The Teacher-Student Chatroom Corpus version 2: more lessons, new annotation, automatic detection of sequence shifts

Andrew Caines[1] **Helen Yannakoudakis**[2] **Helen Allen**[3]
Pascual Pérez-Paredes[4] **Bill Byrne**[5] **Paula Buttery**[1]

[1] ALTA Institute & Computer Laboratory, University of Cambridge, U.K.
{andrew.caines|paula.buttery}@cl.cam.ac.uk
[2] Department of Informatics, King's College London, U.K.
helen.yannakoudakis@kcl.ac.uk
[3] Cambridge University Press & Assessment, University of Cambridge, U.K.
helen.allen@cambridge.org
[4] Departamento de Filología Inglesa, Universidad de Murcia, Spain
pfp23@cam.ac.uk
[5] Department of Engineering, University of Cambridge, U.K.
bill.byrne@eng.cam.ac.uk

Abstract

The first version of the Teacher-Student Chatroom Corpus (TSCC) was released in 2020 and contained 102 chatroom dialogues between 2 teachers and 8 learners of English, amounting to 13.5K conversational turns and 133K word tokens. In this second version of the corpus, we release an additional 158 chatroom dialogues, amounting to an extra 27.9K conversational turns and 230K word tokens. In total there are now 260 chatroom lessons, 41.4K conversational turns and 363K word tokens, involving 2 teachers and 13 students with seven different first languages. The content of the lessons was, as before, guided by the teacher, and the proficiency level of the learners is judged to range from B1 to C2 on the CEFR scale. Annotation of the dialogues continued with conversational analysis of sequence types, pedagogical focus, and correction of grammatical errors. In addition, we have annotated fifty of the dialogues using the Self-Evaluation of Teacher Talk framework which is intended for self-reflection on interactional aspects of language teaching. Finally, we conducted machine learning experiments to automatically detect shifts in discourse sequences from turn to turn, using modern transfer learning methods with large pre-trained language models. The TSCC v2 is freely available for research use.

1 Introduction & Related Work

Caines et al. (2020) introduced the Teacher-Student Chatroom Corpus (TSCC), a collection of 102 online English lessons between 2 teachers and 8 students containing 13.5K conversational turns and 133K word tokens, with the students adjudged to be writing at the CEFR levels of B1, B2 and C1. The lessons contained in the TSCC were anonymised, annotated with grammatical error corrections and discourse analyses, and made freely available to other researchers[1]. The motivation was to collate a dataset with which to study one-to-one interaction and language teaching, to investigate the linguistic skills involved in online chat at different levels of English proficiency, and potentially in the long-term to gather training data for developing a tutoring dialogue manager or chatbot.

In this paper, we report on further development of the corpus into a second version of the TSCC, with new lessons, annotations in the same style as those carried out before, and new annotations within a pre-defined pedagogical framework which we present below. The TSCC 2.0 includes an additional 158 lessons from new and existing students, amounting to 27.9K conversational turns and 230K word tokens. In total the 2$^{\text{nd}}$ version of the corpus features 2 teachers and 13 students, 41.4K conversational turns and 362.9K word tokens. The range of student CEFR levels found in the TSCC now includes C2 as well as B1 to C1.

[1] Visit forms.gle/pKc48WMhnySC8zDk9 to review the licence and submit a data request.

Andrew Caines, Helen Yannakoudakis, Helen Allen, Pascual Pérez-Paredes, Bill Byrne and Paula Buttery. The Teacher-Student Chatroom Corpus version 2: more lessons, new annotation, automatic detection of sequence shifts. *Proceedings of the 11th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2022).* Linköping Electronic Conference Proceedings 190: 23–35.

Turn	Role	Anonymised	Corrected	Resp.to	Sequence
1	T	Hi there ⟨STUDENT⟩, all OK?	Hi there ⟨STUDENT⟩, all OK?		opening
2	S	Hi ⟨TEACHER⟩, how are you?	Hi ⟨TEACHER⟩, how are you?		
3	S	I did the exercise this morning	I did *some* exercise this morning		
4	S	I have done, I guess	I have done, I guess		repair
5	T	did is fine especially if you're focusing on the action itself	did is fine especially if you're focusing on the action itself		scaffolding
6	T	tell me about your exercise if you like!	tell me about your exercise if you like!	3	topic.dev

Table 1: Example of numbered, anonymised and annotated turns in the TSCC (where role T=teacher, S=student, and 'resp.to' means 'responding to'); the student is here chatting about physical exercise. From Caines et al. (2020).

The new lessons, like those in the first release of the corpus, have been annotated for various discourse and classroom properties. These include the 'threading' of conversational turns so that non-sequential responses are connected with their appropriate conversational threads; the delineation of major and minor sequences in the discourse, as well as the labelling of their types; the identification of the pedagogical focus of sequences where applicable, along with any resources referred to; correction of grammatical errors by the student, and an assessment of student CEFR level for each lesson. The corpus and annotation are described in more detail in section 2.

In addition, fifty of the original lessons have been annotated using the Self-Evaluation of Teacher Talk framework (SETT) (Walsh, 2006, 2013), a schema designed for 'reflective practice' by language teachers for the purpose of their continuing professional development (Walsh, 2006). We annotated both teacher and student turns with aspects from SETT which we could identify. This gives us another way of considering the data collected, from a pedagogical and discourse-based perspective, and in section 3 we present the procedure for SETT annotation and the analyses we conducted.

We also describe initial experiments attempting to automatically detect when new discourse sequences are initiated in the lesson transcripts. This involved a 'transfer learning' approach, fine-tuning a large language model pre-trained with transformers on our specific machine learning task (Ruder et al., 2019). We cast the task as one of identifying when a turn in a chat lesson is followed by a new discourse sequence. As such, we are modelling the data collected so far in terms of discourse management by both teachers and students.

Finally in sections 5 and 6, we review the work which has already been done with the first version of the corpus, and we outline our future plans to further expand the corpus, improve our automated lesson manager, and develop teacher and student lesson feedback for self-development purposes for those taking part in the chatroom conversations.

2 Corpus description

The design and collection of data for the original TSCC was described in full in Caines et al. (2020), and we give a brief recap here. Participants arranged to hold one-to-one English language lessons in an online and private chatroom. The lessons were about one hour each, and the structure and content of each lesson was determined by the teachers. The students were recruited by the teachers themselves or through social media, and were located in several different countries around the world. An excerpt from the corpus is shown in Table 1, with selected annotation labels to illustrate the type of data available.

Transcriptions of the lessons were prepared for inclusion in the corpus through several annotation stages: firstly, they were anonymised by replacing any personal names with placeholders such as

	Version 1	Version 2
Lessons	102	260
Conv.turns	13,552	41,484
Words	132,895	362,440

Table 2: Comparative statistics for versions 1 and 2 of the TSCC.

CEFR	Version 1	Version 2
B1	36	36
B2	37	143
C1	29	29
C2	0	52

Table 3: Number of lessons by student CEFR level in versions 1 and 2 of the TSCC.

	Teachers	Students
Conv.turns	22,130	19,342
Words	238,324	124,090
Words/turn	10.8	6.4

Table 4: Comparing teacher and student contributions in version 2 of the TSCC.

⟨TEACHER⟩ and ⟨STUDENT⟩. Next, grammatical errors by the student were identified and corrected in a minimal fashion. Then, various linguistic and pedagogical features were marked up, including any non-sequential conversation threads (where a participant responded to a turn which was not the other participant's previous one), the start of new sequence types within the dialogue, the identification of the skill(s) focused on within that sequence, along with the use of any resources both internal and external to the chatroom.

The timeline of the lessons ranges from November 2019, through the onset of the COVID-19 pandemic to June 2021. We are open to collecting new data, and so the corpus may continue to grow, but this version of the TSCC comes from that twenty-month period.

2.1 New lessons

New data has been collected for the TSCC in the form of 158 new lessons. Now the corpus involves 2 teachers and 13 students, amounting to 41K conversational turns and 362K whitespace-delimited words (Table 2). The bulk of additional data was assessed by an expert to be at CEFR levels B2 and C2 (Table 3). The students' first languages are: Italian, Japanese, Mandarin Chinese, Russian, Spanish, Thai and Ukrainian. Table 4 shows how contributions to chatroom conversations compare for teachers and students.

2.2 Sequence, focus & resource types

Sequence types represent major or minor shifts in conversational sequences – sections of interaction with a particular purpose, whether that purpose is social or educational or a mixture of both. Borrowing key concepts from the CONVERSATION ANALYSIS approach (Sacks et al., 1974), we seek out groups of turns which together represent the building blocks of the chat transcript: teaching actions which build the structure of the lessons.

Teaching focus records which skill or skills were being targeted within a given sequence. **Use of resource** indicates whether any materials or stimuli external to the lesson are referred to

Compared to the original corpus, we have amended the annotation schema in various ways. First, some quality checks led to corrections to labels which were misspelled or in the wrong field. Second, we added new sequence types based on our work with the corpus over a longer time period. Now there is a 'non-English' sequence type, which might occur when the teacher and student switch to a different language (the learner's L1, for instance) either to explore or clarify a concept, or check and discover new vocabulary in English. And there is 'free practice', which relates to the learner being encouraged to make use of target content more freely than they would in a controlled exercise. In addition, it seemed sensible to move 'admin' of the lesson into the set of sequence types, rather than the set of sequence foci/focuses as it was in the original corpus.

We made minor modifications to the set of sequence foci, such that the skills 'writing', 'speaking', 'listening' and 'reading' are added, while the previously existing 'typo' is subsumed by the new 'writing' focus type. 'Exam practice' is renamed 'exam prep' – as in, exam *preparation* – because we found that not only were the teachers setting practice drills for the students but they were also discussing preparation strategies.

Finally, we note that many types of teaching resource emerged through collection of new data: the original list was open-ended, and has been extended in a bottom-up fashion.

In the Appendix, the full list of annotation types and their descriptions are copied from Caines et al. (2020) along with the amendments described above.

3 SETT annotation

The Self-Evaluation of Teacher Talk framework (SETT) was designed for reflective practice by language teachers (Walsh, 2006). This means that it was intended for teachers to review recordings of their lessons, indicate the modes and 'interactures' they were engaged in with their class as the lesson progressed, and reflected on these practices for continuing self-improvement. The focus of reflection is on the interaction between teacher and students, in order to develop 'classroom interactional competence' (Walsh, 2013), and as such the framework is useful and relevant to our own analysis and quest for deeper understanding of the conversations in the TSCC. It was designed for use by teachers but the generic interaction-based aspects of SETT are still applicable to students as well, even if the teacher-driven management aspects are not.

Within the SETT framework, a **mode** is a 'classroom micro-context' and an **interacture** is an 'interactional feature'. Thus, classroom interaction is framed as a series of interactions and micro-contexts, where discourse is co-constructed by teachers and students, and the resulting conversations support and enable student learning (Walsh, 2013). SETT is a way for teachers to reflect on these interactions, in the scenario where their lessons have been recorded, and notice where learning opportunities and a 'space for learning' are created (Walsh and Li, 2012). This is in line with proposals for interactive and engaging learning environments in state school classrooms, which may equally be applied to a language school scenario (Alexander, 2008; Mercer, 2019).

3.1 SETT modes

There are four modes in the SETT framework. These are listed and defined below:

- **Managerial**: to transmit information, refer learners to materials, introduce/conclude an activity, or change from one mode of learning to another;

- **Classroom context**: to enable learners to express themselves clearly, establish a context, and promote oral fluency;

- **Materials**: to provide language practice around a piece of material, to elicit responses in relation to the material, check and display answers, clarify if needed;

- **Skills & systems**: to enable learners to produce correct forms, manipulate target language, to provide corrective feedback, and display correct answers*.

* For reasons explained below in section 3.3 we reduced these four modes to three for our annotation exercise, merging 'skills & systems' with 'materials'.

3.2 SETT interactures

We use the following nine original SETT interactures, and based on our initial experience annotating lesson transcriptions, we augmented these with an additional three interactures which are marked in italics below:

- **Confirmation check (CC)**: the teacher confirms that they have understood the learner's utterance, or vice versa;

- **Display question (DQ)**: a question to which the teacher knows the answer;

- **Direct repair (DR)**: the teacher corrects an error quickly and directly;

- ***Enquiry (EN)***: the learner asks a language question.

- **Extended teacher/learner turn (ExtT)**: a turn containing either more than one substantial main clause, many relative clauses, at least one long relative clause, or a combination of such clauses;

- **Form-focused feedback (FBF)**: the teacher gives explicit feedback on the words or form used by the learner, rather than the perceived intended meaning of their utterance;

- ***Instruction (IN)***: the teacher gives direct instructions;

- **Referential question (RQ)**: a genuine question to which the teacher does not know the answer, which typically encourages extended learner turns;

- **Scaffolding:Extension (S:E)**: the teacher does not accept a learner's first answer, implicitly or explicitly encouraging more output;

- **Scaffolding:Modelling (S:M)**: the teacher provides an example of the target language feature for the learner;

- ***Scaffolding:Presentation (S:P)***: the teacher explains a language point;

- **Seeking clarification (SC)**: the teacher asks a student to clarify something the student has said, or vice versa;

It is apparent that SETT is mainly teacher-focused but does have some capacity for application to student turns: the scaffolding, repair, instruction, question, and feedback interactures are almost certain to be applied to teacher turns, but the clarification, confirmation and extended turn interactures could be on either side, and enquiry is intended for student turns.

3.3 SETT annotation in the TSCC

For this new version of the corpus, we selected 50 lessons for annotation of modes and interactures, in order to investigate the types of teacher-student dialogues and pedagogical observations we could make in our dataset. Based on initial attempts to make annotation decisions in practice, we adapted the existing SETT labels so that the modes were reduced from 4 to 3 different types, and 3 new interactures were appended to 9 of the originals. In terms of modes, we found that it was difficult in practice to distinguish between 'materials' and 'skills & systems', since both relate to affording the opportunity for students to display what they know and to provide feedback accordingly. Therefore these two modes were merged into one for practical purposes.

As an exploratory exercise, we annotated the first 50 lessons in the corpus, for SETT modes and interactures on both the teacher and student side. One annotator carried out the work, based on clear guidelines – in future, it would be beneficial to collect multiple annotations for the same transcriptions, and to cover more lessons from the corpus. Here we report on the results of this initial annotation exercise, finding overall that the distribution of modes and interactures between teachers and students is broadly as expected on the basis of their definitions.

Mode	Teacher	Student
classroom context	18.2	26.2
managerial	42.1	27.1
materials/skills	30.1	41.1
multi-modes	9.6	5.6

Table 5: Proportion of SETT modes for teachers and students in a sample of 50 lessons from the TSCC (%).

Firstly it is worth noting that the proportion of turns between teacher and student is approximately even in the transcriptions as a whole (at a ratio of 53:47 respectively). Nevertheless, three times as many modes are set by the teacher as by the student. This is to be expected because the modes relate to lesson management and pedagogical acts. Table 5 shows how the three modes are distributed for teachers and students. For teachers, most of the modes they set are managerial, whereas the students mostly set modes for materials or skills practice. A small number of turns involved multiple modes at once.

Then in terms of interactures, we found that there were four times as many identifiable interactures by teachers as there were by students. On the one hand this fits with the fact that SETT was developed with teachers in mind, and on the other hand indicates that more of the interactional moves in a one-to-one lesson are made by the teacher, as might be expected. Specifically, instruction, feedback, repair, questions and scaffolding tended to be on the teacher side, whereas enquiry tended to be on the student side. Both teachers and students used extended turns, confirmation checks and sought clarification.

Figure 1 shows how student and teacher interactures differ both in magnitude (the teacher bars tend to reach higher on the y-axis) and type (the distribution of bars on the x-axis is quite different). In future work, we intend to analyse how modes and interactures relate to each other, since they were not devised as independent variables but ones which interplay and depend on each other to some extent. The SETT framework sets out some expected mode-interacture correspondences, and this is something that warrants investigation in our own dataset. The annotation of 50 lessons within the SETT framework is included in this second release of the corpus.

4 Classification experiments

As well as attempting to understand how teacher-student chatroom interactions progress during and across lessons, we can also attempt to apply machine learning techniques to predict features of the data. It is potentially useful to be able to predict when to introduce new sequences, and as such we report on experiments which detect and classify sequence shifts within chatroom transcripts. It is common practice in modern NLP to apply transfer learning methods whereby large language models pre-trained with transformers are 'fine-tuned' to a given task and dataset (Ruder et al., 2019). The BERT model is the best-known example of this, but there are many derivatives and alternative models (Devlin et al., 2019; Rogers et al., 2020).

We apply the transfer learning approach to our problem of classifying sequence shifts in the chatroom transcripts. Using our corpus of chatroom lessons, the classifier is trained to learn when new discourse sequences begin. Given a turn t_i from the corpus, the machine learning task is to predict whether a new discourse sequence begins in the next turn t_{i+1} or not.

4.1 Data preparation

The lesson transcripts in the TSCC need to be prepared for the machine learning task: the reshaped dataset is included with the new corpus release. We cast the text classification task as a binary one of *new sequence detection* – that is, does a new sequence begin after the current turn, or not? The initial input string is therefore turn t_i and the corresponding label comes from t_{i+1} as a 0 or 1.

To exemplify, consider the imagined turns below between teacher (T) and student (S):

turn		**label**
1	T: Does that all make sense?	0
2	S: yes, understood.	0
3	T: Good, time for some revision!	1
4	S: ok	0

If we consider turn 1 here, then the input string is 'does that all make sense?' and its corresponding label comes from turn 2; i.e. 0. With turn 2 on the other hand, the input string is 'yes, understood.', and the label is 1 because turn 3 marks the start of a new discourse sequence relating to revision.

Moreover, we experiment with longer inputs by using the special separator token `[SEP]` available in the BERT-ish vocabulary[2]. Thus, two text strings may be passed to a BERT-ish model, with `[SEP]` between them, and we use this to include the preceding turn t_{t-i} when learning to detect sequence shifts. This takes advantage of the long inputs which large pre-trained models can handle (usually 512 tokens[3]) and models an intuition that the preceding turn is useful context when determining whether a new discourse sequence is needed.

To exemplify these longer input strings, we return to the imagined turns between teacher and student. Looking at turn 2, the input string becomes a concatenation of turn 1, the `[SEP]` token, and turn 2 (lower-cased) –

> does that all make sense? `[SEP]` yes, understood.

– and the label is 1. For comparison, the input string for turn 3 is –

> yes, understood. `[SEP]` good, time for some revision!

– and the label is 0, because turn 4 does not involve a new discourse sequence.

In subsequent variations, we experiment by prefixing the current turn t_i with the *two* previous turns, to incorporate more of the preceding context, and we introduce two new special tokens `[t]` and `[s]` at the start of each turn, to indicate whether it is the teacher's or student's turn. The intuition here is that, since teachers and students play different roles in the discourse, it may be useful to signal which one is chatting when.

4.2 Implementation

We opt to work with the DistilBERT compressed language model rather than a larger language model, because it brings energy savings without compromising greatly on performance (Sanh et al., 2019). In addition, a model which is faster for inference would be beneficial in CALL applications where users do not want to be kept waiting overly long. We use the `transformers` Python library from HuggingFace (Wolf et al., 2020), obtaining the pre-trained model and tokenizer for

[2] The `[SEP]` token exists because one of the original training tasks for BERT was next sentence classification – this can be used to tackle question answering challenges, by concatenating the question and answer with `[SEP]` in between (Devlin et al., 2019).

[3] Note that tokens in the context of transformer language models are 'subword tokens' automatically derived from training corpora via byte-pair encoding or an algorithm such as WordPiece (Gage, 1994; Schuster and Nakajima, 2012).

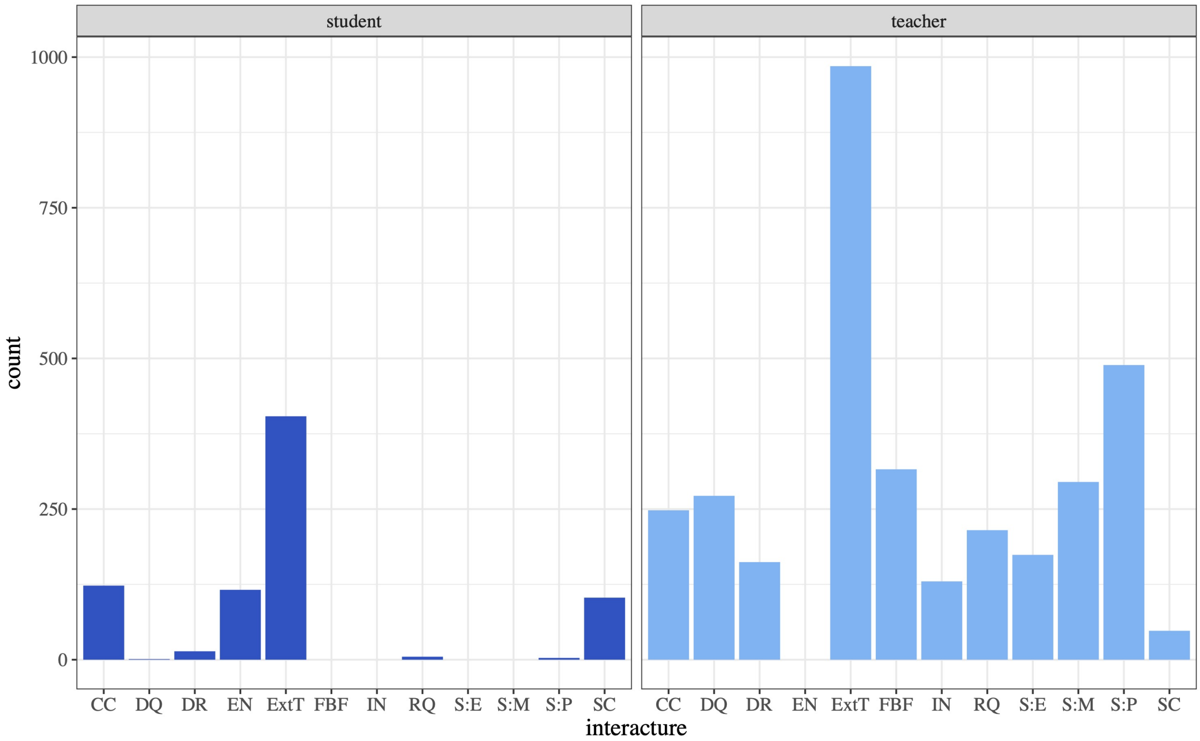

Figure 1: Frequency of interactures by students and teachers in a sample of 50 lessons from the TSCC.

DistilBERT 'base' (smaller than 'large') uncased (the vocabulary is all in lowercase).

We prepare the turns from the 260 chatroom lessons in our corpus in the formats described above. Each data instance is a turn prefixed with 0, 1 or 2 preceding turns. We randomly split these instances into an 80:20 train-test split. The majority of chat turns are not succeeded by a new sequence. Therefore we have a class imbalance whereby approximately 30% of turns bear a positive label, the remainder are negative. To address this issue, we weight the positive instances three times more than the negative ones in the loss function.

To fine-tune DistilBERT on our classification task, we use the built-in `transformers` trainer on 2 GPU for 2 epochs per experiment, with the default batch size of 8, AdamW optimizer (Loshchilov and Hutter, 2019), initial learning rate of $5e$-05 and linear learning rate scheduler.

Our evaluation measures are precision (true positives over true positives and false positives) and recall (true positives over true positives and false negatives). We also report the F_1 scores which are the harmonic mean of precision and recall.

For comparison, we implement two probabilistic baselines based on statistical information in the training data. The first is based on the proportion of new sequences over all the turns in the training set (**overall prob**) – 0.288 – using that probability as a weight in randomly predicting whether a turn is followed by a new discourse sequence or not.

The second baseline uses information from the training data as to the number of turns between new discourse sequences (**sequence length prob**). For each turn in the training data we record the sequence length (in turns) at that point. Thus we can say how many times we have observed a sequence of length l and how many times we see a sequence one turn longer $(l + 1)$. The probability of a new sequence given a sequence of length l is thus the count of sequences of that length (c_l) divided by the sum of c_l and the count of times we see a sequence one longer than l (c_{l+1}). This is a way of stating how probable we think it is that a sequence will stop at length l:

$$p_{new.seq} = \frac{c_l}{c_l + c_{l+1}} \quad (1)$$

Then for each turn in the test set, a prediction of 0 or 1 for a new sequence is generated using $(1 - p_{new.seq})$ and $p_{new.seq}$ as sample weights respectively. We also impose an upper bound on the length of a sequence, given the longest seen in the

Expt	P	R	F_1
Overall prob†	.291	.290	.291
Sequence length prob†	.288	.584	.386
Current turn t_i	.377	.433	.403
+ role tokens	.382	.455	.415
+ 1 previous turn	**.398**	**.636**	**.489**
+ role tokens	.391	.454	.420
+ 2 previous turns	.393	.515	.445
+ role tokens	.395	.447	.420

Table 6: Text classification experiments to automatically detect new discourse sequences in the following turn t_{i+1}: precision, recall, and F_1-measure. † indicates the mean of 100 runs. Best performance in bold.

training data: 32 turns. We run both baselines one hundred times each and report average results in Table 6.

4.3 Results

As shown in Table 6, we find that the best performing model is the one trained on the current turn t_i concatenated with the previous one t_{i-1}, mainly due to much better recall than the other experiment settings. This way of preparing the data outperforms the basic case of only passing the current turn as input to the model, as well as the additional context available from two previous turns. Prefixing each turn with the special teacher and student tokens `[t]` and `[s]` only helped in the basic case of having only turn t_i as input: it did not help when one or two preceding turns were included.

All models outperform the probabilistic baselines, suggesting that a machine learning approach is a good direction for future work. It may be that a hybrid approach involving heuristics, additional features and transfer learning will bring further advances, as discussed below.

4.4 Discussion

There are other variations that could be tried to improve the performance of our models. Among these are pre-trained language models which are larger than DistilBERT, albeit with greater environmental impact (Strubell et al., 2019), or which can take longer inputs (e.g. Big Bird or Longformer (Zaheer et al., 2020; Beltagy et al., 2020)). Different hyperparameters might be trialled, along with different ways of representing the text such as additional features or encodings with the input strings. It might be helpful, for instance, to include grammatical error detection as a pre-processing task, since it may be that certain errors are associated with new sequences such as scaffolding, elicitation or presentation. A temporal feature might help determine when to shift topics or call on management sequences such as homework and lesson closure.

Furthermore, the task could be reformulated as teacher-centric: for CALL, it may only be necessary to model the teacher's shift in discourse sequences rather than both teacher and student shifts as we have done here. This would fit with the perspective of the teacher as manager (Legutke and Thomas, 1991). In future, models could be trained to only predict the teacher side of the discourse and to steer the lesson in an adaptive, orderly and meaningful way.

In addition, human evaluation would be beneficial because our notion of 'ground truth' here is based on a series of teacher-student dyads and the discourses they built on specific occasions, and the judgements of the annotators who identified sequence shifts and sequence types in the lesson transcriptions. Aside from the lesson beginning with an opening sequence and ending with a closing sequence, there is in reality no absolute *truth* as to when new sequences are required. Each lesson could have been constructed in a myriad different ways and still be perfectly good. Therefore, evaluation via precision and recall is a decent indicator, but does not tell the whole story. It may be that we can train a new sequence classifier on such data as the TSCC, but that the best measures of performance will be derived from human-computer interaction.

Beyond the detection of new sequences, it may also be useful to automatically predict which sequence type comes next. So far we have approached the problem as binary classification, but the annotation exists in the TSCC to train a multiclass classifier identifying the types listed in the appendix – a much more challenging proposition. However, decisions would need to be made whether to separate the major and minor sequence types into separate machine learning tasks, or to tackle them both at the same time. Also, many sequences are multi-label in the sense that there can be more than one sequence type associated with a given turn. This makes the machine learning task harder, and has implications for how the data should be prepared and the models evaluated.

5 Related work

Caines et al. (2020) featured a review of other work related to the TSCC, and we refer the interested reader to that section of the paper rather than repeat it here. In the intervening period others have cited the original TSCC paper, and we wish to highlight some of those new publications[4].

A similar dataset has been produced by Yuan et al. (2022) – ErAConD, an Error Annotated Conversational Dialog Dataset – which is intended for research into grammatical error correction in English chat conversations. ErAConD features 186 conversations between crowdworkers and the BlenderBot dialogue system (Roller et al., 2020). Some distinguishing features are that the conversations are between human and machine rather than human-to-human, and the error annotation has been carried out in a manner similar to Náplava et al. (2022). Like the TSCC, ErAConD is available for research use[5].

There has been other research using the Blender chatbot, along with GPT-3 (Brown et al., 2020), to construct AI teachers (Tack and Piech, 2022), using the student turns in the first version of the TSCC and the mathematics Uptake dataset (Demszky et al., 2021) to generate and evaluate chatbot responses. Tack & Piech found that the models performed well on conversational uptake (how well the response expanded on the student input) – especially Blender – but still have some way to go in terms of realism, comprehension and helpfulness. In addition, Tyen et al. (2022) seek to automatically adapt Blender outputs for different levels of English proficiency using a variety of different methods and English language resources. The prompt the adapted models to 'self-chat' and find that a re-ranking approach works best, after evaluating the level of the chats with human examiners.

Filighera et al. (2022) focus on improved feedback systems for language learners, giving short answer feedback to explain scores for German and English exercises. Nguyen et al. (2022) give an assessment of the state-of-the-art for educational technologies and how well they handle code-switching, pointing to future directions and opportunities for research. In this second version of the TSCC, the turns which feature words from languages other than English are labelled as 'non-English' sequences. This does not mean that the turns are entirely in another language – though they may be – but rather that there is at least some non-English present in the turn. It may be fruitful to identify whether those turns tend to be explanatory (the teacher drawing on another language to build knowledge of English) or naturalistic conversational code-switching.

Jain et al. (2022) present EDICA (Educational Domain Infused Conversational Agent), a virtual agent for language teaching. They fine-tune the GPT-2 language model (Radford et al., 2019) on the CIMA dataset of Italian tutoring dialogues collected from crowdworkers role-playing student and tutor roles (Stasaski et al., 2020). CIMA is enriched with conceptual information about the exercises and the actions taken by the students. This kind of meta-information is an approach we could consider for future work with the TSCC.

Two new corpora have been created: the first a corpus of online lessons in Russian as a foreign language (RuTOC; (Lebedeva et al., 2022)), and the second a corpus of Korean task-oriented dialogue data (Seung-Kwon et al., 2022). Notably, the latter states that the aim is to collaborate with human teachers, not replace them; a sentiment we echo.

6 Conclusions & future work

In this paper we have described the second version of the Teacher-Student Chatroom Corpus. The new version adds another 158 hour-long chatroom transcripts to the 102 lessons in version 1 of the corpus. Two teachers and thirteen students are involved, representing seven L1s, and ranging from CEFR proficiency level B1 to C2. The new transcripts have been annotated in the same way as those in the first version, and a subset of 50 transcripts have been annotated for SETT modes and interactures.

We presented some initial experiments to automatically detect new discourse sequences. We showed that a fine-tuned DistilBERT model could outperform probabilistic baselines in detecting new sequences, based on a concatenation of the preceding and current turn. There remains room for improvement through further experimentation and feature-engineering, as well as alternative evaluation methods where we move from the idea of a single ground truth to human ratings of tim-

[4] Citing papers were obtained from Google Scholar (accessed 11 October 2022).

[5] See https://github.com/yuanxun-yx/eracond

ing and appropriateness. In these machine learning experiments we are working towards discourse modelling in pedagogic scenarios; in future, such models could be applied to online tutoring applications where we wish to guide the lesson from sequence to sequence.

Other future plans include further expansion of the corpus, and work to develop teacher feedback systems to aid in teacher training and professional development.

Acknowledgments

This paper reports on research supported by Cambridge University Press & Assessment, University of Cambridge. Additional funding was provided by the Cambridge Language Sciences Research Incubator Fund and the Isaac Newton Trust. We thank Helena Edmondson, Rowan Amber Jones, Olga Cherrington, and Professors Michael McCarthy & Steve Walsh for their help and contributions to this work. We are grateful to the teachers, students and annotators without whose enthusiastic participation this corpus would not have been feasible. We also thank audience members at the Teaching and Language Corpora Conference (TaLC) 2022, and the anonymous NLP4CALL reviewers for their feedback.

References

Robin Alexander. 2008. *Towards Dialogic Teaching: Rethinking Classroom Talk (4th edn)*. York: Dialogos.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document Transformer. *arXiv*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

Andrew Caines, Helen Yannakoudakis, Helena Edmondson, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery. 2020. The teacher-student chatroom corpus. In *Proceedings of Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL)*.

Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori Hashimoto. 2021. Measuring conversational uptake: A case study on student-teacher interactions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Anna Filighera, Siddharth Parihar, Tim Steuer, Tobias Meuser, and Sebastian Ochs. 2022. Your answer is incorrect... would you like to know why? introducing a bilingual short answer feedback dataset. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Raghav Jain, Tulika Saha, Souhitya Chakraborty, and Sriparna Saha. 2022. Domain infused conversational response generation for tutoring based virtual agent. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*.

Maria Yu Lebedeva, Antonina N Laposhina, Natalia A Alksnit, and Tatyana V Lyashenko. 2022. RuTOC: A corpus of online lessons in Russian as a foreign language. *Philological Class*, 27.

Michael Legutke and Howard Thomas. 1991. *Process and Experience in the Language Classroom*. London: Routledge.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Neil Mercer. 2019. *Language and the Joint Creation of Knowledge: the selected works of Neil Mercer*. Abingdon: Routledge.

Jakub Náplava, Milan Straka, Jana Straková, and Alexandr Rosen. 2022. Czech grammar error correction with a large and diverse corpus. *Transactions of the Association for Computational Linguistics*, 10.

Li Nguyen, Zheng Yuan, and Graham Seed. 2022. Building educational technologies for code-switching: Current practices, difficulties and future directions. *Languages*, 7(3).

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot. *arXiv*, 2004.13637.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in Natural Language Processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18.

Harvey Sacks, Emanuel Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696–735.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and Korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. IEEE.

Choi Seung-Kwon, Lee Yo-Han, and Kwon Oh-Wook. 2022. A study on task-oriented dialogue data of a dialogue system for foreign language tutoring: Focusing on Korean dialogue data. *Foreign Languages Education*, 29(1):105–124.

Katherine Stasaski, Kimberly Kao, and Marti A. Hearst. 2020. CIMA: A large open access dialogue dataset for tutoring. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Anaïs Tack and Chris Piech. 2022. The AI Teacher test: Measuring the pedagogical ability of Blender and GPT-3 in educational dialogues. In *Proceedings of the 15th International Conference on Educational Data Mining (EDM)*.

Gladys Tyen, Mark Brenchley, Andrew Caines, and Paula Buttery. 2022. Towards an open-domain chatbot for language practice. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*.

Steve Walsh. 2006. *Investigating Classroom Discourse*. London: Routledge.

Steve Walsh. 2013. *Classroom Discourse and Teacher Development*. Edinburgh: Edinburgh University Press.

Steve Walsh and Li Li. 2012. Conversations as space for learning. *International Journal of Applied Linguistics*, 23:247–266.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Xun Yuan, Derek Pham, Sam Davidson, and Zhou Yu. 2022. ErAConD: Error annotated conversational dialog dataset for grammatical error correction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big Bird: Transformers for longer sequences. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*.

Appendix: annotation types in TSCC 2.0

In this section we provide a full list of sequence types and teaching foci. We also give a list of resources encountered so far, but note that this is an open-ended class, because the labels are data-driven and the possibilities are endless (though slow-growing). For the most part, the labels and their definitions are copied over from the original TSCC paper (Caines et al., 2020), with some amendments as described in section 2.

Sequence types: We indicate major and minor shifts in conversational sequences – sections of interaction with a particular purpose. We define a number of sequence types listed and described below, firstly the major and then the minor types, or 'sub-sequences':

- Opening – greetings at the start of a conversation; may also be found mid-transcript, if for example the conversation was interrupted and conversation needs to recommence.
- Topic ___ – relates to the topic of conversation (minor labels complete this sequence type).
- Exercise – signalling the start of a constrained language exercise (*e.g.* 'please look at textbook page 50', 'let's look at the graph', *etc*); can be controlled or freer practice (*e.g.* gap-filling versus prompted re-use).
- Redirection – managing the conversation flow to switch from one topic or task to another.
- Disruption – interruption to the flow of conversation for some reason; for example because of loss of internet connectivity, telephone call, a cat stepping across the keyboard, and so on...
- Homework – the setting of homework for the next lesson, usually near the end of the present lesson.
- Closing – appropriate linguistic exchange to signal the end of a conversation.
- Admin – lesson management, such as 'please check your email' or 'see page 75' (*compared to version 1: moved from 'teaching focus'*).
- Free practice – ... (*new in version 2*).
- Non-English – ... (*new in version 2*).

  Below we list our minor sequence types, which complement the major sequence types:

  - Topic opening – starting a new topic: will usually be a new sequence.
  - Topic development – developing the current topic: will usually be a new sub-sequence.
  - Topic closure – a sub-sequence which brings the current topic to a close.
  - Presentation – (usually the teacher) presenting or explaining a linguistic skill or knowledge component.
  - Eliciting – (usually the teacher) continuing to seek out a particular response or realisation by the student.
  - Scaffolding – (usually the teacher) giving helpful support to the student.
  - Enquiry – asking for information about a specific skill or knowledge component.
  - Repair – correction of a previous linguistic sequence, usually in a previous turn, but could be within a turn; could be correction of self or other.
  - Clarification – making a previous turn clearer for the other person, as opposed to 'repair' which involves correction of mistakes.
  - Reference – reference to an external source, for instance recommending a textbook or website as a useful resource.
  - Recap – (usually the teacher) summarising a take-home message from the preceding turns.
  - Revision – (usually the teacher) revisiting a topic or task from a previous lesson.

Teaching focus: Here we note what type of knowledge is being targeted in the new conversation sequence or sub-sequence. Note that these do not accompany every sequence type – they are only used where applicable.

- Grammatical resource – appropriate use of grammar.
- Lexical resource – appropriate and varied use of vocabulary.
- Meaning – what words and phrases mean (in specific contexts).
- Discourse management – how to be coherent and cohesive, refer to given information and

introduce new information appropriately, signal discourse shifts, disagreement, and so on.

- Register – information about use of language which is appropriate for the setting, such as levels of formality, use of slang or profanity, or intercultural issues.
- Task achievement – responding to the prompt in a manner which fully meets requirements.
- Interactive communication – how to structure a conversation, take turns, acknowledge each other's contributions, and establish common ground (*does not yet feature in the corpus*).
- World knowledge – issues which relate to external knowledge, which might be linguistic (*e.g.* cultural or pragmatic subtleties) or not (they might simply be relevant to the current topic and task).
- Meta knowledge – discussion about the type of knowledge required for learning and assessment; for instance, 'there's been a shift to focus on X in teaching in recent years'.
- Content – a repair sequence which involves a correction in meaning; for instance, Turn 1: Yes, that's fine. Turn 2: Oh wait, no, it's not correct.
- Writing - a focus on writing skills and orthographic issues such as spelling, grammar, punctuation (*new in version 2, and subsumes 'typo' from version 1*).
- Speaking - a focus on speaking skills (*new in version 2*).
- Listening - a focus on listening skills (*new in version 2*).
- Reading - a focus on reading skills (*new in version 2*).
- Exam prep – specific drills to prepare for examination scenarios, as well as discussions around exam strategy (*updated label and definition for version 2*).

Use of resource: At times the teacher refers the student to materials in support of learning. The resources encountered so far are, `book, chat, dictionary, movie, sample paper, social media account, student's writing, textbook, video, website`.