Tutorbot Corpus: Evidence of Human-Agent Verbal Alignment in Second Language Learner Dialogues

Arabella Sinclair
University of Edinburgh
10 Crichton Street
Edinburgh, Scotland
s0934062@sms.ed.ac.uk

Kate McCurdy
University of Edinburgh
10 Crichton Street
Edinburgh, Scotland
s1841537@sms.ed.ac.uk

Christopher G. Lucas
University of Edinburgh
10 Crichton Street
Edinburgh, Scotland
clucas2@inf.ed.ac.uk

Adam Lopez
University of Edinburgh
10 Crichton Street
Edinburgh, Scotland
alopez@inf.ed.ac.uk

Dragan Gašević
Monash University
Melbourne
Australia
dragan.gasevic@monash.edu

ABSTRACT

Prior research has shown that, under certain conditions, Human-Agent (H-A) alignment exists to a stronger degree than that found in Human-Human (H-H) communication. In an H-H Second Language (L2) setting, evidence of alignment has been linked to learning and teaching strategy. We present a novel analysis of H-A and H-H L2 learner dialogues using automated metrics of alignment. Our contributions are twofold: firstly we replicated the reported H-A alignment within an educational context, finding L2 students align to an automated tutor. Secondly, we performed an exploratory comparison of the alignment present in comparable H-A and H-H L2 learner corpora using Bayesian Gaussian Mixture Models (GMMs), finding preliminary evidence that students in H-A L2 dialogues showed greater variability in engagement.

Keywords

Language learning, chatbot, dialogue, alignment, tutoring, agent, second language, student engagement, assessment

1. INTRODUCTION

This work reports on evidence of alignment within student dialogue to that of an automatic tutor even when both parties are restricted in their capacity to align: the student as an L2 learner may lack the linguistic proficiency to show alignment [5], and the agent aligns only minimally by design. Alignment consists of interlocutor interaction adaptation, resulting in convergence, or in their sharing of the same concept space [13, 8]. Alignment of student to tutor in dialogue has been used as a predictor of both student learning and engagement [20]. A key aspect of dialogue is the speakers' ability to align: to either show engaged, willing behaviour, or display little discernible adaption to their interlocutor. Interestingly, humans have been shown to exhibit greater alignment to agents than to other humans [4, 6]. In an automated L2 tutoring setting, where students have been shown to imitate tutors as part of their learning process [10] it is of great interest to determine whether the user/learner is actively engaged, simply gaming the system, or disengaged, either because of lack of ability or motivation [1]. Modelling alignment of student to tutor as evidence of engagement could serve as a useful tool in the design of tutor intervention or student assessment since there has been limited research into identifying signs of engagement or gaming in the automated L2 tutoring setting.

Given this relevance of alignment in modelling engagement during tutor-student L2 dialogues [20], one key question is whether L2 students demonstrate alignment behavior in conversation with an automated dialogue agent, even when they know the agent is not human. Prior work has established that L2 students display alignment when conversing with a human tutor, in Human-Human (H-H) interactions [17]; however, this work has also demonstrated relatively *symmetric* alignment, as human tutors verbally aligned with their students in turn — this raises the possibility that L2 learners may fail to display alignment if the dialogue is predominantly *asymmetric* , when interacting with a agent whose capacity to align is also limited. Studies of Human-Agent (H-A) dialogues in other domains demonstrate that fluent speakers verbally align with agents [4, 6], but given the unique constraints affecting alignment in L2 dialogue [5], we cannot assume that L2 students will behave similarly. If they do, a second key question arises: do L2 students display similar alignment behavior in H-H and H-A dialogues? Even if students align in both contexts, exploratory analysis may reveal critical differences which could inform educational researchers and practitioners working with dialogue agents. Hence, our work addresses the following research questions: **RQ1** *Do L2 students show alignment to an automated dialogue agent (i.e. H-A alignment)?* and **RQ2** *What is the nature of the alignment found in the H-A corpus and how does it differ from that of H-H dialogues?*

Arabella Sinclair, Kate McCurdy, Adam Lopez, Christopher G. Lucas and Dragan Gasevic "Tutorbot Corpus: Evidence of Human-Agent Verbal Alignment in Second Language Learner Dialogues" In: *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, Collin F. Lynch, Agathe Merceron, Michel Desmarais, & Roger Nkambou (eds.) 2019, pp. 414 - 419

We present a study of student verbal alignment within a new dialogue corpus consisting of transcripts from a language teaching app where students are interacting with a dialogue agent. We contrast this H-A corpus with a comparable H-H L2 learner corpus of tutoring dialogue transcripts. We found that students in H-A interactions align to the agent more so than they would by chance, albeit to a lesser degree than students in H-H dialogues. Our results found that within H-H dialogues, students exhibited greater alignment than tutors. Finally, we compared the distribution of student to tutor alignment within both corpora, revealing more variance in alignment within the H-A dialogues. We hypothesise this was due to either student engagement effects, or different types of student alignment strategy within the H-A dialogues than the more uniform alignment present in the H-H corpus.

2. BACKGROUND

To achieve effective communication within dialogue, speakers typically align, adapting their interaction to their interlocutor. The Interactive Alignment Model (IAM) [13], describes this process as that of speakers agreeing on a shared conceptual space. In educational settings, by contrast, alignment has been found to predict both student learning and engagement [20].Automatic alignment between interlocutors occurs over different linguistic levels, including that of the lexical, syntactic and semantic [13]. *Lexical* alignment consists of speakers beginning to use the same words [21, 17] or phrases [6] as each other. *Syntactic* alignment consists of the use of the same parts of speech patterns, such as similar noun-phrase constructions, or similar adjuncts [14] as the conversation progresses. Finally, *semantic* alignment can range from adaptation to individual differences in personality [11] to convergence at a higher level of representation such as Dialogue Acts [16]Recent research has established a number of metrics for linguistic alignment which can be computed automatically, enabling large-scale corpus analysis based on sequential pattern mining [6]. These methods quantify alignment in terms of the *expressions*, or contiguous sequences of tokens appearing in the utterances of both interlocutors. While these methods have been applied to the analysis of H-A interaction [6] and H-H student-tutor interaction [17], the work presented in this paper is the first to apply this computational methodology to compare H-A and H-H dialogue in an educational L2 setting.

Within an L2 practice setting, we predict alignment to have slightly different properties compared to a fluent conversational setting where speakers tend to have a symmetric contribution and equal status within the dialogue [18], and are equally capable of participating [5]. L2 learners have been found to perform at a higher level when speaking in dialogue with a peer than in a monologue context [15] This suggests students draw from the example language of their interlocutor leading us to expect evidence of alignment. L2 students have also been shown to learn vocabulary through taking part in dialogue [9], suggesting this process of alignment and repetition of their interlocutor's speech produces learning gains. In the case of the tutor, their need to adhere to the ZPD suggests that their alignment patterns will also differ from that of straightforward dialogue. These different factors influence the speakers' convergence to a shared mental state [5]. Vygotsky's theory of ZPD [19] states students will learn best when addressed at the correct level, therefore we also expect to see alignment, in the case of tutors in H-H dialogues, to student ability.

Table 1: Tutorbot dialogue example. Italics indicate Expression Repetition

1.	bot:	What is your *favorite day of the week* ?
2.	user:	My *favorite day of the week* is Friday ...
3.	bot:	Do you play sports ?
4.	user:	yes
5.	bot:	What sport do you *play* ?
6.	user:	I *play* volleyball and I go running
7.	bot:	When do you do that ?
8.	user:	On Monday , Wednesday and Friday
9.	bot:	What time does it start ?
10.	user:	At 4 o'clock in the afternoon

3. CORPORA

We are interested in the comparison between student alignment in H-H and H-A dialogues. The H-A corpus analyzed in this study comprises dialogues drawn from a large-scale commercial platform for L2 learners[1]. In this application, novice learners of English who had completed lessons on relevant topics were offered the possibility to review the material via simple conversations with the automated dialogue agent Tutorbot. Given the focus on relevant learning material, the agent engaged learners in a system-initiative dialogue with extensive guidance, rather than user-initiative [2]; as a result, Tutorbot steered the learner conversations very deliberately, and alignment from the tutor agent to the student was highly limited by design. A sample dialogue from the corpus can be seen in Table 1. The H-H corpus used is the Barcelona English Language Corpus (BELC) [12] which consists of tutor guided conversations with L2 learners of English at varying stages of fluency from absolute beginner to approaching intermediate. The tutor's goal was to elicit as much conversation from the learner as possible while setting them at ease in as natural and conversational a manner as they could. Key differences are shown in Table 2. However, it should also be noted that the Tutorbot corpus only consists of single utterance turns, whereas BELC has multiple. The topics are also more diverse in BELC, as the Tutorbot explicitly guided learners to review practiced material rather than engage in open-ended discussion. Nevertheless, certain main topics (*how are you, where are you from, tell me about your family, hobbies, what time do you do that*) and the beginner/lower-intermediate range of learner ability are common to both, facilitating automated alignment comparison.

4. METHODS

4.1 Alignment

In order to analyse the verbal alignment present in both corpora, which allows us to answer both *RQ1* and *RQ2*, we use the expressions-based measures introduced by [6]. This approach identifies sequences of tokens (*Expressions*) which are used by both dialogue participants (thus *established* as expressions). These expressions allow us to see the fixed expressions established between speakers, called the routiniza-

[1]This data was kindly shared with us by Babbel, https://www.babbel.com/

Table 2: H-A and H-H Corpora Differences

	Tutorbot	BELC
number of dialogues	3689	118
average Num. utterances	20.41	130.69
average Num. tokens	128.99	634.28
average tokens/utterance	6.32	4.85
communication medium	typed	spoken
speakers	H-A	H-H
student L1	German	Spanish
vocabulary overlap	0.085	0.251

tion process in the interactive alignment theory [13], and thus an indication of speaker alignment. We re-define the following in order to discuss our results in the following sections:

Expression Lexicon EL is the set of expressions used by both speakers for a given dialogue.

Expression Variety (EV) is the size of the EL normalised by the total number of tokens in the dialogue. This ratio indicates the variety of the expression lexicon relatively to the length of the dialogue: the higher the EV, the more incidence of established expressions between participants. The EV indicates the routinization between speakers.
$EV = \frac{length(EL)}{numberoftokens}$

Expression Repetition – speaker (ER_S) is the ratio of Expressions to dialogue produced. This is measured in tokens. This value indicates the Expression repetition present in the dialogue, i.e. the higher the ER, the more the speakers dedicate tokens to the repetition of established expressions. This is indicative of speaker alignment.

Initiated Expression (IE_S) are the established expressions initiated by S

Vocabulary Overlap (VO) is the ratio of shared tokens between interlocutors S_1 and S_2. The higher the VO, the more vocabulary is shared between speakers.
$VO = \frac{(Tokens_{S_1} \cap Tokens_{S_2})}{(Tokens_{S_1} \cup Tokens_{S_2})}$

4.2 Baseline

In order to test that the alignment reported was not simply due to corpus-specific vocabulary effects (which would be influenced by the vocabulary overlap defined in the previous section), a 'scrambled baseline' was created for each corpus. This was achieved by creating a 'bag of words' of the tokens produced by each speaker for a specific dialogue, then substituting each token from each speakers utterances with one from the shuffled bag of words. This method retains the turn-taking of the speakers, and the distribution of utterance lengths from the original dialogue, but removes any word ordering present. In the results section for each alignment measure, we report on whether the effects were significantly different from this baseline. This baseline allows us to compare the effects of alignment across corpora, answering *RQ1*.

4.3 Alignment Distribution Clustering

In order to answer *RQ2* investigating student alignment differences within and between the H-H and H-A corpora, we fitted a Gaussian mixture model (GMM)[7] to the student ER_S data for both the H-H and H-A students. GMMs allowed us to detect and characterize distinct sub-populations within a larger group, provided those sub-populations were marked by differences in a parameter of interest, e.g., measured ER_S. To find the number of components which best fitted the data, we used a Bayesian Gaussian mixture model with a Wishart prior of [[0.1]] on the precisions and a scale-1 exponential prior on the number of clusters, and selected the most probable number of clusters given the data (i.e. the posterior mode), assuming that up to seven clusters might be present. We used a Bayesian approach in order to avoid the degeneracies that are common when using maximum-likelihood estimation and information criteria (e.g., AIC or BIC) to estimate cluster counts and parameters [3]. To implement this, we used the toolkit scikit-learn[2], package *BayesianGaussianMixture*; the priors on component means were scikit-learn 0.20 defaults.

5. RESULTS AND ANALYSIS

The following subsections all contribute to answering *RQ1*, through the comparison of H-H to H-A student alignment and corpus statistics. Section 5.5 specifically explores the variation in alignment styles across corpora, allowing us to answer *RQ2*.

5.1 Expression Lexicon

The Expression lexicon is the set of expressions which are shared between speakers. On inspection, the most common multi-word expressions being aligned to in the Tutorbot corpus fell into two main categories: 1) the student using the direct re-form of the question in the creation of their answer: "*bot|4: What **is** your **favorite day of the week** ? user|5: My [favorite day of the week] [is] Friday*". 2) The student reflecting the question back to the tutor-bot. "*bot|4: Where **do you live**? user|5: I live in <LOCATION>, where [do you live]?*". The rephrasing in BELC is different: it is more likely that the tutor will re-phrase the student's single or multi word answer as a form of confirmatory feedback. e.g. *"Tutor: you like going out with your friends, good"* when this is really more repetition/confirmation. The student alignment also consisted of their reflection of tutor questions back to them, and in their repetition of tutor scaffolding moves (something not present in the Tutorbot corpus due to the agent dialogue design) Table 3 contains details of the vocabulary overlap, speaker specific token ratios and the expression lexicon size differences between corpora.

Table 3: Corpora Differences- values represent the average per dialouge

	Tutorbot	BELC
Expression Lexicon Size (ELS)	3.04	48.55
S1/tokens (%)	0.81	0.68
S2/tokens (%)	0.19	0.30
Voc. Overlap	0.085	0.251
Voc. Overlap S1	0.105	0.312
Voc. Overlap S2	0.258	0.613

[2]https://scikit-learn.org/stable/

5.2 Vocabulary Overlap

The vocabulary overlap (VO) between speakers gives us an idea about how likely ''alignment' according to our metric will occur by chance. The results in Table 3 therefore can inform our interpretation of the levels of ER_S reported in section 5.4. Student VO in BELC (HH) is much higher than from the students in Tutorbot (HA) (*0.613 vs. 0.258*) This could be due to the fact that Tutorbot learners were at a lower level of proficiency, so they did not use such extensive vocabulary; alternatively, it could be due to the method of data collection: Tutorbot allows learners a one turn response (a single utterance), limiting their production.

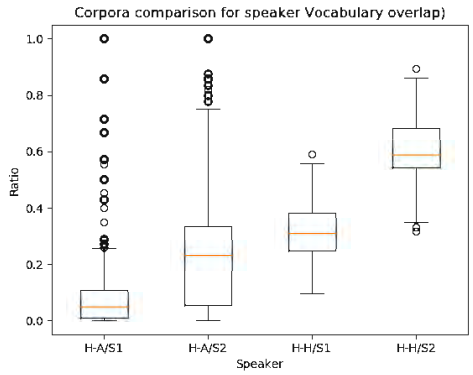


Figure 1: H-H/A corpora Vocabulary Overlap. Speaker difference was significant for H-A ($p < 0.0001$ ($statistic = 6.42, pvalue = num1.4e-10$) **and H-H ($p < 0.001$)** ($statistic = -2.11, pvalue = 0.00036$) **S1 = Tutor/Agent, S2 = Student**

5.3 Expression Variation

We compare the H-H and H-A corpora of real interactions to each other, and to the baseline H-H_R and H-A_R corpora to control for vocabulary effects. Firstly, EV was significantly higher for the H-H corpus ($mean = 0.075, std = 0.025$) than that in the H-A corpus ($mean = 0.032, std = 0.046$). Statistical difference was checked by performing a t-test ($statistic = -10.05, p - value = 1.888 \times 10^{-23}$), indicating H-H interactions result in a richer expression lexicon than H-A interactions. The EV values were much lower than those reported for negotiation dialogues [6], which may be due to dialogue type: routinisation may form a much greater part of negotiation than it does L2 tutoring. Another reason for the low EV in the H-A corpus is that the student cannot establish expressions other than by chance since the Tutorbot corpus is system-initiated and is not designed to align to the student's responses. Neither the EV of the H-H nor the H-A corpus was statistically greater than the H-H_R and H-A_R baselines, which can be in part attributed to the high proportion of single-token expressions in both corpora, leading to greater likelihood of their existence in the scrambled baseline.

5.4 Expression Repetition

Expression repetition (ER_S) is the main indication of speaker alignment measured. Figure 2 shows the different degrees

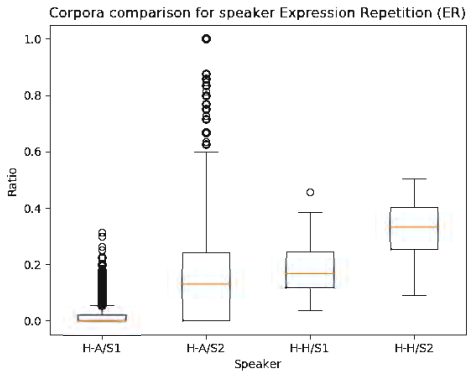
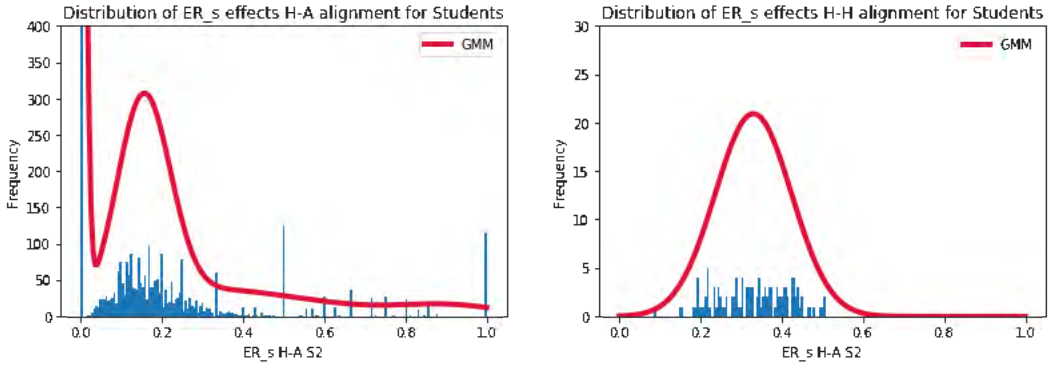


Figure 2: H-H/A corpora ER_Ss. Speaker difference is significant for H-A ($p < 0.0001$) ($statistic = -44.91, pvalue = 0.0$) **and H-H ($p < 0.0001$)** ($statistic = -12.71, pvalue = 1.77 \times 10^{-28}$) **S1 = Tutor/Agent, S2 = Student**

of ERs for both the H-A and H-H corpora. The difference between the ERs of each speaker was significant for both corpora: H-A ($statistic = -44.91, p - value = 0.0$) and H-H ($statistic = -12.71, p - value = 1.770 \times 10^{-28}$). It is interesting to note the asymmetry between speakers for both dialogues. The tutor in the H-H dialogues had a significantly lower proportion of ER than the student, suggesting ER has less to do do with teacher strategy as with learner strategy. We compared each ER_S with its ER_R for both corpora: for the H-A corpus, student ER_{S2} ($mean = 0.192, std = 0.235$) was significantly higher than that of ER_{R_S2} ($mean = 0.134, std = 0.206$) ($statistic = -11.20, p - value = 6.593 \times 10^{-29}$). Meanwhile, tutor ER_{S1} ($mean = 0.016, std = 0.032$) was significantly *lower* than that of their scrambled baseline ER_{R_S1} ($mean = 0.024, std = 0.037$) ($statistic = 9.865, p - value = 8.2012 \times 10^{-23}$) indicating the absence of alignment expected from an agent not designed to do so. For the H-H corpus, student ER_{S2} was not significantly different from their baseline ER_{R_S2} ($statistic = 0.932, p - value = 0.352$), nor was tutor ER_{S1} ($statistic = 2.506, p - value = 0.013$). This can be explained in part by the fact that VO for the H-H corpus ($mean = 0.251, std = 0.061$) was significantly larger than in the H-A corpus ($mean = 0.085, std = 0.146$) ($statistic = -12.32, p - value = 3.089 \times 10^{-34}$).

5.5 Student ER Distribution

In answer to *RQ2*, we compare the distributions of per-dialogue ER_S values between H-A and H-H corpora. Figure 3 shows histograms of ER frequency for each corpus, which suggest there were multiple types of student alignment in the H-A corpus (a), in contrast to a single cluster of ER values for the H-H corpus (b). To quantify these differences in student alignment – and go beyond a comparison of averages which neglects the possibility of differences across individuals and dialogues – we fit a Bayesian Gaussian Mixture Models [7] (described in Section 4.3) to student ER_S values. The results of our model indicate that the most probable number of clusters, given the data (i.e., the posterior mode), was 5 for the H-A corpus (Figure 3a) and 1 for the H-H corpus (Figure 3b). This analysis also reveals a



(a) GMM with number of components = 5
means: (0.0004, 0.15, 0.31, 0.52, 0.90)
weights: (0.30, 0.49, 0.08, 0.09, 0.05)

(b) GMM with number of components = 1
mean: 0.33
weight: 1

Figure 3: Frequency of Expression Repetition values (High ER indicates greater alignment). Gaussian Mixture Models (GMM) which best fitted to the data shown by the red line. Means: centroids of the component clusters. Weight: the proportion of dialogues in a cluster.

Table 4: Qualitative Analysis of H-A dialogues at the 'centroids' of the component clusters

ER	Description and example
0-0.01	**no-response** or **request for help**: students either do not engage with the agent, or demonstrate inability to engage
0.1-0.15	**minimal response**: students respond curtly, appear less engaged *bot: What is your favorite day of the week ?* *user: That 's Sunday .*
0.25-0.4	**high engagement**: dialogues either longer with align and rephrase within longer utterances, without excess repetition, or shorter dialogues consist of more repetition and rephrasing, and the limited vocabulary contributes to alignment *bot: Do [you] _have a_ boyfriend or _a_ girlfriend ? Or _a husband_ or _a_ wife ?* *user: I [have [a] husband] .*
0.5-0.55	**minimal response**: low rate of student production, typical response one high-frequency word, low engagement despite high alignment *'hi', 'bye.'*
0.85-0.9	**high repetition**: all student responses are rephrases, dialogues very short *bot: _Hello_ , _nice to see you_ !* *user: [Hello] [nice to see you] too*

cluster in the H-A corpus which has a qualitatively comparable mean value to the one in H-H (0.310-H-A, 0.330-H-H). Table 4,shows this cluster contains the longest dialogues in Tutorbot, which are qualitatively the most similar to those in BELC.

We hypothesise the other clusters are either, in the case of low level ER, signs of student lack of engagement (alignment being symptomatic of engagement within dialogue) or, in the case of higher ER, signs that the students are in some way conversing in a manner impossible to find in H-H dialogues. We hypothesise either this is due to the communication medium: students can copy, paste and edit the agent utterance to create their response or due to students' desire to learn through continual repetition of the agent's phrases.

Table 4 shows examples and descriptions of the H-A corpus data, corresponding to the component means in Figure 3. Since the H-H corpus was gathered as part of an experiment, we know that there would not be 'outlier' behaviour present, but the upper and lower ranges show some differences in interaction style of the learner.

6. DISCUSSION

In relation to *RQ1*, whether there is evidence of student - agent alignment in L2 dialogues, we find significant H-A alignment. The magnitude of this effect was weaker than that found in H-H dialogues, and we hypothesise that adaptive student support in the form of tutor alignment is essential for students to align to the degree they do in an L2 H-H setting. We found no significant alignment of agent to student, however an agent designed to interact with more explicit alignment may more resemble the alignment found in the H-H corpus. We found asymmetrical alignment within the H-H corpus, which was in keeping with results reported on lexical priming for the same corpus which found the strongest priming effects are those from student to tutor [17]. In relation to *RQ2*, concerning the exploratory analysis of alignment differences across corpora, a particularly salient finding are the differences in alignment across dialogues, suggesting different patterns of student engagement could be detected via their alignment levels. Table 4 shows that there was a clear 'normal range' for interaction, and the outliers showed different signs of student non-engagement. Our key finding is that there was greater variability in H-A compared to H-H alignment (best fit of 5 clusters compared to a single cluster), although role of factors such as dialogue and utterance length in these findings should be investigated in future work. We hypothesise that building a more alignment-focused tutoring agent could increase student engagement and yield results consistent to those within BELC. This could lead to better online L2 tutoring systems which promote student engagement and therefore improve participation and learning. It may be that the nature of an online learning platform will always result in some students who do not fully engage, and need different interven-

tion strategies. Using an alignment metric in the manner of our study could allow for the identification of these students, measurement of their engagement, and prediction of personalised interventions.

7. CONCLUSIONS AND FUTURE WORK

This paper presents a comparative analysis on student to tutor alignment in both an H-A and an H-H dialogue setting. We found students aligned to the agent, although this alignment was not stronger than that present in H-H dialogues which is the case for both negotiation [6] and task-based dialogues [4]. We hypothesise we can better explore this in a setting where the agent is specifically designed to align to the student.A limitation of our study is that both corpora were collected independently and therefore differ in more aspects than the one we wish to explore. In future work it would be desirable to collect data in a controlled setting which is more similar to the Tutorbot corpus to facilitate a more in-depth comparison. Another avenue for future research is the design of adaptive 'alignment' moves for the automated tutor to make. The design could draw on how the ZPD influences alignment and what the common ERs are in the H-H corpus, such as confirmatory rephrasing (e.g. *"Student: I speak Germanish", "Tutor: you speak **German**? Great!"*) or repetition (e.g. *"student: I am 20 years old", "tutor: **20 years old?** good!"*). This research has a number of implications for the educational community, particularly regarding the use of alignment as an indicator of engagement. Furthermore, our method of clustering student ERs to identify 'normal' engagement behaviour for a given domain may inform the detection of outliers and has potential for automating dialogue planning and intervention policies.

8. ACKNOWLEDGMENTS

We are grateful for the helpful discussions had with Nicolas Collignon, Edmund Fincham and Pablo Leon and comments from our anonymous reviewers. We thank the team at Babbel and specifically Zach Sporn and Joel Kiesey for making this collaboration possible.

9. REFERENCES

[1] R. Baker, J. Walonoski, N. Heffernan, I. Roll, A. Corbett, and K. Koedinger. Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research*, 19(2):185–224, 2008.

[2] S. Bibauw, T. FranÃğois, and P. Desmet. Discussing with a computer to practice a foreign language: research synthesis and conceptual framework of dialogue-based call. *Computer Assisted Language Learning*, 0(0):1–51, 2019.

[3] C. Biernacki and S. ChrÃl'tien. Degeneracy in the maximum likelihood estimation of univariate gaussian mixtures with em. *Statistics Probability Letters*, 61:373–382, 02 2003.

[4] H. P. Branigan, M. J. Pickering, J. Pearson, and J. F. McLean. Linguistic alignment between people and computers. *Journal of Pragmatics*, 42(9):2355–2368, 2010.

[5] A. Costa, M. J. Pickering, and A. Sorace. Alignment in second language dialogue. *Language and cognitive processes*, 23(4):528–556, 2008.

[6] G. D. Duplessis, C. Clavel, and F. Landragin. Automatic measures to characterise verbal alignment in human-agent interaction. In *18th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 71–81, 2017.

[7] S. Farrell and S. Lewandowsky. *Computational Modeling of Cognition and Behavior*. 02 2018.

[8] S. Garrod and M. J. Pickering. Alignment in dialogue. *The Oxford handbook of psycholinguistics*, pages 443–451, 2007.

[9] R. Hawkes. *Learning to Talk and Talking to Learn: How Spontaneous Teacher-learner Interaction in the Secondary Foreign Languages Classroom Provides Greater Opportunities for L2 Learning*. PhD thesis, University of Cambridge, 2012.

[10] F. M. Holley and J. K. King. Imitation and correction in foreign language learning. *The Modern Language Journal*, 55(8):494–498, 1971.

[11] A. Isard, C. Brockmann, and J. Oberlander. Individuality and alignment in generated dialogues. In *Proceedings of the Fourth International Natural Language Generation Conference*, INLG '06, pages 25–32, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[12] C. Muñoz. *Age and the rate of foreign language learning*, volume 19. Multilingual Matters, 2006.

[13] M. J. Pickering and S. Garrod. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190, 2004.

[14] D. Reitter and J. D. Moore. Alignment and task success in spoken dialogue. *Journal of Memory and Language*, 76:29–46, 2014.

[15] P. Robinson and R. Gilabert. Task complexity, the cognition hypothesis and second language learning and performance. *IRAL-International Review of Applied Linguistics in Language Teaching*, 45(3):161–176, 2007.

[16] A. Sinclair, R. Ferreira, A. Lopez, C. Lucas, and D. Gasevic. I wanna talk like you: Speaker adaptation to dialogue style in l2 practice conversation. In *Proceedings of Artificial Intelligence in Education - 20th International Conference*, 2019.

[17] A. Sinclair, A. Lopez, C. Lucas, and D. Gasevic. Does ability affect alignment in second language tutorial dialogue? In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 41–50, 2018.

[18] A. Sinclair, J. Oberlander, and D. Gasevic. Finding the zone of proximal development: Student-tutor second language dialogue interactions. In *Proc. SEMDIAL 2017 (SaarDial) Workshop on the Semantics and Pragmatics of Dialogue*, pages 107–115, 2017.

[19] L. Vygotsky. Zone of proximal development. *Mind in society: The development of higher psychological processes*, 5291:157, 1987.

[20] A. Ward and D. Litman. Dialog convergence and learning. *Frontiers in Artificial Intelligence and Applications*, 158:262, 2007.

[21] A. Ward and D. Litman. Measuring convergence and priming in tutorial dialog. *University of Pittsburgh*, 2007.