Using public domain resources and off-the-shelf tools to produce high-quality multimedia texts

Manny Rayner
FTI/TIM
University of Geneva, Switzerland
Emmanuel.Rayner@unige.ch

Belinda Chiera
The University of South Australia
Adelaide, Australia
Belinda.Chiera@unisa.edu.au

Cathy Chua
Independent scholar
Adelaide, Australia
cathyc@pioneerbooks.com.au

Abstract

In the turbulent world of 2022, where mass population movements due to war and disaster are becoming increasingly common, language skills are more relevant than ever. People who wish to achieve a high level of proficiency when learning a new language benefit from reading literary texts, but many learners find this a challenging hurdle. Annotating texts with integrated audio and translations is a popular way to try and make them easier to approach. However, doing this automatically with TTS and machine translation engines produces unengaging results, while human annotation is slow and expensive. Here, we present a method that uses simple scripts and readily available computational resources for speech recognition and sentence alignment to combine public-domain resources from sites like Gutenberg and LibriVox into high-quality annotated multimedia versions of literary texts. Initial results with French texts of up to 80K words in length are promising, with audio/text word error rates under 0.25% and audio/translation word error rates around 1%, producing results that are usable after only minimal postediting.

1 Introduction and motivation

In Anthony Powell's semi-autobiographical WW II novel *The Soldier's Art*, the narrator mentions to his division commander that he can read Balzac in the original French, and is surprised by the response: General Liddament immediately tells him to apply for a job in Military Intelligence. Since 1943, there have of course been some important changes. English has firmly established itself as the world language, and language technology has made enormous progress, but the fundamentals are the same. People with strong language skills are still prized by the security services, who see little prospect of replacing them with Google Translate and related AI/ML-based technologies. Large-scale movements of linguistic communities, driven by war, climate change, and economic disaster, are making these skills increasingly relevant, not just to Intelligence but to many related sectors including immigration, law enforcement and social services. Learning to read complex texts is an essential component in acquiring high level language skills. Duolingo and similar gamified platforms are a popular way to get started with a new language and reach low intermediate level, but they will not give the large vocabulary and grasp of idiom that comes from extensive reading.

Benchmarks for language skills are competency in reading, writing, listening and speaking. A simple but effective technology for supporting the development of reading skills, widely used at least since the days of the Roman Empire (Dickey, 2016), is the bilingual text: the text is divided into segments, each one paired with a gloss/translation in the annotation language. More recently, Reading While Listening (RWL; Woodall, 2010; Isozaki, 2014; Chang and Millett, 2014; Friedland et al., 2017; Pellicer-Sánchez et al., 2018; Schwieter and Benati, 2019) simultaneously supports the development of reading and listening skills. As put forward in Krashen's seminal Input Hypothesis and Reading Hypothesis (e.g. Krashen, 1982, 1989, 2004), reading as a language acquisition technique works best where the learner is presented with comprehensible text in a low-stress situation. This is the basic rationale behind both bilingual texts and RWL.

Although RWL studies support the idea that enjoyment is key and that literature is an answer (Woodall, Chang, Isozaki), there are major obsta-

cles to the implementation. Expense — Woodall's study involved copies of hard copy book and audio for the class; lack of resources — Chang could only reference short news items of video plus transcript online; and variety — anecdotally, Lee (2019) gives a detailed example of what we all know intuitively; it is offputting to have to read something we do not enjoy. Added to this, the process is an often less than ideal user experience, for example, constant rewinding of audio.

Online learning environments are an obvious way to resolve the sorts of problems we see in such studies. There are now many platforms that provide functionality which includes bilingual texts, RWL, and additional features: we will call these "multimodal documents". Examples include the Microsoft Azure Immersive Reader[1], LingQ[2], Learning With Texts[3], the Perseus Digital Library's Scaife viewer[4] and Clilstore[5]. The most common strategy for providing audio is to create it using a Text To Speech (TTS) engine; the most common strategy for including translations or glosses is to integrate machine translation engines and/or electronic dictionaries.

A striking example of this approach is the Azure Immersive Reader. The upside of the platform is immediately apparent. For a large number of reading languages and annotation languages, the learner only has to point the tool to the text they wish to read, and they are immediately presented with a version containing TTS audio and machine-translation generated glosses in the annotation language. Unfortunately, after even an hour of using the tool, the downside is equally apparent; the quality of the annotations is quite low. Many learners will find it fatiguing to listen to TTS audio or read MT-engine generated glosses for more than a short time. A couple of recent studies have systematically compared TTS-generated and human-recorded audio for this kind of document (Akhlaghi et al., 2021, 2022a). For the languages where TTS does best, teachers and native speakers rate it as comparable with non-professional human audio from the point of view of pedagogical adequacy; but even non-professional human voices are rated as much more natural and much pleasanter to listen to. However, though human-created annotations produce multimodal texts of substantially higher quality, the time and effort required to create them is considerable.

In this paper, we explore a possible compromise between the competing alternatives of creating multimodal documents by automatic and human annotation. There is a great deal of high-quality public domain literary content available for free download, in both text and audio form; well known sites include Gutenberg[6] and LibriVox[7]. Given a source-language text, source-language audio, and a target-language text, it is in principle possible to perform automatic or semi-automatic alignment to create an annotated multimedia document.

The question is how well the idea works in practice: what tools are needed, how high the error rates are, and how much manual cleaning up has to be be done afterwards. When we started the work described here, we were in fact fairly pessimistic. In particular, descriptions of the process used to generate the widely used LibriSpeech corpus (Panayotov et al., 2015) suggested to us that the error rates for audio alignment of literary texts would be quite high, maybe between 3 and 5 percent. Another moderately recent paper (Xu et al., 2015) suggested to us that the task of performing translation alignment on literary texts was also challenging. It seemed reasonable to assume that performing both tasks at once would be harder than performing either one separately.

The experiments we present here, carried out using the open source LARA platform[8], suggest that the task is much more tractable than we had originally believed. Work is still at an early stage, but we now think it reasonable to hope that, for many literary texts, error rates of 1 percent or lower can be achieved using readily available off-the-shelf tools to perform speech recognition and translation alignment, with the outputs from these tools combined using straightforward methods.

The rest of the paper is organised as follows. Section 2 briefly presents LARA. Section 3 describes the alignment method, and Section 4 our initial experiments. The final section concludes and suggests further directions.

[1] https://azure.microsoft.com/en-us/services/immersive-reader/
[2] https://www.lingq.com/
[3] https://sourceforge.net/projects/learning-with-texts/
[4] https://scaife.perseus.org/
[5] http://multidict.net/clilstore/
[6] https://www.gutenberg.org/
[7] https://librivox.org/
[8] https://www.unige.ch/callector/lara

2 LARA

LARA (Akhlaghi et al., 2019; Bédi et al., 2020; Zuckerman et al., 2021; Akhlaghi et al., 2022b) is a platform for producing annotated multimodal texts, under development by an international consortium since 2018. Texts can include a variety of annotations, including audio, translations, concordances, interactive images and video; links to many such texts can be found on the LARA examples page[9]. For the purposes of the current paper, the only features that will be of interest are audio and translation annotations attached to text segments.

LARA is a good platform for doing this kind of experiment, since it is open source, supports many languages, and produces attractive results which can immediately be posted on the web. In § 4, we provide links to several examples of LARA documents created using the methods described here.

3 Multimedia documents by alignment

We describe a simple method that combines data produced by readily available online resources to add annotations to a text document. The methods were implemented in Python inside LARA but use no special properties of the framework. We assume that the input consists of a) a text in the reading language, b) a translation of the text in the annotation language, and c) an audio version of the text in the reading language. The desired output is a version of the text in the reading language, segmented into units (typically sentence-length or a bit larger) each of which is associated with a translation in the annotation language and an audio file. Table 1 in the next section includes links to examples.

We assume the existence of the following third-party resources:

Splitting on silences A tool that can take an audio file and split it into segments separated by silences of a designed minimum length and loudness contrast.

Speech recognition A tool that can take an audio file and return a (generally more or less inaccurate) text transcription.

Translation alignment A tool that can take a large text and a translation, and convert them into an ordered sequence of aligned units typically of around sentence size.

For these experiments, we used ffmpeg[10] for splitting on silences, Google Cloud Speech-to-Text[11] for speech recognition, and YouAlign[12] for sentence alignment. The processing steps are as follows:[13]

1. **Resources:** Start with a) source-language text, b) annotation-language text, c) source-language audio.

2. **Translation alignment:** Send the source-language and annotation-language text files to the sentence aligner, to create two parallel sentence-segmented corpora.

3. **Source segmented by translation alignment:** Add markings to the source corpus showing the breaks corresponding to the translation alignment.

4. **Split on silences:** Use the split-on-silences tool to divide up the audio corpus, choosing thresholds that make typical pieces a bit smaller than sentences. In practice it is quick to find such thresholds.

5. **Speech recognition:** Send the pieces of audio generated by the previous step to the speech recogniser.

6. **Make double-aligned text:** Use a beam search to align the sequence of recognition results against the text.[14] Add markings to the source corpus showing the breaks corresponding to the audio alignment. The result is a text that is segmented both by translation alignment and by audio alignment.

7. **Post-process double-aligned text:** Post-process the source corpus, iteratively applying a small set of transformations that reduce differences between the translation alignment and the audio alignment. Most importantly, if a translation alignment marker and an audio alignment marker are separated

[9] https://www.unige.ch/callector/lara-content

[10] https://www.ffmpeg.org/

[11] https://cloud.google.com/speech-to-text

[12] https://youalign.com/

[13] The appendix to this paper gives details on how to obtain and use the code.

[14] In these experiments, the beam width used was 80 tokens.

by text which does not include a word, move this text to the other side of the earlier marker.

8. **Make joint aligned text:** Segment the source text by breaking at the points where the two types of segmentation markers agree. In each segment of the jointly segmented corpus produced by the previous step, concatenate the component audio segments from the audio segmentation and the component translation segments from the translation segmentation.

The result of the above series of operations gives the final annotated corpus. Obviously there is no guarantee of success: in the worst case, there will only be one segment. In practice, however, we have found that the joint segmentation is fine-grained enough that it appears quite useful.

In the next section, we will give examples of what happens with substantial texts. Figure 1 illustrates the processing flow for a passage taken from one of these.

4 Initial experiments

Table 1 summarises the results of initial experiments. We present the texts used, the metrics, and the results, and discuss their significance.

4.1 Texts

We used four French texts with accompanying audio and English translations: Rimbaud's *Les poètes de sept ans* (long poem), Maupassant's *La parure* (short story), Flaubert's *Un cœur simple* (novella), and Proust's *Combray* (novel). All four are well known pieces of French literature. The first three often appear as course reading in advanced French courses; the fourth is generally regarded as difficult even at this level. Our rationale for choosing it was curiosity to try a worst case scenario. If the method gave credible results on something as challenging as Proust (very long text, very long sentences, very complex grammatical structure, very large vocabulary), we postulated that it would probably work on many other texts too. Audio was in all cases taken from the LitteratureAudio site[15], and text from Gutenberg.

4.2 Metrics

The specific task we study in this paper is not well known in the literature, though it has points of contact with well known tasks. We adapt standard metrics in as conservative a way as possible.

We take the hopefully uncontroversial point of view that the quality of a triple alignment of the kind we are interested in here, simultaneous alignment of audio, text and translation, depends on three things: a) the quality of the audio/source-text alignment, b) the quality of the audio/translation alignment, and c) the quality of the segmentation. (a) and (b) are obvious. (c) is slightly less obvious, but a moment's reflection shows that it is essential. In the trivial alignment where the whole text becomes one segment, the error rates for (a) and (b) are zero, but this is clearly a very bad alignment. We need some measure of the extent to which the segmentation divides the text into appropriate pieces.

For (a), audio/source-text alignment, our metric is simple word error rate (WER). For each segment, we compare the aligned text with the reference text and compute WER in the usual way. For (b), audio/translation alignment, we decided that WER was in this case also the most appropriate metric. It is not a common metric for translation quality, but the specific properties of the task suggested to us that metrics like BLEU, METEOR etc (Papineni et al., 2002; Banerjee and Lavie, 2005) would work much less well as error rates are very low, and we are producing translations by the unusual method of extracting segments of an existing translation. It seemed logical to use a metric which measures how many of the correct words had been extracted: in practice, we found that it was virtually always the case that the correct match could be identified.

The least obvious metric is the one for (c). After reviewing the relevant literature, we decided to use the *boundary similarity* metric of (Fournier, 2013), which returns a number between 0 and 1 measuring the similarity of a given segmentation to a gold standard segmentation. As described in the 2013 paper, boundary similarity is the result of substantial work correcting and improving previous segmentation metrics. It has been used by several studies since then (e.g. Özmen et al., 2014; Shaw, 2015; dos Reis Mota, 2019), and is implemented in a readily available Python package.[16]

For the texts used, we created reference segmentations by comparing the text and translation, dividing them into minimal units where there was intuitively a clear text/translation alignment. In

[15] https://www.litteratureaudio.com/

[16] https://pypi.org/project/segeval/

1 (a). SOURCE LANGUAGE TEXT

```
Une porte s'ouvrait sur le soir; à la lampe
On le voyait, là-haut qui râlait sur la rampe,
Sous un golfe de jour pendant du toit.  L'été
Surtout, vaincu, stupide, il était entêté
À se renfermer dans la fraîcheur des latrines:
Il pensait là, tranquille et livrant ses narines.
```

1 (b). ANNOTATION LANGUAGE TEXT

```
A doorway open to evening:  by the light
You'd see him, high up, groaning on the railing
Under a void of light hung from the roof.  In summer,
Especially, vanquished, stupefied, stubborn,
He'd shut himself in the toilet's coolness:
He could think in peace there, sacrificing his nostrils.
```

2. TRANSLATION ALIGNMENT

```
Une porte s'ouvrait sur le soir; à la lampe ->
A doorway open to evening:  by the light

On le voyait, là-haut qui râlait sur la rampe, ->
You'd see him, high up, groaning on the railing

Sous un golfe de jour pendant du toit.  ->
Under a void of light hung from the roof.

L'été -> In summer,

Surtout, vaincu, stupide, il était entêté ->
Especially, vanquished, stupefied, stubborn,

À se renfermer dans la fraîcheur des latrines:  ->
He'd shut himself in the toilet's coolness:

Il pensait là, tranquille et livrant ses narines.  ->
He could think in peace there, sacrificing his nostrils.
```

3. SOURCE TEXT SEGMENTED BY TRANSLATION ALIGNMENT

```
//Une porte s'ouvrait sur le soir; à la lampe//
On le voyait, là-haut qui râlait sur la rampe,
//Sous un golfe de jour pendant du toit.  //L'été//
Surtout, vaincu, stupide, il était entêté//
À se renfermer dans la fraîcheur des latrines:
//Il pensait là, tranquille et livrant ses narines.//
```

Figure 1: Example of processing (passage from *Les poètes de sept ans*). Source text in black, translated text in blue, LARA markup in red. Double slashes (//) mark segments in the translation alignment. [Continued on next page]

5. RECOGNITION RESULTS FOR SPLIT AUDIO FILES

```
"une porte s'ouvrait sur le soir", "à la lampe on le voyait là au
pire aller sur la rampe sous un golf 2 jours pendant du toit", "l'été
surtout", "vaincu stupide", "il était temps tu étais à se renfermer
dans la fraîcheur des latrines", "il pensa est là tranquille", "et
livrant ses narines"
```

6. DOUBLE-ALIGNED TEXT (BEFORE POSTPROCESSING)

```
//||Une porte s'ouvrait sur le soir; ||à la lampe//
On le voyait, là-haut qui râlait sur la rampe,
//Sous un golfe de jour pendant du toit.  //||L'été//
Surtout, ||vaincu, stupide, ||il était entêté//
À se renfermer dans la fraîcheur des latrines:
//||Il pensait là, tranquille ||et livrant ses narines.
//||
```

7. DOUBLE-ALIGNED TEXT (AFTER POSTPROCESSING)

```
Une porte s'ouvrait sur le soir;|| à la lampe//
On le voyait, là-haut qui râlait sur la rampe,
//Sous un golfe de jour pendant du toit.|||// L'été||//
Surtout, vaincu, stupide,|| il était entêté//
À se renfermer dans la fraîcheur des latrines:||//
Il pensait là, tranquille|| et livrant ses narines.|||//
```

8. JOINT ALIGNED TEXT

```
Une porte s'ouvrait sur le soir; à la lampe
On le voyait, là-haut qui râlait sur la rampe,
Sous un golfe de jour pendant du toit.|| L'été||
Surtout, vaincu, stupide, il était entêté
À se renfermer dans la fraîcheur des latrines:||
Il pensait là, tranquille et livrant ses narines.||
```

Figure 1: [Continued from previous page] Example of processing (passage from *Les poètes de sept ans*). Source text in black, translated text in blue, LARA markup in red. Double slashes (`//`) mark segments in the translation alignment. Double vertical bars (`||`) mark segments in the audio alignment and the reconciled alignment.

Text	**Text length**		**Seg lengths (Wds)**				**Error rates (%)**				**Links**	
	Wds	**Hrs**	**Splt**	**Tr-Al**	**J-Al**	**Ref**	**Rec**	**Seg**	**Txt**	**Tr**	**Raw**	**Ed**
Rimbaud	535	0:04	8.6	7.4	12.6	11.7	27.5	7.1	0.8	0.8	☞	☞
Maupassant	2853	0:17	12.7	12.1	15.7	12.8	16.8	18.3	0.2	0.2	☞	☞
Flaubert	11730	1:37	8.7	17.9	18.6	14.0	18.1	24.9	0.0	1.1	☞	☞
Proust	78283	7:52	19.9	45.5	53.7	34.0	23.5	36.7	0.0	0.9	☞	☞

Table 1: Examples of annotated texts produced. "Rimbaud" = *Les poètes de sept ans*, "Maupassant" = *La parure*, "Flaubert" = *Un cœur simple*, "Proust" = *Combray*, **Text length/Wds** = length of source text in words, **Text length/Hrs** = length of source audio in hours, **Seg lengths/Splt** = average lengths of segments produced by splitting on silences, **Seg lengths/Tr-Al** = average lengths of segments produced by translation alignment, **Seg lengths/J-Al** = average lengths of segments produced by joint alignment, **Seg lengths/Ref** = average lengths of segments in gold standard segmentation, **Error rates/Rec** = speech recognition word error rate, **Error rates/Seg** = 1 – `segeval` boundary similarity score, **Error rates/Txt** = joint alignment word error rate for source text, **Error rates/Tr** = joint alignment word error rate for translations, **Link/Raw** = link to final LARA document without postediting, **Link/Ed** = link to final LARA document with postediting. LARA documents should be viewed in Chrome or Firefox.

practice, reference segments are almost always either sentences or parts of sentences delimited by punctuation marks like semi-colons, colons, dashes or parentheses.

To summarise, the quality of a given alignment is given by a triple of numbers between 0 and 1: the WER for audio/text and audio/translation alignment, and the boundary similarity score for the segmentation. It would ideally be good to reduce this to a single number, but as yet it is not clear to us how to do so effectively.

4.3 Results

We processed all four texts through the pipeline described in §3 and manually annotated the results.[17] Annotation on each text was performed as follows. A script converted the final aligned version into a form where each segment was presented in an editable form where the source text and translation appeared under an audio control. The annotator, a native English speaker with a good knowledge of French, listened to the audio and then corrected the audio and translations if they failed to match[18]. For over 90% of the segments, no correction was needed. For nearly all of the remainder, the correction was to move text either to the preceding or the following segment. The annotator also added the gold standard segmentation information. When annotation was complete, a second script was used to calculate error rates and other statistics:

Seg length/Splt: Average length, in words, of segments produced by splitting on silences.

Seg length/Tr-Al: Average length, in words, of segments produced by translation alignment.

Seg length/J-Al: Average length, in words, of segments produced by reconciliation of translation alignment and audio alignment.

Error rate/Rec: Speech recognition word error rate.

Error rate/Seg: Segmentation word error rate, defined as 1.0 minus the boundary similarity score produced by the `segeval` package.

Error rate/Txt: Word error rate for source text segments produced by reconciliation of translation alignment and audio alignment.

Error rate/Tr Word error rate for translation text segments produced by reconciliation of translation alignment and audio alignment.

Finally, we post-edited the resulting multimodal texts as follows. First, we ran each text through a script which applied the corrections to text and translations given by the manual annotations described at the beginning of this section. Second, we made a small number of layout changes to break out titles as separate segments (this allows LARA to add a table of contents in the longer texts), and to divide the text into pages. The last two columns of Table 1 contrast raw and post-edited versions.

4.4 Discussion

Table 1 gives an impression of how well the alignment method works on representative texts ranging in length from a few hundred words to nearly a hundred thousand words. We look at the three components of the metric in turn.

First, audio alignment has worked very well. Looking at the column **Error rates/Txt**, we see that WER is under 1% for all four texts, and under 0.25% for the three longest ones. It is noteworthy that the good result comes despite quite high word error rates, typically on the order of 20%, in the speech recognition (column **Error rates/Rec**). The recognition WER may be misleading, since French has many silent letters, resulting in an abnormally high proportion of homophones; thus the recogniser may for example recognise *grands* (“large”, plural) when the reference word is *grand* (“large”, singular). Since the matching algorithm is character-based rather than word-based, this usually makes no difference; however, changing to word-based matching only degraded performance very slightly. We need to investigate the issues further using a larger sample of texts.

Looking at the column **Error rates/Tr**, we see that translation alignment has also worked quite well, though substantially less well than audio alignment; error rates are around 1%. Examination of translation errors shows that they always result from errors in the third-party translation alignment software. Our impression is that this commercial tool has been optimised for speed rather than accuracy, and that lower error rates are possible.

[17] We have also processed other texts, including a second Proust novel. We will present the results when we have finished annotating the data. Anecdotally, the quality is similar to that obtained in the examples given.

[18] We had hoped to use two annotators, in order to obtain inter-rater reliability figures, but were unable to find a second person willing to take on this demanding task at short notice. We will address the issue in future work.

Je me demandais quelle heure il pouvait être;|| j'entendais le sifflement des trains qui, plus ou moins éloigné, comme le chant d'un oiseau dans une forêt, relevant les distances, me décrivait l'étendue de la campagne déserte où le voyageur se hâte vers la station prochaine;|| et le petit chemin qu'il suit va être gravé dans son souvenir par l'excitation qu'il doit à des lieux nouveaux, à des actes inaccoutumés, à la causerie récente et aux adieux sous la lampe étrangère qui le suivent encore dans le silence de la nuit, à la douceur prochaine du retour.|| J'appuyais tendrement mes joues contre les belles joues de l'oreiller qui, pleines et fraîches, sont comme les joues de notre enfance.

Figure 2: Passage from *Combray* illustrating problems with segmentation, LARA markup in red. Double bars (| |) show segment boundaries from the gold standard segmentation. Only the one in bold (**| |**) is found by the alignment pipeline.

By far the least satisfactory result is the segmentation (column **Error rates/Seg**). The error rate, defined as 1 minus the boundary similarity score, varies considerably across the texts, increasing as the texts become more complex and reaching 36% for the very challenging Proust text. This corresponds to quite often feeling that the segments produced are too long: most commonly, a suboptimal segment consists of two sentences which the aligner has failed to split apart, or a long sentence which has not been divided at semi-colons. Figure 2 illustrates. Comparing the columns **Segment lengths/Tr-Al** and **Segment lengths/Ref** makes it clear that, with the translation aligner used in these experiments, it is impossible to attain a good segmentation score, since the segments produced by the translation aligner are already substantially longer than the gold standard segments.

5 Summary and further directions

The decreasing stability of the world means advanced language skills are of correspondingly greater importance. Acquisition of these skills, in particular large vocabularies, requires extensive reading of complex texts. Many learners find this a difficult step; multimodal texts, which include integrated audio and translation, both smooth the transition and help keep the learner's reading and listening skills in sync. We have described an implemented method for creating high-quality multimodal texts from existing online resources and presented encouraging initial results on representative French texts.

When we started, we were far from certain that automatic alignment methods would do well for this task. Based on the results of the LibriSpeech project (Panayotov et al., 2015) and the literary sentence alignment studies from (Xu et al., 2015) and other work cited there, we expected that a good deal of post-editing would be needed. However, for texts we have tried so far, the error rates are much lower than we had anticipated, and the results appear usable with very light post-editing.

It is not clear to us why our results are so much better than expected. The processing pipeline from §3 is an almost minimal recipe for producing a joint alignment using a beam search; the only non-obvious step is (7), post-processing of the double-aligned corpus. Removing it degrades the **Seg** score by a few percent and has almost no effect on the other two metrics, so this is not the explanation.

A more plausible hypothesis is that the LibriSpeech team were simply trying to solve a different problem, producing a large corpus of reliably aligned sentences, and paid no attention to the question, uninteresting to them, of how accurately they could align a complete literary text. Another is that the quality of readily available speech recognition engines and sentence aligners has substantially improved since 2015. We are impressed with the robustness of Google Cloud Speech-to-Text. For example, we discovered that litteratureaudio include background music in some of their offerings, using it at the starts and ends of sections and to underline key passages; also, the voice talents interpret the material in an imaginative way, rendering direct speech dramatically in different voices. We were concerned that both of these aspects might cause problems for speech recognition performance, but in fact there were none. The bottom line is that the task of automatically creating audio- and translation-annotated texts out of pub-

lic domain corpus resources appears considerably more tractable than we had thought. Our main purpose in the current paper is to communicate this discovery to other members of the community who may also find it interesting and useful.

The data presented here suggests three priorities for continued investigation. First, the method should be tested on more texts, in several languages; second, we require user feedback for the resulting multimedia versions; third, we need to further systematise the post-editing process. We have already begun work on all of these. We briefly outline two specific threads of work initiated during the period Oct–Nov 2022 in collaboration with other LARA partners.

First, together with Ivana Horváthová of the Constantine the Philosopher University, Nitra, Slovakia, we are using the alignment methods to construct a LARA version of A.A. Milne's *Winnie-the-Pooh* with Slovak glosses. As an initial proof-of-concept experiment, we processed the first few pages and obtained excellent results; we are now negotiating with the copyright-holders to obtain the permissions needed to use the Slovak translation of the whole book. If we are able to do this, our plan is to perform an experiment, probably starting in Q1 2023, where we would contrast user perceptions of the resulting LARA document with a version of the same text run on the Azure Immersive Reader.

Second, we are working together with Neasa Ní Chiaráin and Harald Berthelsen of Trinity College Dublin, Ireland, to investigate the idea of performing alignment with a different recogniser, specifically the Kaldi-based ASR platform for Irish developed by the Trinity College group (ABAIR-ÉIST; https://www.abair.ie/; Lonergan et al. 2022). We have again only got as far as a proof-of-concept experiment, where we aligned a short Irish text corresponding to about five minutes of audio. Results were encouraging, with error rates similar to those we obtained on the French texts from §4. We hope to be able to progress this work further in the near future.

Ethics Statement

Methods like those described here naturally raise issues involving copyright. To the best of our knowledge, we have appropriate copyright permissions for all the text and audio materials used in the experiments.

Acknowledgements

We would very much like to thank Lieve Macken for pointing us to the YouAlign tool. Many people in the LARA community have directly or indirectly contributed to the development of the alignment method. We would particularly like to thank Branislav Bédi, Harald Berthelsen, Catia Cucchiarini, Ivana Horváthová, Christèle Maizonniaux, Neasa Ní Chiaráin, Chadi Raheb and Rina Zviel-Girshin.

A Appendix: using the scripts

People interested in using the Python scripts we refer to here should consult the online LARA documentation (Rayner et al., 2019–2022), which describes how to download, install and invoke the relevant software. Details can be found in the sections headed "Using the Python code: prerequisites" and "Automatic cutting-up and alignment with audio and translation".

References

Elham Akhlaghi, Ingibjörg Iða Auðunardóttir, Anna Baczkowska, Branislav Bédi, Hakeem Beedar, Harald Berthelsen, Cathy Chua, Catia Cucchiarini, Hanieh Habibi, Ivana Horváthová, Junta Ikeda, Christèle Maizonniaux, Neasa Ní Chiaráin, Chadi Raheb, Manny Rayner, John Sloan, Nikos Tsourakis, and Chunlin Yao. 2022a. Using the LARA Little Prince to compare human and TTS audio quality. In *Language Resources and Evaluation Conference*, pages 2967–2975. European Language Resources Association.

Elham Akhlaghi, Ingibjörg Iða Auðunardóttir, Branislav Bédi, Hakeem Beedar, Harald Berthelsen, Cathy Chua, Catia Cucchiarini, Brynjarr Eyjólfsson, Nedelina Ivanova, Christèle Maizonniaux, Neasa Ní Chiaráin, Manny Rayner, John Sloan, Sigurður Vigfússon, and Ghil'ad Zuckermann. 2022b. Reading assistance through LARA, the learning and reading assistant. In *2nd Workshop on Tools and Resources for REAding DIfficulties (READI)*, page 1.

Elham Akhlaghi, Branislav Bédi, Matthias Butterweck, Cathy Chua, Johanna Gerlach, Hanieh Habibi, Junta Ikeda, Manny Rayner, Sabina Sestigiani, and Ghil'ad Zuckermann. 2019. Overview of LARA: A learning and reading assistant. In *Proc. SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education*, pages 99–103.

Elham Akhlaghi, Anna Bączkowska, Harald Berthelsen, Branislav Bédi, Cathy Chua, Catia Cuchiarini, Hanieh Habibi, Ivana Horváthová, Pernille Hvalsøe, Roy Lotz, Christèle Maizonniaux, Neasa Ní Chiaráin, Manny Rayner, Nikos Tsourakis, and Chunlin Yao.

2021. Assessing the quality of TTS audio in the LARA learning-by-reading platform. In *CALL and professionalisation: short papers from EUROCALL 2021*, pages 1–5.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Branislav Bédi, Matt Butterweck, Cathy Chua, Johanna Gerlach, Birgitta Björg Guðmarsdóttir, Hanieh Habibi, Bjartur Örn Jónsson, Manny Rayner, and Sigurður Vigfússon. 2020. LARA: An extensible open source platform for learning languages by reading. In *Proc. EUROCALL 2020*.

Anna C.S. Chang and Sonia Millett. 2014. The effect of extensive listening on developing L2 listening fluency: Some hard evidence. *ELT journal*, 68(1):31–40.

Eleanor Dickey. 2016. *Learning Latin the ancient way: Latin textbooks from the ancient world*. Cambridge University Press.

Chris Fournier. 2013. Evaluating Text Segmentation using Boundary Edit Distance. In *Proceedings of 51st Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aaron Friedland, Michelle Gilman, Michael Johnson, and Abera Demeke. 2017. Does reading-while-listening enhance students' reading fluency? preliminary results from school experiments in rural uganda. *Journal of Education and Practice*, 8(7):82–95.

Anna Husson Isozaki. 2014. Flowing toward solutions: literature listening and L2 literacy. *The Journal of Literature in Language Teaching*, 3(2):6–20.

Stephen Krashen. 1982. *Principles and practice in second language acquisition*. Pergamon Press.

Stephen Krashen. 1989. We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *The modern language journal*, 73(4):440–464.

Stephen Krashen. 2004. *The power of reading: Insights from the research*. Greenwood Publishing Group.

Sy-Ying Lee. 2019. A fulfilling journey of language acquisition via story listening and reading: A case of an adult scholar. *Language Learning and Teaching*, 8(1):1–9.

Liam Lonergan, Mengjie Qian, Harald Berthelsen, Andy Murphy, Christoph Wendler, Neasa Ní Chiaráin, Christer Gobl, and Ailbhe Ní Chasaide. 2022. Automatic speech recognition for irish: the abair-éist system. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 47–51.

Can Özmen, Alexander Streicher, and Andrea Zielinski. 2014. Using text segmentation algorithms for the automatic generation of e-learning courses. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (* SEM 2014)*, pages 132–140.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ana Pellicer-Sánchez, Elsa Tragant, Kathy Conklin, M Rodgers, A Llanes, and R Serrano. 2018. L2 reading and reading-while-listening in multimodal learning conditions: An eye-tracking study. *ELT Research Papers*, 18(01):1–28.

Manny Rayner, Hanieh Habibi, Cathy Chua, and Matt Butterweck. 2019–2022. *Constructing LARA content*. https://www.issco.unige.ch/en/research/projects/callector/LARADoc/build/html/index.html. Online documentation.

Pedro José dos Reis Mota. 2019. *BeamSeg: a Joint Model for Multi-Document Segmentation and Topic Identification*. Ph.D. thesis, Carnegie Mellon University.

John W. Schwieter and Alessandro Benati. 2019. *The Cambridge Handbook of Language Learning*. Cambridge University Press.

Ryan Shaw. 2015. Segmenting oral history transcripts. In *International Conference on Theory and Practice of Digital Libraries*, pages 326–329. Springer.

Billy Woodall. 2010. Simultaneous listening and reading in ESL: Helping second language learners read (and enjoy reading) more efficiently. *TESOL journal*, 1(2):186–205.

Yong Xu, Aurélien Max, and François Yvon. 2015. Sentence alignment for literary texts: The state-of-the-art and beyond. In *Linguistic Issues in Language Technology, Volume 12, 2015-Literature Lifts up Computational Linguistics*.

Ghil'ad Zuckerman, Sigurður Vigfússon, Manny Rayner, Neasa Ní Chiaráin, Nedelina Ivanova, Hanieh Habibi, and Branislav Bédi. 2021. LARA in the service of revivalistics and documentary linguistics: Community engagement and endangered languages. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, pages 13–23.