ASR-based CALL systems and learner speech data: new resources and opportunities for research and development in second language learning

Catia Cucchiarini[a], Stephen Bodnar[a], Bart Penning de Vries[a], Roeland van Hout[b] and Helmer Strik[a,b]

[a] Centre for Language and Speech Technology, Radboud University, Nijmegen, The Netherlands
[b] Department of Linguistics, Radboud University, Nijmegen, The Netherlands

E-mail: C.Cucchiarini@let.ru.nl; s.bodnar@let.ru.nl; b.penningdevries@let.ru.nl; r.vanhout@let.ru.nl; h.strik@let.ru.nl

Abstract

In this paper we describe the language resources developed within the project "Feedback and the Acquisition of Syntax in Oral Proficiency" (FASOP), which is aimed at investigating the effectiveness of various forms of practice and feedback on the acquisition of syntax in second language (L2) oral proficiency, as well as their interplay with learner characteristics such as education level, learner motivation and confidence. For this purpose, use is made of a Computer Assisted Language Learning (CALL) system that employs Automatic Speech Recognition (ASR) technology to allow spoken interaction and to create an experimental environment that guarantees as much control over the language learning setting as possible. The focus of the present paper is on the resources that are being produced in FASOP. In line with the theme of this conference, we present the different types of resources developed within this project and the way in which these could be used to pursue innovative research in second language acquisition and to develop and improve ASR-based language learning applications.

Keywords: second language acquisition, automatic speech recognition, computer assisted language learning

1. Introduction

Initiatives such as Speech and Language Technology in Education (SLaTE, www.sigslate.org), CALICO (https://www.calico.org/), and BEA (Workshops on Innovative Use of Natural Language Processing (NLP) for Building Educational Applications) aim at stimulating cross-fertilization between the fields of language acquisition and education on the one hand, and Human Language Technologies (HLT) on the other. In most cases, the information flows from education research to technology development, in the sense that knowledge and expertise on education and pedagogy inform the development of Computer-assisted Language Learning (CALL) systems (Ellis & Bogart, 2007).

Another possibility for fruitful cross-fertilization is to implement CALL systems to advance research in second language acquisition (SLA). So far, this kind of research has been mainly limited to written language (Sauro, 2009; Heift & Rimrott, 2012). This can partly be ascribed to the scepticism surrounding the technology that analyzes spoken output from language learners, namely Automatic Speech Recognition (ASR) technology, especially when applied to non-native speech (Benzeghiba et al. 2007).

However, recent developments in ASR, and in particular in ASR of non-native speech (Eskenazi, 2009;Van Doremalen et al. 2009; 2010), open up new avenues of research for SLA by making it possible to accurately analyze non-native speech output at different levels of detail. In turn, this allows a variety of studies on second language speaking development. One of the advantages is then that the effects of different factors influencing second language speaking can be investigated in a controlled and systematic way through the use of ASR-based CALL systems.

In Penning de Vries et al. (2010) we argued that such a form of interaction between CALL and SLA research could be useful to study the role and effectiveness of corrective feedback (CF) in oral proficiency. We provided a review of various studies that addressed the topic of CF in SLA and showed that many problems still need to be solved to clarify the role of CF. In addition, we suggested that an ASR-based CALL system could be used to study the effect of CF on second language speaking under near-optimal conditions. This is precisely what we have been investigating in the project "Feedback and the Acquisition of Syntax in Oral Proficiency" (FASOP). In this project an ASR-based CALL system has been developed and is being used to carry out research on the role of CF in SLA and on the complex relationship between (types of) CF, practice, language proficiency and motivation. We have already reported on the background and the results of this project related to use of the ASR-based CALL system, the role of CF in the acquisition of oral syntax and the complex relationship between CF, motivation and acquisition in other papers (Bodnar et al. 2011; Penning de Vries et al., 2010; 2013; 2014). The focus of the present paper is not so much on the research that is being carried out within this project, but rather on the resources that are being produced as a result of the research being carried out. In line with the theme of this conference, we intend to present the different types of resources developed within this project and the way in which these could be used to pursue innovative research in SLA and to develop and improve ASR-based language learning applications..

2. Feedback and the Acquisition of Syntax in Oral Proficiency: the FASOP project

In the SLA literature the role of CF still constitutes an important topic of research and debate. Although many studies show that CF can be useful (Norris & Ortega,

2000; Russell & Spada, 2006; Lyster & Saito, 2010), it is not clear under which conditions CF is most effective (Lyster et al. 2013). Many factors are emerging as mediating CF effectiveness such as educational setting, type of CF (Lyster & Saito, 2010), and learner differences (Dornyei, 2005). Controlled research is required to further investigate these issues (Russel & Spada, 2006; Goo & Mackey, 2013), in particular in on-line processing like oral L2 learning. As a result, an improved method to research CF could benefit the field of SLA, and language education in general.

The FASOP project aims at investigating the effectiveness of various forms of practice and feedback on the acquisition of syntax in L2 oral proficiency, as well as their interplay with learner characteristics such as education level, learner motivation and confidence. For this purpose use is made of a CALL system that employs ASR technology to allow spoken interaction and to create an environment for experiments that guarantees as much control over the language learning setting as possible. An ASR-based CALL system has the capability to provide CF that is systematic, consistent, intensive, and clear enough to be perceived as such, and that prompts self-repair and modified output. In addition, complementary instruments such as questionnaires can be integrated into the environment and employed to measure important learner characteristics such as motivation, attitude and confidence.

The aspect of oral proficiency that is the focus of this project is grammatical accuracy. This is an important component of oral proficiency (Housen & Kuiken, 2009) that cannot be sufficiently practiced in the oral modality in traditional classroom settings because it requires much time from students and teachers. The grammatical feature specifically addressed in the project is subject-verb inversion in Dutch main clauses. In Dutch the finite verb appears in second position irrespective of the first constituent. If the finite verb is preceded by a constituent other than the subject NP, the verb remains in second position and the subject comes after the finite verb. This feature appears to be problematic for learners of Dutch (Jordens, 1988). We designed a system that provides learners with the opportunity to practice grammatical structures and internalize grammatical rules, receiving corrective feedback whenever required, based on ASR. The ASR system has to deal with non-native, learner speech at the beginner level, which is of course challenging. In addition, the system has to provide accurate CF, so as not to confuse the learners.

3. The GREET system in the FASOP project

In this section we describe how the GREET system has been designed (3.1), how it is employed to log all sorts of events that might be relevant for our research purposes (3.2), and how it has been used so far to conduct experiments in the FASOP project (3.3).

3.1 The architecture of GREET system

In Figure 1 we give a schematic overview of the GREET practice system taken from Penning de Vries et al. (2014). Learners interact with the system through a graphical user interface (GUI) and begin by logging into the system. Subsequently, they receive tasks to work on from the courseware database. They first have to watch short video clips (30-45 s) and after each video they are asked three to five questions about the clip. The responses to these questions are to be formed by using given words or segments of a sentence which are shown on the screen, in random order, and the learner has to mentally arrange these “word blocks” to build and then record a sentence. Providing sentence segments as building blocks constrains the number of possible responses and contributes to higher ASR accuracy.

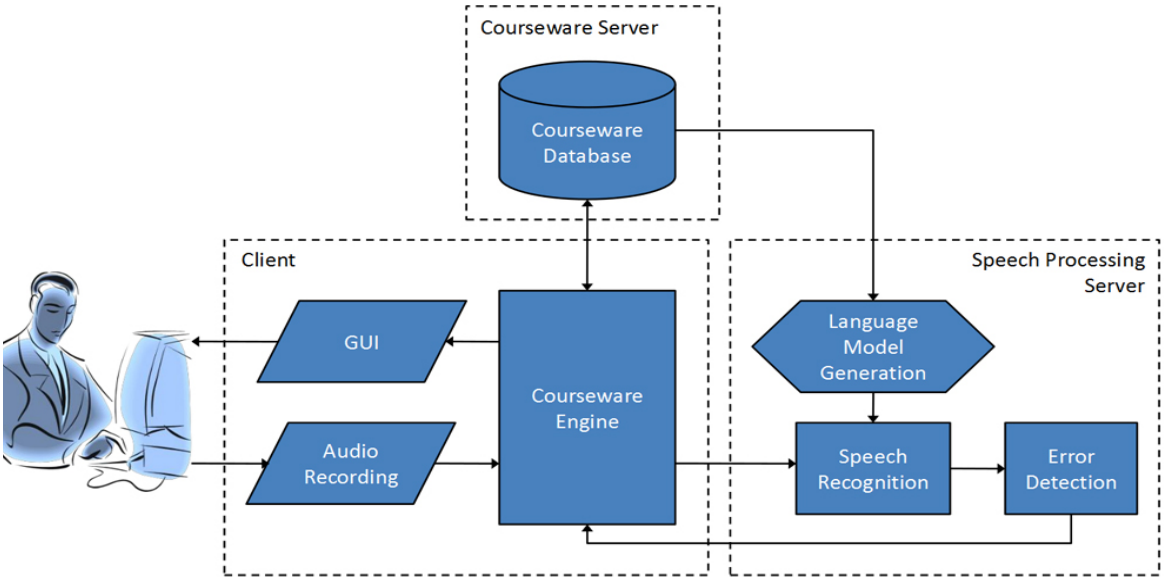

Figure.1. A schematic overview of the GREET system

For each question in the exercise, the language model contains all possible permutations (i.e. all possible sentences) that can be created with the word blocks. These are labelled with meta-information that indicates whether that answer-sentence is grammatically correct or incorrect. After the learner has recorded an utterance, the speech recognizer computes a recognition result and a confidence level. In the next step, the recognizer maps the utterance to an answer-sentence in the language model. If the confidence level is below a preset threshold, the system does not accept the recognition result and instead assumes that the learner did not record a valid attempt. In this case the system will not attempt to detect errors, but will ask the learner to try again. Alternatively, if the confidence level is sufficiently high, the recognition result is mapped to an answer sentence, and the corresponding message regarding the sentence's grammaticality is sent to the courseware engine.

The final step in this procedure is the presentation of a feedback message to the learner. This is produced based on the result of the error detection stage. The courseware engine determines what type of feedback is presented, and how it is presented. As will be explained below this can vary depending on experimental design requirements.

Although in our design we took the limitations of ASR into account, it is still necessary to test whether the system is effective in improving learners' proficiency. In Penning de Vries et al. (2014) we showed that this is indeed the case.

3.2 Log capabilities in the GREET system

An important aspect of the GREET system is that it logs all system-learner interactions. In this way various indicators of learner behavior and preferences can be calculated based on the events recorded. This allows for more insight into learner behavior and learning process.

All utterances by the users are recorded and the ASR component makes it possible to check and analyze the speech produced. The interactions are stored in a database and this allows us to look in detail at learner behavior and inspect the logs for irregular behavior. We store interaction data that can be relevant for research purposes, in our case to study the impact of corrective feedback or the relationships between feedback, proficiency and motivation. These data include page views, number of video clips viewed, number of questions viewed, time on different types of pages, number of reformulations (which we define as the second, third, or further attempts to answer a question), time spent on producing reformulations, number of recordings, ASR recognition results, and type of feedback returned. A complete overview of events that are recorded and of measures that are calculated is presented in the Appendix.

When a learner begins an activity, the system creates a practice session object to store the events that occur. In later analyses these sessions serve as records of the interactions that took place during practice.

3.3 Experiments with the GREET system

Experiments are run through a website and the learner interacts with the system through a web browser. All recognition is performed on the web server. This allows us to run more experiments at the same time and at different locations and ensures that the speech materials produced are automatically stored in a central repository.

So far the GREET system has been used to carry out various experiments in which different configurations have been adopted to implement different experimental conditions. These include pilot experiments and experiments proper. In all of these cases a variety of data have been collected and stored and these can be used for various purposes. For instance, even if pilot experiments may be of limited use for research purposes, they are in any case valuable in terms of language resources, because speech recordings have been made that add to the database.

Thus far two pilot experiments and six experiments have been carried out. These differ along different dimensions such as the target group, the research focus and the forms of CF employed. With respect to the target group, we distinguish between high-educated (university students) and low-educated learners (basic education). As to the topic, we investigate the impact of CF on both language proficiency and learner motivation. Finally, we have experimented with different forms of feedback varying from no immediate feedback with the promise of a final score at the end to immediate feedback in the form of either prompts or recasts (c.f. Lyster, Saito & Sato, 2013).

To carry out these different experiments, some features of the system had to be adapted. For instance, an interesting question is whether grammar practice with CF has an advantage over practice without immediate feedback (Penning de Vries et al. 2014). For this purpose two different versions of the systems were used, one that provided immediate feedback in the form of prompts as described in Section 3.1 and one that did not provide CF after each utterance, but notified the learner they would receive a score at the end. In a different experiment aimed at testing the effectiveness of different forms of feedback, the system was configured to provide either prompts or recasts, i.e. reformulations in which the error has been corrected.

As explained in Section 3.2, during all these experiments all learner-system interactions and all relevant events that take place are logged making it possible to relate these behavioral data to data concerning proficiency, motivation and metadata. This results in a rich database that can be exploited to get more insight in the learning process (Penning de Vries. et al. 2014) .

4. Language resources in the FASOP project

The information provided in the previous sections has made it clear that the GREET system, as a language resource, has at least two important functions: It can be used as an environment for conducting innovative research on language learning, but it can also be used as a CALL system for facilitating language learning. In Strik et al. (2012) we presented another ASR-based system that had been developed for practicing oral skills in Dutch L2, the DISCO system, and we showed that such a system can actually be used also for collecting other valuable language resources, namely speech recordings of L2 learners. This of course also applies to the GREET system, as will be discussed below.

4.1 GREET as a system for conducting research

In Section 3.2 we have explained how the GREET system has been employed to conduct various experiments in which specific system features could be varied to create different experimental conditions. In addition, we have indicated that during the experiments a range of additional data can be collected varying from the speech recordings to data on learner behaviour and learner motivation. Since all features can be controlled and monitored, this provides a powerful testing environment.

This testing environment can easily be adapted to address research questions other than those investigated in FASOP. For instance, one could study whether it is better to provide feedback in the written or in the oral modality, or whether practice is more effective in the written or in the oral modality (Drozdova et al. 2013). Other research possibilities include investigations of other grammatical features and other aspects of oral proficiency such as the acquisition of morphology, pronunciation or vocabulary.

GREET simultaneously gathers a multitude of log-data that contain detailed information on learner behaviour and preferences. This makes GREET a flexible research instrument and rich data source that can successfully be employed to gain more insight into the processes underlying language learning and to inform the development of new language learning applications.

4.2 GREET as a system for language learning

Research so far (Penning de Vries et al. 2014) has shown that the Dutch L2 learners that participated managed to profit from the training in the various conditions. This is of course an interesting finding that indicates the potential of such systems for language teaching in general. One of the problems in foreign language teaching is how to provide sufficient practice and feedback on oral skills in an efficient and effective way. Our studies show that GREET can be employed for successful language learning. In other words, the research environment we have created to carry out experiments is ecologically valid: the optimal conditions created for research purposes can be transferred to the practice of language teaching.

4.3 GREET as a system for collecting L2 speech

All the speech produced by the subjects participating in FASOP experiments has been recorded, analyzed by the speech recognizer and stored in a database. So far we have collected speech from 180 speakers for a total of 120 hours of recordings. Speaker metadata comprise usual data such as age, gender, proficiency level, educational level, languages spoken, age of arrival, length of L2 instruction, and frequency of computer and language use.

The speech collected is comprised of pre-test and post-test recordings in a Discourse Completion Task (DCT) (Van de Craats, 2009) and oral productions as a result of interactions with the system elicited in GREET practice (Section 3.1). In the DCT elicited oral production, participants saw the beginning of a sentence which they were required to complete. To establish some context for the task, they were given a lead-in sentence, one or two hint words, and a picture. To answer, the participant pressed the record button and spoke a full sentence.

The speech produced by the learners during GREET practice started after the learner had watched a short (30-45s) clip of an ongoing story. A virtual teacher character displayed on the screen then asked the learner questions about the content. To answer, the participants had to construct sentences using 'word-blocks': parts of a sentence that need to be combined in the correct order to form the answer to the question.

As explained in Strik et al (2012) an important advantage of speech material collected through an ASR-based CALL system is that it comes with all the relevant information for further processing. For instance, the speech comes with annotations, alignments, segmentations and confidence measures. In addition, the speech is enriched with log data about what appeared on the screen, how the user responded, how long the user waited, what action was performed (e.g., whether they spoke an utterance, moved the mouse, etc.), the feedback provided by the system, how the user reacted to this feedback (whether they corrected their answer following feedback (or not), whether they skipped the question or tried again). As a result, when language learners use GREET to practice oral skills, their utterances are recorded in such a way that it is possible to know in which context the utterance was spoken by using the information in the database logs mentioned above.

Such an enriched corpus can be used for research on language learning from various perspectives. In addition, it can contribute to developing new, improved language learning systems. After completion of the

project we would like to make our resources available. This could be done through international networks such as CLARIN or META-Net or local repositories such as the HLT Agency of the Dutch Language Union. However, this also requires that the data be organized and structured according to general standards and protocols so as to guarantee interoperability, which is beyond the scope of the present project.

5. Discussion

In the previous sections we have shown how a project that was not originally intended for developing language resources, but rather for conducting research on language acquisition, can produce a wide range of interesting resources that might be useful not only to other researchers, but also to developers of speech technology and CALL applications. Since in projects of this nature the resources are a by-product rather than the main goal, no funding is specifically allotted for designing the databases and making them available for further research. Nevertheless this is a point that deserves attention in future research programmes for different reasons.

First, if data are not accessible and available for external inspection by other researchers, any form of accountability is absent, meaning that the empirical basis of a given claim, theoretical or not, is lacking. The availability and accessibility of empirical research data is a rapidly increasing demand in academic publishing, a development that applies to language data as well. It implies that any research project has to invest more time in the future to make data accessible, auditable and exchangeable. Sometimes this is even considered a prerequisite for obtaining project funding. This is of course a positive development that will undoubtedly benefit the whole field, as it increases the possibilities for linguistic research and for meta-analyses. At the same time, because of this criterion, additional funding needs to be allotted to new research projects to make all of this possible.

Second, considering the developments in HLT, the trends in CALL and the demands on data accessibility, it is to be expected that projects like FASOP will become more frequent in the near future. This means that considerable amounts of language resources could be created by these future research projects, which is of course a welcome development in terms of efficiency and sustainability.

Against this background the various stakeholders (funding agencies, researchers and developers) should probably start thinking about how to assist researchers in making their resources available to the whole community. It is not just a question of having a central repository for storing the data, the point is how the data should be structured and made available in the form of standardized corpora and databases so that other researchers and developers can easily access them and use them for research and innovation. Making the data available requires specific expertise and time that researchers usually do not have, but that should be made available to them to the benefit of the whole language research community.

6. Conclusions

In this paper we have presented the resources that have been developed as a by-product of a project primarily focused on conducting research on second language acquisition, the FASOP project. We have explained how these resources might be exploited for other research agendas and for developing advanced applications for computerized second language learning.

Since it is very likely that projects such as these will become more common in the future, it seems that the research community as a whole should not only begin to adopt the general practice of asking individual researchers or teams to make their data available, but also to start thinking about how to facilitate them in such initiatives, since this requires expertise and efforts that go well beyond their normal work and competencies.

7. Acknowledgements

We would like to thank our colleague Joost van Doremalen for developing the ASR component of the CALL system used in this experiment. This work is part of the research program FASOP, which is funded by the Netherlands Organization for Scientific Research (NWO).

8. References

Benzeghiba, M., R. D. Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, C. Wellekens, (2007) Automatic speech recognition and speech variability: a review, *Speech Communication* 49, pp. 763--786.

Bodnar, S., Penning de Vries, B., Cucchiarini, C. Strik, H., Van Hout, R. (2011). Feedback in an ASR-based CALL system for L2 syntax: A feasibility study. *Proceedings SLaTE-2011*, Venice, Italy.

Dörnyei, Z. (2005). *The Psychology of the Language Learner: Individual Differences in Second Language Acquisition*. Lawrence Erlbaum Associates: New Jersey.

Drozdova, P. Cucchiarini, C. & Strik, H. (2013) L2 syntax acquisition: the effect of oral and written computer assisted practice, *Proceedings Interspeech 2013*, Lyon, France.

Ellis, N.C., Bogart, P.S.H., (2007). Speech and Language Technology in Education: the perspective from SLA research and practice, *Proceedings SLaTE 2007*, Farmington PA, pp. 1--8.

Eskenazi , M. (2009). An overview of spoken language technology for education. *Speech Communication*, 51, pp. 832--844.

Goo, J., Mackey, A. (2013). The case against the case against recasts. *Studies in Second Language Acquisition*, 35, pp. 127--165.

doi:10.1017/S0272263112000708.

Heift, T., Rimrott, A. (2012). Task-related Variation in Computer-assisted Language Learning. *Modern Language Journal*, 96(4), pp. 525-543.

Housen, A., Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30, pp. 461--473. doi:10.1093/applin/amp048

Jordens, P. (1988). The acquisition of word order in Dutch and German as L1 and L2. *Second Language Research*, 4, pp. 41--65. doi:10.1177/026765838800400103.

Lyster, R., Saito, K. (2010). Oral feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition*, 32, pp. 265–302. doi:10.1017/S0272263109990520.

Lyster, R., Saito, K., Sato, M. (2013). Oral corrective feedback in second language classrooms. *Language Teaching*, 46, pp. 1--40. doi:10.1017/S0261444812000365.

Norris, J., Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, 50, 417-528. doi: 10.1111/0023-8333.00136.

Norris, J., Ortega, L. (2003). Defining and measuring SLA. In C. Doughty & M. Long (Eds.) *The Handbook of Second Language Acquisition* (pp. 717--761). Malden: Blackwell.

Penning de Vries, B., Cucchiarini, C. Strik, H & Van Hout, R. (2010). The Role of Corrective Feedback in Second Language Learning: New Research Possibilities by Combining CALL and Speech Technology In *Proceedings of L2WS*, Japan.

Penning de Vries, B., Bodnar, S., Cucchiarini, C., Strik H., Van Hout, R. (2013). Spoken grammar practice in an ASR-based CALL system, *Proceedings of SLaTE 2013*, Grenoble, France, pp. 60--65.

Penning de Vries, B., Cucchiarini, C. Strik, H., Van Hout, R. (2014). Spoken grammar practice and feedback in an ASR-based CALL system, *Computer Assisted Language Learning*, DOI: 10.1080/09588221.2014.889713

Russell, J., Spada, N. (2006). The effectiveness of corrective feedback for the acquisition of L2 grammar: A meta-analysis of the research. In J. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 133-164). Amsterdam: John Benjamins Publishing.

Sauro, S. (2009). Computer-mediated corrective feedback and the development of L2 grammar, *Language Learning & Technology*, 13 (1), pp. 96--120

Strik, H., Colpaert, J., van Doremalen, J., Cucchiarini, C. (2012). The DISCO ASR-based CALL system: practicing L2 oral skills and beyond. *Proceedings of the Conference on International Language Resources and Evaluation (LREC 2012)*, Istanbul.

Van de Craats, I. (2009). The Role of 'is' in the acquisition of finiteness by adult Turkish learners of Dutch. *Studies in Second Language Acquisition*, 31, pp 59--92. doi: 10.1017/S0272263109090032

Van Doremalen, J., Strik, H., Cucchiarini, C., (2009). Utterance Verification in Language Learning Applications. *Proceedings SLATE 2009*, Wroxall Abbey.

Van Doremalen, J., Cucchiarini, C., Strik, H., (2010). Optimizing automatic speech recognition for low-proficient non-native speakers. *EURASIP Journal on Audio, Speech, and Music Processing*.

Appendix

Variable	Practice descriptors calculated from the logged events
v	The number of video pages visited
q	The number of question pages visited
redos	The number of question pages that were visited a second time or more
redoPcntSlope	The percentage of question pages that were visited a second time or more out of all the questions visited
a	The number of attempts made
OK	The number of OK feedback instances
WR	The number of wrong sequence feedback instances
DNU	The number of 'did not understand' feedback instances
skip	The number of times a learner skipped a question
aq	The number of attempts divided by the number of questions completed
pcntFTryOKs	The percentage of correct attempts out of all attempts made
pcntFTryOKFirstPassOnly	The percentage of correct attempts out of all attempts made per practice session, excluding attempts that were made on a question that had been seen one or more times earlier in practice
skipsP10Q	The number of skips that occurred for every 10 questions
tt	The total amount of time spent practising
tv	The total amount of time spent watching videos
tq	The total amount of time spent working on questions
tprepq	The total amount of time spent preparing to answer questions
treca	The total amount of time spent recording utterances
twait	The total amount of time spent waiting for system feedback
tskip	The total amount of time before receiving a system response (CF or save notification) and pressing the skip button
tprep1	The total amount of time preparing for a first utterance
treform	The total amount of time preparing for second, third, fourth or later utterances
treca1	The total amount of time recording a first utterance
trecareform	The total amount of time recording a second, third, fourth or later utterances
tpq	The mean amount of time spent per question (time on question pages / questions)
tp1Try	The mean amount of time spent preparing for a first attempt at a question
tp234Try	The mean amount of time spent preparing for a second or later attempt at a question
tRec1TryPQ	The mean amount of time recording a first attempt at a question
tRec234TryPQ	The mean amount of time recording a second or later attempt at a question
tAttempting	The mean amount of time attempting a question
tReformPQ	The mean amount of time spent reformulating (time preparing + time recording) per question
t1Try	The mean amount of time spent on a first attempt (time preparing + time recording) at a question
tWaitPQ	The mean amount of time spent waiting for a system response per question