Published in Xiong Deyi, Eric Castelli et al. (eds.), *Proceedings of IALP 2012 (2012 International Conference on Asian Language Processing)*, pp. 169–172. Hanoi, Vietnam: MICA Institute, Hanoi University of Science and Technology.

Combining Documentation And Research: Ongoing Work On An Endangered Language

Alexis Michaud
MICA* & LACITO**
*CNRS/Hanoi Univ. of Science and Technology **CNRS
*Hanoi, Vietnam **Villejuif, France
alexis.michaud@vjf.cnrs.fr

Andrew Hardie
University Centre for Computer Corpus Research on Language (UCREL)
Lancaster University, United Kingdom
a.hardie@lancaster.ac.uk

Séverine Guillaume
LACITO
CNRS
Villejuif, France
severine.guillaume@vjf.cnrs.fr

Martine Toda
Laboratoire de Phonétique et Phonologie (LPP)
CNRS
France
martinetoda@gmail.com

***Abstract*—This paper is intended for an audience of speech technology specialists who believe that "automatic processing of under-resourced languages is a way to study language diversity with a multi-disciplinary view" (L. Besacier, keynote speech at this conference). It aims (i) to provide an illustration of the way in which data are collected in fieldwork on endangered languages, bringing attention to the quality of the transcriptions and annotations created by linguists; (ii) to present the contents and format of a set of endangered-language documents synchronizing sound and text, which are currently available online; and (iii) to sketch out some of the research purposes and applications to which these documents lend themselves, and which we intend to pursue in future work.**

Keywords—multimedia corpora; language documentation; endangered languages; spontaneous speech; interlinear glossing; online databases; long-term preservation; Yongning Na; Sino-Tibetan

I. Introduction

A. *Endangered linguistic documents: the dismal current state of documentation of the world's languages*

The need to document the world's languages is now well-known to linguists and the general public. Fewer people, however, are aware of the dismal current state of linguistic documentation in many research institutions. Looking back at a century of speech recording, the legacy is not as extensive – and nowhere as tidy – as the layman would think. "Enormous amounts of data – often the only information we have on disappearing languages – remain inaccessible both to the language community itself, and to ongoing linguistic research" [1], and eventually disappear.

"The data that we create (…) should be reusable, both by ourselves and by others. First because any claims that we make based on that data must themselves be replicable and testable by others, and second, because the effort of creating a digital representation of the data should not be duplicated later by others, but used as a foundation that can be built on. (…). This is all the more important when a linguist makes the only recordings for an endangered language–one that may no longer be spoken in the near future" [1].

In our view, one way of curbing this great waste of linguistic resources lies in a closer association of documentation, research, and technological applications. This paper presents a set of documents collected in the course of linguistic research on a Sino-Tibetan language spoken in China: Yongning Na. These documents are freely available online, with multilingual time-aligned annotations. This presentation aims (i) to provide an illustration of the way in which data are collected in fieldwork on endangered languages, bringing attention to the quality of the transcriptions and annotations created by linguists; (ii) to present the contents and format of a set of documents synchronizing sound and text that are currently available online; and (iii) to sketch out some of the research purposes and applications to which these documents lend themselves, and which we intend to pursue in future work.

B. *The Pangloss Collection, developed by the CNRS-LACITO laboratory*

The set of resources described here is deposited in the Pangloss Collection developed by the CNRS-LACITO laboratory.

1) *Objectives*

Beginning in the 1990s, when the storage of such data in digital form became practical, the LACITO (Langues et Civilisations à Tradition Orale) research group of the French CNRS (Centre National de la Recherche Scientifique) has undertaken the digital archiving of data collected by its members, and proposed this service to colleagues from other centres. The aim is to broaden the range of possibilities for research on these materials, as well as to ensure their availability to the speech communities (and their long-term conservation) [2]. Over the last ten years, the number of languages in the archive has increased from 20 to more than 70. Currently, the Pangloss Collection contains 142 hours of recordings: over 1300 recordings, of which about one third (400) have a transcription.

2) *Pangloss: "We must cultivate our garden"*

The name "Pangloss" (meaning "all languages") reflects the project's scope: data sets from any language may be deposited. More importantly, the name is inspired by Voltaire's novel *Candide, or the Optimist*, in which Candide cuts short the rambling discourses of the philosopher Pangloss by a reminder that "we must cultivate our garden". Documents in the Pangloss Collection are not collected for the sake of building a language archive: they are recorded and transcribed by linguists as part of their research activity, and their preparation for online publication requires the hands-on involvement of researchers. The Pangloss Collection grows when a researcher understands its advantages for research and other purposes, and decides to engage in the formatting of her or his data. The guiding principle of the Pangloss Collection is that a close association between documentation and research is highly profitable to both. We now turn to a detailed example.

II. Documenting Yongning Na: An overview

Yongning Na is an unwritten Sino-Tibetan language spoken in and around the plain of Yongning, at the border between the Chinese provinces of Yunnan and Sichuan. There exist substantial publications about the language [see 3 and references therein]. On the other hand, there were no recordings available online when the work reported here began.

The first author of this paper stayed in the village of Yongning about two months a year from 2006 to 2009. From August 2011 to November 2012, he is based in China, working with his main consultant on a day-to-day basis.

A. *The choice of recorded materials is guided by the research topics*

Modern-day linguists typically make extensive audio and video recordings. An individual researcher often has many hours of recordings. *Word lists* are elicited first, to work out the phonemic system. *Narratives* (folk tales, life stories, explanations about traditional techniques…) and *dialogues* are the backbone of linguistic documentation. They illustrate the use of words in context, and constitute a reliable base for an endless range of research purposes.

In addition to these two basic text genres, recordings are guided by the issues encountered in research, and reflect the diversity of linguists' interests. In the case of the documents presented here, the balance is clearly tilted towards phonetic/phonological topics, which the researcher had the interest to pursue in greatest detail. Specifically, the tone system of Yongning Na is an area which calls for in-depth description: the language has a complex system of morphologically conditioned tone change. Numerous elicitation sessions have therefore been devoted to an investigation of the tone system, each focussing on a specific issue, such as tone changes that take place when an object is associated with a verb.

There is therefore no headlong conflict between a documentation agenda and a research agenda: all documents recorded in the course of research are relevant additions to the online collection, gradually enriching the record.

B. *Advantages of fieldwork conditions for collecting abundant and reliable data*

Data collection is an underestimated challenge, and perhaps a weak spot of some current linguistic studies. It is obvious that the empirical basis of one's research is of paramount importance for all later stages. On the other hand, the importance of good communication with language consultants is not always recognized. The consultants' perception of the investigator's intentions exerts considerable influence on their behaviour [4–6]. In this respect, the fieldworker's experience may be useful to the "laboratory worker". Documents collected in fieldwork compare favourably in many respects with those collected in the lab. In fieldwork, the investigator's familiarity with her or his consultants allows for the thoughtful design of materials to be recorded. In a nutshell, fieldwork materials should not be dismissed as less refined than recordings done in laboratories.

For some languages, adding complete glosses at word level can be done semi-automatically. However, semi-automatic treatment is painstaking in languages that contain numerous homophonous words, making it sometimes more appropriate to do all the glossing by hand than to wade through lists of homophonous lexical entries. The annotation of the Yongning Na data was entirely manual.

Creating reliable, fine-grained transcriptions and annotations for documents in less-documented languages can only be a labour of love, into which investigators and their consultants put great amounts of time and effort. The quality of these hand-made annotations is usually excellent. This is a good start in life for these resources, which can then be further enriched, and used for a variety of purposes. But before looking at potential applications, here is an overview of the data set in Yongning Na.

III. Current state of the online collection

A. *Presentation of the documents*

1) *Overall amount of data*

The set of Yongning Na documents is in the process of being enriched, so the figures below are indicative only.

11 hours of narratives have been recorded since 2006. Eight texts with complete Chinese and French annotations are now available online (one of them with English annotation as well), corresponding to 1.3 hour. Eight more (1.5 hour, also available online) are transcribed and translated but without word-level glosses. Phonetic/phonological elicitation sessions with full Chinese and English (and French) annotations amount to more than two hours (over 40 documents). This is above the average volume of data for languages in the Pangloss Collection: the collection hosts data from 70 languages; the average is 2 hours per language (20 documents per language, 6 of them transcribed and annotated). On the other hand, the proportion of annotated data is comparable with the average for the whole Pangloss Collection in its present state: below 33%. This important point will be returned to below, when discussing the potential of speech processing technologies.

2) *Technical quality*

About one fifth of the documents were recorded as stereo WAV files (24-bit, 44,100 Hz) with an audio channel and an electroglottographic channel. (Electroglottography is the ultimate reference for measuring fundamental frequency; it also allows for the evaluation of other glottal parameters such as the open quotient.) Another fifth consists of mono audio files. The rest are stereo audio files comprising a signal from a headworn microphone.

The annotation is logically structured text, in XML format. Narratives have the following structure: a TEXT is divided into S (sentences in a loose sense), which are divided into W (words). It is possible to indicate a further division into M (morphemes). Each level can have translations into any number of languages. The sound and its annotation are synchronized at the S level [2].

B. *Online browsing*

The resources (recordings and annotations) are freely available under a CreativeCommons licence. The Pangloss Collection is hosted in a broader repository named Cocoon (formerly CRDO-Paris), via which the resources are referenced by various search engines including OLAC and OAIster. The metadata follow Dublin Core/Open Language Archives Community standards. The documents can be accessed via the following page:

http://lacito.vjf.cnrs.fr/archivage/index_en.htm

It is also possible for depositors to customize a web page presenting the data for the language at issue. For Yongning Na, the web page is:

http://lacito.vjf.cnrs.fr/archivage/languages/Na_en.htm

C. *Long-term conservation*

A technical description of archiving and web hosting falls outside the scope of this paper; let us simply mention that long-term conservation is guaranteed through a partnership with a perennial archiving institution: CINES [7], which also hosts other multimedia archives (in particular that of the Aix-en-Provence speech research laboratory [8], whose aims are very close to those of the Pangloss Collection).

IV. Perspectives for Research and Multi-lingual Applications

A. *Collaborative enrichment of resources*

Once a resource is available online, its potential for research can be increased through collaborative enrichment. Beyond manual annotations, this includes possibilities for semi-automatic and automatic additions.

1) *Manual additions: transcription and translations*

It is the linguist's job to provide a transcription (in International Phonetic Alphabet, IPA) and a sentence-level translation into English or another national language. As described above, about one fourth of the linguistic resources contained in the Pangloss Collection are transcribed, glossed and translated into one or more languages. The enrichment of resources should not be viewed as a finite process, ending when the resource is put online. Putting resources online is a great way of drawing colleagues' attention to the data and improving their annotation. A marginal but real example concerns Ubykh, an extinct language of the Caucasus. Raw recordings had been put online for their undisputable historical value, although scanned manuscript notes of the original researcher, Georges Dumézil, were available as the sole documentation. Later on, further annotations, transcriptions and English translations were contributed by Brian Fell and others (for details, see http://lacito.vjf.cnrs.fr/archivage/languages/Ubykh_en.htm).

Documents on endangered languages may be further enriched for purposes that were not necessary envisaged at the time of collection. This is happening for the Na language: Roselle Dobbs, a highly motivated student of Chinese and Na, agreed to contribute English translations to one of the Na documents available online. The document is significantly enriched by the addition of the English annotations, and several mistakes in the original annotation were pointed out by Ms. Dobbs. For an orthography development project, the automatic addition of an orthographic tier to the documents would be an easy process; this would constitute a highly significant enrichment of the data set.

This brings us to the topic of information-technology techniques.

2) *The enrichment of resources through speech-processing technology*

Endangered-language data sets such as that described in Section III have strikingly similar characteristics in research institutions worldwide: a portion of the data to the order of 10 to 30% has a fine-grained annotation; another 10 to 30% has a transcription/coarse annotation; and a third part (sometimes well over half of the recordings) is not transcribed at all. Since human operators obviously find it hard to process all the data they collect, it would seem logical to tap the potential of natural-language processing techniques to add annotation. In the process, the resources' potential for use in speech technology applications could be greatly increased.

Phoneme-level alignment with the speech signal

Phonetic transcriptions (which are the norm in languages with no writing system) mapped to the speech signal at sentence or paragraph level can be refined to the phoneme level with the help of phonetic alignment softwares, such as EasyAlign [9] or SailAlign [10]. These programs typically use acoustic models trained on particular languages, but preliminary tests by the fourth author suggest that they are robust enough to be applied to other languages as well, by operating a simple phoneme-to-phoneme mapping based on similarity between the language of the resource being processed and the language on which the algorithm was initially trained. The accuracy of those alignments is not as good as human annotation, but further algorithms may be applied to obtain finer results, and final human verification of this output is less time-consuming than doing a complete manual alignment.

Adapting models trained on large data sets to under-documented languages

State-of-the-art spoken language processing techniques now offer the opportunity to adapt acoustic and language models trained on huge data sets to particular sets of data, such as foreign-accented speech or under-documented languages, using much smaller data samples, of which only a fraction needs to be annotated [11–13]. These techniques make it possible to perform automatic speech recognition and machine translation of under-documented data.

Putting these technological functions together, it would be possible, in principle, to improve greatly the annotation of raw or little-annotated resources, allowing them to be further processed along with the manually-annotated subset of data.

B. Application in multilingual acoustic modeling

One of the major interests of the linguistic resources in the Pangloss collection from the point of view of spoken language processing resides in their rich annotation, especially phonetic transcriptions coded in the IPA standard. The development of language-independent or multilingual acoustic models for speech recognition of under-resourced languages/non-native accents is a research topic of growing popularity [14–16, to cite a few]. In building such acoustic models, the variety of speech sound covered by the training data constitutes a significant factor in the recognition performance [14]. The highly varied phonetic material contained in this kind of language archive can be seen as a potential resource for the improvement of phone and phone sequence coverage.

C. Browsing the linguistic content

The choice of XML markup in the Pangloss Collection allows for a high degree of interoperability with various software. XML is a pivotal format: being fully explicit, it makes conversions easy. In addition, the use of explicit conventions for glossing allows the automatic processing of linguistic content.

An example is the integration of subsets of the Pangloss Collection into CQP-Web [17], an online Corpus Query Processor which allows for queries over texts (including linguistic annotations): typically, it builds concordances and looks for patterns of co-occurrence. The Na data set will be integrated to CQP-Web by the end of the year 2012. Uploading these data into the database is a straightforward operation using conversion scripts, which can be renewed on a roughly half-yearly basis if the researcher has produced an improved version of the transcription and annotation in the meantime. This often happens for less-documented languages: the transcription and glosses are based on a description done by the researcher, which typically evolves during the researcher's lifetime. CQP-Web already hosts another data set from the Pangloss Collection: documents in the Limbu language (Nepal).

V. Conclusion

It has been said that "the study of endangered languages has the potential to revolutionize linguistics", and that "the vanguard of the revolution will be those who study endangered languages" [18]. We hope that this is so, not only because of the considerable cultural interest of those resources, but for their high potential in multilingual applications as well, where their extreme diversity and richness may be put to good use. In his keynote talk at this conference, Laurent Besacier [19] suggests that "automatic processing of under-resourced languages is a way to study language diversity with a multi-disciplinary view". The present paper simply described a specific data set; our future work will consist in following up on the perspectives outlined here, by developing software tools to bridge "the gap between the language experts and the technology experts" [19] and to tap the great potential of data sets from endangered languages.

Acknowledgments

Many thanks to Boyd Michailovsky, Bernard Bel and Eric Castelli for advice and encouragement, to Roselle Dobbs for editorial expertise, and (last but not least) to the Na language consultants and friends.

Fieldwork on Yongning Na is funded through grants PASQi (2007-2012) and HimalCo (2012-2014) from the Agence Nationale de la Recherche (ANR, France). The present work is a contribution to the "Evolutionary Approaches to Phonology/Panchronic Phonology" operation within the LabEx project "Empirical Foundations of Linguistics" (http://www.labex-efl.org).

References

[1] N. Thieberger and R. Nordlinger, "Doing Great Things with Small Languages (Australian Research Council grant DP0984419)," 2006. http://linguistics.unimelb.edu.au/research/projects/greatthings.html

[2] M. Jacobson, B. Michailovsky, and J. B. Lowe, "Linguistic documents synchronizing sound and text," *Speech Communication*, vol. 33 [special issue: "Speech Annotation and Corpus Tools"], pp. 79–96, 2001.

[3] L. Lidz, "A descriptive grammar of Yongning Na (Mosuo)," University of Texas, Department of linguistics, Austin, 2010.

[4] C. Bowern, *Linguistic fieldwork: a practical guide*. Basingstoke [England]; New York: Palgrave Macmillan, 2008.

[5] P. Newman and M. Ratliff, *Linguistic fieldwork*. Cambridge: Cambridge University Press, 2001.

[6] U. Mosel, "Field work and community language work," in *Essentials of language documentation*, J. Gippert, N. P. Himmelmann, and U. Mosel, Eds. Berlin/New York: de Gruyter, 2006, pp. 67–83.

[7] "CINES: Centre Informatique National de l'Enseignement Supérieur." http://www.cines.fr

[8] "SpLanDR -Speech and Language Data Repository." http://sldr.org

[9] J.-P. Golman, "EasyAlign: an automatic phonetic alignment tool under Praat," in *Proceedings of InterSpeech 2011*, Florence, 2011.

[10] A. Katsamanis, M. Black, P. Georgiou, L. Goldstein, and S. Narayanan, "SailAlign: Robust long speech-text alignment," in *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, Philadelphia, 2011.

[11] Z. Wang, T. Schultz, and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," in *Proceedings of ICASSP 2003*, Hong Kong, 2003, I:540–543.

[12] M. Bacchiani and B. Roark, "Unsupervised language model adaptation," in *Proceedings of ICASSP 2003*, Hong Kong, 2003, I:224–227.

[13] T. N. D. Do, E. Castelli, and L. Besacier, "Mining Parallel Data from Comparable Corpora via Triangulation," in *Proceedings of International Conference on Asian Language Processing - IALP 2011*, Penang, Malaysia.

[14] T. Schultz and A. Waibel, "Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition," *Speech Communication*, no. 35, pp. 31–51, 2001.

[15] J. Köhler, "Comparing Three Methods to Create Multilingual Phone Models for Vocabulary Independent Speech Recognition Tasks," in *Proceedings of Multi-lingual Interoperability in Speech Technology (MIST)*, Leusden, The Netherlands, 1999.

[16] D. Imseng, H. Bourlard, P. Dines, N. Garner, and M. Magimai-Doss, "Improving non-native ASR through stochastic multilingual phoneme space transformations," in *Proceedings of Interspeech 2011, Florence, Italy, 2011*, Florence, 2011.

[17] A. Hardie, "CQPweb - combining power, flexibility and usability in a corpus analysis tool," forthcoming.

[18] D. H. Whalen, "How the study of endangered languages will revolutionize linguistics," in *Linguistics today: Facing a greater challenge*, P. van Sterkernburg, Ed. Amsterdam/Philadelphia: John Benjamins, 2004, pp. 321–344.

[19] L. Besacier, "A multi-disciplinary approach for processing under-resourced languages," in *Proceedings of International Conference on Asian Language Processing (IALP 2012)*, Hanoi, 2012.