Finding the Zone of Proximal Development: Student-Tutor Second Language Dialogue Interactions

Arabella Sinclair
University of Edinburgh
10 Crichton St
EH8 9AB
aj.sinclair-2@sms.ed.ac.uk

Jon Oberlander
University of Edinburgh
10 Crichton St
EH8 9AB
jon@ed.ac.uk

Dragan Gasevic
University of Edinburgh
10 Crichton St
EH8 9AB
dragan.gasevic@ed.ac.uk

Abstract

The goal of dialogue practice for a second language learner is to facilitate their production of dialogue similar to that between native speakers. This paper explores the characteristics of student and tutor dialogue in terms of their differences from classic conversational and task-oriented corpora. Interlocutors have the tendency to align to the language of the other in conversational dialogue, creating a symmetry between speakers which learners of a language may be unable at first to achieve. Our hypothesis is that as a learner's competence increases, symmetry between learner and tutor language increases. We investigate this at both a surface and a deeper level, using automatic measures of linguistic complexity, and dialogue act analysis. The data supports our hypothesis.

1 Introduction

Alignment and entrainment are phenomena of dialogue present to varying degree depending on the nature of the interaction. For second language learners,[1] aligning with their interlocutor allows them to bootstrap their knowledge from the more competent linguistic example being given to them (Robinson, 2011). Their constrained fluency, however, limits their ability to achieve this in all areas. This leads us to predict differences in alignment and symmetry between learner and native dialogue, whether conversational or task based, due to this difference in speaker status.

Our goal was to understand the patterns and dynamics of student and tutor interaction and the level of synchronisation between the two actors in these dialogues. Likewise we want to compare L2 with native dialogues, in both conversational and task-based styles. To that end, we analyse and compare transcribed dialogues between L2 learners and tutors (an excerpt of which is shown in Figure 1), to key characteristics observed in dialogues between native English speakers. We posit that both task-oriented and conversational dialogue corpora are relevant for comparison because on the one hand L2 learner dialogue can be viewed as both a learning or a teaching task, and on the other, the student is trying to participate in and gain conversational skill, while the tutor encourages it. Our assumption is that tutors monitor students' convergence and use this to identify when the student is capable of learning more. This task of pushing the student, yet reassuring them, to promote their production, involves a tutor's constant adaption to remain within the Zone of Proximal Development.[2]

Example Dialogue
INV: what time did you arrive today in the morning?
PAR: when arrive in the.
INV: yes when did you arrive today?
PAR: hmm seven-eight+half half+past+eight.
INV: uhhuh good.
INV: and what time will you finish?
PAR: hmm three.
INV: at three uhhuh.

Figure 1: Example of Learner-Tutor dialogue from the BELC corpus, where INV stands for interviewer and PAR participant.

[1] Here we use second language (L2) in the broad sense, to include any language additional to the speaker's native language.

[2] The Zone of Proximal Development (ZPD) is "the distance between actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance, or in collaboration with more capable peers" (Vygotsky, 1978, p. 86). In other words, the ZPD is a space between the learner's current level of development, and their potential development when supported by an interlocutor.

The overarching goal of our work is to obtain a better understanding of the patterns of L2 learner dialogues at different levels of expertise in order to inform work in the field of Computer Assisted Language Learning (CALL), specifically dialogue agents for L2 tutoring. This differs from existing work in this domain (Ferreira et al., 2007) as it focusses on one to one tutoring dialogues, and uses automatic measures of complexity in addition to dialogue act analysis. Dialogue agents for tutoring science and engineering subjects, as in Auto Tutor (Graesser et al., 2005) or BEETLE (Dzikovska et al., 2014) have achieved some successes, however dialogue agents for one-to-one L2 conversational learning are less well explored. L2 agents' goals are to practice conversational English as well as to both implicitly and explicitly correct the learner in order to scaffold[3] new vocabulary or grammatical constructs. Examples of dialogue agents for one on one L2 learning are CLIVE (Zakos and Capper, 2008), an agent which allows learners to practice basic conversation and fall back on their native language for clarification and more teaching oriented work, which varies the explicitness of corrective feedback (Wilske and Wolska, 2011). Immersive games-based dialogue tutoring has been proven an effective environment for language learning (Johnson and Valente, 2009) and dialogue agents for facilitating collaborative learner dialogue in the context of online courses also exist (Kumar et al., 2007). None of these expressly focus on adapting the complexity of an agent's language to the learner.

Objectives

This paper is an initial study to compare aspects of L2 learner dialogue across levels, and between native dialogue corpora, both conversational and task based. Our objectives comprise comparing these three dialogue types over the following dimensions:

O1 Linguistic Complexity
a) Per speaker
b) Over the course of a dialogue
c) Across levels
d) Between dialogue corpora type
(learner/conversational/task-based)

O2 Dialogue Act (DA) distribution
a) Speaker's own DAs per level
b) DA share per dialogue (speaker labelled)
c) Cross corpora, regardless of speaker
d) DA bigrams to inspect turn taking
(such as speaker-statement/question turn bigrams)

O3 Complexity of specific Dialogue Acts characteristic of L2 learning
a) Statements
b) Questions

We want to compare multi-level L2 dialogue with that of native speakers, covering different dialogue types. Section 2 describes the choice of corpora to achieve these objectives. The measures with which we will compare these aspects are addressed in section 3. These draw from the fields of Readability Analysis, Automatic Assessment of text, Second Language Acquisition research; and from the Dialogue analysis literature. We present the results of these comparisons in Section 4. Sections 5 and 6 discuss the implications of these findings and propose future work which will build on these conclusions.

2 Corpora

Corpus	Type	English	Size
Map Task (MT)	task based	native/fluent	128
Switchboard (SB)	conversational	native/fluent	1155
BELC	learner practice	non-native (level 1-4)	118

Table 1: Corpora types and details

The L2 dialogues used consist of a section of The Barcelona English Language Corpus (BELC) (Muñoz, 2006), containing transcripts from 118 semi-guided interviews conducted over the course of 4 sessions; over a long period of time, with the same participants each session. The participants had received each on average about 200 hours of English instruction before the start of the study and between each session. The interviewer's role was that of an encouraging tutor where "Interviewers attempted to elicit as many responses as possible from the learners, and accepted learner-initiated topics in order to create as natural and interactive a situation as possible". The interviews were *semi-guided* in that the interviewer "began

[3]Scaffolding refers to one of the roles of an L2 tutor: providing contextual supports for meaning through the use of simplified language. First introduced by Wood et al. (1976).

with a series of questions about the subjects family, daily life and hobbies. This constituted a warming-up phase that helped students feel more at ease.".

Transcripts of one-to-one L2 learner-tutor dialogues do not exist in great quantity and BELC includes the kinds of scaffolding and backchannel acknowledgement aspects of L2 tutoring we want to model. Figure 1 contains a short example of this.

In order to contrast the task element of L2 dialogue with its conversational goal, we use the Map Task corpus (Anderson et al., 1991) and Switchboard corpus (Godfrey et al., 1992) (Table 1). The MapTask corpus consists of dialogues between two participants, the *Giver* and the *Follower*. They are tasked with describing or marking a route on a map that is marked on only the giver's map, the follower has to follow their partner's instructions and mark the same path on their own copy of the map. This task based dialogue was chosen for its leader and follower dynamic, which we contrast to L2 learner conversation where the learner is much less fluent than their interlocutor. The Switchboard corpus is a large corpus collected from telephone conversations between native speakers on one of a set of pre-defined conversational topics. The speakers did not necessarily know each other, had equal status, and the aim was to produce largely unconstrained conversation.

3 Comparison Methods

Existing methods for grading dialogue of students and tutors within science tutoring involve latent semantic analysis between student response and documents consisting of relevant syllabus (Graesser et al., 2000). The challenge in assessing L2 learner dialogue is that the language itself is the syllabus, and although students responding in a relevant manner is important, the main aspects are: a) the level of complexity of the language which they can produce; and b) the level of complexity of the language of their interlocutor to which they are capable of successfully responding. In the latter case, successfully responding means not just responding to a question with silence or signalling they do not understand.

3.1 Linguistic Complexity

Existing measures of text complexity developed to predict the readability of discourse have been applied to dialogue in the form of subtitles from television shows of varying age of audience (Vajjala and Meurers, 2014), successfully differentiating between subtitles aimed at young children, children of school age and adults in terms of the complexity of the language shown. We use the same feature set to train a simple Linear regression model as a way to 'grade' the transcribed dialogue text in order to compare the complexities of language used between the corpora.

The main feature types used by Vajjala and Meurers (2016) to measure readability are described below:

Lexical Lexically complex words are those for which a simpler synonym exists, diversity and density are measured by type-token and part-of-speech ratios

Morphological Morpho-syntactic properties of lemmas, estimated from the Celex (Baayen et al., 1993) database.

Psycholinguistic Concreteness, meaningfulness and Age of Acquisition measures (Kuperman et al., 2012)

Simple Counts Average sentence length, word lengths and occurrence frequencies, n-grams, "difficult" words from frequency lists, syllables per word and other weighted combinations such as (Farr et al., 1951)

To train our model, we use the graded hand-simplified collection of simple discursive articles provided in the Newsela corpus (Xu et al., 2015). We chose this corpus for two main reasons, firstly the corpus is written for learners (not by learners) at a known level of competence. Secondly, it has a wide and varied vocabulary, large size, and number of distinct level labels (grades 3-12) which will allow us to best deal with the sparse nature of dialogue text.

3.2 Dialogue Act Patterns

Dialogue Act (DA) modelling can tell us a lot more about the dynamics of a dialogue such as whether participation is equal, whether certain DAs are more prevalent in particular dialogues, and what the strategy of the individual speakers is. In order to gain this deeper look at the structure of the dialogue, utterances were automatically labelled with a subset of DA labels from Stolcke et al. (1998) selected for their relevance to the dialogues in question, and whether they were simple enough to be captured with a regular expression rule. The resulting utterances for each DA label were manually inspected and found to conform to the pattern specified by the regular expression rule. The regular expression tags were also compared to the gold standard labels of the Switchboard corpus, achieving an F1 score of 0.82 although these labels were not used. Table 2 contains a description of the DAs applied.

Tag	**Example**
YES-NO-QUESTION	*do you XX, are you XX*
DECLARATIVE YES-NO-QUESTION	*so XX ?*
BACKCHANNEL-QUESTION	*yes?/ oh yeah? / no? / really?*
WH-QUESTION	*ok and wh*... / wh* .. / uhhuh ok wh* ..*
GENERAL-OTHER-QUESTION	*Any other question*
YES ANSWERS	*yes .*
NO ANSWERS	*no / nope / uh no*
SIGNAL-NON-UNDERSTANDING	*hmm. / ah. / [-spa] no se/ silence*
BACKCHANNEL-ACKNOWLEDGE	*uhhuh*
RESPONSE ACKNOWLEDGEMENT	*ok. / good. / right ok*
REPEAT-PHRASE	*XX ok/ ah XX: when XX is in previous utterance*
STATEMENT	*Any other utterance*

Table 2: Dialogue Acts selected from the 42 labels used in (Stolcke et al., 2000) with their accompanying reg-ex recognition examples. Labels *general statement* or *general question* are bucket labels, for any utterance not falling into other categories.

In order to achieve the best quality of labels, the existing hand labelled DAs available in both Switchboard and MapTask were grouped into categories aligning to those we chose to use for our rule based labelling. The alignment is shown in Table 3 and these final tags are compared in the following sections.

4 Results

To address the aspects of linguistic complexity analysis (Objective *O1*), we separately analyse the first and second halves of the dialogue, divided by speaker. We then use our complexity model to assign the resulting text a 'grade' in order to compare the surface level linguistic complexity (Figure 2). We observe that for learners at L1, the tutor and student tend towards convergence of complexity, and at a higher level they diverge. Switchboard (SB) has a complexity a little above that of the most advanced of the BELC dialogues, and there is neither significant difference between half nor speaker. MapTask (MT) has a similar difference in complexity between speakers as the L1 & L2 of BELC, although both are more complex. There is no convergence of complexity between speakers, nor significant change over their dialogue. Additionally, a simple word-per-utterance count per speaker across levels and corpora shows

Rule based	**Map Task**	**Switchboard**
yes-no-question	query-yn	yes-no-Question
declarative yes-no-question	check	declarative yes no question
backchannel-question	–	backchannel question tag question
wh-question	–	wh-question
general-other-question	query-w *(other q)*	open question rhetorical question declarative wh question or-clause (or question)
yes answers	reply-y	yes answer
no answers	reply-n	no answer reject
signal-non-understanding	–	signal non understanding
backchannel-acknowledge	–	backchannel ack
response acknowledgement	acknowledge	response ack
repeat-phrase	–	repeat phrase
statement	instruction explanation clarify ready align reply-w	statement opinion agreement/accept appreciation conventional closing hedge other quotation affirmative non-yes A action directive collab. completion hold before A/agree **

**The remaining switchboard dialogue acts each make up 0.1% or less of the switchboard utterances and would also fall within the STATEMENT label when classified with our rules: *negative non no answers, other answers ,dis-preferred answers, 3rd party talk, offers, options and commits, self talk, downplayer, maybe/accept part, apology, thanking*

Table 3: Mapping of our rule based dialogue act labels to those used in the Switchboard and Map task corpora.

the symmetry of Switchboard, asymmetry of MapTask and a trend from asymmetry to symmetry as level increases for BELC in terms of speaker contribution.

Dialogue Act Tags	BELC	MT	SB
yes_answers:	5.2%	11.3%	1%
no_answers:	1.7%	4.8%	1%
backchannel_ack:	3.3%	↓	19%
response_ack:	2.3%	24.2%	1%
sig_non_understand:	8.0%	0%	.1%
repeat_phrase:	1.9%	–	.3%
yes_no_Q:	3.5%	6.5%	2%
declarative_yes_no_Q:	6.8%	5.2%	1%
backchannel_Q:	2.7%	↓	1.1%
wh_Q:	9.3%	↓	1%
general_other_Q:	25.0%	11.6%	.8%
statement:	36.4%	32.3%	68%

Table 4: Dialogue Act distribution across utterances with *SB* for Switchboard, *MT* for MapTask and *Q* for Question. The ↓ means that the act is grouped and this is the percentage for the previous act combined. There are on average a greater proportion of *statements* in SB, more *questions* and *sig_non_understand* in BELC, and comparatively more *yes* and *no_answers* in both BELC and MT than in SB.

Following Objective *O2*, we firstly look at the average distribution of DAs, regardless of speaker, in Table 4. This shows there is a significantly greater ratio of *statements* to *questions* in SB, and the inverse is found in BELC. Continuing this cross-corpora view, Figure 3 shows the distribution of DAs for the average dialogue split by speaker. This shows a general asymmetry of *statement* contribution in BELC and MT (between student and follower) and a very symmetrical share between speakers in SB. Comparing BELC levels, Figure 3 also shows that learners at a higher level make a more similar proportion of *statements* to their tutor than at mid level. The proportion of *gen_other_question* increases for students as it decreases for tutors. This becomes closer to the symmetrical contribution of native speakers in SB, as does a student's percentage of *yes_answers*, which increases with level.

The distribution of individual speakers' DAs is shown in Figure 4. This shows that a student's *questions, statements, response acknowledgement* and *yes_answers* increase, and their *signal_non_understanding*, and *no_answers* decrease with the student level. The tutor's *general_questions* decrease with student level, as their *statements* increase slightly along with *WH-questions*, and *signal_non_understanding*.

Bigram	BELC				MT	SB
Speaker	L1	L2	L3	L4		
TT/AA/GG	30.6	21.1	18.3	19.8	22.7	47.6
TS/AB/GF	34.3	39.3	39.3	39.4	35.2	2.5
ST/BA/FG	34.7	38.8	39.9	39.1	34.9	2.5
SS/BB/FF	0.3	0.8	2.5	1.7	7.2	47.3
Statement	L1	L2	L3	L4		
TT/AA/GG	1.73	1.18	2.44	2.08	11.62	40.61
TS/AB/GF	2.64	2.92	3.61	3.60	2.76	1.47
ST/BA/FG	6.04	6.71	7.34	6.98	2.22	1.45
SS/BB/FF	0.00	0.38	1.11	0.52	0.84	31.44
Question	L1	L2	L3	L4		
TT/AA/GG	**0.573**	**0.289**	**0.190**	**0.138**	0.017	0.129
TS/AB/GF	0.054	0.068	0.101	0.076	0.013	0.001
ST/BA/FG	0.027	0.031	0.044	0.038	0.012	0.001
SS/BB/FF	0.000	0.001	0.003	0.002	0.010	0.061

Table 5: Dialogue Act bigrams for speakers, statements and questions. T=Tutor, S=Student for BELC corpus, A=speakerA, B=speakerB for Switchboard corpus, F=Follower, G=Giver for MapTask corpus. e.g. TS/AB/GF = tutor-student/speakerA-speaker-B/giver-follower average bigram percentages.

Table 5 shows the average percentage of DA bigrams for the utterances in each dialogue. This shows a symmetrical contribution of SB speakers. The first bigram type, *Speaker*, can be interpreted as a higher incidence of single utterance speaker turns in all levels of BELC, compared to the opposite in native SB & MT where multi-utterance turns are most common, particularly for the instruction giver in MT.

Finally, to address *O3*, Figure 5 shows the average 'grade' of the text in only the *Statements* and *Questions* of each type of dialogue. In order to better understand the constant distance in level between the tutor and the student within the *question* 'grades', we examined the bigrams for *statements* and *questions* alone, which can be seen in the bottom two segments of Table 5. These show an increase in tutor *statement* bigrams at L3 & 4, and a steady decrease in tutor *question* bigrams approaching L4.

5 Analysis and Discussion

From the results discussed in Section 4, it is clear that tutors adapt their conversation strategy to the level of the learner in all dimensions we explored.

In terms of surface level complexity (*O1*), Figure 2 suggests that it is only when the tutor and student start the dialogue at a similar enough 'grade'

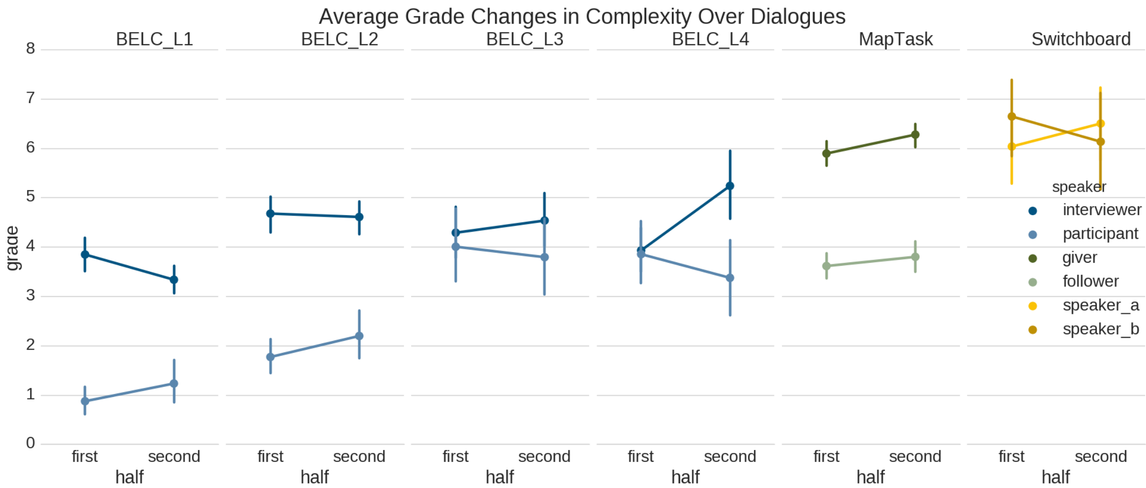

Figure 2: Average Student tutor complexities for first and second halves of dialogues by level. In the BELC results, the convergence and divergence of the *tutor's* complexity grade in relation to the student's in level (L)1 and 4 is significant ($t = 6.25$, $p = 1.60\text{e-}08$, $t = -4.18\text{e+}00$, $p = 2.95\text{e-}04$), as is the divergence of complexity between speakers in the second half of the dialogue in L4 ($t = 3.18$, $p = 2.47\text{e-}03$). There is no significant difference between any grade complexity in the Switchboard corpus, and although the speakers in the MapTask are at a significantly different grade level ($t = 6.52$, $p = 1.12\text{e-}10$), their dialogue has no significant increase in complexity.

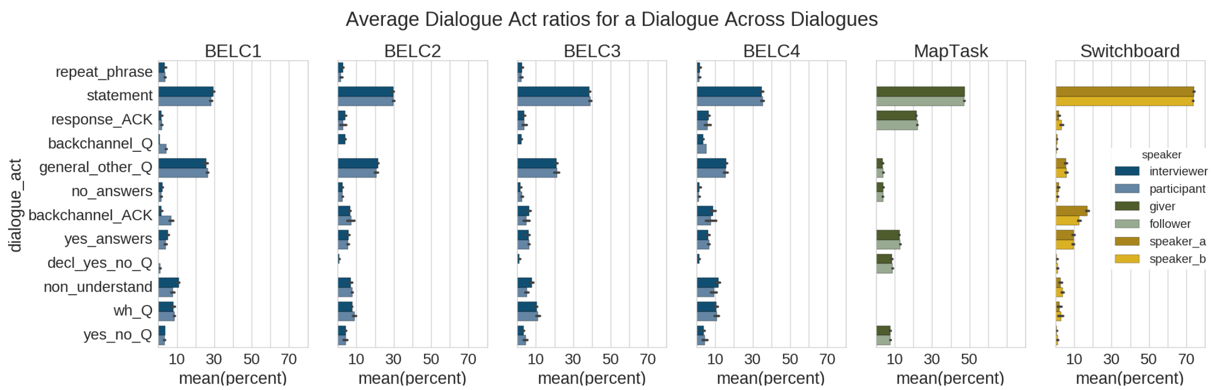

Figure 3: Dialogue Act percentages by corpus for the average dialogue.

that the tutor changes their strategy and increases the complexity of their input, to push the learner as their 'task' is tutoring not conversation. The difference in complexity of student and tutor in *L1* & *2* is similar to task based speakers in MT, in *L3* it becomes more symmetrical as in the native speakers in SB, and at *L4* the tutor changes their complexity to increase this distance once more. We interpret this as the tutor adhering to the zone of proximal development. Additionally, we interpret the change in L2 dialogue from an asymmetrical speaker complexity balance like MT, to a more symmetrical contribution like SB, as a phenomena of tutoring dialogue: to shift from a task-like structure to a conversational one as student competence increases.

Analysis of the DAs (*O2*) show the general increase in the students' share of the dialogue, not only in terms of *statements*, but also *questions*; the production of which takes greater cognitive task than simply responding to them. This increase in asking questions can be seen as the student's taking a more active role in the conversation, which demonstrates an additional dimension to their acquisition of skills. Not only do they proportionally contribute a greater share of the questions and statements to the dialogue at a higher level (Figure 4), but within their own share of the dialogue

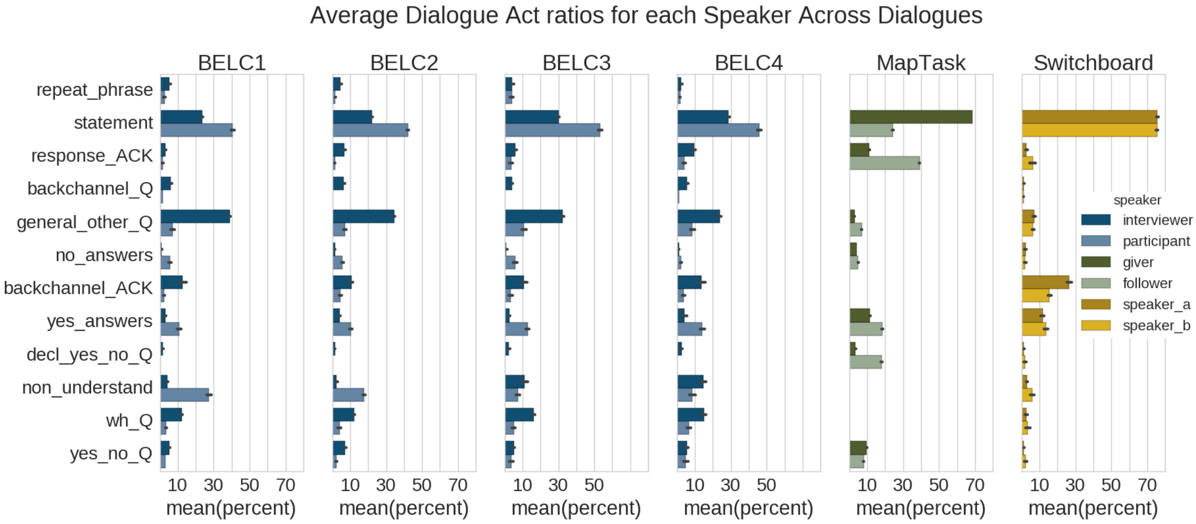

Figure 4: Average Dialogue Act percentages per dialogue by corpus: for an individual speaker's average share of the dialogue.

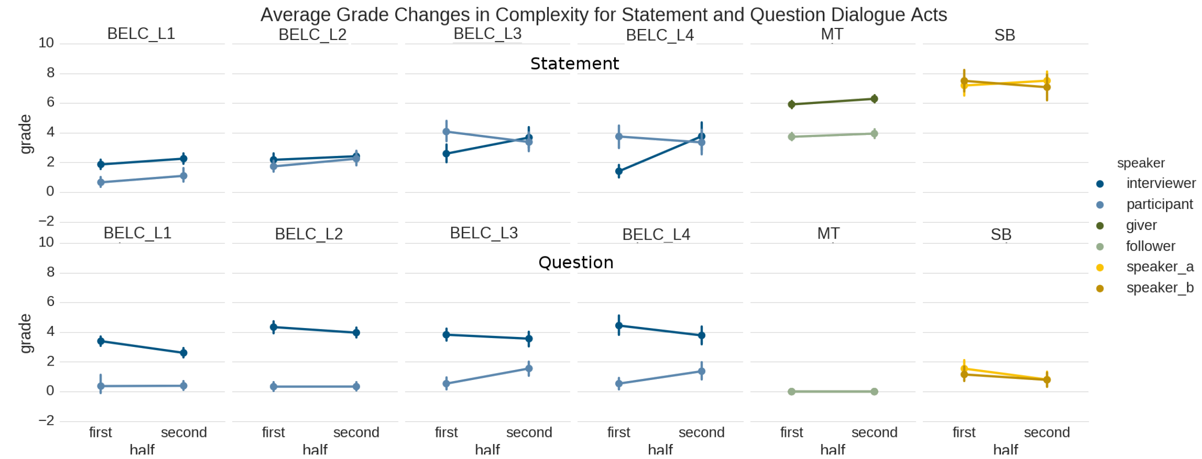

Figure 5: Complexities within Statement and Question dialogue acts in the three corpora.
For the Statements (upper row), the interviewer's statements between the first and second half of Levels 3 and 4 significantly (t = -2.28, p = 2.72e-02, t = -4.18, p = 2.95e-04) increase in grade complexity. In Levels 3 and 4, the convergence from different grades to a similar grade between speakers is significant (t = -3.08, p = 3.51e-03, t = -5.10, p = 2.58e-05). For the Questions (lower row), the difference between interviewer and participant grade is significant across levels: at Level 1, the interviewer's trend to converge is significant (t = 3.24, p = 1.82e-03), as is the student's at Levels 3 and 4. (t = -3.13, p = 3.01e-03, t = -2.26, p = 3.26e-02).

the proportion of their utterances signalling *non-understanding* (as defined in Table 2) decreases, with their participation in question and statement acts increasing (Figure 3).

The final objective, *O3*, of this work was to explore whether examining the complexity within certain dialogue acts can better inform us of the patterns of student tutor dialogues. Figure 5 allows us to see at a finer grained level what happens when the tutor changes strategy at Levels 3 & 4. We hypothesise first that although tutor *questions* tend to align to the complexity level of the students and vice versa in levels 3 & 4, they never converge; and secondly that the tutor adapts their *statements* to match the complexity of the student. We suggest that this is evidence of the tutors monitoring students' convergence, using this to identify when the student is capable of learning more. These shifts in our view, are signs of the tutor observing the Zone of Proximal Development.

On analysing DA bigrams in order to further investigate the patterns of statement and complexity changes, we note differences in terms of both turn taking and types of turn taking (Table 5). Our interpretation is that the single-utterance turn taking is a tutoring strategy (as evidenced in BELC), as this is the only aspect where there is no trend towards the symmetry of SB. Our interpretation is that tutor *question* bigrams are evidence of scaffolding, a key strategy of the Zone of Proximal Development. We see their decrease a sign that the tutor no longer needs to paraphrase themselves to be understood. This helps illuminate Figure 5, that although the questions asked may not be significantly more complex, it is likely that a lot fewer of them go unanswered at L4 than at L1.

6 Future Work

As this was an initial study, DAs for the BELC corpus were not annotated by hand, resultantly, our analysis of DAs has to be at a relatively coarse grained level. The algorithmic annotations were developed on the judgement of a single annotator; further work will recruit additional annotators and establish inter-annotator agreement. In future work, we aim to annotate the BELC dialogues with the full 42 DA tag set of the Switchboard corpus, in order to more thoroughly investigate whether there are level-specific sequences we can observe. It would be interesting to work with the tag *collaborative completion* so as to further examine the use of *scaffolding* in the tutor's dialogue. In the future, we also plan to expand our comparisons to include that of participant topic introduction and measures of semantic relevance of questions to answers. We also plan to compare the patterns in BELC's L2 dialogues with those of science tutoring dialogues and other spoken and text based language tutoring corpora, to better model tutoring strategy. Our observations will be applied to the task of predicting "good" tutor utterances given a dialogue history window and a student utterance. In other words, we will work towards developing a tutoring language model, constrained both by dialogue act and linguistic complexity.

Acknowledgments

Thanks to Vajjala and Meurers (2016) for sharing their feature set and Xu et al. (2015) for sharing their corpus.

References

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The hcrc map task corpus. *Language and speech*, 34(4):351–366.

R Harald Baayen, Richard Piepenbrock, and Rijn van H. 1993. The {CELEX} lexical data base on {CD-ROM}.

Myroslava Dzikovska, Natalie Steinhauser, Elaine Farrow, Johanna Moore, and Gwendolyn Campbell. 2014. Beetle ii: Deep natural language understanding and automatic feedback generation for intelligent tutoring in basic electricity and electronics. *International Journal of Artificial Intelligence in Education*, 24(3):284–332.

James N Farr, James J Jenkins, and Donald G Paterson. 1951. Simplification of flesch reading ease formula. *Journal of applied psychology*, 35(5):333.

Anita Ferreira, Johanna D Moore, and Chris Mellish. 2007. A study of feedback strategies in foreign language classrooms and tutorials with implications for intelligent computer-assisted language learning systems. *International Journal of Artificial Intelligence in Education*, 17(4):389–422.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE.

Arthur C Graesser, Peter Wiemer-Hastings, Katja Wiemer-Hastings, Derek Harter, Tutoring Research Group Tutoring Research Group, and Natalie Person. 2000. Using latent semantic analysis to evaluate the contributions of students in autotutor. *Interactive learning environments*, 8(2):129–147.

Arthur C Graesser, Patrick Chipman, Brian C Haynes, and Andrew Olney. 2005. Autotutor: An intelligent tutoring system with mixed-initiative dialogue. *Education, IEEE Transactions on*, 48(4):612–618.

W Lewis Johnson and Andre Valente. 2009. Tactical language and culture training systems: Using ai to teach foreign languages and cultures. *AI Magazine*, 30(2):72.

Rohit Kumar, Carolyn Penstein Rosé, Yi-Chia Wang, Mahesh Joshi, and Allen Robinson. 2007. Tutorial dialogue as adaptive collaborative learning support. *Frontiers in artificial intelligence and applications 158*.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 44(4):978–990.

Carmen Muñoz. 2006. *Age and the rate of foreign language learning*, volume 19. Multilingual Matters.

P. Robinson. 2011. *Second Language Task Complexity: Researching the Cognition Hypothesis of Language Learning and Performance*. Task-based language teaching : issues, research and practice. John Benjamins Publishing Company.

Andreas Stolcke, Elizabeth Shriberg, Rebecca Bates, Noah Coccaro, Daniel Jurafsky, Rachel Martin, Marie Meteer, Klaus Ries, Paul Taylor, and Carol Van Ess-Dykema. 1998. Dialog act modeling for conversational speech. In *In AAAI spring symposium on applying machine learning to discourse processing*.

Andreas Stolcke, Noah Coccaro, Rebecca Bates, Paul Taylor, Carol Van Ess-Dykema, Klaus Ries, Elizabeth Shriberg, Daniel Jurafsky, Rachel Martin, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Sowmya Vajjala and Detmar Meurers. 2014. Exploring measures of readability for spoken language: Analyzing linguistic features of subtitles to identify age-specific tv programs.

Sowmya Vajjala and Detmar Meurers. 2016. Readability-based sentence ranking for evaluating text simplification. *arXiv preprint arXiv:1603.06009*.

Lev Vygotsky. 1978. Interaction between learning and development. *Readings on the development of children*, 23(3):86.

Sabrina Wilske and Magdalena Wolska. 2011. Meaning versus form in computer-assisted task-based language learning: A case study on the german dative. *JLCL 26, no. 1*.

David Wood, Jerome S Bruner, and Gail Ross. 1976. The role of tutoring in problem solving*. *Journal of child psychology and psychiatry*, 17(2):89–100.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

John Zakos and Liesl Capper. 2008. Clive–an artificially intelligent chat robot for conversational language practice. In *Artificial Intelligence: Theories, Models and Applications*, pages 437–442. Springer.