LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 12 (2018), pp. 393–429
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/24793
Revised Version Received: 5 July 2018

Integrating Automatic Transcription into the Language Documentation Workflow: Experiments with Na Data and the Persephone Toolkit

Alexis Michaud
CNRS-LACITO

Oliver Adams
Computing and Information Systems, The University of Melbourne

Trevor Anthony Cohn
Computing and Information Systems, The University of Melbourne

Graham Neubig
Language Technologies Institute, Carnegie Mellon University

Séverine Guillaume
CNRS-LACITO

Automatic speech recognition tools have potential for facilitating language documentation, but in practice these tools remain little-used by linguists for a variety of reasons, such as that the technology is still new (and evolving rapidly), user-friendly interfaces are still under development, and case studies demonstrating the practical usefulness of automatic recognition in a low-resource setting remain few. This article reports on a success story in integrating automatic transcription into the language documentation workflow, specifically for Yongning Na, a language of Southwest China. Using Persephone, an open-source toolkit, a single-speaker speech transcription tool was trained over five hours of manually transcribed speech. The experiments found that this method can achieve a remarkably low error rate (on the order of 17%), and that automatic transcriptions were useful as a canvas for the linguist. The present report is intended for linguists with little or no knowledge of speech processing. It aims to provide insights into (i) the way the tool operates and (ii) the process of collaborating with natural language processing specialists. Practical recommendations are offered on how to anticipate the requirements of this type of technology from the early stages of data collection in the field.

E-ISSN 1934-5275

1. Introduction[1] As noted by Thieberger (2017), "the promise offered by automated speech recognition suggests that the currently time-intensive task of transcription may be greatly aided in the future by novel methods for building transcripts for many more hours of recordings than has previously been possible". This article reports on one such method: automatic phonemic transcription. Within the language documentation workflow, automatic phonemic transcription promises to allow for the creation of a "rough draft" transcription which could be useful as a canvas for the linguist, who corrects mistakes and produces the translation in collaboration with language consultants. However, as a result of technical and logistic hurdles, this new and rapidly evolving technology is not yet commonly integrated in the linguistic documentation workflow.

With a view to contributing to ongoing efforts to make language documentation and language technology meet (Blokland et al. 2015),[2] we provide a case study of using automatic phonemic transcription for the documentation of Yongning Na, a Sino-Tibetan language of Southwest China. After 20 months of fieldwork (spread over a period of 12 years, from 2006 to 2017), 14 hours of speech had been recorded, of which 5.5 hours were transcribed: 200 minutes of narratives and 130 minutes of morphotonology elicitation sessions. These data were used as training data for Persephone, an open-source toolkit.[3] A single-speaker speech transcription tool was trained over the transcribed materials, in order to perform phoneme recognition on the remaining untranscribed audio files. The error rate is low: on the order of 17% for vowels and consonants. The automatic transcriptions were found to be useful in the process of linguistic documentation, reducing the manual effort required for creating transcripts, and also allowing for new insights that might not have been gained by the linguist alone.

The Yongning Na experiment can be considered a proof of concept, confirming that technology is now mature for attempting the deployment of (single-speaker) automatic transcription software as soon as a few hours of transcribed recordings are available. This raises the question of how to replicate this success for the widest possible range of languages. Detailed technical information on the application of Persephone to two languages (Yongning Na and San Juan Quiahije Chatino, a language of Mexico) is provided in a conference paper (Adams et al. 2018) intended for a

[1]The authors are grateful to Martine Adda-Decker, Laurent Besacier, Alexandre François, Nathan Hill, Guillaume Jacques, Zihe Li (李子鹤), Liberty Lidz, Annie Rialland, Mandana Seyfeddinipur, Felix Stahlberg, Jacqueline Vaissière and Guillaume Wisniewski for precious discussions. Many thanks to Afonso Xavier Canosa Rodrigues, Eric Jackson, Johann-Mattis List, Yoram Meroz, and Stephen Morey for participating in a highly stimulating online comments session in December 2017. Many thanks to the two anonymous reviewers for comments on a draft version. Many thanks to the Language Documentation and Conservation editorial team for felicitous suggestions. This work is related to the research strand "Phonetics and Phonology" of the Paris-based LABEX "Empirical Foundations of Linguistics" (funded by the ANR/CGI).

[2]In 2017, a Special Interest Group on Under-resourced Languages (SIGUL) was set up jointly by the International Speech Communication Association (ISCA) and the European Language Resources Association (ELRA). A recent initiative specifically targeted at transcription is the Transcription Acceleration Project (TAP), initiated in 2016, which aims at identifying the workflow of linguists during transcription, and developing assistive tools to accelerate that process (project leaders: Janet Wiles and Nick Evans).

[3]The repository for the software source and documentation is https://github.com/persephone-tools/persephone.

readership of computer scientists; by contrast, the present paper is intended primarily for a readership of field linguists and language documentation specialists. The article explains what we did: the process of establishing interdisciplinary collaborations for the Yongning Na data set (§2), the method of preprocessing the materials to serve as input to the Persephone toolkit (§3), and the influence of the tool on the workflow of linguistic documentation (§4). Finally, perspectives for future work are set out (§5). Overall, the article aims to convey a feel for the usefulness of this technology, its limitations, and the requirements on the data that serves as training corpus. Practical recommendations are offered on how to anticipate the requirements of this type of technology from the early stages of data collection in the field.

2. Chronology of the work In this section we first describe our experiences, both tribulations and successes, with applying language processing tools to Yongning Na. We start with describing the process of collecting data for the language, which was done with particular care to create data that could be used with automatic speech processing down the road (§2.1). We then explain how this data set was brought to the attention of computer scientists through inclusion in a standardized database, and interaction in computer science-based venues (§2.2). Next, we describe some difficulties faced in initial attempts to use automatic transcription technology (§2.3). Finally, we describe some attempts at partnership between linguists and computer scientists, differences in the ways of thinking and priorities of these two groups that we discovered in the process of our interaction, and also (perhaps most importantly) how we overcame these differences to reach a mutually valuable relationship (§2.4).

2.1 Data collection with the prospect of automatic speech processing in mind

Yongning Na is a Sino-Tibetan language of Southwest China, spoken in and around the plain of Yongning, located in the province of Yunnan (Glottolog: YONG1270, ISO 639-3: NRU; for short, the language is called "Na" below). Resources about the language include a reference grammar (Lidz 2010), a book-length description and analysis of the tone system (Michaud 2017a), and a Na-English-Chinese-French dictionary (Michaud 2015). The total number of speakers is estimated at 47,000 in the Ethnologue database (Lewis et al. 2016), but this figure includes other languages such as Laze (Huáng 2009; Michaud & Jacques 2012), not to mention that cross-dialect variation within Na is high (Dobbs & La 2016; Ā Huì 2016), so that the number of people to whom the Na dialect of the Yongning plain is intelligible is likely to be much lower. Moreover, the figure of 47,000 speakers is identical to the "Ethnic population" figure provided by the same source (Lewis et al. 2016), but the assumption of a neat match between ethnicity and language is questionable, given that there is a gaping generational imbalance in proficiency: language shift (to Southwestern Mandarin) clearly seems under way, due to mandatory Mandarin-medium education in boarding schools starting at age 7, prevalence of Mandarin media, and socioeconomic pressure to speak Mandarin (Lidz & Michaud 2008).

Alexis Michaud's fieldwork on Na began in 2006. A phonetician by training, he aimed for high-quality audio. The main motivation was to have crisp and clear audio

signals that allow for phonetic study of fine acoustic details (in the spirit of Niebuhr & Kohler 2011; Smith & Hawkins 2012; see also Blevins 2007). Another reason for paying attention to the quality of audio recordings was the prospect of applying speech processing tools. Two technologies seemed especially interesting. One was forced alignment at the phoneme level.[4] Transcriptions of Na materials were manually time-aligned at the sentence level; addition of phoneme-level alignment would open many possibilities for semi-automated analysis of phonetic-phonological phenomena (see DiCanio et al. 2013, and the study of clicks in !Xung by Miller & Elsner 2017).

A further perspective that seemed promising consisted of creating a tool for automatic speech recognition by starting from available acoustic and language models trained on huge data sets, and adapting them to a new language (in this instance: Na) using much smaller data samples, of which only a fraction needs to be annotated (Wang et al. 2003; Bacchiani & Roark 2003; Do et al. 2011). These techniques open possibilities for performing automatic speech recognition and machine translation. Putting these technological functions together, it seemed possible to greatly improve the annotation of raw or little-annotated audio resources.

Speech recognition tools for the most richly-resourced languages are designed to perform well even on audio signals that are noisy, including voices mixed with background noise (Seltzer et al. 2013), sound with reverberation (Kinoshita et al. 2016), and overlapping speech turns (Ma & Bao 2017). But these technological feats require massive training sets: transcribed audio along with a substantial amount of textual data for language model training. By contrast, when training is done on the basis of a small amount of training data, best results are obtained when data quality is high. To train an acoustic model for phoneme recognition, it appeared important to have materials in which the pronunciation is clear, the audio signal clean, and the (hand-made) transcription faithful and consistent.

The objective of collecting low-noise, low-reverberation audio was not always achieved: in the field, cultural appropriateness is a paramount criterion, and compromises need to be made accordingly. "The importance of good recording needs to be balanced against the importance of keeping the participants at ease" (Souag 2011:66). For instance, evenings would be a good time for recording in the village of Yongning, as all is quiet after the day's work is over and farm animals go to sleep, but nowadays evenings are a time for watching television, so recording at that time of day would conflict with the hosts' family life. Recordings were therefore conducted in the daytime, when the ambiance is less quiet. Despite these compromises, the overall quality of the Na recordings is good, by self-report (pending more formalized procedures for assessing corpora: see in particular the perspectives set out by Thieberger et al. 2016). This is probably part of the reason why application of the Persephone tool for automatic transcription (the object of this article) gave good results. This raises the issue of to what extent our automatic transcription experiments for Na can be replicated

[4]Tools for automated alignment include EasyAlign (Goldman 2011), SailAlign (Katsamanis et al. 2011), and Web-MAUS (Strunk et al. 2014). The state of the art is presented in Johnson et al. (2018), who also confirm the feasability and usefulness of automated alignment in language documentation settings.

for other data sets: not all field linguists place that much emphasis on audio quality. This is an empirical question, which is being investigated by applying Persephone to a range of different fieldwork data sets (from various languages). To preview results from §3.4 (Figure 2), an encouraging finding is that, for a given amount of training data, error rates are on the same order of magnitude for Na and for Chatino, a tonal language of Mexico.

For the Na materials, all the collection and transcription work was carried out by one person (Alexis Michaud). Great care was exercised when transcribing the materials, choosing to verify transcriptions several times with the main consultant rather than cover more ground (transcribe more narratives). Since these materials constitute the empirical basis for all of the following tasks, from fundamental linguistic research to the development of multimedia teaching materials (as an example: Hirata-Edds & Herrick 2017), it seemed reasonable to invest the (considerable) amount of time it takes to produce high-quality transcriptions.[5]

The format chosen was that of the Pangloss Collection (Michailovsky et al. 2014), an open archive of "rare" languages developed since 1994 at the *Langues et Civilisations à Tradition Orale* (LACITO) research group of the French *Centre National de la Recherche Scientifique*. The Pangloss Collection uses an extremely simple hierarchical structure, encoded in XML. Primary documents are categorized as `texts` (typically: spontaneous narratives, such as traditional stories, life stories or procedural texts such as recipes)[6] or `word lists`. A `text` is made up of *sentences* made of *words* made up of *morphemes*; a `word list` is made up of *words* made up of *morphemes*.

2.2 Bringing the data to the attention of computer scientists The Na data set was brought to the attention of computer scientists through inclusion in a standardized database (the Pangloss Collection), and interaction in computer science-based venues. In principle, resources in the Pangloss Collection are easy to locate. Pangloss is hosted in a broader repository, CoCoON (for *Collection de Corpus Oraux Numériques*), which follows the metadata standard defined by the Open Language Archives Community (OLAC): the metadata are presented on the web in a form compatible with the OAI Protocol for Metadata Harvesting, and made available through OLAC's consolidated catalogue of linguistic documents. Thus, a web search for resources in Na will lead to the OLAC page listing – among other items – all the Na documents in the Pangloss Collection (http://www.language-archives.org/language/nru). There are pointers to this list of resources from other websites, such as the page about Na in the Ethnologue database (https://www.ethnologue.com/language/nru). The Pangloss Collection offers unrestricted access to the multimedia files and their annotation. Thus, data in the Pangloss Collection seem as FAIR as can be, to use an acronym that

[5]Thoughts and recommendations on "data collection, the underestimated challenge" are set out in Niebuhr & Michaud (2015).

[6]Referring to an entire document (audio file plus transcription) as a "text" is at variance with common usage; so is the use of "word list" to refer to various phonological materials. These two expressions are therefore typeset in the character style used for computer code, as `text` and `word list`, to remind the reader of their status as technical vocabulary.

summarizes key concerns for research data and tools: Findability, Accessibility, Interoperability, and Reusability (Wilkinson et al. 2016).

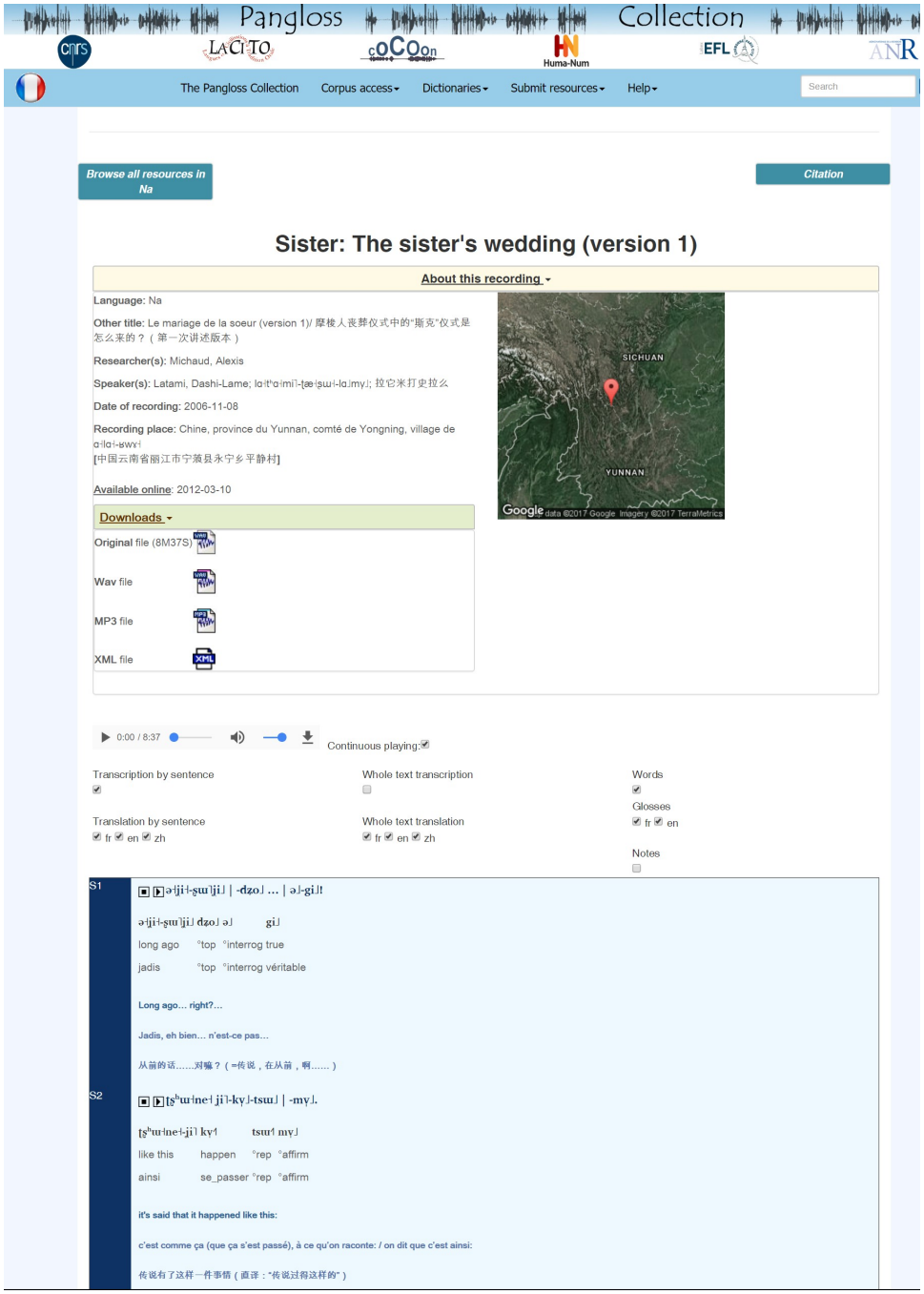

Figure 1. Web interface of the Pangloss Collection, designed for consulting documents one at a time.

In practice, however, language archives focusing on less-widely studied languages do not seem to be much used by specialists of automatic language processing at present. Bringing data to the attention of natural language processing specialists can take a little more than good indexing. Language archives have been encouraged to initiate a dialogue with the field of user-centered design (Wasson et al. 2016): a key objective is to attend better to the wishes of members of the language communities, but dialogue with computer science specialists about interface design would be another worthwhile objective, to make the archives "programmer-friendly". To take an example, from a computing perspective it is desirable to be able to download data sets in one click, not one document at a time. It is impractical for someone doing computer science research to have to scrape a website to download the full set of files available for a particular language. The current interface of the Pangloss Collection is designed for visitors who wish to consult documents one after the other, as shown in Figure 1; the "Downloads" menu offers the files for a particular document. Surprising as it may seem to the linguist who has painstakingly recorded, transcribed, annotated, and uploaded the files one by one, the requirement of writing a script to download all the files at once is enough friction to turn some computing people off. It seems advisable to meet the needs of these users by offering the option of downloading entire data sets in a few clicks.[7]

In 2012, after one year of fieldwork during which about three hours of transcribed and translated resources had been prepared, a presentation of the data set was tailored for the 2012 International Conference on Asian Language Processing (Michaud et al. 2012). The aims were: (i) to provide an illustration of the ways in which data are collected in fieldwork, drawing 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. This led to the initiation of collaborations with researchers in computer science.

2.3 First attempt: Automatic transcription tests that did not reach the level of practical usefulness This section describes some difficulties faced in initial attempts to use automatic transcription technology. Speech recognition technology requires large amounts of transcribed audio recordings to serve as training data for the acoustic model, and large amounts of texts to train the language model. Data sets in archives such as the Pangloss Collection are not sufficiently large for the training of software for speech recognition and synthesis according to the usual workflow. In view of this situation, a suggestion that has been made in the literature is to improve "the portability of speech and language technologies for multilingual applications, especially for under-resourced languages" (Besacier et al. 2014; see also Schultz & Waibel 2001). A pilot study on Na (Do et al. 2014) aimed to produce automatic transcrip-

[7]Needless to say, the situation is different when consultants or speaker communities wish to set access restrictions on certain materials. Currently, all the data sets in the Pangloss Collection are offered online in open access.

tions by combining two approaches: (i) developing a "lightweight" acoustic model of the target language ("lightweight" in the sense that it is based on a small training corpus) and (ii) applying automated speech recognition software from five national languages (English, Mandarin Chinese, French, Vietnamese, and Khmer) to identify those sounds that are acoustically similar in Na and in one or more of these five languages, on the basis of rule-of-thumb phoneme mapping: Na /p, b, t, d, k, g/ are roughly similar to French /p, b, t, d, k, g/, Na /ɯ, ɤ/ to Vietnamese /ɯ, ɤ/, Na /ɻ/ to Mandarin /ɚ/, and so on. The "lightweight" model would offer a complementary perspective: it was expected to do well on phonemes that are unattested in the five other languages, such as the syllable transcribed /hī˧/, which is nasalized throughout (about Na phonemes and phonotactics, see Michaud 2017a:447–479).

The "lightweight" acoustic model was created by using transcribed narratives in Na as training data for CMU Sphinx (https://cmusphinx.github.io/), an open-source toolkit for automatic speech recognition (Lamere et al. 2003). Tone was left out of the scope of this preliminary work, for the technical reason that it was not handled in the CMU Sphinx toolkit. This is a major shortcoming, because tone has a high functional yield in Na.

The tool was then tested on a previously untranscribed narrative by the same speaker. Errors were so numerous that the automatically generated transcription could not serve as a useful canvas for the linguist. For instance, lengthened vowels found at tone-group boundaries were often misinterpreted as sequences of two syllables, the second with an initial continuant such as /l/. After the pilot study, work came to a halt because the computer scientist on the team had the opportunity to focus on her main field of expertise (automatic translation) and stopped doing speech recognition. At that point, the prospect of arriving at a speech recognition tool that would help linguistic documentation of Na seemed very distant.

On a much more positive note, the following section (§2.4) describes attempts at partnership between linguists and computer scientists that finally proved successful. In the process of our interaction, we discovered differences in the ways of thinking and priorities of these two groups, and (most importantly) we overcame these differences to reach a mutually valuable relationship.

2.4 Finding common ground between computer scientists' interests and linguists' interests, and testing automatic transcription with Persephone Although the pilot study presented at the 4th International Workshop on Spoken Language Technologies for Under-resourced Languages remained inconclusive, it attracted the attention of an internationally leading team doing research in the field of speech recognition technology. Felix Stahlberg, then an M.A. student at the Karlsruhe Institute of Technology, asked whether he could use the Na data for research purposes. He explained[8] that the goal was not to develop a high-quality and highly tuned Automatic Speech Recognition system for the Na language, but to try to develop new methods for building an Automatic Speech Recognition system for a new language from scratch with minimal effort. The scenario was that nothing was known about the language, and the

[8]Information from e-mail exchanges is used with the authors' consent.

data set available consisted solely of audio recordings of sentences plus translations into other languages. Thus, the challenge was to automate all the following tasks: (i) establishing the phoneme inventory, (ii) generating phoneme-level alignments for the audio data, (iii) training an acoustic model, and (iv) identifying words and their pronunciation in the target language. In short, the aim was to make a language accessible for speech technology by only using audio recordings and written translations, bypassing the need for transcriptions, pronunciation dictionaries, and even phoneme set definitions. From the point of view of computer science, this ambitious objective was much more interesting than the creation of a high-quality automatic speech recognition tool for Na. Its scope was more general: attempting to make new workflows possible in computer-aided linguistic documentation. The technological challenge was much greater, because the development of automatic speech recognition tools on the basis of transcribed and annotated audio is essentially considered a solved issue, and hence belongs to *applied* computer science, rather than computer science *research*.

Two years later, Alexis Michaud wrote to ask about results and possible collaborations. The answer was that Felix Stahlberg had moved on to other topics. To the linguist, this sounded like a failure, and a sign that the approach needed to be changed: why not set reasonable goals and create useful tools, instead of setting over-ambitious goals and failing to reach them?

There is much to learn from such stories: they underline that computer scientists tend to have a different relation to research than linguists. Computer science researchers often try to push the boundaries of what is possible, even though they are almost certain that they are attempting the impossible.

> The idea is that even if a certain direction turns out to be too ambitious for the moment, there are usually still lessons to be learnt from the attempt. Progress in computer science often arises from many people trying crazy ideas and failing, but still setting the foundation for other crazy ideas one of which eventually turns out to be useful in practice. Core ideas in computer science were often premature at the time of publishing (neural networks are one famous example of that). This is probably due to the fact that our field is moving so rapidly. A new (or newly discovered) technology like neural networks can make entire lines of research almost obsolete within a few years, so people do not feel too restricted by the technical possibilities right now. I think that many linguists are quite different in that regard. The research object is the language itself, and linguists tackle this research object with well-established techniques to discover and analyze these phenomena. Computer science feels like the techniques change significantly every few years, making analyses you made a few years ago outdated.
>
> (Felix Stahlberg, personal communication to Alexis Michaud, 2017)

The third computer scientist to try his hand at Na data was Oliver Adams, and this finally led to significant practical realizations. Oliver Adams got in touch with Alexis Michaud in August 2015. Oliver's initial plan was to build on the work of Stahlberg et al. (2014) to apply alignment models to the Na corpus, and to apply automatic speech recognition technology to the audio files in order to get automatic phoneme transcriptions. We discussed how we could work in a way that would help both computer science experts and linguists make progress towards their respective objectives, without trying to make the objectives converge 100%. It is natural that the perspectives should be different. Collaboration requires slight compromises in terms of schedule, giving away time for tasks that are not a priority in one's own agenda. At the time when our collaboration started, the linguist (Alexis) was interested in attempting to build a high-quality Automatic Speech Recognition system fine-tuned for the main Na consultant. This would speed up the transcription of the entire set of audio files recorded by that speaker; it could also bring new facts to attention, for instance by highlighting variant usages, those that depart from the statistical "norm". The creation of a language-specific tool is not an exciting computer science challenge in itself, but the specific issues encountered can be interpreted by the computer science partners in the most general terms: for instance, the limited size of the Na corpus raises the challenge of limited data to train models on. Thus, addressing specific issues can yield valuable insights, resulting in improvements to generic tools, and maybe even leading to unexpected technological breakthroughs.

Over the months, Oliver and Alexis kept each other up to date on their plans and watched out for opportunities to have projects "click together": achieving the right blend of practical usefulness for language documentation, on the one hand, and on the other hand technological challenge (to make a contribution to research in computer science).

Various scenarios were considered, then abandoned. One was respeaking of the data (Sperber et al. 2013): this additional step in the workflow consists in asking a native speaker to listen to the original recording over headphones and repeat it chunk by chunk, in order to achieve greater homogeneity and cleaner audio. Respeaking is facilitated by the AIKUMA application (Bird et al. 2014; Blachon et al. 2016; Adda et al. 2016). But this was not carried out because no other speakers of Na with a comparable degree of proficiency were available to do the respeaking (as research assistants). The limited amounts of time that the main consultant can give to the linguist are best devoted to recording additional original materials and discussing linguistic issues, rather than to the mechanical task of going through a set of audio files and repeating each sentence.[9]

[9]A related concern was that materials obtained through respeaking may not constitute consistent and reliable documents from a linguistic point of view. If the "respeaker" uses a flat, unemotional voice (somewhat like conference interpreters do), the signals might be more suitable than the original for today's phonemic transcription software, but the mismatch between their intonation and that of the original may cause issues a decade or two down the line, by the stage when speech recognition systems handle not only vowels, consonants and tone, but also more and more components of intonation (phrasing, prominence, and the conveyance of attitudes and emotions). As for the option of imitating the original as closely as possible during respeaking, trying to play the part fully (including the rhythm and the expression of attitudes and emotions) requires *acting* skills, which not all language consultants possess, and the expressiveness defeats

Another plan was automatic phoneme-level time alignment of existing Na transcriptions. (The usefulness of forced alignment for phonetic studies was mentioned in §2.1.) This was not carried out because Oliver explored technology for multilingual acoustic models that bypasses the need for phoneme-level alignment in the training data: deep learning algorithms operating over stretches of audio of up to 10 seconds.

During this phase of our exchanges, it was hard for the linguist to fathom to what extent one or another of the strands of research explored by Oliver (such as Adams et al. 2015; Adams et al. 2017) would lead to achievements of practical use in language documentation. But mutual confidence was building up nonetheless, based on our shared commitment to language documentation. In the same way as some linguists feel more strongly than others about the value of language diversity, and come to identify language documentation and language description as priorities, some computer scientists consider *language processing for under-resourced languages* as their field of specialization. A computer scientist working for the first time on real data on a newly documented endangered language can be as excited as a linguist on a first trip to the field. The sense of a common goal fosters mutual interest between linguists and computer scientists, itself conducive to mutual understanding.

On April 10th, 2017, Oliver reported having trained an automatic transcription tool on the Na data, with a phoneme error rate of about 17%. Since that figure by itself was not very evocative for the linguist, Oliver and Alexis agreed to have the tool tested during the following field trip, scheduled for April–May 2017. One month later, as Alexis's fieldwork began, Oliver sent in sample transcriptions of previously untranscribed audio, produced through a phonemic transcription tool that covered tone as well as vowels and consonants. Progress in comparison to the pilot study (Do et al. 2014) was spectacular. The automatically generated transcription for the first 10 seconds of audio (the beginning of the narrative "Benevolence") is shown in (1).

(1)

ə ˧ ʝ i ˧ ʂɯ ˥ ʝ i ˩ z o ˩ n o ˧ n ɑ ˩ ʈʂʰ ɯ ˥ dʑ o ˩ z o ˩ n o ˧ l e ˧ ʐ̩ w ɤ ˩ z o ˩ n o ˧ **ɲ i ˧** b i ˩ ʈʂʰ ɯ ˧ l ɑ ˩ ɲ i ˩

The passage in (1) only contains one error: identification of the tone of the syllable highlighted in bold, /**ɲi˧**/, as Mid instead of Low. This is not a surprising error: phonetically, the fundamental frequency of the first in a sequence of L tones following a H or M tone is intermediate between that of the preceding (H or M) tone and that of the following L tones (Michaud 2017a:368).

The automatic transcriptions proved useful as a canvas for the linguist, who proofreads the transcription, as it were, and produces the translation (with the help of language consultants). The first sentence in (1) is shown in (2) after the linguist's revisions: correction of the tone of /ɲi˩/ from mid to low, adding tone breaks "|" as well as white space and other punctuation, and addition of comments and of translations in French, English and Chinese.

to a great extent the initial purpose of respeaking (to obtain audio that is more homogeneous, and hence easier to transcribe, than the original). On the issue of repetitions, see also Winter (2015) and Niebuhr & Michaud (2015).

(2)

ə˧ʲji˧-ʂɯ˥ʲji˩, | zo˩no˧, | nɑ˩ ʈʂʰɯ˥-dʑo˩, | zo˩no˥, | le˧-ʐwɤ˩, | zo˩no˧ ɲi˩-bi˩ | ʈʂʰɯ˧ lɑ˩ ɲi˩.

Comment: /-bi/ is the ADVERSATIVE.

Comment: did the consultant say /zo˩no˧ ɲi˩/ ('today'+COPULA) or /zo˩no˧ ne˩/ ('like today')? The answer is: /zo˩no˧ ɲi˩/.

Autrefois, eh bien, les Na, eh bien, ce dont je vais parler, ça s'applique aussi aujourd'hui / ça vaut aussi bien pour aujourd'hui. (Littéralement: Parler des Na autrefois, même si on l'applique à aujourd'hui / même si c'est d'aujourd'hui qu'il s'agit, c'est pareil / il en va également ainsi.)

In the old times, well, the Na…well…what I'm going to say applies to the present (=to present-day Na customs) just as well.

从前，那么，摩梭人……（我）讲的（这些），现在也一样（=我讲的虽然是从前摩梭习俗， 可是今日摩梭人的行为和做人标准也照样就是那样）。

Alexis wrote an enthusiastic 3,000-word report on May 12th, 2017 (available on a research blogging platform: Michaud 2017b). The initial success (producing a transcription that was useful to the linguist as a canvas) was all the more impressive as the first version of the tool was still fairly off-the-shelf, and there was much room for further improvement. In January 2018, the tool was enriched by adding detection of prosodic boundaries: the tone-group boundaries, indicated by a vertical bar | in transcriptions (see (2) for an example). This constitutes a breakthrough for the automatic transcription of Na prosody, as tone groups constitute all-important units in Na prosody (see Chapter 7 of Michaud (2017a) for details). Work on error analysis and improvements to the tool continues in a dialogue between the coauthors of this article.

To sum up, the exchange between the linguist and the computer scientists was often illuminating, and we have plans to continue this collaboration, which we find productive and enjoyable. Without ever meeting in person, or even hearing one another's voices, we achieved a common goal, reaping the harvest of a common choice to follow Open Science principles (open access to research data, tools and publications, open-source software, and so forth). We should acknowledge, however, that there was initially a degree of uncertainty about where we were going and how far we would go together. The conclusion that we draw from this experience is that a specific research community needs to grow at the intersection of both fields, comprising linguists who are interested in investing time to learn about natural language processing and computer scientists who want to achieve "great things with small languages" (Thieberger & Nordlinger 2006) and to do great things *for* endangered languages.

3. How automatic transcription with Persephone works The speech recognition tool presented here is named after the goddess who was abducted by Hades and must spend one half of each year in the Underworld. Which of linguistics or computer science is Hell, and which the joyful world of Spring and light? For each it's the other, of course. The name Persephone is thus intended as a tongue-in-cheek al-

lusion to the sense of estrangement and uncanniness that one can occasionally feel in interdisciplinary collaboration. Interdisciplinary work has difficulties and rewards of its own – like fieldwork, in a way, but to some fieldworkers the difficulties can be perceived as "the last straw", coming on top of a heavy work load. It is clear that linguists engaged in language documentation cannot spend half their lives learning about language processing, any more than computer scientists would want to spend half their lives reading books of linguistics. But some understanding of how the software works is nevertheless extremely useful (probably indispensible) for anyone who wishes to make use of the potential of these new technologies. "Digital Humanities drive the scholarly tradition towards a more uniform approach or an approach where the divide between two cultures – humanities on the one side, and computing on the other – is no longer constructive" (Collins et al. 2015:10; see also the classical reflections by Fillmore 1992).

This section provides information on the materials that were used, and about the preprocessing that is necessary to transform a corpus into a training set for Persephone. These pieces of information can also be read as recommendations to linguists who are interested in the creation of automatic speech recognition tools for the languages they document.

3.1 Materials The materials used for training the acoustic model were not recorded specifically for this purpose: instead, the entire set of transcribed materials available was used as training data, with the exception of two held-out subsets used for validation and testing. These materials were collected as part of classical linguistic fieldwork (as described in textbooks by Bouquiaux & Thomas 1971; Newman & Ratliff 2001; Dixon 2007). As pointed out by Woodbury (2003:47), "good corpus production is ongoing, distributed, and opportunistic", and is thus unlike scenarios in which data acquisition is tailored to meet the requirements of a certain type of speech processing.

All the materials are available online in the Pangloss Collection. The resources (recordings and annotations) are freely available for browsing and download, under a Creative Commons licence (about Creative Commons licenses in their broader legal, social, and epistemological context, see e.g. Bourcier & de Rosnay 2004:85–94). Readers are invited to check out these materials (currently at http://lacito.vjf.cnrs.fr/pangloss/corpus/list_rsc_en.php?lg=Na).

The corpus available for training and testing the acoustic model consists of 200 minutes of narratives (24 transcribed `texts`) and 130 minutes of morphotonology elicitation sessions (`word lists`). For narratives, the annotations are synchronized with the recordings at a level loosely referred to as the sentence (<S> level), with an average duration on the order of three seconds. Phonological materials (`word lists`) are made up of phrases or short sentences, and the time-alignment is based on these units, which are generally less than three seconds in length. 80 more narratives (over nine hours of recordings) await transcription, and constitute the materials for which we wished to obtain an automatic transcription.

The acoustic model is speaker-specific: it is trained exclusively on data from the main language consultant. Speaker-independent ASR (automatic speech recognition)

is much more difficult than speaker-specific ASR, even in cases where the dialect under study is relatively homogeneous. This is a point that linguists interested in experimenting with ASR tools need to take into account. Lively dialogues with overlap of speakers are richer than monologues in some respects, but the resulting corpus will be a hard nut to crack for ASR software. If the transcriptions indicate the identity of the speaker for each sentence or speaker turn, a speaker-specific acoustic model could be built for each speaker. The software could in principle be set up so as to incorporate speaker diarization (identification of who spoke when; during which intervals each speaker is active) and apply the right acoustic model for each interval, even disentwining speaker overlaps, but automated diarization remains a difficult problem (Garcia-Romero et al. 2017; Liberman 2018). Speaker-independent automatic phonemic transcription will require more development work than was necessary for the single-speaker Na data set that we used.

Finally, an important characteristic of the training data is that there should be the best possible fit between the audio and its transcription. If a word is repeated five times in a row in a narrative ("they rowed, rowed, rowed, rowed, rowed!"), the linguist may choose to write the word only twice in the transcript ("they rowed, rowed…!") because it does not appear necessary to write the same word many times: the ellipsis "…" is sufficient. But what is obvious to a human listener is not so to the software, which is trained in a purely phonetic/phonological mode, without any higher-level processing. For instance, when an acoustic model was first trained on the Na data (in April–May 2017), the audio files for phonological materials (such as nouns in carrier sentences, numeral-plus-classifier phrases and object-plus-verb combinations) contained occasional repetitions that were not indicated in the transcription. Thus, the first token in the `NounsInFrame` document (nouns recorded in a frame that brings out their phonological tone pattern) was 'This is a mountain' in Na, /ʈʂʰɯ˥ | ʁwɤ˥ ɲi˩/ (demonstrative plus target noun plus copula). In the audio, the expression is repeated: /ʈʂʰɯ˥ | ʁwɤ˥ ɲi˩. | ʈʂʰɯ˥ | ʁwɤ˥ ɲi˩./ But the transcription, intended for human readers, was simply /ʈʂʰɯ˥ | ʁwɤ˥ ɲi˩/. In view of such cases of mismatch between audio and transcription, all the phonological materials (which amounted to about one-third of the data set) were left out from the training corpus, which was thus restricted to narratives. To address this issue, the linguist spent a few hours going through all these materials (in September 2017) to adjust transcription so as to match the audio. Thereafter, the acoustic model was re-trained, using the full set of transcribed Na data as training corpus.

Some linguists may not want to take the trouble to go through their corpus to check that the transcriptions match the audio word for word: this requirement is at variance with common practice. The Na data were recorded and transcribed by a persnickety phonetician; from our experience, this is not the majority case among the corpora in language archives. But in a scenario where the training corpus is extremely small (on the order of one hour, or even less), linguists interested in using Persephone (and other speech recognition tools) should aim for exhaustive transcriptions that are faithful to the audio. Preliminary tests with another corpus (also from the Pangloss Collection) in which there are numerous mismatches between audio and phonemic

transcriptions suggest that these mismatches result in high error rates down the line. Linguists preparing transcriptions for use in automatic speech recognition can make a virtue of necessity and consider that exhaustive transcription draws attention to significant details that might otherwise be overlooked. They can also explore the option of respeaking (Sperber et al. 2013): a native speaker is asked to listen to the original recording over headphones and repeat it. This additional step in the workflow aims to ensure greater homogeneity and cleaner audio. In addition to the work on Na data, Oliver Adams also trained a phonemic recognition tool for Eastern Chatino as spoken in San Juan Quiahije, Oaxaca, Mexico (Cruz & Woodbury 2006; Cruz 2011); the Chatino corpus used for this test, described in Ćavar et al. (2016), was prepared with a view to experimenting with automatic processing, and it includes materials recorded through respeaking.

Persephone can be used to help bring transcriptions closer to the audio. The transcription tool can be run over the training dataset, finding instances where the recognition probability for the transcription is low, and flagging these for manual verification. Or the user can look at the edit distance between the predicted transcript and the manual transcript. As an example, (3) shows an excerpt from a working document produced in January 2018: a parallel view of an entire text, allowing for comparison of the linguist's transcription with the output of Persephone. This document is generated by setting aside one of the transcribed texts (in this instance: BuriedAlive2), training a model on the rest of the corpus, then applying that model to the extracted text (this is referred to technically as "cross-validation"). This parallel view is intended for error analysis (to identify aspects of the tool that can be improved), but it also brings out errors in the original transcriptions. *Ref* (3a) is the reference (the linguist's transcription) and *Hyp* (3b) is the model's best hypothesis (the automatic transcription). Glosses are provided in (4).

(3) a. *Ref:* zo˩no˥ | njæ˧˩sɯ˩kɣ̩˩ | nɑ˩ʈʂʰɯ˥ | ʈʂʰɯ˧˩ne˧˩ɟi˥ pi˩kɣ̩˩mæ˩ | ə˩gi˩ |
b. *Hyp:* zo˩no˥ | njæ˧˩sɯ˩kɣ̩˩ | nɑ˩ʈʂʰɯ˥ | ʈʂʰɯ˧˩ne˧˩ɟi˥ | pi˧˩kɣ̩˩mæ˩ | ə˩gi˩ |

(4) zo˩no˥, | njæ˧˩sɯ˩kɣ̩˩ | nɑ˩ ʈʂʰɯ˥ | ʈʂʰɯ˧˩ne˧˩ɟi˥ | pi˧˩-kɣ̩˩ mæ˩ | ə˩-gi˩ |
zo˩no˥ njæ˧˩sɯ˩kɣ̩˩ nɑ˨˩ ʈʂʰɯ˧˩ ʈʂʰɯ˧˩ne˧˩ɟi˥ pi˥ -kɣ̩˨˩ mæ˧˩ ə˩-
well 1PL.EXCL Na TOP thus to.say ABILITIVE PTCL INTERROG
gi˩
true
'Well…we Na, this is what we say, isn't it! / This is how our story goes!' (BuriedAlive2.2)

In this example, two stylistic options are open to the speaker concerning the verb 'to say' and the two following morphemes (the ABILITIVE, and a discourse particle conveying obviousness). They can be integrated to the same tone group as the previous morpheme (the adverb 'thus'), in which case their tones are all lowered to L (/pi˩-kɣ̩˩ mæ˩/) by a phonological rule (whereby all tones following H are lowered to L), as in

(3a). Or these three syllables can constitute a tone group of their own, in which case the surface tone sequence is M.L.L (/pi˦-kɣ̩˩ mæ˩/), as in (3b). (For detailed explanations about such cases, see Michaud 2017a:335–337.) In the automatic transcription, there is a tone-group boundary before the verb 'to say', and its tone is detected as Mid (˦). Returning to the audio, it seems clear that this transcription (3b) is right, and the reference transcription (3a) was wrong in this respect. The difference is acoustically small, and since both variants are well-formed, the consultant would not correct the investigator when checking the transcription. Comparison of manual transcriptions with automatically generated transcriptions allows an opportunity for fine-grained confrontation with the data.

3.2 Preprocessing To serve as a training set for an acoustic model, the transcriptions need to be preprocessed. Part of the Persephone toolkit is functionality to preprocess transcribed data in the Pangloss XML format. (A goal of ongoing improvements to the Persephone tool is to increase the range of input formats supported.) Preprocessing is done by scripts applied to the XML files containing the documents' annotation ("annotation" is used as a cover term for an audio or video file's transcription, translation, glosses, and notes). Translation, word-level glosses, and the various notes present in the annotation are left aside, only retaining the phonemic transcription and the time codes of the sentence. Thus, for sentence 2 of the narrative "The sister's wedding", only the following two elements are retained:

```
<AUDIO start="2.1473" end="3.6244"/>
<FORM>ʈʂʰɯ˦ne˦ ʝi˥-kɣ̩˩ tsɯ˩ -mɣ̩˩.</FORM>
```

Syllables are parsed into *initial consonant* and *vowel*, on the basis of an inventory of phonemic units. This allows the syllable /ʈʂʰɯ/ to be parsed into /ʈʂʰ/+/ɯ/: / ʈʂʰ/ is a trigraph for an aspirated postalveolar affricate, not a sequence of three phonological units. Parsing is computationally trivial in Na due to the language's simple syllable structure (Michaud 2017a:448–451).

The information that is fed into the Persephone toolkit is thus limited to (i) the time code of the sentence: the information that it starts at 2.15 seconds and ends at 3.62 seconds in the audio file, and (ii) the surface-phonological representation of the sentence: /ʈʂʰ ɯ ˦ n e ˦ ʝ i ˥ k ɣ̩ ˩ ts ɯ ˩ m ɣ̩ ˩/. On the basis of this information, the Persephone toolkit learns statistical patterns of association between each of the phonemic units (ʈʂʰ, ɯ, n, e, ʝ, i, k, ɣ̩, ts, m, as well as the tones) and specific properties of the acoustic signal. The quality of the acoustic model is assessed by the precision with which it identifies phonemes in a test corpus (composed of data which are not part of the training corpus, of course).

To facilitate preprocessing, an important recommendation for linguists interested in using the Persephone tool (as well as any other type of data processing) is to use pivotal formats for their data: formats that are fully consistent (logically structured text), and hence machine-readable.

All available transcribed materials were admitted into the training corpus used to produce the acoustic model for Na: we did not attempt to exclude passages that

contain non-canonical realizations. The tedious, time-consuming task of culling supposedly non-canonical passages would prove pointless: going through each of the extant transcribed documents to mark audio passages for exclusion from the training corpus, the linguist would realize that this amounts to drawing a line in the sand between "canonical" and "non-canonical" pronunciations, where no such binary distinction exists in the data. Variation is a universal of spoken language: in narratives, speaking rate (tempo), speaking style and other parameters keep changing according to communicative and expressive needs. Retaining only passages that are most neatly articulated – tending towards the *hyperarticulated*, along the continuum from *hypo-* to *hyperarticulated* (Lindblom 1990) – would not only decrease the size of an already minimal training corpus (remember that training sets used to build acoustic models for "big languages" are cherry-picked from hundreds or thousands of hours of data); it could also result in an inappropriate training corpus, which would not resemble the materials to which the software needs to be applied. Since the aim is to transcribe "spontaneous speech" (phoneticians' label for all types of speech that is not read or otherwise "scripted"), it makes sense for the training corpus to embrace the stylistic diversity of this type of speech. It was therefore considered acceptable to let some borderline materials into the training corpus. For instance, when the speaker suggests or requests corrections, those are indicated using the following conventions:

```
<pronounced text> [consultant's suggested rephrasing]
```

As an example, here is the passage of XML code corresponding to sentence 80 in the narrative "Healing":

```
<S id="HealingS080">
        <AUDIO start="360.2002" end="361.9495"/>
        <FORM>hĩ˧, | ə˧ɻi˧-ʂɯ˧ɻi˩, | ʈʂʰɯ˧ne˧-ʝi˥ | le˧-ʂɯ˧-hĩ˧, | <qʰɑ˩ne˩
        ʝi˥-bi˩> [qʰɑ˩jɤ˩ dʑoʌ!] |</FORM>
        <TRANSL xml:lang="en">People who died like that, in the old
        times, <what could one do about it?> [there were lots and
        lots!] </TRANSL>
        <TRANSL xml:lang="fr">Des gens qui mouraient comme ça (=de
        maladie), autrefois, <qu'est-ce qu'on y pouvait?> [qu'est-
        ce qu'il y en avait!] </TRANSL>
</S>
```

The materials in angle brackets correspond to the speech (the audio file) in a one-to-one way, whereas the materials in square brackets do not. The angle brackets delimit text that the consultant said at recording but wished to correct at transcription; replacement text is delimited by square brackets. For instance, in the above example, at recording the speaker said "what could one do about it?" but at transcription she wished to correct the text to "there were lots and lots!" This can concern various stretches of speech, from one phoneme (phonemic substitution) to rephrasing of several words.

During preprocessing of the data in Persephone, passages between square brackets, which are not heard on the recording, are simply deleted. As for passages between angle brackets (which match the words said in the audio recording), they were included in the training corpus. This is somewhat at variance with the principle to have a high-quality corpus. But excluding them from the training corpus would be complicated. Excluding entire sentences whenever they contain materials in angle brackets would be computationally straightforward, but would lead to the exclusion of about one-fifth of the materials, severely decreasing the size of the training corpus. Excluding only those portions of audio signal that correspond to passages in angle brackets would be complicated, because time alignment is only available for the sentence level. So the linguist would need to go back to each example to mark (manually) the passage in the audio that corresponds to the words inside angle brackets. This is precisely the type of task that we do not wish to include in the workflow: scientifically uninteresting tasks that add to the work load. The tool is devised in such a way as to require as little added effort as possible from the linguist: Persephone should take as input a set of extant documents, essentially *as is*.

A related though somewhat different situation is that of "mistaken" tokens in `word lists` (elicited lexical or morphophonological materials). Erroneous forms are especially frequent in the recordings documenting numeral-plus-classifier phrases: the consultant was asked to record long sequences of phrases such as "one year, two years, three years" and so on; in this unfamiliar task, mistakes were made in the tone patterns of quite a few of these phrases (which constitute a highly irregular corner of Na morphotonology). When the pronunciation is fumbled (making it an issue how to transcribe phonemically), the token is left untranscribed, and hence not included in the training corpus that serves as input to Persephone. But when the pronunciation is clear and unambiguous, the token is transcribed (duly recording the fact that it is mistaken, by marking it with a double dagger ‡). From a linguistic point of view, this allows for some statistical observations, such as that the proportion of mistaken realizations in the recordings correlates with the degree of morphotonological complexity: mistakes are most frequent for those forms that have the most irregular patterns (Michaud 2017a:180–181). From the point of view of the automatic transcription tool, it was considered acceptable to use these mistaken tokens for training, as if someone said "a lamb of leg" in English and this token were let into a training corpus. This illustrates the purely phonetic/phonological nature of the training task: the acoustic model associates sounds (transitions between acoustic states) with phonemes, abstracting away from any higher-level information.

Our choice to include all available materials in training the acoustic model for Na should not be taken to imply that "cherry-picking" materials for the training corpus is necessarily a bad idea. On the contrary, it is good to be able to try various combinations: speech recognition specialists select their training corpus with the same care as cider brewers combine varieties of cider apples. The key point is that the corpus is too small at present for us to afford to be selective. In future work, we would like to try a range of different combinations of training materials. It would be interesting to test whether the lowest error rates on narratives (spontaneous speech) are achieved

by using 100% spontaneous speech in the training corpus (i.e., materials that are identical in genre to the fresh materials to be transcribed), or by blending in a certain percentage of morphophonological elicitation materials (those that are classified as `word lists` in the Pangloss Collection, as opposed to `texts`), which are less variable and contain materials which (overall) are articulated with the greatest clarity. But those issues do not really arise in the resource-scarce setting of a newly documented language – the scenario which we focus on here.

3.3 An overview of the model architecture This section, the most technical in the article, gives an overview of the architecture underlying the computational model used. Unlike traditional speech recognition frameworks, which include a number of separate components working in tandem to arrive at a *word-based* transcription, this model uses a single neural network to predict the *phonemic* transcription of an utterance from the speech signal.

Data preparation: As a first step, the log filterbank features for 25-millisecond frames in the audio are computed as a preparatory step. These features can be considered a sort of spectrogram representing the energy at different frequencies throughout the recording.

For evaluation of the model training, the data must be split into three sets: training, validation, and test sets. These are typically randomly selected, either at the story level or the utterance level.

The neural network: The core of the system is a neural network that takes as input the filterbank features (our spectrogram) and produces a phonemic transcription of the entire utterance. By feeding the model many training instances (input features paired with a reference translation), the model learns to provide label predictions for each of the frames. These labels might be phonemes, tones, or orthographic characters.

The specific model used is similar to that of Graves et al. (2013). The neural network consists of several layers, which are used to learn progressively more abstract representations of the data. These layers consider information both in the frame to be labeled as well as in frames nearby and far away, thereby capturing properties of the acoustic input corresponding to morphemes, words and other linguistic context surrounding the current frame. After the final layer, the neural network outputs the probability of alternative choices of labels for each frame. A second algorithm then takes information about the probability of each label for each frame from the neural network and then compares it to the reference transcription in order to determine how the network should change in order to provide better labelings. The network is changed slightly and it is fed another example.

After the model has witnessed many utterances over and over, it begins to learn useful patterns. After the model has demonstrated its robust performance on a held-out validation set (to ensure the model is not simply learning idiosyncrasies of the

training set – a phenomenon known as "overfitting"), it can be quantitatively evaluated on this held-out test set, or applied to untranscribed data.

3.4 Evaluation: Specificities of newly documented languages A central piece of information in technical reports in the field of speech processing is the quantitative evaluation of the tool. For a phonemic transcription tool such as the one developed for Na, there are separate variables for phonemes, tones, and prosodic boundaries: phoneme error rate, tone error rate, and boundary error rate.[10] The phoneme error rate was about 17% when the tool was first created in April 2017. The tone error rate was on the order of 25%; further work since then has brought it down to about 20%. Figure 2 shows phoneme error rate and tone error rate as training data are scaled from 10 to 150 minutes for Na (circles ○). The results can be contrasted with those obtained for the San Juan Quiahije dialect of Eastern Chatino (multiplication sign ×). The phoneme system of Chatino is somewhat simpler than that of Na, and the *Phoneme error rate* curve for Na (left-hand side of Figure 2) is higher than that for Chatino, indicating that, for the same amount of training data, the phoneme error rate is lower on Chatino materials (i.e., performance is better). Conversely, Chatino tones are phonetically much less straightforward than Na tones, and accordingly, the *Tone error rate* curve for Na is below that for Chatino (right-hand side of Figure 2), indicating that, for the same amount of training data, the tone error rate is lower on Na materials.

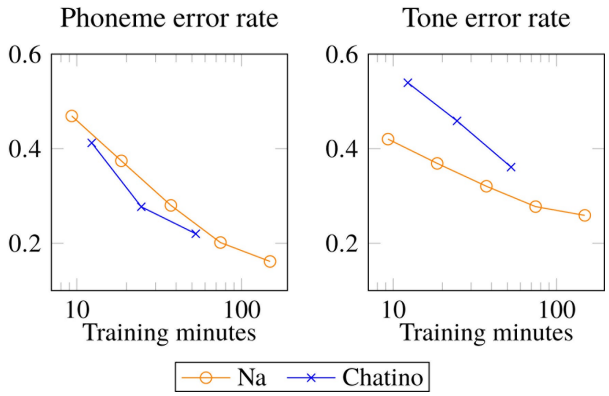

Figure 2. Phoneme error rate and tone error rate on test sets as training data are scaled from 10 to 150 minutes for Na, and from 12 to 50 minutes for Chatino. (The x axis has a logarithmic scale.)

For humans, speech recognition in purely phonemic mode is an unnatural task, because speech understanding is a holistic process of interpretation of the speaker's communicative intentions. It is not easy to determine the performance ceiling for

[10]The labels "phoneme error rate" (for vowels and consonants) and "tone error rate" (for tones) are used for convenience, despite their phonological incorrectness: the tones are contrastive phonological units, and thus *phonemic* in the same sense as vowels and consonants. The distinction is really between error rates for vowels and consonants on the one hand, and tones on the other.

a recognition system operating in phonemic mode, but in view of the plasticity of human speech, the ceiling for "clean" audio of spontaneous speech is very unlikely to be higher than 90% (or perhaps 95%). Said differently, it is unlikely that an automatic phonemic transcription system can ever get lower than a 10% (perhaps 5%?) error rate on the Na test set (or on similar data from other languages). Seen in this light, the model for Na is already close to the highest performance that may be expected of a recognition system operating in phonemic mode over spontaneous speech.

Specificities of corpora of newly documented languages need to be taken into account, however. Error rates are calculated by putting aside a sample of the texts transcribed by the linguist (not including them in the training set), for use as test set. The result of automatic transcription for the test set is then compared with the linguist's transcription, which serves as "gold standard" against which to appraise the software's performance. But this assumes that the input corpus is error-free. That is being very generous with the linguist: when a linguist transcribes materials in the field, there is bound to be a less than perfect match between the audio and the transcription. Here is an example: at the beginning of sentence 19 in the "Mountains" narrative, Persephone identified a vowel, which was absent from the linguist's transcription because it was deliberately omitted. When the narrative was recorded, the investigator, playing the role of respondent, repeated the key part of a sentence (S18) as a token of attention: /tsʰe˩hõ˩-pʰo˥/ 'eighteen head (of cattle)'. The speaker, knowing that the investigator was simply playing as respondent and had no intention of saying anything further, immediately went on, but she added (at the beginning of S19) a quick 'yes,' /ĩ˧/, as a response to the investigator's sign of attention: see (5).

(5) a. *Consultant:* tʰi˩˥, | ɖɯ˧-pʰo˧ ɖɯ˥-pʰo˩-hĩ˩ dʑo˩, | tsʰe˩hõ˩-pʰo˥ ɲi˩-ze˩ mæ˩, | ə˩-gi˩! |
tʰi˩˥ ɖɯ˧ pʰo˧ – -hĩ˥ dʑo˥ tsʰe˩hõ˩ ɲi˩ -ze˧ mæ˧
well one CLF.cattle RED.iteration REL TOP eighteen COP PFV OBVIOUSNESS
ə˩- gi˩
INTERROG true
'If one counts head by head, that's eighteen head of cattle, isn't it!'

b. *Investigator:* tsʰe˩hõ˩-pʰo˥! |
tsʰe˩hõ˩˥ pʰo˧
eighteen CLF.cattle
'Eighteen head!!' *(Intended to show appreciation of how much that is.)*

c. *Consultant:* ĩ˧, | gẹ˧-wo˥-hĩ˩ ʈʂʰɯ˩ dʑo˩ | tʰi˩˥, | tsʰe˩hõ˩-pʰo˥ ɲi˩-ze˩ mæ˩! |
ĩ˧ gẹ˧ wo˩ -hĩ˥ ʈʂʰɯ˧ dʑo˥ tʰi˩˥ tsʰe˩hõ˩ pʰo˧
yes nine CLF.teams of oxen REL FOCALIZER TOP well eighteen CLF.cattle
ɲi˩ -ze˧ mæ˧
COP PFV OBVIOUSNESS
'Yes: nine teams (of oxen), that's eighteen head of cattle, isn't it!' (sentences 18–19 of the "Mountains" narrative)

At transcription, the investigator, out of dislike for his own voice, deliberately left out the embryonic dialogue (what he had said, and the consultant's ensuing "yes"). When testing the automatic transcription tool, the "yes" that is present in the recording was faithfully transcribed; the resulting mismatch between the automatic transcription and the manual transcription that serves as "gold standard" counted as an error (raising the phoneme error rate), unfairly underrating the software's performance.

3.5 The linguist's criterion for assessment: Is the transcription useful as a canvas?

The low error rate in the automatic transcription is a piece of good news, of course. But in the proposed workflow, the automatically generated phonemic transcription is not an end product: it is used by the linguist as a canvas, somewhat in the same way as in Computer Aided Translation, where human translators correct automatic translations (Green et al. 2013). The linguist finalizes the transcription at the same time as the process of enrichment of the document begins: the linguist adds translations and various other types of annotation that suit the objectives of the documentation project, in collaboration with language consultants. So the main concern is: at which point does the automatic speech recognition tool become good enough to be useful as part of the language documentation workflow? Is the automatically generated canvas a useful starting point, or does it contain too many errors – or errors of a particularly harmful type because they are difficult to spot and correct?

It is for each linguist to test and provide an answer. The viewpoint expressed in work on Chatino by a team that includes a native speaker is that "even a relatively inaccurate speech recognition system can assist a researcher with the command of the language in fast annotation of the heritage resources" (Ćavar et al. 2016:4009). In the case of Na, the output of the acoustic models tested for the pilot study (Do et al. 2014) was too poor to be useful as a starting point for transcription. By contrast, the output of Persephone proved a pleasant and useful starting point: as soon as he began using the transcription, the linguist found that it changed the nature of the work, from "inputting from scratch" (typing the full transcription, as was done earlier) to "proofreading mode". The following section goes into some detail on the benefits of adopting the automatically generated transcription as a canvas.

4. Influence on the workflow of linguistic documentation

4.1 Revising the automatically generated transcription: Is it faster than inputting?

Revising the automatically generated transcription involves correcting the transcription and adding a sentence-level translation, as well as various comments. Transcription of the "Housebuilding2" narrative, a recording of 30 minutes, was completed using the automatically generated transcription. The first 22 minutes had been transcribed between 2014 and 2016, over three field trips whose main focus was on other topics (the study of tonal processes, and the creation of a dictionary). The ratio of transcription time to real time was about 60 to 1 (one hour's work with the consultant to transcribe and translate one minute of speech). The last eight minutes were

done in May 2017 on the basis of the automatic transcripts sent by Oliver Adams, in about seven hours of work sessions. The quantitative difference in the speed of work with the earlier method (typing the documents "from the ground up") is not striking, but it needs to be kept in mind that about half of transcription sessions is spent discussing various topics, adding new example sentences to the dictionary, and writing comments of all sorts. What matters is the quality of the exchange between investigator and consultant, and the quality of the documentation produced; from this perspective, Alexis Michaud considers use of Persephone a great success, and is a fully convinced user.

4.2 Types of errors in the automatically generated phonemic transcription Overall, errors are relatively few. Technical successes include the treatment of vowel lengthening: vowel length is not contrastive in Na, but phonetically, there are considerable differences in vowel length, contributing to conveying prominence and phrasing. For phonemic transcription purposes, lengthening needs to be overlooked, and the Persephone software successfully overlooks it.

Loanwords are a source of errors in the automatically generated phonemic transcription. In the "Benevolence" narrative there are a few loanwords, such as 'Taiwan' and 'Japan', borrowed from Southwestern Mandarin. 'Japan', pronounced /ʐɯ˩pe˧/, has a /pe/ syllable that contradicts Na phonotactics: in Na, /e/ and /i/ only contrast after dental fricatives and affricates (Michaud 2017a:457); after bilabials such as /p/, there only exists one possibility, transcribed as /pi/, and realized phonetically as [pi]~[pe]. But impressionistic listening suggests that some loanwords consistently have a [pe] pronunciation, calling for phonemicization as /pe/. Clearly, it would be asking too much of a transcription tool trained over a few hours of data to deal with these complexities: the /pe/ syllable is quasi-absent from the training set, and Persephone bases itself on the patterns to which it was exposed in the training set.

Another set of errors comes from reduced forms: weak realizations of grammatical items. This is a well-documented limitation of automatic phoneme recognition (examples concerning recognition for French are analyzed by Adda-Decker 2006:881); here are some examples in Na.

- The noun 'person' /hĩ˥/ and the relativizer suffix /-hĩ˥/ are homophonous (segmentally and tonally identical), but the latter is articulated much more weakly than the former, often without a fricative portion in the acoustic signal; accordingly, the relativizer is often recognized as /ĩ/ in automatic transcription, without an initial /h/.

- The initial consonant /ʈʂʰ/ in the demonstrative /ʈʂʰɯ˥/ is often strongly hypoarticulated, losing both aspiration and affrication; this results in its frequent transcription as a fricative /ʂ/, /ʐ̩/, or /ʑ/.

- The negation /mɤ˧/ appears as /mõ˧/ in the automatically generated transcript for Housebuilding2.290. This highlights that the vowel in that syllable is probably nasalized, and acoustically unlike the average /ɤ/ vowel for lexical words.

The extent to which a word's morphosyntactic category influences the way it is pronounced is known to be language-specific (Brunelle et al. 2015); the phonemic transcription tool indirectly reveals that this influence is considerable in Na. Cases of mismatch between phonemic transcription and acoustic realization are also present in the training data. An extreme example is the word 'really, actually': its full form is /ʈʰæ˧mi˧-ɳɯ˩/, but it is most often pronounced in a highly reduced form, as a monosyllable with a lengthened vowel and some nasalization, which can be approximated in narrow (phonetic) transcription as [ʈʰæ̃˧]. When Persephone builds an acoustic model, the mismatch between the transcription /ʈʰæ˧mi˧-ɳɯ˩/ and the phonetic realization as [ʈʰæ̃˧] is detrimental to the statistical model's accuracy, as the tool does not have any means to tease apart content words and grammatical words.

Facing this situation, the linguist may realize that /ʈʰæ˧mi˧-ɳɯ˩/ is an etymological notation rather than a synchronically appropriate phonemic transcription, and choose to change transcription to [ʈʰæ̃˧] or some other solution that is closer to the synchronic reality. An English equivalent would be transcription of *Mrs.* as *Mistress*: this is etymologically correct, but synchronically the term of address is /ˈmɪsɪz/, whereas the noun is /ˈmɪstrəs/ (~/ˈmɪstrɪs/).

Changing transcription of the Na word 'really, actually' to [ʈʰæ̃˧] in the training corpus will remove a source of noise. But such piecemeal adjustments can only deal with the most extreme cases of mismatch. A phonemic transcription tool has the limitations inherent in phonemic theory (as opposed to exemplar-based models: Johnson 2007): an acoustic model cannot handle the gap between phonemic transcription, on the one hand, and pronunciation habits for individual words, on the other hand. Adequate treatment of this issue would require a leap to a different approach to speech recognition: a full-fledged ASR (automatic speech recognition) system, based on a language model, making use of morphosyntactic information to identify *words*, instead of *phonemes*. ASR requires a considerably larger corpus than automatic phonemic transcription, "because models should estimate the probability for all possible word sequences" (Kurimo et al. 2017:961); by facilitating the creation of an enlarged corpus, automatic transcription has the potential to serve as a stepping-stone towards ASR, as discussed in §5.3.

4.3 Practical benefits for the linguist Using the automatic transcription as a starting point for manual correction was found to confer practical benefits to the linguist. In casual oral speech, there are repetitions and hesitations that can create trouble for the transcribing linguist, as they make it harder to find one's bearings when navigating the audio signal. When using an automatically generated transcription as a canvas, there can be full confidence in the linearity of transcription, because the model produces output that is faithful to the acoustic signal. The linguist's attention can therefore focus on the key tasks: understanding the text, translating, annotating, and communicating with the language consultants.

Another benefit is that inputting errors are avoided. Typographic errors are not uncommon when juggling with several keyboard layouts, some of which make use of combinations of keys. These errors can be difficult to identify down the line, when the

consultant is not available. By providing a high-accuracy first-pass automatic transcription, much of this manual data entry is entirely avoided. The linguist improves the transcription in "proofreading mode", as it were, without the distracting burden of data entry.

4.4 Psychological benefits for the linguist Linguistic fieldwork can be a fabulous human experience, broadening one's linguistic and cultural horizons in spectacular ways. But fieldwork also tends to isolate one from one's family, friends, and language community. The amount of work and the sense of responsibility can be overwhelming: if one is the only linguist working on an endangered language, one's errors may never be corrected later. Feeling that one will receive support from computing tools that build on the latest advances in information science curbs the feeling of loneliness: one is in good company among pioneers of the Digital Age, the likes of those who founded the World Wide Web not so long ago.

A love of exploration is part of the motivation of many "diversity linguists": newly described languages constitute unknown land to explore. Linguists are not unlikely to find a similar thrill in exploring new methods which take them to the frontiers of natural language processing. The psychological support is in proportion to the technology's potential: considerable.

The prestige of cutting-edge technology also matters for members of the speech community. The low levels of prestige associated with "minority" languages can act as a (semi-unconscious) deterrent for university students who have an interest in their native language. For instance, in East/Southeast Asian villages, a bright student who makes it to university level is expected to look for stable, well-paid employment after graduation. The prospect of documenting and studying languages that are on the way out, interviewing elderly people in remote villages, tends to be frowned upon: promising students reaching the M.A. stage may feel that they are under family pressure not to continue to the Ph.D. stage. Increased use of reliable, efficient, and user-friendly computer technology can enhance the social prestige of language documentation, leading to more acceptance and support of documentary vocations.

The perspective of using linguists' data as a training set for automatic phonemic transcription also sets a long-term agenda, with rewards along the way. An incentive for transcribing two or three hours of audio is that this allows for training an automatic phonemic transcription tool, facilitating work at later stages. An incentive to continue all the way to the stage when one has a dictionary and large amounts of fully transcribed and glossed texts is that this may allow for training a full-fledged ASR system (based on a language model: see §5.3) which identifies words, not just phonemes and tones, and thus allows for automatic glossing (and potentially, automatic translation). More data collection means more training data for better performance of the computer tools; this snowball effect is a powerful encouragement. A further positive consequence of this perspective is to (re-)whet the linguist's appetite for fresh documentation. Linguistic diversity consists not only in the diversity of linguistic codes, but also of the uses and potentialities of those codes (Woodbury 2003:37), so recording a wealth of data is essential. But the linguist's greediness for data can tend

to decrease as it becomes clear that a lifetime will not suffice to transcribe all of the materials. There is of course the possibility that another linguist will take up one's work later, learn the language, and continue the tasks of transcription and analysis, but this is an uncertain prospect. The existence of tools for automatic phonemic transcription (and for other tasks further down the line) is an encouragement to continue recording new data without being too concerned about the possibility that they may never be put to good use.

5. Perspectives for further work The wishes set out below are not directly relevant to the discussion of the Persephone toolkit that constitutes the core object of the present article; this section may, however, fulfill the function of highlighting the characteristics (and the limitations) of Persephone in its present state, thereby indirectly helping readers understand the tool.

5.1 Beyond mono audio signals: Multichannel audio data and other types of signals All the files in the Na training corpus have audio from a microphone placed about 50 centimeters from the speaker's mouth. Additionally, some files have a second audio channel, from a head-worn microphone. An advantage of the head-mounted microphone is that, being located close to the speaker's mouth, it captures weaker signals, spoken in a low voice. Signal processing specialists can use stereo audio for various treatments such as source separation and noise reduction. This potential of the Na data has not been exploited yet. Also, some audio files are accompanied by an electroglottographic signal. Electroglottography (Fabre 1957; Fourcin 1971) is the ultimate reference for calculating fundamental frequency; the electroglottographic signals could be used for tone recognition.

5.2 Computer implementation of morphotonology Errors for tone are more numerous than for phonemes. Tone was initially treated by Persephone in much the same way as vowels and consonants; this did not yield the best results, because the phonetic realization of tones depends a great deal on their position inside the sentence, due to intonational processes such as declination (see Michaud & Vaissière 2015, intended as a beginner-friendly review) and to morphotonological processes taking place inside tone groups. In January 2018, tone-group recognition was added to the Persephone tool for Na, greatly improving tonal recognition. To make further progress, tone will need to be modeled in a more explicit and coherent way. We would like to build on linguistic analyses of tone in Na, adding, for instance, linguistic constraints at decoding time to enforce valid tone sequences. For example, the contrasts between H and M are neutralized in tone-group-initial position, and neutralized in tone-group-final position following a L tone (Michaud 2017a).

5.3 Towards automatic glossing and translation The ultimate aim, in terms of speech processing tools, would be to arrive at a full-fledged automatic system which outputs

not only transcripts, but glossed and translated texts. Progress in automatic translation has benefits for language documentation, because "[t]he primary object of statistical translation models, bilingual aligned text, closely coincides with interlinear text, the primary artefact collected in documentary linguistics" (Bird & Chiang 2012). A dictionary of Na (Michaud 2015) in computer-readable format is available (Bonnet et al. 2017), and a few glossed texts. Some tools are specifically designed so as to sidestep the requirement of large training sets, such as LATTICELM.[11] How much progress can be made in this direction will largely depend on the extent to which collaborations between linguists and computer scientists can continue. Various pieces of software can facilitate the creation of an adequate training corpus: for instance, in the process of revising the output of Persephone, the software could guide the linguist's attention to passages identified as requiring correction (Sperber et al. 2017).

5.4 Extracting knowledge from the acoustic model Another technologically ambitious perspective for further work consists in extracting information from the acoustic model to learn more about the language's phonetic/phonological system. Intuitively it seems clear that if the machine gets the right answers (correct identification of phonemes), it "knows" something about the target language's acoustics. This knowledge seems well worth spelling out, to arrive at a characterization of phonemes in terms of their defining acoustic properties. (As an example of phonetic research aiming to arrive at acoustic characterization of phonemes: Vaissière 2011a; Vaissière 2011b.) This also applies to prosodic boundaries: prosodic boundaries in the audio signal are known to be cued by fundamental frequency, duration, phonation type, and fine detail in the articulation of vowels and consonants (Georgeton & Fougeron 2014); it would be interesting to find out which cues are used by Persephone to identify the boundaries of tone groups in Na.

Due to the nature of the statistical models, known as "neural-network models", it is not easy to look under the hood and retrieve "knowledge" from the model (in the case of Persephone, spelling out which acoustic properties are associated with which phonemes). Software based on a neural-network architecture is generally used as a black box.[12] Theoretically, however, it should be possible to devise ways to open the box. Recent research focuses on *post-hoc interpretability*: the goal is to relate what

[11]The idea behind LATTICELM (which may initially seem distinctly odd to linguists) is to learn a lexicon and language model directly from the phoneme lattices. Thus, model training is based directly on speech, instead of the usual workflow of training a language model on text. No dictionary is required either, because phoneme lattices are used without word segmentation. For further detail, readers are referred to a computer science paper: "Learning a language model from continuous speech" (Neubig et al. 2010).

[12]Paradoxically, it seems as if the tools became less and less intelligible as their performance (their predictive power) improves. Neural-network models tend to be cut off from "human knowledge", or scientific knowledge as embodied in the phonetic and linguistic literature: for instance, in the field of speech acoustics, classical work by Fant (1960) and Stevens (1998). This paradox links up with worries about de-humanizing threats associated with the broad and diverse set of tools loosely grouped under the terms "machine learning" and "artificial intelligence" (Scherer 2016; Helbing et al. 2017), and with distrust of "Digital Humanities" as "an escape from the messiness of the traditional humanities to the safety of scripting, code, or interface design" (Grusin 2014:81). In this context, we want to emphasize that experiments in "computer-aided language documentation" by no means aim to replace human expertise by computer tools, but to improve the leverage of field linguists and ultimately speed up the language documentation process (Adams 2018:1–9).

the model predicts (in the case of Persephone: the phonemes, tones, and tone-group boundaries) to input variables that are readily interpretable, and which humans can make sense of (see in particular the tutorial by Montavon et al. 2017). This has potential interest for both computer science and linguistics. Computer scientists are interested in the problem of opening the black box *per se*. Improved methods for understanding how the black box "ticks" may prove relevant for linguists, too, because the machines have a perspective that differs from that of linguists, and reflections on this perspective could bring to light new knowledge about fields such as acoustic phonetics and phonology.

6. Conclusion Automatic transcription is considered mature technology by the computer science community, but this technology remains unfamiliar to many people engaged in language documentation, so we wanted to spread the word, using our work on Na as an example. We hope that the present article, read in conjunction with the more technical presentation by Adams et al. (2018) on tests carried out on Na and Chatino data, will convey a feel for possibilities of using Persephone for new languages, contributing to ongoing efforts to harness speech recognition technology to aid linguists. Our focus here was on the "low-hanging fruit": the performance that can be achieved in automatic phonemic transcription on the basis of a few hours of carefully transcribed audio. So it appeared useful to write the present report as early as possible after the creation of the automatic transcription tool for Na, while our experience of the early stages was still fresh in our heads. We are keenly aware that exploration of the full range of scenarios for using automatic transcription in language documentation has barely begun.

Last but not least, there is the issue of how collaborations between linguists and computer scientists can be set up to deploy tools such as Persephone. The Persephone software is available as open-source code from GitHub (https://github.com/-persephone-tools/persephone) under an open-source license. But deploying such a tool in research centers focusing on language documentation is not easy because these centers often have limited computer engineering staff.

For those interested in participating in research about the use of language technology for language documentation, there is the scenario of mid- to long-term interdisciplinary collaboration between computer scientists and linguists. Interdisciplinary dialogue is interesting for all concerned. The tools' designers and developers are researchers with a strong background in computer science and an interest in linguistics. They get feedback from the linguists, ranging from analyses of the errors made by a transcription tool to copious "wish lists" for new developments. The linguists gain an additional perspective into their research topics: for instance, new linguistic insights can be gained in the process of error analysis, because automatic phonemic transcription yields a "signal-shaped" view on the audio recordings which can be compared with the linguist's "brain-shaped" transcription. The authors of the present article plan to continue this type of collaboration.

A second scenario would consist in empowering the linguists to use the tool on their own. The computer scientists in the team of authors are working on a "linguist-

friendly" interface designed to allow a simple and efficient workflow of training-recognizing-proofing. When a linguist has enough data to serve as a training set, (s)he could run Persephone by clicking a few buttons, without requiring help from a computing expert. The interface would allow the linguist to feed the training data into Persephone, and to run the software on new audio recordings to get transcriptions. However, linguists' documents tend to have formatting complexities (adjustments made for the target language, for the genres in the corpus, and for specific research aims) which make preprocessing nontrivial. For instance, if the linguist's transcriptions are orthographic, conversion to a phonemic representation will need to be carried out: input to Persephone does not have to be in a phonetic alphabet, but it needs to be designed on phonemic principles. Moreover, tests using new materials will not only involve covering more data formats at preprocessing, but also addressing new technological challenges, such as the contrastive phonation types (breathy voice, glottal constriction, etc.) encountered in East/Southeast Asia (Ferlus 1979; Gao 2015 and references therein) and in Nilotic (Remijsen et al. 2011). The development and testing of a front-end (user interface) for Persephone is currently progressing in parallel with improvements to the back-end for maximal extensibility: improving the computational core so that it deals better with the widest possible range of languages. Until plans for creating a "linguist-friendly" interface come to fruition, the help of a computer scientist remains necessary for using Persephone (a user guide for computer scientists is available from the tool's online repository).

A third scenario consists in setting up short-term collaborations between a linguist and a computer scientist. An M.A. student in computer science (or even an undergraduate) could do the job of applying Persephone to a new language. The linguist would need to understand the requirements for training data; ideally, (s)he would have automatic processing in view when planning and conducting data collection and transcription. The computer science student would do the preprocessing for the data set, run the Persephone tool to train an acoustic model, and teach the linguist how to re-train the model as more data become available. It is difficult to indicate how much time is required, as this depends on data and technical competence. In the best cases, a week's work may be enough; two to four weeks if the data require significant preprocessing. That time could inflate to much longer if the linguist needs to revise manual annotations. From the perspective of computer science students and their supervisors, what is at stake in such internships is that differences between languages in terms of phoneme inventory, phonotactic complexity, prosodic structure, morphological complexity, and other dimensions are certain to raise interesting issues and to open perspectives for improvements to the tool.

Finally, a fourth scenario would involve the creation of a computer science unit for the deployment of automatic speech recognition tools for language documentation on an international scale. Linguists would entrust their data to the computer science specialists, whose task would consist in (i) explaining to linguists how to get the materials in shape for the machine training, and, where indispensible, advising linguists to record additional data, (ii) training an optimal acoustic model for the target language, and (iii) improving the tools in a dialogue with linguists. A team

of as few as one or two permanent members could get a lot of great research and development done, in a partnership with language documentation stakeholders (linguists, language communities, language archives, etc.).[13] The transcription, translation, glossing, and archiving of endangered-language multimedia data matters no less to language sciences than the publication of research monographs, and could justify similar efforts to set up innovative, collaborative infrastructures. This workflow is probably ideal from the fieldworkers' point of view, as it minimizes the additional work on their part (so they can focus on their core tasks) and maximizes the tools' efficiency (as the software is operated by experts, who are interested in improving the tools to address the widest range of linguistic challenges). Linguists need to learn more about computer science, but they also need technical support. This scenario requires strong institutional support: there needs to be an organizational turn to accompany the "computational turn", or uptake of state-of-the-art tools for collecting, annotating, and archiving language data may continue to be regrettably slow. Researchers in phonetics laboratories often collaborate closely with engineers; it does not seem unreasonable to consider that, in this day and age, research centers where language documentation is conducted have similar needs to phonetics laboratories in terms of information-technology staff.

References

Ā, Huì 阿慧. 2016. 永宁摩梭话"le+V+se"结构的声调实验分析 – 阿拉瓦和舍垮的–对比 [An experimental analysis of the tones of "le+V+se" in the Yongning Mosuo language: A comparison between Alawa and Shekua dialects]. Kunming: Yunnan University. Master's thesis.

Adams, Oliver. 2017. Automatic understanding of unwritten languages. Melbourne: The University of Melbourne. Doctoral dissertation.

Adams, Oliver, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird & Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. *Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018)*, 3356–3365. Miyazaki, Japan. https://halshs.archives-ouvertes.fr/halshs-01709648.

Adams, Oliver, Adam Makarucha, Graham Neubig, Steven Bird & Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 1, 937–947. Stroudsburg, PA: The Association for Computational Linguistics.

[13]Possible business models include collaboration between several institutions and funding agencies, or crowdfunding: joint support from a great number of institutions (typically, universities), following the model of the publishing house Language Science Press (http://langsci-press.org/), a sort of "cooperative" which was successfully set up within a few years and currently operates with only two permanent staff.

Adams, Oliver, Graham Neubig, Trevor Cohn & Steven Bird. 2015. Inducing bilingual lexicons from small quantities of sentence-aligned phonemic transcriptions. In Federico, Marcello, Sebastian Stüker & Jan Niehues (eds.), *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2015)*. Da Nang, Vietnam.

Adda, Gilles, Sebastian Stüker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, et al. 2016. Breaking the unwritten language barrier: The BULB Project. *Procedia Computer Science* 81. 8–14. https://doi.org/10.1016/j.procs.2016.04.023.

Adda-Decker, Martine. 2006. De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux. In *Actes des XXVIe Journées d'Etude de la Parole*, 389–400. Dinard.

Bacchiani, M. & Brian Roark. 2003. Unsupervised language model adaptation. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 224–227. Hong Kong: IEEE.

Besacier, Laurent, Etienne Barnard, Alexey Karpov & Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication* 56. 85–100.

Bird, Steven & David Chiang. 2012. Machine translation for language preservation. In Kay, Martin & Christian Boitet (eds.), *Proceedings of COLING 2012: Posters*, 125–134. Mumbai: The COLING 2012 Organizing Committee. http://aclweb.org/anthology/C/C12/C12-2013.pdf.

Bird, Steven, Florian R. Hanke, Oliver Adams & Haejoong Lee. 2014. Aikuma: A mobile app for collaborative language documentation. In Good, Jeff, Julia Hirschberg & Owen Rambow (eds.), *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 1–5. Baltimore: The Association for Computational Linguistics.

Blachon, David, Elodie Gauthier, Laurent Besacier, Guy-Noël Kouarata, Martine Adda-Decker & Annie Rialland. 2016. Parallel speech collection for under-resourced language studies using the LIG-AIKUMA mobile device app. *Procedia Computer Science* 81. 61–66.

Blevins, Juliette. 2007. Endangered sound patterns: Three perspectives on theory and description. *Language Documentation & Conservation* 1(1). 1–16. http://hdl.handle.net/10125/1721.

Blokland, Rogier, Marina Fedina, Ciprian Gerstenberger, Niko Partanen, Michael Rießler & Joshua Wilbur. 2015. Language documentation meets language technology. In Pirinen, Tommi, Francis M. Tyers & Trond Trosterud (eds.), *Proceedings of the First International Workshop on Computational Linguistics for Uralic Languages*, 8–18. Tromsø: UiT The Arctic University of Norway. doi:10.7557/scs.2015.2.

Bonnet, Rémy, Céline Buret, Alexandre François, Benjamin Galliot, Séverine Guillaume, Guillaume Jacques, Aimée Lahaussois, Boyd Michailovsky & Alexis Michaud. 2017. Vers des ressources électroniques interconnectées: Lexica, les dic-

tionnaires de la collection Pangloss. In *Actes des 9èmes Journées Internationales de la Linguistique de corpus*, 48–51. Grenoble.

Bouquiaux, Luc & Jacqueline Thomas. 1971. *Enquête et description des langues à tradition orale. Volume I: l'enquête de terrain et l'analyse grammaticale*. 2nd edition 1976. Paris: Société d'Études Linguistiques et Anthropologiques de France.

Bourcier, Danièle & Melanie Dulong de Rosnay. 2004. *International commons at the digital age*. (Droit et Technologies). Paris: Romillat.

Brunelle, Marc, Daryl Chow & Thụy Nhã Uyên Nguyễn. 2015. Effects of lexical frequency and lexical category on the duration of Vietnamese syllables. In *Proceedings of ICPhS XVIII*. Glasgow.

Ćavar, Małgorzata, Damir Cavar & Hilaria Cruz. 2016. Endangered language documentation: Bootstrapping a Chatino speech corpus, forced aligner, ASR. In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, et al. (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 4004–4011. Portorož, Slovenia: European Language Resources Association.

Collins, Sandra, Natalie Harrower, Dag Trygve Truslew Haug, Beat Immenhauser, Gerhard Lauer, Tito Orlandi, Laurent Romary & Eveline Wandl-Vogt. 2015. Going digital: Creating change in the Humanities. Berlin: ALLEA.

Cruz, Emiliana. 2011. Phonology, tone and the functions of tone in San Juan Quiahije Chatino. Austin: University of Texas at Austin. Doctoral dissertation. http://hdl.handle.net/2152/ETD-UT-2011-08-4280.

Cruz, Emiliana & Tony Woodbury. 2006. El sandhi de los tonos en el Chatino de Quiahije. In *Las Memorias del Congreso de Idiomas Indígenas de Latinoamérica-II*. Austin. http://www-ailla.lib.utexas.edu/site/cilla2/ECruzWoodbury_CILLA2-_sandhi.pdf.

DiCanio, Christian, Hosung Nam, Douglas H. Whalen, H. Timothy Bunnell, Jonathan D. Amith & Rey Castillo García. 2013. Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment. *Journal of the Acoustical Society of America* 134(3). 2235–2246.

Dixon, Robert M. 2007. Field linguistics: A minor manual. *Sprachtypologie und Universalienforschung* 60(1). 12–31.

Do, Thi Ngoc Diep, Eric Castelli & Laurent Besacier. 2011. Mining parallel data from comparable corpora via triangulation. In Dong, Minghui, Bin Ma & Tien Ping Tan (eds.), *Proceedings of the International Conference on Asian Language Processing (IALP 2011)*, 185–188. Penang, Malaysia: IEEE.

Do, Thi Ngoc Diep, Alexis Michaud & Eric Castelli. 2014. Towards the automatic processing of Yongning Na (Sino-Tibetan): Developing a "light" acoustic model of the target language and testing "heavyweight" models from five national languages. In *Proceedings of the 4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2014)*, 153–160. St Petersburg: SPIIRAS. http://halshs.archives-ouvertes.fr/halshs-00980431/.

Dobbs, Roselle & Mingqing La. 2016. The two-level tonal system of Lataddi Narua. *Linguistics of the Tibeto-Burman Area* 39(1). 67–104.

Fabre, Philippe. 1957. Un procédé électrique percutané d'inscription de l'accolement glottique au cours de la phonation: Glottographie de haute fréquence. *Bulletin de l'Académie Nationale de Médecine* 141. 66–69.

Fant, Gunnar. 1960. *Acoustic theory of speech production, with calculations based on X-ray studies of Russian articulations*. The Hague: Mouton.

Ferlus, Michel. 1979. Formation des registres et mutations consonantiques dans les langues mon-khmer. *Mon-Khmer Studies* 8. 1–76.

Fillmore, Charles J. 1992. Corpus linguistics or computer-aided armchair linguistics. In Svartvik, Jan (ed.), *Directions in corpus linguistics: Proceedings of Nobel Symposium* 82, 35–60. Berlin: De Gruyter Mouton.

Fourcin, Adrian. 1971. First applications of a new laryngograph. *Medical and Biological Illustration* 21. 172–182.

Gao, Jiayin. 2015. Interdependence between tones, segments and phonation types in Shanghai Chinese. Paris: Université Paris 3-Sorbonne Nouvelle. Doctoral dissertation.

Garcia-Romero, Daniel, David Snyder, Gregory Sell, Daniel Povey & Alan McCree. 2017. Speaker diarization using deep neural network embeddings. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4930–4934. New Orleans: IEEE.

Georgeton, Laurianne & Cécile Fougeron. 2014. Domain-initial strengthening on French vowels and phonological contrasts: Evidence from lip articulation and spectral variation. *Journal of Phonetics* 46. 128–146.

Goldman, Jean-Philippe. 2011. EasyAlign: An automatic phonetic alignment tool under Praat. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*, 3233–3236. Florence.

Graves, Alex, Abdel-rahman Mohamed & Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, 6645–6649. Vancouver: IEEE.

Green, Spence, Jeffrey Heer & Christopher D. Manning. 2013. The efficacy of human post-editing for language translation. In Grinter, Rebecca, Thomas Rodden, Paul Aoki, Ed Cutrell, Robin Jeffries & Gary Olson (eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 439–448. Paris: ACM.

Grusin, Richard. 2014. The dark side of Digital Humanities: Dispatches from two recent MLA conventions. *Differences* 25(1). 79–92.

Helbing, Dirk, Bruno S. Frey, Gerd Gigerenzer, Ernst Hafen, Michael Hagner, Yvonne Hofstetter, Jeroen van den Hoven, Roberto V. Zicari & Andrej Zwitter. 2017. Will democracy survive Big Data and Artificial Intelligence? *Scientific American*.

Hirata-Edds, Tracy & Dylan Herrick. 2017. Building tone resources for second language learners from phonetic documentation: Cherokee examples. *Language Documentation & Conservation* 11. 289–304. http://hdl.handle.net/10125/24737.

Huáng, Bùfán 黄布凡. 2009. 木里水田话概况 (A survey of Muli Shuitian). *Journal of Sino-Tibetan Linguistics* 汉藏语学报 3. 30–55.

Johnson, Keith. 2007. Decisions and mechanisms in exemplar-based phonology. In Speeter Beddor, Patrice, Maria-Josep Solé & Manjari Ohala (eds.), *Experimental approaches to phonology*, 25–40. Oxford: Oxford University Press.

Johnson, Lisa M., Marianna Di Paolo & Adrian Bell. 2018. Forced alignment for understudied language varieties: Testing Prosodylab-Aligner with Tongan data. *Language Documentation & Conservation* 12. 80–123. http://hdl.handle.net/10125/24763.

Katsamanis, A., M. Black, P. Georgiou, Louis Goldstein & S. Narayanan. 2011. SailAlign: Robust long speech-text alignment. In *Proceedings of the Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*. Philadelphia.

Kinoshita, Keisuke, Marc Delcroix, Sharon Gannot, Emanuël AP Habets, Reinhold Haeb-Umbach, Walter Kellermann, Volker Leutnant, Roland Maas, Tomohiro Nakatani & Bhiksha Raj. 2016. A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP Journal on Advances in Signal Processing* 2016(1). 7.

Kurimo, Mikko, Seppo Enarvi, Ottokar Tilk, Matti Varjokallio, André Mansikkaniemi & Tanel Alumäe. 2017. Modeling under-resourced languages for speech recognition. *Language Resources and Evaluation* 51(4). 961–987.

Lamere, Paul, Philip Kwok, William Walker, Evandro B. Gouvêa, Rita Singh, Bhiksha Raj & Peter Wolf. 2003. Design of the CMU sphinx-4 decoder. In *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH 2003 – INTERSPEECH 2003)*. Geneva.

Lewis, M. Paul, Gary F. Simons & Charles D. Fennig (eds.). 2016. *Languages of China: An Ethnologue country report* 19th edition. Dallas: SIL International. http://www.ethnologue.com/.

Liberman, Mark. 2018. The first DIHARD speech diarization challenge. *Language Log*. http://languagelog.ldc.upenn.edu/nll/?p=36699.

Lidz, Liberty. 2010. A descriptive grammar of Yongning Na (Mosuo). Austin: University of Texas. Doctoral dissertation.

Lidz, Liberty & Alexis Michaud. 2008. Yongning Na (Mosuo): Language documentation in the Sino-Tibetan borderland. Presented at the International Conference on Sino-Tibetan Languages and Linguistics (ICSTLL 41), London, Sept. 18–21, 2008.

Lindblom, B. 1990. Explaining phonetic variation: A sketch of the H&H theory. In Hardcastle, W.J. & Alain Marchal (eds.), *Speech production and speech modelling*, 403–439. Dordrecht: Kluwer.

Ma, Yong & Changchun Bao. 2017. 基于高层信息特征的重叠语音检测 (Overlapping speech detection using high-level information features). *Journal of Tsinghua University (Science and Technology)* 57(1). 79–83.

Michailovsky, Boyd, Martine Mazaudon, Alexis Michaud, Séverine Guillaume, Alexandre François & Evangelia Adamou. 2014. Documenting and researching endangered languages: The Pangloss Collection. *Language Documentation & Conservation* 8. 119–135. http://hdl.handle.net/10125/4621.

Michaud, Alexis. 2015. Na-English-Chinese dictionary. https://halshs.archives-ouvertes.fr/halshs-01204638.

Michaud, Alexis. 2017a. *Tone in Yongning Na: Lexical tones and morphotonology*. (Studies in Diversity Linguistics 13). Berlin: Language Science Press. http://langsci-press.org/catalog/book/109.

Michaud, Alexis. 2017b. Speech recognition for newly documented languages: Highly encouraging tests using automatically generated phonemic transcription of Yongning Na audio recordings. *HimalCo - Himalayan Corpora*. https://himalco.hypotheses.org/285.

Michaud, Alexis, Andrew Hardie, Séverine Guillaume & Martine Toda. 2012. Combining documentation and research: Ongoing work on an endangered language. In Deyi, Xiong, Eric Castelli, Dong Minghui & Pham Thi Ngoc Yen (eds.), *Proceedings of IALP 2012 (2012 International Conference on Asian Language Processing)*, 169–172. Hanoi, Vietnam: MICA Institute, Hanoi University of Science and Technology.

Michaud, Alexis & Guillaume Jacques. 2012. The phonology of Laze: Phonemic analysis, syllabic inventory, and a short word list. Yuyanxue Luncong 语言学论丛 45. 196–230.

Michaud, Alexis & Jacqueline Vaissière. 2015. Tone and intonation: Introductory notes and practical recommendations. *KALIPHO – Kieler Arbeiten zur Linguistik und Phonetik* 3. 43–80.

Miller, Amanda & Micha Elsner. 2017. Click reduction in fluent speech: A semi-automated analysis of Mangetti Dune !Xung. In Arppe, Antti, Jeff Good, Mans Hulden, Jordan Lachler, Alexis Palmer & Lane Schwartz (eds.), *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 107–115. Honolulu: Association for Computational Linguistics.

Montavon, Grégoire, Wojciech Samek & Klaus-Robert Müller. 2017. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* 73. 1–15.

Neubig, Graham, Masato Mimura, Shinsuke Mori & Tatsuya Kawahara. 2010. Learning a language model from continuous speech. In *Proceedings of the Eleventh Annual Conference of the International Speech Communication Association (Interspeech 2010)*, 1053–1056. Baixas, France: International Speech Communication Association.

Newman, Paul & Martha Ratliff. 2001. *Linguistic fieldwork*. Cambridge: Cambridge University Press.

Niebuhr, Oliver & Klaus J. Kohler. 2011. Perception of phonetic detail in the identification of highly reduced words. *Journal of Phonetics* 39(3). 319–329.

Niebuhr, Oliver & Alexis Michaud. 2015. Speech data acquisition: The underestimated challenge. *KALIPHO – Kieler Arbeiten zur Linguistik und Phonetik* 3. 1–42.

Remijsen, Bert, Otto G. Ayoker & Timothy Mills. 2011. Shilluk. *Journal of the International Phonetic Association* 41(1). 111–125.

Scherer, Matthew U. 2016. Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harvard Journal of Law & Technology* 29(2). 354–400. doi:10.2139/ssrn.2609777.

Schultz, Tanja & A. Waibel. 2001. Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication* 35. 31–51.

Seltzer, Michael L., Dong Yu & Yongqiang Wang. 2013. An investigation of deep neural networks for noise robust speech recognition. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, 7398–7402. Vancouver: IEEE.

Smith, Rachel & Sarah Hawkins. 2012. Production and perception of speaker-specific phonetic detail at word boundaries. *Journal of Phonetics* 40(2). 213–233.

Souag, Lameen. 2011. Review of: Claire Bowern. 2008. Linguistic fieldwork: A practical guide. *Language Documentation & Conservation* 5. 66–68. http://hdl.handle.net/10125/4489.

Sperber, Matthias, Graham Neubig, Christian Fügen, Satoshi Nakamura & Alex Waibel. 2013. Efficient speech transcription through respeaking. In *Proceedings of Interspeech 2013*, 1087–1091. Lyon: ISCA.

Sperber, Matthias, Graham Neubig, Jan Niehues, Satoshi Nakamura & Alex Waibel. 2017. Transcribing against time. *Speech Communication* 93. 20–30. doi:10.1016/j.specom.2017.07.006.

Stahlberg, Felix, Tim Schlippe, Stephan Vogel & Tanja Schultz. 2014. Towards automatic speech recognition without pronunciation dictionary, transcribed speech and text resources in the target language using cross-lingual word-to-phoneme alignment. In *Proceedings of the International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2014)*.

Stevens, Kenneth. 1998. *Acoustic phonetics*. Cambridge: MIT Press.

Strunk, Jan, Florian Schiel & Frank Seifart. 2014. Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 3940–3947. Reykjavik: European Language Resources Association (ELRA).

Thieberger, Nick. 2017. LD&C possibilities for the next decade. *Language Documentation & Conservation* 11. 1–4. http://hdl.handle.net/10125/24722.

Thieberger, Nick, Anna Margetts, Stephen Morey & Simon Musgrave. 2016. Assessing annotated corpora as research output. *Australian Journal of Linguistics* 36(1). 1–21. doi:10.1080/07268602.2016.1109428.

Thieberger, Nick & Rachel Nordlinger. 2006. Doing great things with small languages (Australian Research Council grant DP0984419). https://arts.unimelb.edu.au/soll/research/past-research-projects/great-things-small-languages.

Vaissière, Jacqueline. 2011a. On the acoustic and perceptual characterization of reference vowels in a cross-language perspective. In *Proceedings of ICPhS XVII*. Hong Kong.

Vaissière, Jacqueline. 2011b. Proposals for a representation of sounds based on their main acoustico-perceptual properties. In Hume, Elizabeth, John Goldsmith & W. Leo Wetzels (eds.), *Tones and features*, 306–330. Berlin: De Gruyter Mouton.

Wang, Zhirong, Tanja Schultz & A. Waibel. 2003. Comparison of acoustic model adaptation techniques on non-native speech. In *Proceedings of the 2013 IEEE In-*

ternational Conference on Acoustics, Speech and Signal Processing (ICASSP 2003), 540–543. Hong Kong: IEEE.
Wasson, Christina, Gary Holton & Heather S. Roth. 2016. Bringing user-centered design to the field of language archives. *Language Documentation & Conservation* 10. 641–681. http://hdl.handle.net/10125/24721.
Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3. 160018.
Winter, Bodo. 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*. Glasgow: The University of Glasgow.
Woodbury, Tony. 2003. Defining documentary linguistics. In Austin, Peter (ed.), *Language documentation and description*, vol. 1, 35–51. London: School of African and Oriental Studies.

Alexis Michaud
michaud.cnrs@gmail.com
https://orcid.org/0000-0003-1165-2680

Oliver Adams
oliver.adams@gmail.com
https://orcid.org/0000-0002-3622-5119

Trevor Anthony Cohn
trevor.cohn@unimelb.edu.au
https://orcid.org/0000-0003-4363-1673

Graham Neubig
gneubig@cs.cmu.edu

Séverine Guillaume
severine.guillaume@cnrs.fr
https://orcid.org/0000-0003-1772-2600