Linguistic Analysis, Ethical Practice, and Quality Assurance in Anonymizing Recordings of Spoken Language for Deposit in Digital Archives

Diana Sofia Ovalle Lopez
Department of Spanish Immersion
Fremont Christian School
Fremont, USA
slopez@fremontchristian.org

Robert E. Vann
Department of Spanish
Western Michigan University
Kalamazoo, USA
robert.vann@wmich.edu

***Abstract*— This report considers linguistic analyses as matters of ethical practice and quality assurance in the anonymization of recordings of spoken language for deposit in a digital language archive. Ethically, researchers must be committed to protecting the identities of primary data providers. Accordingly, conducting pragmatic analyses before initiating technical anonymization procedures can aid in determining exactly what discourse, in what contexts, might constitute identifying information in need of anonymization. Qualitatively, one of the main goals of language documentation is to preserve as much primary data as possible for future research. Accordingly, conducting phonotactic analyses with the help of computer software can aid in determining precise chronometer readings for each tonal insertion to excise as little primary data as possible during anonymizations. These findings warrant further research on anonymization protocols in digital language archive projects.**

Keywords*— *language documentation, digital language archives, spoken language corpora, anonymization practices, linguistic analysis, pragmatics, phonotactics, research ethics, quality assurance, Spanish, DARDOSIPCAT

I. Introduction and Statement of Problem

Language documentation is a growing field of study that continues to evolve with the advancement of technology. As in most research with human subjects, participant identities must be protected. Typically, participant names are usually left out of written publications. However, in the world of digital language archives, language data can be found in the form of audio recordings in which, potentially, the uniqueness of participants' own voices or their ways of speaking could identify them. Given this potential, in practice, audio recordings on deposit in digital language archives can never truly be completely anonymous. Nevertheless, to protect the privacy of research participants, reasonable efforts can and should be made to minimize identifying information in such recordings. In many cases, names and any other identifying information may be "bleeped" out to protect an individual's identity. This anonymization practice involves replacing spoken language in the soundwave with an audible tone. As a matter of professional ethics, the anonymization of audio recordings was one of the quality assurance steps taken in the development of the ***D***igital ***AR***chive to ***DO***cument ***S***panish ***I***n the ***P***aïsos ***CAT***alans, henceforth DARDOSIPCAT, a language documentation project that aims to preserve and disseminate spoken language corpora of Spanish from The Països Catalans. This report addresses linguistic analyses involved in DARDOSIPCAT anonymization practices. Pragmatic, phonetic, and phonological analyses were crucial in developing principled anonymization practices. These practices involved (a) determining exactly what could be potentially identifying information and (b) separating coarticulated sounds across word boundaries.

II. Review of Relevant Literature

Following [1], while the act of collecting data should be seen as distinct from that of analyzing it, language description can indeed inform language documentation. In this sense, "descriptive techniques are part of a broad set of techniques applied in compiling and presenting a useful and representative corpus of primary documents of the linguistic practices found in a given speech community" [1, p. 2]. Accordingly, linguistic analyses can be a key component of language documentation. In DARDOSIPCAT, such analyses have been nothing less than necessary to ethically archive anonymized access copies of the primary data collected.

Moreover, the process of anonymization is, by nature, an editing process that may compromise the accountability of the work in question, leading to problems of interpretation. According to [2, p. 563], "heavy editing of recorded materials may give an artificial or even misleading impression of the original linguistic event." Therefore, as a matter of quality assurance, the process of anonymization must be meticulous yet considered in order to preserve as much primary data as

This research was partially funded by Western Michigan University.

possible. Among others, [3] and [4] have both pointed out the importance of distributed and redundant collaboration in this regard. It would be exceedingly difficult for one researcher alone to carry out all the tasks involved in anonymizing the individual recordings of multiple corpora on deposit in a given documentation project. Because we need to account for human error, best practice is for multiple individuals to revise the work that others have done.

III. Methodology

Our work with digital recordings on deposit in DARDOSIPCAT mainly concerned interviews that were originally recorded on analog cassette tape in Barcelona, Spain in 1995. In 2015, a previous research assistant (henceforth RA) digitized the tapes, creating digital audio files in AIFF format. In 2018, another RA listened carefully to the AIFF files, cataloging potentially identifying information on an Excel spreadsheet that included approximate chronometer readings for each stretch of discourse that might possibly contribute to the identification of research participants.

The anonymization process started by using the suggested chronometer readings to isolate each individual stretch of potentially identifying discourse for each audio recording using Audacity software. At times the approximate chronometer readings turned out to be spot-on and no fine tuning was necessary. In most cases, however, the approximate readings needed further specification to be precise. Determining accurate timing involved both an ear for detail and subsequent visual analysis of soundwaves to precisely identify the starting points and endpoints of identifying information. When human hearing alone was not reliable enough to discern an acceptable split point for a diphthong, for example, use of Audacity, Praat, and ExpressScribe software facilitated finding the most accurate starting points, endpoints, and volumes for the tones to be inserted. This process required careful linguistic analysis, which RAs carried out with the help of the Principal Investigator (henceforth PI) and the software mentioned above.

During the anonymization process, we found that some audio recordings contained potentially identifying information that had not been initially included on the Excel spreadsheet. As well, the PI determined that some of the stretches of potentially identifying information initially included on the Excel spreadsheet did not actually correspond to identifying information. For example, in one instance, the stretch of discourse “Hola, hola” was initially mistaken for “Hola, Laura”; such entries were removed from the spreadsheet.

For the purposes of quality assurance and research ethics, the PI determined that, before depositing anonymized access recordings in the archive, an RA should review each audio recording a second time to search for additional potentially identifying information that might previously have been missed. Thus, the phase of the anonymization process during which we cataloged potentially identifying information was recursive.

This added attention to detail was intended as a measure to help safeguard the anonymity of individuals whose spoken language is on deposit in DARDOSIPCAT. Nevertheless, because one must be very focused while listening to each audio recording in order to “catch” any potentially identifying information, names in particular, one may begin “fishing” for names where there are none as in the example above. Any human error in this regard that were to lead to unnecessary bleeping, though well-intentioned, could hinder the authenticity or richness of the primary data.

A. *Pragmatic Analysis*

One of the gaps in language documentary literature concerns best practice recommendations for exactly what (and how much) to anonymize. In DARDOSIPCAT, we turned to pragmatic analyses to determine whether or not certain information was identifying. As more identifying information was discovered, the PI established anonymization policies for certain cases.

One of these policies concerned the names of places in which the speakers and their parents had been raised. We determined that, with the gender and age of each speaker given in resource metadata, if users of the access recordings were to learn from the recordings themselves that speakers and their families were from particular places outside Barcelona, speaker identities could potentially be ascertained. Consequently, our best practice policy was to anonymize the name of any place of speaker or family origin that was not located within the Barcelona metropolitan area. Importantly, these same place names were fine to leave un-anonymized when mentioned in discourse contexts other than those of speaker/family origin.

The second anonymization policy concerned the speakers' majors and the universities they attended. Again, because resource metadata include the gender and age of each speaker, if users of the access recordings were to learn from the recordings themselves both the university that speakers attended and the major they pursued there, speaker identities could potentially be ascertained. Sometimes, there was mention of just the major but not the university; however, because in Barcelona in 1995 some majors were offered at only one university, the PI determined that in such cases the mention of the major should be anonymized. Subsequently, it became policy to always anonymize the major if the university was previously mentioned or if that specific major was only offered at one university. When the major was not specific to one university, we decided not to anonymize the major, but rather the university. Although at times we questioned whether anybody would ever purposely analyze such information just to identify a participant, we determined that our ethical duty as researchers requires we do everything we can reasonably do to protect the anonymity of the participants.

B. *Phonological and Phonetic Analysis*

Isolating identifying information sometimes proved difficult in context due to the formation of diphthongs and synalephas across word boundaries in Spanish. In such cases, the anonymization process required further phonological and phonetic analysis for quality control. Such analysis was often required when the suggested chronometer readings included a word ending in a vowel before the utterance to be anonymized. For example, one of the anonymizations was for the name “Elizabeth”, and the chronometer readings included the stretch of discourse “La Elizabeth” ‘The Elizabeth’. In order to anonymize the name but leave the article, chronometer readings had to be set to insert a tone exactly where the /a/ in “La” ends, but before the beginning of the /e/ in Elizabeth.

Dealing with consonantal coarticulations and nasalized vowels was another challenge we encountered when isolating identifying information. For example, in the phrase “en Molins de Rei” ‘in Molins de Rei’, as the hometown of one of the speakers' parents, “Molins de Rei” represented potentially identifying information. Due to an obligatory process of nasal assimilation in Spanish, the initial nasal consonant was bilabial before the nasal stop. Moreover, the articulation of the vowel was nasalized before the initial nasal consonant.

In complex cases like those described above, we used Praat to visually analyze the spectrogram of the contextualized audio fragment. In the case of “La Elizabeth”, via Pratt we were able to discern the speech formants, which represent concentrations of energy based on frequency. In Spanish, the second formant represents the highest amplitude that a soundwave reaches, and each vowel reaches a different amplitude in the wave. Looking at the spectrogram of “La Elizabeth”, we were able to see the rising of the second formant from [a] to [e] and thus identify the precise chronometer reading at which to insert the anonymization tone. In the case of “en Molins the Rei”, we used Praat to determine the onset of nasality in this sequence based on acoustical measures.

IV. Findings and Significance

One finding of our research is that pragmatic analyses are needed to discern potentially identifying information in a contextually-appropriate manner. This finding is significant to the accountability of documentary work; without such analyses, researchers have no principled way of knowing the extent of potentially identifying information that a recording may include. Accordingly, pragmatic analyses of spoken language corpora should be a prerequisite to the insertion of anonymization tones in digital recordings on deposit in language archives. Once such analyses have been completed, researchers can implement principled anonymization policies uniformly throughout the documentation to deal with context-dependent identifying information that might otherwise remain undetected.

Another finding of our research was that phonotactic analyses conducted with the help of computer software are needed to accurately isolate potentially identifying information in the phonetic phrase for later anonymization. Given that Spanish syllabifies discourse irrespective of word boundaries, without computer-mediated phonotactic analyses, researchers have no principled way of determining precise chronometric readings at which to begin and end anonymization tones in digital recordings of spoken Spanish. This finding is significant to documentary accountability insofar as it improves quality in anonymization processes. Once such analyses have been completed, researchers can be assured of high-quality tonal insertions in the development of anonymized language resources.

V. Conclusion

Language documentation is a science composed of various types of linguistic analysis. This report has described practices and protocols utilized in DARDOSIPCAT to demonstrate how and why linguistic analyses may be useful in the anonymization of spoken language resources for deposit in digital language archives. Given its goal of preserving as much primary data as possible for future research, the anonymization process is arduous, meticulous, and iterative. Accordingly, we have discussed pragmatic and phonotactic analyses as matters of both professional ethics and quality assurance.

As technology continues to advance and digital language archives grow richer in content, it is important for everyone involved in language documentation to stay committed to protecting participant identities. Because language documentation is an ongoing, ever-evolving process best achieved in continued collaboration, further research on anonymization procedures is warranted. Such research could improve ways of maintaining high-quality recordings of spoken language while also protecting the privacy of primary data providers.

References

[1] N. P. Himmelmann, "Documentary and descriptive linguistics," Linguistics, vol. 36, no. 1, pp. 161-195, 1998, doi:10.1515/ling.1998.36.1.161.

[2] S. Bird and G. Simons, "Seven dimensions of portability for language documentation and description," Language, vol. 79, no. 3, pp. 557-582, 2003, doi: 10.1353/lan.2003.0149.

[3] N. P. Himmelmann, "Language documentation: What is it and what is it good for?," in Essentials of Language Documentation, J. Gippert, N. P. Himmelmann, and U. Mosel, Eds. Berlin: Mouton de Gruyter, 2006, pp. 1-30.

[4] A. C. Woodbury, "Language documentation," in The Cambridge Handbook of Endangered Languages, P. K. Austin and J. Sallabank, Eds. Cambridge, UK: Cambridge University Press, 2011, pp. 159-186.