's Languages*. For the present survey, we are only summarizing papers that were done subsequently to the depositing in the archive, while most of the articles just mentioned were done by those who collected, analyzed and then archived the data.

Most of the archives listed here were designed for general use, not for phonetics specifically, but they have proven to be quite useful in that regard. Acoustic recordings are the most common means of accessing the phonetics of languages, providing evidence for many features. Recordings of various languages began soon after Edison's cylinder recording became available. A. L. Kroeber and J. P. Harrington, for example, recorded dozens of Native American languages of California. Some of the language material was in songs, which are challenging for phonetic analysis, but stories and conversations, often spanning several cylinders, were also recorded. Recent advances in laser technology have allowed the recovery of previously unusable cylinders and many archives include recordings of both elicited and narrative speech. The following section will outline many of the studies that have been performed to date.

Many of the studies to be summarized here make use of both archival and newly obtained material. This is due both to the gaps that are bound to occur in the sampling of speech sounds in spontaneous speech and to the typical use of highly controlled contexts for phonetic analysis. Indeed, Ladefoged (2003) downplays the usefulness of recorded texts for phonetic documentation: "From a phonetician's point of view there is no point in making lengthy recordings of folk tales, or songs that people want to sing. Such recordings can seldom be used for an analysis of the major phonetic characteristics of a language, except in a qualitative way. You need sounds that have all been produced in the same way so that their features can be compared" (p. 9). However, recordings of texts are exactly what we have for many endangered languages. In an explicit comparison of elicited and narrative vowel spaces, DiCanio et al. (2015) (summarized more fully below) found that the formants of elicited speech were somewhat more extreme than those in texts, but that both showed good separation between the categories. But combining the two techniques is common.

### *Challenges in using acoustic archives.*

Progress has been made in the textual analysis of large corpora (e.g., Kübler & Zinsmeister, 2015), but progress in phonetics remains slow. Despite the existence of a very useful handbook on procedures for doing large-scale acoustic analysis (Harrington & Cassidy, 1999), implementation of the techniques relies on the usefulness of the transcriptional data. If a corpus is annotated in practical orthography (as mentioned before) at the word or sentence level, a great deal of parsing remains to be done before acoustic analysis can take place. A range of challenges thus face a researcher approaching such a corpus.

Orthographies vary widely in their phonological transparency, but fortunately, most recently devised orthographies are relatively shallow, allowing for fairly consistent grapheme to phoneme translation. One large challenge for even the most consistent orthography is allophonic variation. Underlying forms are often transcribed even though highly elided forms are the ones present in the speech signal. In narrative speech in Mixtec, for instance, voiceless stops (/p, t, k, kw/) are often produced as voiced and lenited variants ([b, d, ɣ, ɣw]) (C. T. DiCanio, Amith, Castillo García, & Lilley, 2016). The orthographic system utilizes only the former, underlying forms.

Transcription errors are bound to occur in any large corpus, and they can also lead to sizable misalignments. (Examining such misalignments can be of use as a check on the transcription itself, however.) How to treat such errors is a matter of researcher choice, and certainly the number of errors is an important factor; the smaller the percentage, the less likely errors will influence the results. Some issues that introduce misalignments are not technically errors in the transcription. Hesitation sounds are often left out of transcriptions, but generally are interpreted as speech by forced aligners. In elicitation tasks, the matrix language is often left untranscribed, which leads to unusable alignments if no corrections are made.

## Results from acoustic archives

This section presents summaries of the articles that we were able to find that analyze archives phonetically. We did not include analyses of material collected by the authors and then archived at a later date. Although this is the ideal approach to maintaining and sharing data, it is not our focus here. The list also excludes work on majority languages, of which there is a substantial amount. There are several works that make some use of the UCLA archive (e.g., Bradfield, 2014; Hsieh, 2007; Simas Frazão, 2013) or Pangloss (Schiering, Bickel, & Hildebrandt, 2010), but not to a large enough extent to count as phonetic studies. No doubt we have missed others, and we hope that many more will be added.

*UCLA: Blankenship (2002)*

Languages that contrast voice quality in their vowel systems have been understudied, in part because the acoustic measurement techniques are not terribly robust. Blankenship (2002) studied four languages from the UCLA Phonetics Lab Archive (Mazatec (ISO 639-3 code maj), Tagalog (ISO 639-3 code tgl), Chong (ISO 639-3 code cog), and Mpi (ISO 639-3 code mpi)) to evaluate not just the presence of laryngealization but the time course over the duration of the vocalic segment as well. Shifting from [ʔ] to [h] to modal could take place within 25 ms, and the various acoustic signatures of breathiness could occur singly or in one of many different combinations.

*ANLA: Tuttle (2003)*

The existence and discovery of archival material for the Salcha dialect of Tanana (ISO 639-3 code taa) made possible a phonetic comparison with the Minto dialect even though Salcha is no longer spoken (Tuttle, 2003). (The materials were in the ANLC collection (not ANLA at that time) and in the Oral History collection in the Alaska and Polar Regions Archives at the Elmer E. Rasmuson Library at the University of Alaska, Fairbanks.) Acoustic analysis revealed that Minto had a difference in fundamental frequency (F0) between "marked" syllables (syllables that are reconstructed as having constricted vowels in Proto-Athabaskan) while Salcha had no difference. Stress, on the other hand, was marked by F0 in Salcha but not in Minto. (Duration signaled stress in both dialects.) Stress and intonation interacted:

The frequent placement of verbs in sentence final position coupled with F0 marking of stress (which is prominent on verbs) led to a reduction (to a nonsignificant level) of the intonational drop in F0 at the end of sentences for Salcha but not Minto. Given the state of these dialects, only archival data could have shown these patterns.

*UCLA: Esposito (2010)*

Material from the UCLA Phonetics Lab Archive was used to test perception of voice quality distinctions in non-native speakers (Esposito, 2010) (see also Keating & Esposito, 2007). One experiment used 10 languages while a second used Mazatec (ISO 639-3 code maj). Of their three listener groups, Spanish- and English-speaking listeners were rather inconsistent, while Gujarati-speaking listeners were consistent but used three categories rather than the two used in the languages. This was confirmed with the larger sample from Mazatec, and it was further found that the best acoustic predictor of the Mazatec distinction was not used by any of the three listening groups.

*CLA: Yu (2008)*

Phonetic analysis revealed that quantity alternation in Washo (ISO 639-3 code was) is much more pervasive in the language than it was first described in the 1960s (Yu, 2008). Surprisingly, the current generation of Washo speakers retains subtle phonetic alternations, despite the fact they mostly grew up bilingual, if not English-dominant.

*UCLA: Keating, Esposito, Garellek, Khan and Kuang (2010)*

Contrastive phonation types of four languages, one from the UCLA Phonetics Lab Archive were examined (Keating, Esposito, Garellek, Khan, & Kuang, 2010); the other three were archived in “Production and Perception of Linguistic Voice Quality”, http://www.phonetics.ucla.edu/voiceproject/voice.html. When phonation categories with the same label (e.g. “breathy”) were compared across languages, each category was found to differ from language to language on multiple acoustic measures. For example, breathy in Hmong (ISO 639-3 code mww) is distinct from breathy in Gujarati (ISO 639-3 code guj). This unexpected result suggests that language/speaker differences in voice quality are larger than the phonation category differences themselves.

*UCLA: de Carvalho (2010)*

The vowel characteristics of Pirahã (ISO 639-3 myp) were examined in the recordings of the UCLA archive (de Carvalho, 2010). This language has a three-vowel system, which allows for a test of the vowel dispersion theory (Liljencrants & Lindblom, 1972). Surprisingly, the dispersion in this three-vowel system was as large as has been found for larger inventories, putting limits on the dispersion theory. Some puzzling data on intrinsic fundamental frequency (Whalen & Levitt, 1995) were also presented.

*UCLA: de Boer (2011)*

Thirty languages from the UCLA archive were examined to determine whether the first formant for /i/ and /u/ were the same or not (de Boer, 2011). Both are described as being equally high in the vowel space, leading to the expectation that they would have the same F1 value. However, /u/ consistently had a higher F1, indicating a lower tongue position. A possible explanation based on the capabilities of the vocal tract was proposed. There was a trend for a smaller difference in females than in males, possibly due to differences in the size of the epilaryngeal tube.

*UCLA: Garellek and Keating (2011)*

San Felipe Jalapa de Díaz (Jalapa) Mazatec (ISO 639-3 code maj) has an unusual three-way phonation contrast crossed with a three-way level tone contrast that is independent of phonation (Garellek & Keating, 2011). Acoustic analysis found that the tone categories differ in F0 over the course of the vowel, but that for laryngealized phonation, the beginning of the vowel loses the F0 aspect of the tone contrast. This study shows that the acoustics support the three-way phonation and tone contrasts, even though individual parameters may merge.

*UCLA: Coombs (2013)*

Downstep in Ibibio (ISO 639-3 code ibb) was explored using previously recorded material (Coombs, 2013). These recordings were first found to be alignable with a tool trained on English (C. T. DiCanio et al., 2013). Ibibio shows a complex interaction of prosodic downdrift, phonological downstep, and a possible raising of high tones for enhancing contrast with low tones.

*UCLA: Gordon (2015)*

Using data from four languages, three in the UCLA archive, Gordon (2015) found that the effect of laryngeal setting, posited to be a major source of tonogenesis (Hombert, Ohala, & Ewan, 1979), does not have a consistent effect on F0. The languages were Hupa (ISO 639-3 code hup), Western Apache (ISO 639-3 code apw), Pirahã (ISO 639-3 code myp), and Banawá (ISO 639-3 code bnh). Even with the limited sample of four languages, many interactions between laryngeal setting and segmental organization were found. Surprisingly, the effects that were found were not smaller in magnitude for languages that had distinctive tone already.

*ATM: Lawyer (2015)*

Patwin (ISO 639-3 code pwi), a member of the small and underdescribed Wintuan language family, had aspects of its phonetics outlined from archival sources (Lawyer, 2015). Measures voice onset time (VOT), vowel space and duration provide phonetic detail to the phonological description of the paper.

*ELAR: DiCanio and Whalen (2015)*

The vowel formants in an Arapaho (ISO 639-3 code arp) speech corpus were found to differ in an expected direction with distinctive vowel length, with short vowels being more centralized and long vowels more peripheral (C. T. DiCanio & Whalen, 2015). However, the effect of speech style (elicited vs. narrative) was asymmetrical, with long vowels undergoing greater durational compression in narrative speech than short vowels, but with short vowels showing greater changes in quality. This was an unusual finding of speaking style affecting articulation beyond the operation of durational undershoot.

*AILLA: DiCanio, Nam, Amith, Whalen and Castillo García (2015)*

Measurements of vowels in Yoloxóchitl Mixtec (ISO 639-3 code xty) (C. DiCanio, n.d.) indicated that vowel spaces can be recovered from narratives as well as elicited speech, even though there are differences in overall acoustic space between the two (C. T. DiCanio et al., 2015). Indeed, further separating the elicited tokens (by a median split) into short and long, and a three-way split of spontaneous tokens by duration showed that there was a fairly continuous expansion of the vowel space as duration increased, even for the elicited tokens. These findings have implications for the interpretation of measurements in majority languages as well.

## Discussion and Recommendations

The 13 papers cited have given us solid and often unexpected results, but this level of productivity is less than one might expect, given the great emphasis now put on archiving of linguistic data. We have already discussed some of the challenges to making use of archives, but we will outline those and others here along with recommendations for means of improving the use of these valuable resources. The first three would require resources that are not currently available, but we propose that they are worth considering for future initiatives. In particular, changes that return something of value to depositors would be an added incentive for the depositing of existing material.

### *Lack of phonetic transcription*

A major bottleneck in the processing of language recordings is the labor-intensive, time-consuming process of transcription and annotation. Without at least a translation into a matrix language, recordings are of very limited value (Woodbury, 2003). Depending on the number of annotation levels and the amount of detail and rechecking employed, this process can take between 20 and 100 times real time (e.g., Auer et al., 2010). In general, transcriptions will be done using a practical orthography. This is useful both because such orthographies (when adequately designed) represent most of the important phonological content and because they are of immediate use to the communities using the language. Even the shallowest orthography, however, obscures some phonetic detail, making it difficult to use this level of description as a basis for phonetic analysis. The ideal would be to have, in addition, a phonological and a phonetic annotation level; this is seldom achieved, so other options need to be made available.

Recommendation: All recordings should, ideally, be transcribed. Further, phonetic use of archives would be enhanced if automatic creation of phonemic and phonetic layers from an orthographic layer were performed.

Benefits: Depositors receive something of value from depositing, both in being able to use the phonemic description and in seeing error in the transcription and/or the orthographic rules. Phonetic analysis will be easier.

Drawbacks: Relatively few such systems exist, and their development and implementation would require a substantial investment.

### *Coarse alignment of transcription to audio*

The granularity of the alignment between a transcription and its audio signal varies greatly across corpora. Perhaps the most common is alignment at the sentence level, although even larger units (e.g. paragraph or entire text) are not unheard of. Alignment at the word level is reasonably common and allows some automatic phonetic analysis. Aligning to the phone level or the acoustic segment is the most useful for automatic phonetic analysis but is also the least common. We emphasize that only automatic analysis enables the major benefit of corpora, the ability to analyze a substantial number of examples. Producing these annotation layers by hand is itself time-consuming, though it can be automated to some extent (e.g. C. T. DiCanio et al., 2013).

Recommendation: Archived material would be more useful to phonetic analysis if they had an automatically generated alignment between phonemes and/or phones and the acoustic signal created upon deposit.

Benefits: The alignment will make phonetic analysis feasible with relatively little involvement of the depositor. Errors in transcription often stand out when alignment fails, allowing correction and thus improvement of the dataset.

Drawbacks: Alignment depends on the previously recommended construction of a phonemic level. Aligners have been made to work for many languages, even those with small corpora, but currently, "dictionaries" (lists of forms that occur in the data) must be made. Universal aligners are being developed (e.g., Adda et al., 2016; Strunk, Schiel, & Seifart, 2014), but refinements are desirable. Archives are not currently able to provide this service, and it would require additional work by archive staff, which is typically a substantial cost.

*Limited archiving of physiological measurements*

Despite the long tradition of taking physiological measurements of endangered languages, there is virtually no tradition of archiving such data. Even for major languages, there are fewer such archives than we might expect at this point. The most heavily used is the X-Ray Microbeam dataset (Westbury, 1994). X-ray data from English and French (Munhall, Vatikiotis-Bateson, & Tohkura, 1995) and various languages (Sock et al., 2011) are also available but less frequently analyzed. Electromagnetic articulometry and real-time magnetic resonance imaging data are becoming available for English (Narayanan et al., 2014).

Recommendation: Begin to provide the data structures that archives would need to store physiological data in a way that is useful for further analysis.

Benefits: Aspects of phonetics that cannot be addressed by acoustics alone could be examined in greater detail.

Drawbacks: The design of metadata and data standards is more complex than those for audio signals, thus the amount of work needed to make the data interoperable is substantial. Migration to future platforms may also present a challenge. Storage requirements for physiological data are typically larger than those for audio, potentially requiring an increase in funding.

*Require depositors to collect material for phonetic analysis.*

Elicited material for phonetic analysis is useful even when other speech samples are extensive. Phonetic results can be obtained from narratives and other continuous speech samples, as has been shown in several of the results summarized here, but targeted, elicited material can often be more easily analyzed. The phonemes of a language are relatively few and can be sampled with a relatively small wordlist (Ladefoged & Maddieson, 1996a). Sequences of phonemes are more numerous and also of interest, and sampling them requires a longer list. Prosodic and other larger contexts are similarly more numerous and require an even larger set of stimuli. In recommending this policy, we are, of course, thinking in terms of new collections; legacy material should be accepted in whatever state it exists.

When researchers are planning to deposit their material in a particular archive (and many funders now require designation of the archive to be used), then they will have to address the elicitation requirements before making their recordings.

Recommendation: Require depositors to obtain an elicited set of stimuli for phonetic analysis.

Benefits: Phonetic analysis would be enhanced. Aspects of the phonology come to light even in the process of making the stimulus list.

Drawbacks: Archives do not have any means of enforcing this policy. Some archives do not allow multiple, small files, which are ideal for phonetic examples.

## Conclusion

Archives hold great promise for future phonetic research. Substantial challenges exist in using archives as they are currently configured. Changes in the way that annotation tiers are implemented in archives can improve the accessibility of the data for phonetic research. Creating archives of articulatory and other physiological data is a large task that would have further sizable benefits; the costs are larger as well. The slow start to using archival data can be expected to seem less severe as greater and greater use is made of the archives. The ongoing expenses for digital archives, being greater than those for print archives, make for urgency in using digital archives. Further improvements, while desirable, generally would require even more financial support. Nonetheless, phonetic results can be obtained from these materials. We hope that our survey of those results will be hopelessly out of date in the near future.

Acknowledgements: We thank Siri Tuttle, Patricia A. Keating, Mandana Seyfeddinipur, Alexis Michaud, Susan Kung, Grace Kim-Lambert, and Nick Thieberger for helpful comments; mistakes remain our own.

References:

Adams, O., Cohn, T., Neubig, G., & Michaud, A. (2017). Phonemic transcription of low-resource tonal languages. In S.-M. J. Wong & G. Haffari (Eds.), *Proceedings of the Australasian Language Technology Association Workshop 2017* (pp. 53-60). Brisbane, Australia: Queensland University of Technology.

Adda, G., Adda-Decker, M., Ambouroue, O., Besacier, L., Blachon, D., Ene Bonneau-Maynard, H., . . . Zerbian, S. (2016, 2016-05). *Innovative technologies for under-resourced language documentation: The BULB Project.* Paper presented at the Workshop CCURL 2016 - Collaboration and Computing for Under-Resourced Languages - LREC, Portoroz, Slovenia.

Amery, R. (1995). It's ours to keep and call our own: reclamation of the Nunga languages in the Adelaide region, South Australia. *International Journal of the Sociology of Language, 113*, 63–82.

Anderson, V. B. (2008). Static palatography for language fieldwork. *Language Documentation and Conservation, 2*(1), 1-27.

Auer, E., Russel, A., Sloetjes, H., Wittenburg, P., Schreer, O., Masnieri, S., . . . Tschöpel, S. (2010). ELAN as flexible annotation framework for sound and image processing detectors. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)* (pp. 890-893). Valletta, Malta: European Language Resources Association (ELRA).

Beagrie, N. (2008). Digital curation for science, digital libraries, and individuals. *International Journal of Digital Curation, 1*(1), 3-16.

Becker-Kristal, R. (2010). *Acoustic typology of vowel inventories and Dispersion Theory: Insights from a large cross-linguistic corpus.* (Ph.D. dissertation), University of California, Los Angeles.

Bird, S., & Simons, G. F. (2001). *The OLAC metadata set and controlled vocabularies*. Paper presented at the Proceedings of the ACL 2001 Workshop on Sharing Tools and Resources - Volume 15, France.

Blankenship, B. (2002). The timing of nonmodal phonation in vowels. *Journal of Phonetics, 30*, 163-191.

Bowern, C. L., McDonough, J. M., & Kelliher, K. (2012). Bardi. *Journal of the International Phonetic Association, 42*, 333-351. doi:10.1017/S0025100312000217

Bradfield, J. (2014). Clicks, concurrency and Khoisan. *Phonology, 31*, 1-49.

Chodroff, E., & Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics, 61*, 30-47. doi:http://dx.doi.org/10.1016/j.wocn.2017.01.001

Clement, T. (2014). The ear and the shunting yard: Meaning making as resonance in early information theory. *Information and Culture, 49*(4), 401-426.

Clements, G. N., & Ridouane, R. (Eds.). (2011). *Where do phonological features come from?: Cognitive, physical and developmental bases of distinctive speech categories*. Amsterdam, Philadelphia: John Benjamins Publishing.

Coombs, A. L. (2013). High tone processes in Ibibio. *Proceedings of Meetings on Acoustics, 19*(1), 060232. doi:10.1121/1.4800734

Cottingham, K. (2008). Turning data graveyards into gold mines. *Journal of Proteome Research, 7*(1), 22.

de Boer, B. (2011). First formant difference for /i/ and /u/: A cross-linguistic study and an explanation. *Journal of Phonetics, 39*, 110-114. doi:http://dx.doi.org/10.1016/j.wocn.2010.12.005

de Carvalho, F. O. (2010). Vowel acoustics in Pirahã. *Revista de Estudos da Linguagem, 18*(1), 11-33.

DiCanio, C. (n.d.). Itunyoso Triqui Collection of Christian DiCanio. Retrieved 7/1/2018 https://www.ailla.utexas.org/islandora/object/ailla%3A243683

DiCanio, C. T. (2009). The phonetics of register in Takhian Thong Chong. *Journal of the International Phonetic Association, 39*, 162-188.

DiCanio, C. T. (2012). The phonetics of fortis and lenis consonants in Itunyoso Trique. *International Journal of American Linguistics, 78*, 239-272.

DiCanio, C. T., Amith, J. D., Castillo García, R., & Lilley, J. (2016). *Obstruent lenition and voicing in a Yoloxóchitl Mixtec corpus*. Paper presented at the Satellite meeting on Reduction, 15th Conference in Laboratory Phonology.

DiCanio, C. T., Nam, H., Amith, J. D., Whalen, D. H., & Castillo García, R. (2015). Vowel variability in elicited versus running speech: Evidence from Mixtec. *Journal of Phonetics, 48*, 45-59. doi:http://dx.doi.org/10.1016/j.wocn.2014.10.003

DiCanio, C. T., Nam, H., Whalen, D. H., Bunnell, H. T., Amith, J. D., & Castillo García, R. (2013). Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment. *Journal of the Acoustical Society of America, 134*, 2235-2246.

DiCanio, C. T., & Whalen, D. H. (2015). The interaction of vowel length and speech style in an Arapaho speech corpus. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences* (Vol. Paper number 513, pp. 1-5). Glasgow: University of Glasgow.

Digital Preservation Coalition. (2015). *Digital preservation handbook* Retrieved from http://handbook.dpconline.org/

Dwyer, A. M. (2006). Ethics and practicalities of cooperative fieldwork and analysis. In J. Gippert, U. Mosel, & N. P. Himmelmann (Eds.), *Essentials of language documentation* (pp. 31-66). Berlin: Mouton de Gruyter.

Errington, J. (2007). *Linguistics in a colonial world: A story of language, meaning and power*. New York: Blackwell Publishing.

Esposito, C. M. (2010). The effects of linguistic experience on the perception of phonation. *Journal of Phonetics, 38*, 306-316. doi:http://dx.doi.org/10.1016/j.wocn.2010.02.002

Farrar, S., & Lewis, W. (2007). The GOLD Community of Practice: an infrastructure for linguistic data on the Web. *Language Resources and Evaluation, 41*, 45-60. doi:10.1007/s10579-007-9016-x

Farrar, S., Lewis, W. D., & Langendoen, D. T. (2002). An ontology for linguistic annotation *Semantic Web Meets Language Resources: Papers from the AAAI Workshop, Technical Report WS-02-16* (pp. 11-19). Menlo Park, CA: AAAI Press.

Flemming, E., Ladefoged, P., & Thomason, S. (2008). Phonetic structures of Montana Salish. *Journal of Phonetics, 36*, 465–491.

Fletcher, J., & Evans, N. (2002). An acoustic phonetic analysis of intonational prominence in two Australian languages. *Journal of the International Phonetic Association, 32*(2), 123-140. doi:10.1017/S0025100302001019

Garellek, M., & Keating, P. A. (2011). The acoustic consequences of phonation and tone interactions in Jalapa Mazatec. *Journal of the International Phonetic Association, 41*, 185-205. doi:10.1017/S0025100311000193

Gick, B., Bird, S., & Wilson, I. (2005). Techniques for field application of lingual ultrasound imaging. *Clinical Linguistics and Phonetics, 19*, 503-514.

Goddard, P. E. (1905). Mechanical aids to the study and recording of language. *American Anthropologist, 7*, 613-619. doi:10.1525/aa.1905.7.4.02a00050

Goddard, P. E. (1907). *The phonology of the Hupa language*. Berkeley: University of California Press.

Gordon, M. (2015). Consonant-tone interactions: A phonetic study of four indigenous languages of the Americas. In H. Avelino, M. Coler, & L. Wetzels (Eds.), *The Phonetics and Phonology of Laryngeal Features in Native American Languages* (pp. 129-156). Leiden/Boston: Brill.

Hale, K., Krauss, M., Watahomigie, L. J., Yamamoto, A. Y., Craig, C., Masayesva Jeanne, L., & England, N. (1992). Endangered languages. *Language, 68*, 1-42.

Hargus, S. (2007). *Witsuwit'en grammar: Phonetics, phonology, morphology*. Vancouver: University of British Columbia Press.

Harrington, J., & Cassidy, S. (1999). *Techniques in speech acoustics*. Dordrecht: Kluwer Academic Publishers.

Hinton, L. (2001). Sleeping languages: Can they be awakened? In L. Hinton & K. Hale (Eds.), *The Green Book of language revitalization in practice* (pp. 413-417). London: Academic.

Hombert, J.-M., Ohala, J. J., & Ewan, W. G. (1979). Phonetic explanations for the development of tones. *Language, 55*, 37-58.

Hsieh, F.-f. (2007). *Relational correspondence in tone sandhi.* (Ph.D. dissertation), Massachusetts Institute of Technology.

Keating, P. A., & Esposito, C. (2007). Linguistic voice quality. *UCLA Working Papers in Phonetics, 105*, 85-91.

Keating, P. A., Esposito, C. M., Garellek, M., Khan, S. u. D., & Kuang, J. (2010). Phonation contrasts across languages. *UCLA Working Papers in Phonetics, 108*, 188-202.

Kempton, T., & Moore, R. K. (2014). Discovering the phoneme inventory of an unwritten language: A machine-assisted approach. *Speech Communication, 56*, 152-166. doi:http://dx.doi.org/10.1016/j.specom.2013.02.006

Kuang, J. (2013). The tonal space of contrastive five level tones. *Phonetica, 70*, 1-23.

Kübler, S., & Zinsmeister, H. (2015). *Corpus linguistics and linguistically annotated corpora*. London: Bloomsbury Publishing.

Ladefoged, P. (2003). *Phonetic data analysis: An introduction to fieldwork and instrumental techniques*. Malden, MA: Blackwell.

Ladefoged, P., & Maddieson, I. (1996a). Recording the phonetic structures of endangered languages. *UCLA Working Papers in Phonetics, 93*, 1-7.

Ladefoged, P., & Maddieson, I. (1996b). *The sounds of the world's languages*. Oxford, UK ; Cambridge, Mass. : Blackwell.

Lawyer, L. C. (2015). Patwin phonemics, phonetics, and phonotactics. *International Journal of American Linguistics, 81*, 221-260. doi:doi:10.1086/680310

Leonard, W. Y. (2007). *Miami language reclamation in the home: A case study.* (3367487 Ph.D. dissertation), University of California, Berkeley, Ann Arbor. ProQuest Dissertations & Theses Full Text database.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*, 431-461.

Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language, 48*, 839-862.

Lindblom, B. E. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modeling* (pp. 403–439). Dordrecht: Kluwer Academic Publishers.

Littledoe, J. (1998). *Wampanoag language reclamation project: First steps to healing the circle.* Paper presented at the Thirtieth Algonquian Conference, Boston, MA.

Maddieson, I. (1984). *Patterns of sounds*. New York: Cambridge University Press.

McCrae, J. P., Labropoulou, P., Gracia, J., Villegas, M., Rodríguez-Doncel, V., & Cimiano, P. (2015). One Ontology to Bind Them All: The META-SHARE OWL Ontology for the Interoperability of Linguistic Datasets on the Web. In F. Gandon, C. Guéret, S. Villata, J. Breslin, C. Faron-Zucker, & A. Zimmermann (Eds.), *The Semantic Web: ESWC 2015 Satellite Events: ESWC 2015 Satellite Events, Portorož, Slovenia, May 31 – June 4, 2015, Revised Selected Papers* (pp. 271-282). Cham: Springer International Publishing.

McDonough, J. M. (2003). *The Navajo sound system*. Dordrecht: Kluwer.

McDonough, J. M., & Tucker, B. V. (2012). Replicating Goddard: A contemporary airflow and EGG study of Dene SųŁiné *University of Rochester Working Papers in the Language Sciences, 7*, 1-17.

Miller, A. L. (2007). Guttural vowels and guttural co-articulation in Ju|'hoansi. *Journal of Phonetics, 35*, 56-84.

Miller, A. L. (2016). Posterior lingual gestures and tongue shape in Mangetti Dune !Xung clicks. *Journal of Phonetics, 55*, 119-148. doi:http://dx.doi.org/10.1016/j.wocn.2015.12.001

Munhall, K. G., Vatikiotis-Bateson, E. S., & Tohkura, Y. i. (1995). X-ray film database for speech research. *Journal of the Acoustical Society of America, 98*, 1222-1224. doi:10.1121/1.413621

Narayanan, S., Toutios, A., Ramanarayanan, V., Lammert, A., Kim, J., Lee, S., . . . Proctor, M. (2014). Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC). *Journal of the Acoustical Society of America, 136*, 1307-1311. doi:doi:http://dx.doi.org/10.1121/1.4890284

Niebuhr, O., & Michaud, A. (2015). Speech data acquisition: the underestimated challenge. *KALIPHO - Kieler Arbeiten zur Linguistik und Phonetik, 3*, 1-42.

Pike, K. L. (1947). Grammatical prerequisites to phonemic analysis. *Word, 3*(3), 155-172.

Pulgram, E. (1954). Accent and ictus in spoken and written Latin. *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete der Indogermanischen Sprachen, 71*(3/4), 218-237.

Remijsen, B. (2013). Tonal alignment is contrastive in falling contours in Dinka. *Language, 89*, 297-327.

Remijsen, B., & Ayoker, O. G. (2014). Contrastive tonal alignment in falling contours in Shilluk. *Phonology, 31*, 435-462.

Research Libraries Group. (2002). *Trusted digital repositories: Attributes and responsibilities*. Retrieved from Mountain View, CA:

Rousselot, P.-J. (1897-1908). *Principes de phonétique expérimentale*. Paris: H. Welter.

Schiering, R., Bickel, B., & Hildebrandt, K. A. (2010). The prosodic word is not universal, but emergent. *Journal of Linguistics, 46*, 657-709.

Schmidt, T., Chiarcos, C., Lehmberg, T., Rehm, G., Witt, A., & Hinrichs, E. (2006). *Avoiding data graveyards: From heterogeneous data collected in multiple research projects to sustainable linguistic resources.* Paper presented at the 6th E-MELD workshop, Ypsilanti.

Schwartz, J.-L., Boë, L.-J., Vallée, N., & Abry, C. (1997). The Dispersion-Focalization Theory of vowel systems. *Journal of Phonetics, 25*, 255-286. doi:http://dx.doi.org/10.1006/jpho.1997.0043

Sen, R. (2015). *Syllable and segment in Latin*. Oxford: Oxford University Press.

Simas Frazão, K. (2013). *A sílaba no Akwẽ-Xerente (Jê).* (M.A. thesis), Universidade de Brasília.

Sock, R., Hirsch, F., Laprie, Y., Perrier, P., Vaxelaire, B., Brock, G., . . . Ma, L. (2011). *An X-ray database, tools and procedures for the study of speech production.* Paper presented at the 9th International Seminar on Speech Production (ISSP 2011).

Strunk, J., Schiel, F., & Seifart, F. (2014). Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of LREC* (pp. 3940-3947). Reykjavik, Iceland: European Language Resources Association.

Tuttle, S. G. (2003). Archival phonetics: Tone and stress in Tanana Athabaskan. *Anthropological Linguistics, 45*, 316-336.

Westbury, J. R. (1994). *X-ray microbeam speech production database user's handbook*. Retrieved from

Whalen, D. H., DiCanio, C. T., & Shaw, P. A. (2011). Phonetics of endangered languages. *Acoustics Today, 7*(4), 35-42.

Whalen, D. H., & Levitt, A. G. (1995). The universality of intrinsic F0 of vowels. *Journal of Phonetics, 23*, 349-366.

Whalen, D. H., & McDonough, J. M. (2015). Taking the laboratory into the field. *Annual Review of Linguistics, 1*, 395–415.

Whalen, D. H., & Simons, G. F. (2012). Endangered language families. *Language, 88*, 155-173.

Wieling, M., Tomaschek, F., Arnold, D., Tiede, M. K., Bröker, F., Thiele, S., . . . Baayen, R. H. (2016). Investigating dialectal differences using articulography. *Journal of Phonetics, 59*, 122-143. doi:http://dx.doi.org/10.1016/j.wocn.2016.09.004

Winter, B. (2015). The other N: The role of repetitions and items in the design of phonetic experiments. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences* (Vol. Paper number 181, pp. 1-5). Glasgow: University of Glasgow.

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

Yu, A. C. L. (2008). The phonetics of the quantity alternation in Washo. *Journal of Phonetics, 36*, 508-520.