Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6613–6620
Marseille, 11–16 May 2020

FAB: The French Absolute Beginner Corpus for Pronunciation Training

Sean Robertson[1], **Cosmin Munteanu**[2], **Gerald Penn**[1]
[1]Department of Computer Science, University of Toronto
40 St. George St., Toronto, ON, Canada
[2]Institute of Communication, Culture, Information and Technology, University of Toronto Mississauga
3359 Mississauga Rd., Mississauga, ON, Canada
sdrobert@cs.toronto.edu, cosmin.munteanu@utoronto.ca, gpenn@cs.toronto.edu

Abstract

We introduce the French Absolute Beginner (FAB) speech corpus. The corpus is intended for the development and study of Computer-Assisted Pronunciation Training (CAPT) tools for absolute beginner learners. Data were recorded during two experiments focusing on using a CAPT system in paired role-play tasks. The setting grants FAB three distinguishing features from other non-native corpora: the experimental setting is ecologically valid, closing the gap between training and deployment; it features a label set based on teacher feedback, allowing for context-sensitive CAPT; and data have been primarily collected from absolute beginners, a group often ignored. Participants did not read prompts, but instead recalled and modified dialogues that were modelled in videos. Unable to distinguish modelled words solely from viewing videos, speakers often uttered unintelligible or out-of-L2 words. The corpus is split into three partitions: one from an experiment with minimal feedback; another with explicit, word-level feedback; and a third with supplementary read-and-record data. A subset of words in the first partition has been labelled as more or less native, with inter-annotator agreement reported. In the explicit feedback partition, labels are derived from the experiment's online feedback. The FAB corpus is scheduled to be made freely available by the end of 2020.

Keywords: L2 Corpus, Non-Native, French, Beginner, Ecological Validity, Pronunciation Training, CAPT, CALL, ASR

1. Introduction

Computer-Assisted Language Learning has the potential to greatly ease the burden of educators. There has been considerable effort put into applying speech processing technologies to Computer-Assisted Pronunciation Training (CAPT) (Eskenazi, 2009), of which speech recognition is an especially important component. There are a number of difficult challenges involved in building these systems — chiefly the need for learner data collected in an appropriate manner.

In this paper, we introduce the *French Absolute Beginner* (FAB) corpus. Speech data from the FAB corpus were collected over the course of two experiments (Robertson et al., 2016; Robertson et al., 2018). These experiments were designed around co-operative role-playing tasks similar to those found in language-learning classrooms. The results of the experiment in Robertson et al. (2018) highlighted the importance of careful consideration and adherence to the environments in which CAPT systems are to be deployed. In the same spirit, FAB has three distinguishing factors that motivate its use in the development of CAPT systems:

- FAB was collected over the course of two ecologically valid experiments, which more closely resemble a classroom CAPT intervention.
- One partition of data is labelled according to the native vs. non-native paradigm; another is labelled according to teacher feedback.
- FAB is primarily composed of beginner learners of French, many with no prior experience learning French.

In the following sections, we describe these qualities as "motivating factors" behind the development of this corpus. Later, we provide an overview of the types of data in the corpus as well as some of the distinctions made between its partitions.

The FAB corpus is almost complete; only one partition remains to be segmented. We expect the completed corpus to be freely available on the University of Toronto's *Dataverse*[1] by the end of 2020.

2. Motivations

2.1. Ecological Validity

On the way to error detection, CAPT systems must first perform utterance verification, i.e. dictation, to ensure that speakers are actually saying what we are evaluating them on. To perform utterance verification, CAPT needs ASR. It is well known that ASR systems are sensitive to the type of data they are trained on, which is especially true for non-native speech (Wang et al., 2003). Partly in order to decrease the difficulty of non-native ASR (Cucchiarini and Strik, 2018) and perhaps because most non-native databases already have this type of data (Raab et al., 2007), many CAPT systems, like *Duolingo* (von Ahn, 2013), have learners merely read aloud predefined sentences to be evaluated.

While convenient, read-and-record tasks do not adequately reflect the type of activities or usages of CAPT that are found in classrooms (Thomson and Derwing, 2014). We have previously found that traditionally trained CAPT systems do not yield the same benefits in ecologically valid settings as expected (Robertson et al., 2018). In order to build robust ASR for such settings, which can act as the foundation for CAPT systems, it is necessary to train these systems on appropriate utterances. Specifically, the data must reflect the interaction between learners and the CAPT system. This excludes both read-and-record corpora as well

[1] `https://dataverse.scholarsportal.info/dataverse/toronto`

as direct recordings from traditional classrooms, since the quality and flexibility of a teacher's feedback far exceeds that of a CAPT system's.

Of the corpora that collect the spontaneous speech of learners, to the best of our knowledge, none reflects the feedback of an online CAPT system. Cucchiarini et al. (2008) recorded non-native speakers of Dutch communicating with a Wizard-of-Oz spoken dialogue system, but not in a language-learning context. Likewise, Jurafsky et al. (1994) collected a small corpus containing non-native English speech from communications with a Wizard-of-Oz spoken dialogue system. Sanders et al. (2014) produced a non-native corpus of Dutch speech with some task-based activities, but again without a CAPT intervention.

FAB consists of data that reflect a realistic classroom CAPT intervention. During a paired role-playing task, participants must ask a CAPT system for feedback after every utterance. Each utterance is strongly impacted by prior feedback from the CAPT system. As such, FAB provides a better inventory of the types of utterances that a CAPT system would be exposed to in the wild.

2.2. Teacher-Driven Labels

Just as it is important that the data are collected in realistic scenarios, it is also important that CAPT systems judge segments according to a realistic criterion. Such a criterion would presumably match what a teacher expects to be useful feedback.

To date, the debate on CAPT feedback has hovered around two criteria: *nativeness* and *intelligibility*. The former distinguishes between native and non-native speech - a criterion tacitly endorsed by engineers (Cucchiarini and Strik, 2018), probably because of the comparative ease of its ascertainment. The latter is preferred by educators (Levis, 2005), but its definition - roughly, how easily the utterance is understood by the listener - has a number of contributing factors that make it an unwieldy objective. In either case, segments - be they phonemes, words, utterances, syllables, etc. - are judged according to some latent, passive criterion. In the later experiment from which FAB was collected (Robertson et al., 2018), we found that fine-grained teacher feedback could be more readily predicted by heuristics based on dialogue history and known sources of difficulty for learners than by a CAPT system, even when tuned to model that same teacher's feedback offline on a per-segment basis. This suggests that the teacher feedback was based less on a latent criterion somehow embedded into each segment, and more on an active, context-sensitive learning process that the participants were engaged in.

To that end, while almost half of FAB has been labelled according to the traditional nativeness criterion, another half labels utterances using the word-level feedback provided online to participants during the experiment. The latter label set forsakes explicit, theoretical criteria in favour of mimicking the feedback that a teacher would provide in the given situation. Such an approach is bound to the idiosyncrasies of that teacher, but this risk is no greater than that made when entrusting a student to a single teacher in a traditional classroom. We believe that this alternative label set provides an avenue for training CAPT systems offline in a way more faithful to the environment in which they will be deployed. These labels can be juxtaposed to those from the former half of the data set to determine when and how realistic feedback differs from nativeness judgements.

2.3. Absolute Beginners

CAPT is more impactful to beginner and intermediate learners than to advanced ones (Mahdi and Al Khateeb, 2019). It follows that those who will most clearly benefit from CAPT interventions are those with little to no experience speaking the target language. We suspect that this group is often overlooked because no specialized instruction is necessary for the egregious mistakes made by learners at this stage. However, ignoring absolute beginner CAPT could lead to a lack of feedback in learner utterances (especially in self-driven learning), widening the gulf between written comprehension and production.

There are a variety of non-native speech databases, such as *PF_STAR* (Batliner et al., 2005), *CorAIt* (Combei, 2018), *IFCASL* (Trouvain et al., 2016), *JASMIN-CGN* (Cucchiarini et al., 2008), and *LLESLA* (Sanders et al., 2014), that offer beginner learner utterances. In all of these databases, however, the learners have had some non-negligible prior experience in the target language. The *CBFC* (Yoo and Kim, 2018) contains data from a speaker with just a month's experience - but only one speaker. The *Young Learners Corpus* (Myles, 2012) does feature absolute beginners, but since the learners were all very young and not as subject to phonetic fossilization as adult learners, it is unlikely to match the demographics of a second-language CAPT system.

While suitable corpora may exist for absolute beginners in educator-driven CAPT research, they are difficult to find, may suffer annotation problems, or are not public (O'Brien et al., 2018). To the best of our knowledge, FAB is the first corpus of absolute beginner speech for second language learning. With word-level segmentations, FAB is sufficient to train an ASR-based CAPT system. In addition, FAB may serve in a systematic exploration of the learning strategies of absolute beginners, though such an exploration is beyond the scope of this project.

3. Corpus Development and Description

Data were collected over two CAPT experiments performed at the University of Toronto (Robertson et al., 2016; Robertson et al., 2018). Though there are many similarities in the data collected across the experiments, the differences in experimental design and the collected supplementary data warrant distinction within the corpus. Thus, we partitioned the data into three tranches: one for the first experiment, another for the second, and a third for supplementary read-and-record data. We initially describe the commonalities across partitions in sections 3.1. to 3.3., followed by a description of the unique aspects of each.

To aid in this description, fig. 1 provides a high-level perspective of the data in the corpus and how their derivation differs partition-wise. The major data components of FAB are indicated by white boxes, with examples in typewriter font. Important intermediate processes are indicated with grey boxes. Labelled arrows in the flow chart indicate where partitions differ in processing. The dashed arrow signifies

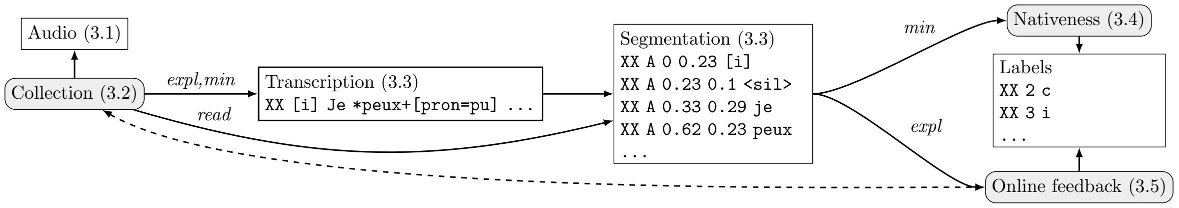

Figure 1: Flow chart detailing the corpus creation and constitution.

Speakers	121
Segmented utterances	approx. 9 thousand
Word segments	approx. 25 thousand
Word segment duration	approx. 3 hours
Recording format	16kHz PCM16 mono WAV
Transcription format	*CTM*, *TextGrid*, *Transcriber*

Table 1: Information on entire FAB corpus.

that collection co-occurs with online feedback. Finally, parentheticals indicate what section of this paper each process or data source is discussed in.

3.1. Overview

Table 1 provides aggregate information across the FAB corpus's partitions. Though 62 experimental sessions were run, one participant did not give her consent to release her recordings, bringing the total number of speakers to 121.

All recordings were captured through the built-in microphone of an iPad Mini 2 at 16kHz with PCM16 linear encoding. While experimentation occurred in an office environment and participants were instructed to avoid extraneous noise, the recordings are not always free of noise. Rustling of coats, laughter, interruptions, and soft background noise are common. We expect this sort of noise to be commonplace in most realistic applications.

All audio are transcribed and segmented at the word level. Manually segmented transcriptions are available in both *NIST SCTK*'s time-marked transcription format (Fiscus, 2008) and *Praat*'s *TextGrid* format (Boersma and Weenink, 2019). There are roughly 25 thousand word segments in the entire dataset excluding noise and silence. This number may change on release as the second experiment has yet to be manually segmented. The average utterance length across about 9000 utterances is only 3 words. The audio segmented as French words has a projected cumulative duration of about 3 hours (*Word segment duration*) based on the similarity between the sizes of segmented and yet-to-be-segmented data. The actual duration of the database will be longer due to noise and silence.

M: Bonjour ***madame****.*
F: Bonjour ***monsieur****.*
M: Je peux vous aider?
F: <points>
M: ***Une pomme****?*
F: ***Non, non.*** *<points>*
M: ***Une banane****?*
F: Oui, s'il vous plaît.

Figure 2: Transcription from model video.

3.2. Collection

Participants were recruited with posters in University of Toronto facilities. Participants were required to be proficient in English, be at least 18 years of age, and speak very little to no French, in the case of the first experiment. For the second experiment, we required in addition that they had never attended French classes. An experimental session always involved two participants, each of which was recruited individually.

Experiments were designed to resemble paired role-play tasks, following the pedagogy of task-based language learning (Skehan, 2003). Each experimental session consisted of a number of dialogues designed in concert with an industry professional who was experienced in French second language curricula. They were tailored to introduce vocabulary, grammar, semantics, and pragmatics without explicit meta-linguistic instruction. Videos presented a prototypical dialogue in Parisian French that participants would be expected to engage in. Where piloting determined that it would be necessary, additional instruction on how to enact the dialogue was provided by an administrator in English (though feedback on the French content was forbidden). As brand-new learners were unlikely to generate spontaneous dialogue in French, video modelling provided participants the necessary scaffolding to engage in a facsimile. Participants then took turns recording dialogue with an iPad application which, in turn, told them whether the sentence was accepted or rejected, optionally providing feedback. The application was in fact secretly controlled by a confederate, the *Wizard* in a *Wizard-of-Oz* experimental design, who was an experienced French teacher.

In order to avoid merely listening and repeating the prototypes, participants were expected to make small changes to the blueprint according to context. Figure 2 provides a transcription of one of the videos near the end of the experiment. This video shows a customer asking for a fruit from a vendor. Boldface indicates the necessary changes that participants were expected to make according to context, namely: using the correctly gendered title; naming the appropriate fruit; and recognizing that the vendor had made a mistake without explicitly repeating that mistake when recording.

Unlike the activities in existing CAPT applications, here participants were not immediately given a transcript to read. The combination of listening, speaking, recall, and linguistic inference put a much greater cognitive load on participants.

3.3. Transcription

As was mentioned, the realistic role-play task was more difficult for participants than merely reading aloud sentences. Because of this, participants were often unable to perceptually segment target words. Idiosyncratic substitutions were commonplace. Unlike traditional ASR transcriptions, which would simply label all substitutions according to whatever was actually said (produced), transcriptions for CAPT must decide whether the word should be labelled according what was produced or what the speaker intended. For consistency and faithfulness to the scenario, we eventually decided to teach transcribers to guess at participants' intentions. While ascribing intent to speakers is a subjective task, it is far more consistent than the alternative. Most words were mispronounced at some level and, when we attempted to transcribe according to what was produced, transcribers noted considerable difficulty choosing words closest to the mispronunciation.

We provided transcripts of all the video dialogues to transcribers, assuming that, because participants had little knowledge of French, they would cleave closely to what was presented to them. As a concrete example, "três" would be transcribed as "trois" whenever uttering "trois" could reasonably follow from the dialogue (a common substitution for our Portuguese participants). Phonetically dissimilar substitutions (e.g. "une pomme" → "App*wah*") were more likely to be labelled as unintelligible or foreign.

We had two fluent speakers of French transcribe each utterance with a mark-up language. Transcriptions are stored in tab-delimited lists of word tokens, utterance identifiers, and some optional utterance-level tags. The mark-up language is an adaptation of the standard[2] proposed for the *Transcriber* tool (Barras et al., 2001). For this corpus, a critical feature of the standard is its ability to mark up mispronounced words with orthographic transcriptions and to indicate which words are foreign. Both occur frequently in the *min* (Section 3.4.) and *expl* (Section 3.5.) partitions of FAB. We modified the standard to use non-speech tokens consistent with standards for speech recognition (Deléglise et al., 2005). Speech data from the *read* partition (Section 3.6.) are from readings and thus did not require direct transcription. In order to improve consistency, transcribers met for a subset of overlapping sessions and came to an agreement on how they could be merged.

Speakers	58
Segmented utterances	4009
Word segments	12126
Word segment duration	1.36
More/less native labels	5767
First language	English(26), Portuguese(13), Mandarin(4), Cantonese(4)
Fluency in French (1-5 asc.)	1(52), 2(8), 3(3)
French experience	None(16), Formal(16), Incomplete(4), Informal(3)
Median age	23
Gender	Female(30), Male(28)
Number languages fluent	2(26), 1(19), 3(11)

Table 2: Demographic and recording information from the *min* partition.

Transcribing mispronunciations is already a difficult task — one exacerbated by the frequency of mistakes made by beginners. A rigorous phonemic transcription would involve an understanding of multiple phonemic inventories spanning the various mother tongues of participants, more training for our transcribers, and a more careful means of ensuring consistency among annotators. Further, it is unlikely that CAPT systems would be able to leverage an open phonemic inventory, as even the CAPT systems that provide fine-grained feedback, e.g. (Harrison et al., 2009), are restricted to the target language inventory and focus on specific language pairs. The French orthographic transcriptions were quick to transcribe, are easily bootstrapped into phonemes (described below), and do not apply the same degree of rigour.

Word-level segmentations were derived from these transcriptions. Very noisy recordings were not segmented. Initial segmentations were generated by force-aligning transcripts to speech using an off-the-shelf speech recognizer (Deléglise et al., 2005). The pronunciation lexicon was augmented with alternative pronunciations derived from feeding the orthographic transcriptions above into a grapheme-to-phoneme transducer (Novak et al., 2012). Those segments were then manually adjusted by the first author.

3.4. Minimal Feedback Role-Play Partition

Data from the *Minimal Feedback Role-Play Partition*, stored in FAB under the subdirectory *min*, were collected during the first experiment (Robertson et al., 2016).

Table 2 provides demographic and recording information for this partition[3]. For discrete-valued entries, bracketed values indicate the count of that unique response. Only

[2] `http://trans.sourceforge.net`, version 1.22, last accessed November 19, 2019.

[3] This information supersedes that originally reported in (Robertson et al., 2016). Here, the demographics of 4 participants thought to be missing have been recovered. In addition, one participant who mentioned "adult courses" has been relabeled as having *Formal* experience.

unique values with a count of at least 2 are included. Filling out demographic information was optional, and some participants did not fill out the full survey. Some participants wrote in more than one first language; we encoded these separately. We code French experience as *Formal* if they have reported finishing some form of French class, *Incomplete* if they dropped their first class, and *Informal* if they mentioned software, travels, etc. Some speakers mentioned casual exposure to the language through music or packaging, but we did not categorize such responses. In accordance with university ethics regulations, the above information will not be directly mapped onto speakers in the release of FAB.

Admittedly, many speakers were not absolute beginners. The participation requirements for the first experiment were considerably more relaxed than in the second. This is primarily due to the grades 4-8 core French programme in Ontario, Canada. However, the vast majority also self-reported very low fluency in French.

The wizard recruited for this experiment was a professionally trained second-language teacher with past experience teaching Parisian French and English. She was instructed by our industry partner in some of the pedagogical underpinnings of the experiment.

For this experiment, feedback from the application (wizard) was limited to accepting or rejecting utterances. The level of feedback mirrored that of a planned language learning video game, focusing on the implicit feedback preferred by communicative pedagogies (Savignon, 1987). There was also considerably less instruction provided by the experiment administrator than in the following experiment.

The lack of pointed feedback was a source of frustration for the wizard. When a participant did not know why the application rejected her utterance, he or she would often repeat the same phrase with no adjustment, possibly in an attempt to make the application give up. The wizard would, assuming the issue was not critical.

One of the goals of our experimentation was to train and evaluate state-of-the-art pronunciation error detectors. This classification task is often framed as distinguishing between *native* and *non-native* speakers. Since the database is full of beginners, classification was not about whether a word sounded non-native but to what degree it did not. Therefore, roughly five thousand word segments across 92 word types from the experiment (excluding contracted determiners, foreign words, non-words, and words with less than 30 instances) were labelled in a binary fashion as *more* or *less* native.

Four native French speakers were hired, each with experience teaching second-language French professionally in Ontario. No attempt was made to control for the native dialect of the annotators, nor did we attempt to vary it. Nativeness was defined to our annotators as: *accentedness*; sounding like a native speaker; how easy it is to detect an accent; and how close a word is to that of a native speaker of French. The description was intentionally left vague so as to better align with annotators' own perspectives.

To build the label set, a modified version of pairwise comparisons was employed. In pairwise comparisons, a large set of relative rankings of pairs of word instances of the same word type are used to determine a full ranking of the set. If N instances of a word are to be fully ranked, an annotator would need to make $N(N-1)/2$ judgements, which is quadratic with respect to N. The number of times a word instance is judged to be more native than its paired instance determines its overall rank with respect to its word type.

ann.	κ_1	κ_2	more	less	contr.	prop.
A	.16	.48(168)	.23	.36	.12	.55
B	.20	.52(184)	.29	.41	.9	.10
C	.17	.50(174)	.29	.37	.13	.10
D	.16	.51(154)	.31	.23	.14	.25

Table 3: Label and average agreement statistics for *min*'s classification task.

Because we only needed a binary label for each instance (more or less native), we could afford to simplify the task so that the order of comparisons was linear. We had two annotators fully rank 10 randomly selected instances per word type. Whenever possible, those sets were non-overlapping. Then, the instances ranked fourth and sixth per annotator were taken as lower and upper boundaries. New instances of a word were compared to both a lower and upper boundary, randomly selected from the two upper and two lower boundaries per word type (one of each boundary each of two annotators). If the new instance was judged less native than both the lower and upper boundary, it was labelled *less*. If the new instance was judged to be more native than both boundaries, it was labelled *more*. If more than the lower boundary and less than the upper boundary, the instance was middling and thus labelled *unsure*. Finally, if less than the lower boundary and more than the upper boundary, the point was labelled as *contradictory*.

To determine inter-annotator agreement, an overlap set of 418 segments was drawn. The overlap set consists of a hand-selected subset of 15 word types, each with a hand-selected number of instances. The chosen number of instances per word type was much smaller than their partition totals, but the ranking of words by instance count was maintained. The actual samples of each word were drawn randomly.

Table 3 provides the proportion of segments labelled *more*, *less*, and *contradictory*, as well as the proportion of the whole database that was labelled by each annotator (*prop.*). Annotators (*ann.*) *D* and *B* are the wizards for the first and second experiment, respectively. κ_1 measures one-versus-rest inter-annotator agreement using Cohen's κ over all points in the overlap set. κ_2 is Cohen's κ on only the points labelled more or less native by both the one annotator and the "rest" annotator. Comparing κ_1 and κ_2 lends support to the notion that labels reflect an underlying ordinality: it is easier for annotators to confuse more/less labels with unsure/contradictory labels than more with less, or *vice versa*. Roughly half of the labels are more or less, corresponding with the expected proportion of instances above and below the boundary points.

Only the points labelled *more* or *less*, 5767 of 9435, then participate in the classification task. Because not all segments have labels, gold-standard labels are stored in a tab-delimited master list. Each entry contains the utterance

Speakers	63
Segmented utterances	approx. 5 thousand
Word segments	approx. 13 thousand
First language	English(28), Mandarin(7), Chinese(9), Cantonese(3), Russian(3), Spanish(2), Vietnamese(2), Farsi(2), Korean(2), Hindi(2), Malayalam(2)
Fluency in French (1-5 asc.)	1(59), 2(4)
French experience	None(37), Informal(8), Full(4)
Median age	23
Gender	Female(32), Male(31)
Number languages fluent	2(36), 3(14), 1(12)

Table 4: Demographic and recording information from the *expl* partition. Segment counts are approximate.

identifier, the index of the segment within the utterance, and the label itself. The labels are partitioned into rough quadrants either by annotator, speaker, or randomly (Robertson et al., 2016).

3.5. Explicit Feedback Role-Play Partition

Data in the *Explicit Feedback Role-Play Partition*, stored under the *expl* subdirectory of FAB, were collected in the second experiment (Robertson et al., 2018), including a large number of additional pilot sessions.

Table 4 lists the demographic and recording information for the *expl* partition. 9 participants entered "Chinese" as a first language rather than a specific variety.

As of writing, the *expl* partition has yet to be manually segmented. The word segment count listed in table 4 is based on word-level feedback provided to participants during experimentation. Given that the figure correlates well with the exact figure from table 2 on the *min* partition, which has slightly fewer participants, we are confident of the order of magnitude of word segments. We expect a similar total word segment cumulative total duration as in *min*.

The second experiment was much more strict when it came to prior experience in French learning. 12 participants did not indicate how much exposure they had had to French teaching. This might be due to a flaw in the survey: there was no specific checkbox for "None", so it had to be entered manually into the "Other" entry.

A new wizard was recruited for this experiment. She also had prior experience teaching French. Though video dialogues were still presented in Parisian French, she herself spoke a Mauritian dialect. The mismatch was no obvious source of confusion given the aptitude of the participants. Unlike the first wizard, she was not instructed in an underlying pedagogy.

The explicit feedback condition had much more feedback per utterance than the minimum feedback partition. In addition to the per-utterance accept/reject, the wizard could choose to provide text-based feedback to illustrate insertions, deletions, and mispronunciations in the recorded utterance. Words could be tapped to hear a native pronounce them. To make sure the participant knew enough of the phrase before word-level feedback was provided, the wizard could make *full rejections,* which would provide no feedback outside a rejection.

The wizard was not instructed on how or whether to label words. There was no guarantee that a word labelled as correct actually was correct, or whether, for example, the wizard was refraining from mentioning a mistake in order to bolster confidence or to emphasize some other aspect of the feedback. As another example, the wizard could choose to present the transcript to participants, even if not a single word contained in it had been uttered, if they were stuck. The wizard often adapted her feedback to the speaker's performance. One of the CAPT systems built in (Robertson et al., 2018) achieved moderate agreement with wizard labels ($\kappa = 0.344$), greater than the state-of-the-art condition, by cycling through more difficult words for Anglophone learners of French. We assigned greater probability to mispronouncing certain word types by bootstrapping the entries in vowel confusability matrices from the speech perception literature (Gottfried, 1984; Flege, 1987; Best, 1995; Levy and Strange, 2008; Levy and Law, 2010). This suggests that wizard feedback is moderately well predicted by the literature on Anglophone French learners. Though we can hypothesize that participants' perceptions of French were coloured by English phonetics — the experiment was administered in English, which implies some proficiency in the language — we cannot make strong conclusions on the matter without a formal analysis of the corpus. Such an analysis is outside the scope of this project.

Progress through the dialogues was more guided than in the first experiment. Participants would be required to first figure out what to say, then, over a series of recordings, learn to say it correctly. While searching for what to say at the beginning of a dialogue turn, mistakes were similar to those from *min*. After written feedback, utterances were clearer, though more prone to errors due to orthography (e.g. reading as if the text were in English). Participants would often stress words that were marked as mispronounced.

Due to the experimental design, the corrections would not always align with what the wizard expected. Two thirds of the time, wizard feedback on correct and mispronounced words (not insertions, deletions, or the decision to accept or reject) was swapped with that of two CAPT systems. The wizard was unaware of the changes, which increased the chance for miscommunication with participants.

We have included the wizard's (and the two CAPT systems') transcriptions and word-level labels in the *expl* partition. We have discussed how this label set could be beneficial to CAPT in section 2.2.. We note here that the feedback provided by the wizard was sufficient to significantly improve participants' pronunciations over the course of the experiment (Robertson et al., 2018). Further ecologically valid experimentation will nevertheless be necessary in order to determine whether a CAPT system trained offline with this new objective will provide adequate feedback to learners.

Speakers	19
Segmented utterances	617
Word segments	2039
Word segment duration	0.28 hours
Total segments	3686
First language	English(8), Mandarin(3)
Fluency in French (1-5 asc.)	1(18), 2(1)
French experience	None(7), Informal(4), Formal(2)
Median age	23
Gender	Male(11), Female(8)
Number languages fluent	2(11), 1(5), 3(3)

Table 5: Demographic and recording information from the *read* partition.

3.6. Read-and-Record Partition

The *Read-and-Record Partition*, located under the *read* subdirectory of FAB, contains supplementary speech data recorded after the second experiment (Robertson et al., 2018).

Table 5 gives demographic and recording information on the partition. Recordings were voluntary and time-permitting. Speakers were a strict subset of those found in the *expl* partition.

The purpose of these recordings was to acquire more speech data for future training of speech systems. As such, there is no classification task associated with the partition. Participants were asked to record sub-sentence chunks from a paragraph of a sample online reading-comprehension test from the website of the *Ministère de l'Education Nationale et de la Jeunesse* of France[4]. The paragraph was slightly modified to cover the French phonemic inventory. Though an English translation was provided to participants beforehand, participants were not expected to understand the text, but merely to repeat it. Both the text and sample native recordings were provided.

Participants tended to mimic the suprasegmental structure of the sample native recordings. An administrator being present during recordings meant that participants maintained some base of effort during recordings. Without understanding, practice, or assessment by the application, recordings were not always as clear as their *expl* counterparts. Unlike the other partitions, however, segments almost always resembled the target phrase, making their transcriptions unambiguous.

4. Conclusions

We have presented FAB, a corpus intended to aid in the development of CAPT systems for French language learners. FAB was collected through two ecologically valid experiments. Its constitution reflects the importance of matching data to the situations in which they will be deployed. It (largely) forgoes standard read-and-record collection schemes or even traditional classroom speech, instead focusing on speech from a realistic CAPT intervention. Though part of the database has been labelled to satisfy a nativeness criterion, another part is labelled with online teacher feedback. We believe the latter set can be used to mimic the context-sensitive feedback provided by teachers in classroom settings. Finally, FAB targets a group of learners often ignored in the literature but, arguably, in the most need of pronunciation feedback: absolute beginners. We are freely releasing FAB in the hopes that it will be used to build CAPT systems that can be more readily integrated into the classroom.

5. Acknowledgements

This research was funded by an Ontario Centres of Excellence Technical Problem Solving grant in partnership with Speax Inc., and a Canada Graduate Scholarship from the Natural Sciences and Engineering Research Council of Canada.

6. Bibliographical References

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (2001). Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1):5–22, January.

Batliner, A., Blomberg, M., D'Arcy, S., Elenius, D., Giuliani, D., Gerosa, M., Hacker, C., Russell, M., Steidl, S., and Wong, M. (2005). The PF_STAR children's speech corpus. In *Interspeech 2005*. International Speech Communication Association (ISCA).

Best, C. T. (1995). A direct realist view of cross-language speech perception. *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, pages 171–204.

Boersma, P. and Weenink, D. (2019). Praat: Doing phonetics by computer.

Combei, C. R. (2018). CorAIt – a non-native speech database for Italian. In *Proceedings of the Fourth Italian Conference on Computational Linguistics*, CLiC-It '17, pages 113–118, Torino. Accademia University Press.

Cucchiarini, C. and Strik, H. (2018). Automatic speech recognition for second language pronunciation training. In *The Routledge Handbook of Contemporary English Pronunciation*, Routledge Handbooks in English Language Studies, pages 556–569. Routledge, Abingdon, Oxon.

Cucchiarini, C., Driesen, J., Van hamme, H., and Sanders, E. (2008). Recording speech of children, non-natives and elderly people for HLT applications: The JASMIN-CGN corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, LREC '08, Marrakech, Morocco. European Language Resources Association (ELRA).

Deléglise, P., Estève, Y., Meignier, S., and Merlin, T. (2005). The LIUM speech transcription system: A CMU Sphinx III-based system for French broadcast news. In *Interspeech 2005*, pages 1653–1656. International Speech Communication Association (ISCA).

Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech Communication*, 51(10):832–844. Spoken Language Technology for Education Spoken Language.

[4] `https://www.ciep.fr/en/tcf-tout-public/`, Sample 9, last accessed November 19, 2019.

Fiscus, J. (2008). SCTK, the NIST scoring toolkit.

Flege, J. E. (1987). The production of "new" and "similar" phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of Phonetics*, 15(1):47–65.

Gottfried, T. L. (1984). Effects of consonant context on the perception of French vowels. *Journal of Phonetics*, 12(2):91–114, April.

Harrison, A. M., Lo, W.-K., Qian, X.-j., and Meng, H. (2009). Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training. In *SCA International Workshop on Speech and Language Technology in Education*, SLaTE '09, pages 45–48, Warwickshire, England.

Jurafsky, D., Wooters, C., Tajchman, G., Segal, J., Stolcke, A., Fosler, E., and Morgan, N. (1994). The Berkeley restaurant project. In *Proceedings of the 1994 International Conference on Spoken Language Processing*, ICSLP '94, pages 2139–2142, Yokohama, Japan. International Speech Communication Association (ISCA).

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, 39(3):369–377.

Levy, E. S. and Law, F. F. (2010). Production of French vowels by American-English learners of French: Language experience, consonantal context, and the perception-production relationshipa). *The Journal of the Acoustical Society of America*, 128(3):1290–1305.

Levy, E. S. and Strange, W. (2008). Perception of French vowels by American English adults with and without French language experience. *Journal of Phonetics*, 36(1):141–157.

Mahdi, H. S. and Al Khateeb, A. A. (2019). The effectiveness of computer-assisted pronunciation training: A meta-analysis. *Review of Education*, 7(3):733–753, October.

Myles, F. (2012). Learning French from ages 5, 7 and 11: An investigation into starting ages, rates and routes of learning amongst early foreign language learners. *ESRC End of Award Report*, RES-062-23-1545.

Novak, J. R., Minematsu, N., and Hirose, K. (2012). WFST-Based grapheme-to-phoneme conversion: Open source tools for alignment, model-building and decoding. In *Proceedings of the Tenth International Workshop on Finite State Methods and Natural Language Processing*, FSMNLP '12, pages 45–49, Donostia, San Sebastián. Association for Computational Linguistics (ACL).

O'Brien, M. G., Derwing, T. M., Cucchiarini, C., Hardison, D. M., Mixdorff, H., Thomson, R. I., Strik, H., Levis, J. M., Munro, M. J., Foote, J. A., and Levis, G. M. (2018). Directions for the future of technology in pronunciation research and teaching. *Journal of Second Language Pronunciation*, 4(2):182–207.

Raab, M., Gruhn, R., and Noeth, E. (2007). Non-native speech databases. In *Proceedings of the 2007 IEEE Workshop on Automatic Speech Recognition & Understanding*, ASRU '07, pages 413–418. Institute of Electrical and Electronics Engineers (IEEE).

Robertson, S., Munteanu, C., and Penn, G. (2016). Pronunciation error detection for new language learners. In *Interspeech 2016*, pages 2691–2695. International Speech Communication Association (ISCA).

Robertson, S., Munteanu, C., and Penn, G. (2018). Designing pronunciation learning tools: The case for interactivity against over-engineering. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 356:1–356:13, New York, NY, USA. Association for Computing Machinery (ACM).

Sanders, E., Craats, I. V. D., and Lint, V. D. (2014). The Dutch LESLLA corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, LREC '14, Reykjavik, Iceland. European Language Resources Association (ELRA).

Savignon, S. J. (1987). Communicative language teaching. *Theory Into Practice*, 26(4):235–242.

Skehan, P. (2003). Task-based instruction. *Language Teaching*, 36(1):1–14, January.

Thomson, R. I. and Derwing, T. M. (2014). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics*, 36(3):326–344, December.

Trouvain, J., Bonneau, A., Colotte, V., Fauth, C., Fohr, D., Jouvet, D., Jügler, J., Laprie, Y., Mella, O., Möbius, B., and Zimmerer, F. (2016). The IFCASL corpus of French and German non-native and native read speech. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, LREC '16, Portorož, Slovenia. European Language Resources Association (ELRA).

von Ahn, L. (2013). Duolingo: Learn a language for free while helping to translate the web. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, IUI '13, pages 1–2, New York, NY, USA. Association for Computing Machinery (ACM).

Wang, Z., Schultz, T., and Waibel, A. (2003). Comparison of acoustic model adaptation techniques on non-native speech. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP '03. Institute of Electrical and Electronics Engineers (IEEE).

Yoo, H. and Kim, I. (2018). CBFC: A parallel L2 speech corpus for Korean and French learners. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, LREC '18, Miyazaki, Japan, May. European Language Resources Association (ELRA).