Large-scale text collection for unwritten languages

Florian R. Hanke and Steven Bird
Department of Computing and Information Systems, University of Melbourne
florian.hanke@gmail.com, sbird@unimelb.edu.au

Abstract

Existing methods for collecting texts from endangered languages are not creating the quantity of data that is needed for corpus studies and natural language processing tasks. This is because the process of transcribing and translating from audio recordings is too onerous. A more effective method, we argue, is to involve local speakers in the field location, using an audio-only translation interface that is portable and easy to use. We present encouraging early results of an experimental investigation of the efficiency of creating translations using this method, and report on the quality of the resulting content.

1 Introduction

Language documentation aims to "provide a comprehensive record of the linguistic practices characteristic of a given speech community" (Himmelmann, 1998). In a typical language documentation workflow, a linguistic event is recorded, then metadata is added concerning participants, location, language, and so forth. Later, the recording is transcribed, glossed at the word or morpheme level, and then a translation is provided. Not all of these activities occur in the field: usually recording, metadata capture, and some transcription work take precedence over word-level glosses and phrasal translations (Thieberger, 2011).

Woodbury (2007) argues that for an archive entry to be analysable for a future linguist, multiple kinds of translations are needed, for example audio recordings of UN-style simultaneous oral translations, or sentence by sentence translations. The typical workflow of documentary linguistics does not produce the amount of data required for large-scale corpus-based analysis of the language once it is no longer spoken (Abney and Bird, 2010). In addition, the typical workflow necessitates the creation of transcriptions before any annotations can be made, for example in ELAN,[1] a popular software tool for linguistic annotation (Berez, 2007).

We propose to add a new path to this workflow (see Figure 1) to facilitate crowdsourcing of translations, whether by local (typically village-based) speakers or by geographically distributed speakers from the diaspora (Reiman, 2010; Bird, 2010; Zaidan and Callison-Burch, 2011). To this end, we have developed mobile phone software with easy-to-use interfaces for collecting oral translations.

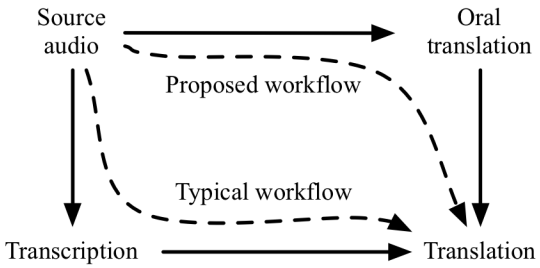

Figure 1: Current and proposed workflows for early oral annotation.

In this paper, we focus on a single activity, oral translation, via one of the mentioned interfaces, a "phonecall" interface. We have developed software that runs on mobile phones, and which has been successfully deployed in off-grid indigenous village settings. It can be used by linguists and native speakers to rapidly collect a substantial amount of high-quality, time-aligned bilingual audio. We report on an experimental investigation of the efficiency of creating translations, and the quality of the translated content.

[1] http://tla.mpi.nl/tools/tla-tools/elan/

International Joint Conference on Natural Language Processing, pages 1134–1138,
Nagoya, Japan, 14-18 October 2013.

2 Oral translation using Aikuma

Oral translation is the process of listening to a segment of audio in a source language and spontaneously producing a spoken translation in a second language. Typically, a body of recordings has already been collected in the (unwritten) source language, and the content of these recordings is to be made accessible to speakers of a more widely spoken language.

Aikuma is an Android application which supports the recording of audio sources, along with phrase-by-phrase oral translation.[2] Aikuma aims to make the recording of high quality translations a simple and natural process. To achieve naturalness, we adopt the metaphor of a phone call. The process does not use the touch screen or any buttons, but relies exclusively on audio and proximity sensors for control.

For example, let us assume we have an original recording of someone telling a story. As soon as this user holds the phone up to his or her ear, the original recording will start playing. This is achieved by using the proximity sensor present in all Android phones. The recording will continue to play as long as the user holds the phone up to their ear. At any moment during the recording, the user is free to speak their translation of what they have heard. The phone is continually monitoring the microphone input and as soon as the user starts speaking, the phone stops playback and begins recording. This enables a wide range of translations: UN-style oral free, phrase-by-phrase, or literal translations (Woodbury, 2007). When the user stops speaking for two seconds, the phone stops recording. It then rewinds the source recording by 650 ms, to ensure that the user does not miss any speech that overlapped with the segment boundary. Finally, it resumes playing the next part of the source. The process is repeated until the end of the source.

The phone's storage now contains the source audio file, along with the translation file and a mapping file. The translation file contains the concatenated recordings of the oral translations. The mapping file specifies how each segment of oral translation corresponds to the source audio. Users can listen to the original, or the translation, or interleaved playback of the original with the translation.

To evaluate our approach, we performed an experiment, then made improvements, then performed a second experiment. Both experiments were conducted in March 2013 in the Interface Design Laboratory at the University of Melbourne.

3 Experiment 1

3.1 Subjects and Materials

The participants of experiment 1 were seven Brazilian university students, aged 19 to 31. All had received four years of instruction in English as a second language.

The following procedure was carried out. Participants were given a one-minute demonstration of the Aikuma app. Then they were free to try it for up to two minutes on a test recording. We used low-end Huawei phones with a touchscreen. As an original source recording, we used an interview of Brazilian Tom Jobim, dating from the 1980s.[3] The participants then used the interface to translate the original source recording.

3.2 Results

The efficiency of the system was surprisingly high: on average, a translation of the 6:19 min long original required 6:38 min. Total length ranged from 12:05 min to 14:31 min, with an average of 12:57 min, a factor of 2.07 times the original's length.

This was a far lower duration than we expected, as the provided translations included mid-speech pauses, speech disfluencies and repaired utterances. It also included the 2 second pause detection times of the system, roughly adding 2-3 minutes to the translation. We could not reasonably expect the translation to be similar in size to the original.

Regarding quality, while we are aware that BLEU scores (Papineni et al., 2001) cannot be used for evaluating the absolute quality, we nevertheless tried to get an impression of the quality by running a single-reference BLEU against the translations.

3.3 Problems and improvements

What caused the short durations and low translation scores?

During the experiment, we noted that participants were struggling to translate parts of the interview. After transcribing and analysing the translations (cf. section 5.2), we discovered that many sentences were simply not translated at all, or only

[2] http://github.com/langtech/aikuma

[3] http://www.youtube.com/watch?v=iEofKzw7ZUg

partially translated. Out of 85 sentences, the participants on average had not fully translated 36.3 sentences (Table 1).

Using our observations of participants and their BLEU scores, we analysed the problems with the approach. The original recording quality was too low, leading to many missing sentences. This in turn resulted in very low BLEU scores.

Particip.	A	B	C	D	E	F	G
Missing	36	32	35	38	41	40	32
BLEU	6.6	11.3	7.5	9.1	13.5	11.8	10.9

Table 1: Missing sentences and BLEU scores.

The participants remarked that hearing the interview for the first time was distracting: Jobim was a popular Brazilian musician who had a gift for storytelling. Participants simply got carried away by the story itself.

This feedback resulted in the following improvements. To mitigate problems with the missing context, we added an additional step to the procedure: before translating, participants would listen once to the entire recording. To avoid problems with poor audio quality, we decided to use a more recent recording which was of perfect audio quality and upgrade to slightly more expensive, but still entry-level HTC phones.[4]

4 Experiment 2

4.1 Subjects and Materials

Ten native speakers of Brazilian Portuguese, aged 20 to 32, from all areas of Brazil participated. One of the participants was a professional interpreter with a NATII accreditation.[5]

All had achieved a TOEFL score of at least 90 points,[6] the requirement to study at the University of Melbourne. They had at least four years of English lessons in high school. Most had only arrived recently and had two or more months of recent experience in speaking the language. Some have had more intensive exposure to English.

We used a high quality audio recording of a recent interview by Celsinho Cotrim with the first Brazilian female judge, Luislinda Valois[7]. The speakers are from the state of Bahia, speak relatively clearly and with a more neutral accent. This is representative of a realistic recording from the Aikuma system, data from a language archive, or a linguistic field recording.

The recording is spoken in Portuguese, has a total duration of 5:06 min and contains 90 sentences or phrases and 806 words in total. We selected this interview for various reasons: the content is of moderate complexity; the recording contains dialogue; the two speakers are not of the same gender, making it easier to distinguish their voices.

Some expressions can not be translated literally but have to be translated idiomatically. One example of this is the Brazilian idiom: 'Meus pais nunca abriram o mão do educação', literally 'My parents would never open the hand of education', which means 'My parents would never drop education'.

4.2 Method

Given the feedback in the pilot experiments, we improved the process as follows:

For the training run, we used the same recording as used in the pilot experiment. We also demonstrated the newly introduced concept of removing the phone to stop playback if they needed to rehear a particular segment. Removing the phone and putting it back on the ear would rewind the recording to the beginning of the last segment. During training, as soon as they seemed to have grasped the concept of translation, we stopped the training. On average, this took 1-2 minutes.

To provide context for the following translation, we asked the participants to listen to the entire original recording once, without performing any translation. None of the translators noted having problems understanding the audio or content.

We then instructed the participants to translate the original carefully without omitting any content. In case they encountered words where they did not know the English translation, we asked them to simply repeat the Portuguese word. If the English translation for a given word or expression was not known, we asked them to paraphrase. We instructed them to decide themselves where to segment the text, we specifically did not tell them to segment on sentence boundaries. Participants were then asked to translate the original recording a second time.

[4] Priced at US$ 160.

[5] http://www.naati.com.au/accreditation.html

[6] TOEFL scores, http://www.ets.org/toefl/ibt/scores

[7] http://www.youtube.com/watch?v=oYK6uoyNGqA

5 Results

We obtained 20 oral English translations, two per speaker, of the same Brazilian Portuguese source recording. All of these translations were carefully transcribed. From the audio recordings, we extracted a few key metrics from the recordings themselves.

5.1 Efficiency

To measure efficiency, we calculated the total time it took to listen to the original plus the time to translate it twice. The silences that are necessary for the interface to work are included in the duration of the task. Translation of the 5:06 min original took on average 15:39 min, a factor of roughly 3. In total, including preparation and listening, the process took slightly more than 35 minutes on average (Figure 2).

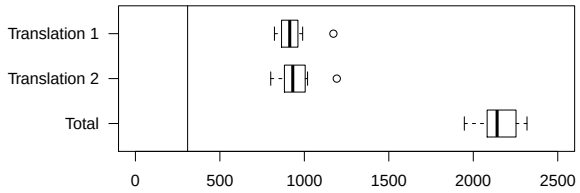

Figure 2: Durations of the recordings (s). The vertical line denotes the duration of the original.

5.2 Quality

5.2.1 Preparation of recorded translations

To analyse and compare the translations, we prepared transcriptions of the original Brazilian Portuguese audio and all 20 English translations.

To evaluate translation quality, we used the Human-targeted Translation Error Rate (HTER), which requires a comparison between the resulting hypothesis translations and a number of reference translations (Snover et al., 2006). For this purpose, we prepared a parallel translation (Table 2). Due to speech disfluencies and repaired utterances, such as 'um' and 'green I mean blue' (Levelt, 1983), the process itself, and varying sentence segmentation the resulting transcriptions needed to be processed further for evaluation.

5.2.2 HTER

HTER uses human annotators to create a specific targeted reference sentence for each translated hypothesis sentence. Each hypothesis sentence is edited by a bilingual editor until it is fluent and has the same meaning as the source sentence.[8] As a targeted reference sentence has to be created for every hypothesis sentence, HTER is very resource intensive.

Brazilian Portuguese	*English*
Eu sou mulher mais feliz do mundo	I am the happiest woman in the world
No importa por que	It does not matter why
Mas eu acho que sou	But I think I am

Table 2: Example reference translation.

As we did not have the necessary resources to perform this analysis on all recordings (100 hours for 22 translations), we selected three example translations based on their BLEU scores (not mentioned) the lowest and highest result, and the expert's translation.

For each of the selected translations, a bilingual annotator created targeted reference sentences. Then, we performed a TER on the six translations (Table 3).

Participant	*Best*	*Worst*	*Expert*
Run 1	0.16	0.23	0.18
Run 2	0.10	0.24	0.08

Table 3: Participant HTER scores in runs 1 and 2.

We found that the changes in HTER scores between translation runs agree with the BLEU score changes: both the "best" user and the expert receive an improved (numerically lower) score, while the "worst" user does not.

Regarding absolute scores, the expert comes out ahead of the "best" user. We assume that this is a result of the expert's sophisticated use of English which was not present in the BLEU references.

6 Conclusions

We have investigated a new method for rapid translation of spoken language materials. The method can be used by amateur translators and offers a faster method for preserving endangered language data while there is still time. Our experiments indicate that the resulting translations are of sufficient quality to be useful in downstream NLP tasks.

[8] The standard HTER process only uses an untargeted reference in the target language to enable editing by monolingual editors.

Acknowledgments

We gratefully acknowledge support of the Swiss National Science Foundation (Hanke) and the Australian Research Council (Bird).

References

Steven Abney and Steven Bird. 2010. The Human Language Project: building a universal corpus of the world's languages. In *Proceedings of the 48th Meeting of the Association for Computational Linguistics*, pages 88–97. Association for Computational Linguistics.

Andrea Berez. 2007. Review of EUDICO linguistic annotator (ELAN). *Language Documentation & Conservation*, 1:283–289.

Steven Bird. 2010. A scalable method for preserving oral literature from small languages. In *Proceedings of the 12th International Conference on Asia-Pacific Digital Libraries*, pages 5–14.

Nikolaus Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36:1–34.

Willem Levelt. 1983. Monitoring and self-repair in speech. *Cognition*, 14:41–104.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. *Science*, 22176:1–10.

Will Reiman. 2010. Basic oral language documentation. *Language Documentation and Conservation*, 4:254–268.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the Association for Machine Translation in the Americas*, pages 223–231.

Nicholas Thieberger. 2011. *The Oxford handbook of linguistic fieldwork*. Oxford University Press.

Anthony Woodbury. 2007. On thick translation in linguistic documentation. *Language Documentation and Description*, 4:120–135.

Omar F. Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1220–1229.