



# THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

# Modelling Speaker Adaptation in Second Language Learner Dialogue

*Arabella Jane Sinclair*



Doctor of Philosophy  
Institute for Language, Cognition and Computation  
School of Informatics  
University of Edinburgh  
2020



# Abstract

Understanding how tutors and students adapt to one another within Second Language (L2) learning is an important step in the development of better automated tutoring tools for L2 conversational practice. Such an understanding can not only inform conversational agent design, but can be useful for other pedagogic applications such as formative assessment, self reflection on tutoring practice, learning analytics, and conversation modelling for personalisation and adaptation.

Dialogue is a challenging domain for natural language processing, understanding, and generation. It is necessary to understand how participants adapt to their interlocutor, changing what they express and how they express it as they update their beliefs about the knowledge, preferences, and goals of the other person. While this adaptation is natural to humans, it is an open problem for dialogue systems, where managing coherence across utterances is an active area of research, even without adaptation.

This thesis extends our understanding of adaptation in human dialogue, to better implement this in agent-based conversational dialogue. This is achieved through comparison to fluent conversational dialogues and across student ability levels. Specifically, we are interested in how adaptation takes place in terms of the *linguistic complexity*, *lexical alignment* and the *dialogue act usage* demonstrated by the speakers within the dialogue. Finally, with the end goal of an automated tutor in mind, the student alignment levels are used to compare dialogues between student and human tutor with those where the tutor is an agent.

We argue that the *lexical complexity*, *alignment* and *dialogue style* adaptation we model in L2 human dialogue are signs of tutoring strategies in action, and hypothesise that creating agents which adapt to these aspects of dialogue will result in better environments for learning. We hypothesise that with a more adaptive agent, student alignment may increase, potentially resulting in improved engagement and learning.

We find that In L2 practice dialogues, both student and tutor adapt to each other, and this adaptation depends on student ability. Tutors adapt to push students of higher ability, and to encourage students of lower ability. Complexity, dialogue act usage and alignment are used differently by speakers in L2 dialogue than within other types of conversational dialogue, and changes depending on the learner proficiency. We also find different types of learner behaviours within automated L2 tutoring dialogues to those present in human ones, using alignment to measure this. This thesis contributes new findings on interlocutor adaptation within second language practice dialogue, with an emphasis on how these can be used to improve tutoring dialogue agents.

## Lay Summary

Speakers in conversations dialogue adapt to one another in different ways, changing things like their choice of words or phrasing depending on what they believe about the person they are speaking with. For example, if someone believes they are talking to someone who is not a fluent speaker of the language, they might choose simpler words or ask questions to see that the other person understands them. We might expect to see this kind of adaptation in dialogues where learners of a second language (L2) are practising conversation with a tutor. This thesis claims that in L2 practice dialogues, tutors and students adapt to each other in different ways depending on learner ability, consistent with the idea that tutors seek to strike a balance between language that is simple enough for the learner to follow the conversation while still challenging the learner, and that learners take advantage of the examples that tutors are providing. We show that over the course of a single dialogue, the tutor and student adapt to each other and this changes with student ability level.

We demonstrate *Linguistic Complexity* adaptation. We show that how hard or easy the language of each speaker is will change over the course of a dialogue. For students with a low ability level, if a tutor's language is much more complex than the learners, tutors make their language easier towards the end of the dialogue. For students who are more fluent speakers, if the tutor's language is of a complexity too similar to the student, the tutor will increase the difficulty of their language, which we think is a sign of them pushing the student to learn.

We demonstrate *Lexical Alignment*, which means we show that over the course of an interaction, tutors and students begin to use the same words as each other. In fluent conversational dialogue between native speakers (such as a typical telephone conversation between acquaintances), alignment is stronger than it is in second language dialogues. We find that with students with higher ability align more to their tutor than lower ability students, becoming more like fluent conversational speakers. We also find that students align more to harder words, which we think may be a sign of them learning vocabulary from the interaction.

We demonstrate *Dialogue Style* adaptation, which we measure by comparing the types of things speakers say, which we call *Dialogue Acts (DAs)*. For example, whether they ask many *questions*, make more *statements*, or just make simple *yes* and *no answers*. We show that low ability learners have a less similar pattern of dialogue interaction to their tutors than high ability learners do. We also show that over the course of a dialogue, the

types of DA speakers use become more similar. For low ability learners, tutors adapt their DAs to be more similar to the student. For high ability learners, it is the student's DA use that changes, becoming more similar to the tutor's. We compare dialogue style in L2 dialogue to fluent conversation, finding DA usage in L2 is very different.

Finally, we compare student alignment when they talk to a human tutor to student alignment when they are talking to a tutor-bot (an automatic L2 tutoring dialogue agent). We show that the types of words and phrases of the tutor that the student uses are different when the dialogue is with an agent. We find that when learning from an agent, they show different behaviour patterns which we can see in the words they use.

We think it is important to understand tutor adaptation and how learners interact with both human tutors and agents. We hope that our work can be used to develop more personalised adaptive L2 dialogue tutoring agents.

# Acknowledgements

They say it takes a village, and in the case of my PhD experience this has been true. I have been truly lucky in the people who have met and inspired me during my years in Edinburgh and who have been crucial to my development as a researcher.

Firstly I would like to acknowledge my supervisors: Jon Oberlander showed me that my interests were worthwhile exploring, and encouraged me to ask questions. He would give rambling, half-baked theories and ideas his full attention, then reflect them back at you as if you were the most interesting person, while rephrasing and adding to them to make them sound like concrete, intelligent, well-formed ideas. I miss him, particularly during the write up. I would not be writing up if he hadn't had the kindness and empathy to take me on as a student. Dragan Gašević has provided me with constant encouragement and has helped me enormously with my confidence as a researcher. He is also to thank for making me think of writing as programming, splitting up sections to make them modular, defining variables before you use them, writing an outline like pseudocode. Both Jon and Dragan took a chance on me as a PhD student changing topics at the end of my first year, doing something I had very little prior experience in, and I will be forever grateful for this.

When Jon passed away and Dragan moved to Australia (although our long-distance supervision works extremely well especially at paper deadlines when I am awake writing at 3am and can message my supervisor for feedback at a reasonable time of day for him), Adam Lopez and Chris Lucas were kind enough to become my Edinburgh-based supervisors, although in practice they also became a large part of my support network. Adam was responsible for encouraging me to participate in his reading group, giving me confidence to expose my half-baked research to my peers and talk about them earlier, which has made my work better, and me a better researcher. Hopefully our meetings will have been enough for me to properly internalise his catchphrase "*maybe you should demonstrate this idea with an example*". Chris has really made me feel comfortable with not knowing how things work, and figuring them out collaboratively in conversations with a notebook, pen, many diagrams, and maths that I can only understand in context. I am very grateful for his patience and enthusiasm to discuss ideas for far longer than our scheduled meeting times. Thank you all from the bottom of my heart for helping me believe that I can become a researcher.

Thanks to Bonnie Webber, Johanna Moore and Helen Pain for kindly giving me feedback in my yearly reviews. Thanks to Mirella Lapata, Alex Lascaredes and Maria Wolters: at pivotal points in my PhD, you helped me take ownership of my work,

becoming a more independent researcher because of this. Thanks to my co-authors Rafael Ferreira and Kate McCurdy, you have both taught me how much I enjoy collaborating.

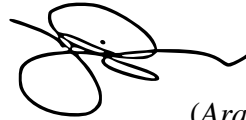
I have been very lucky to get to know the other amazing PhD students who surround me in the informatics forum, I love how international, multidisciplinary and inspiring you all are. You provided an opportunity to *not* talk about our work, it was comforting that we were all struggling along together. Akash, Ursula, Maria, Rui, Kate, Luda, Nico, Pablo, Jenny, Joe, Martino, Yota, Natalia, Natalie, Stan, Andrew you made Edinburgh begin to feel like home for me. Thanks. I was also incredibly lucky to have such nice office-mates: Akash Stef, Siva, Philip, Bharat, Uche, Nellie and David, thanks for making 3.39 the cool office, thanks for the after work beers and coffee breaks. Akash, you became one of my best friends, and I probably would have left my PhD if it wasn't for you telling me to keep my head down and run forwards. Thanks to my PhD siblings Amy, Arlene, Ed, Nico, Pablo, Kate, Naomi for talking about work, being in the same boat, helping figure out where our supervisors were at any given point, and proofreading my work, I really appreciate it. Thanks to all of Adam's Agora group for being so friendly, providing a low-risk atmosphere for problem solving, and for giving me such useful feedback and discussion on my work. I have also had much feedback and support from too many members of the ILCC to name, but the random bits of advice in the corridor, chats by the coffee machine, lunch on the third floor, 4pm cookie time, and coffee dates were all very influential and are all hugely appreciated. Thanks especially to Maria, Nelly, and Kate for keeping me sane this year as I write up. Thanks to Duygu and Maria for making long distance PhD friendship seem easy and for our post PhD career conversations. Thanks Alexander, Avashna, and Kate for coffee with gossip and brainstorming. Thanks to Jessie, Christine, Luke, Flic, Anna, Fergal, Max, Amanda and Rowan for listening to me talk about my thesis endlessly.

Finally, thanks to my family and friends who have provided me with love, emotional support, meals and snacks, long phone calls and glasses of wine through a pretty hectic 4 years of my life. Cicely, thanks for hyping me up, you're the best sister a girl could hope for. Thanks mum, dad, sis, granny and granddad for the proof-reading and being my experiment guinea-pigs. I am grateful to family members past and present for giving me an interest in how things work, why things happen, how people from other places see the world, how language is important and how we can learn from the past. We are a product of our experiences, thanks to everyone who has shaped mine.



# Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

A handwritten signature in black ink, consisting of several loops and a long horizontal stroke extending to the right.

11/4/2020

*(Arabella Jane Sinclair)*

# Table of Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Thesis Outline . . . . .	4
1.1.1	Linguistic Complexity . . . . .	4
1.1.2	Lexical Alignment . . . . .	5
1.1.3	Dialogue Act Usage . . . . .	6
1.1.4	Student to Agent . . . . .	7
1.2	Contributions . . . . .	8
<b>2</b>	<b>Background</b>	<b>11</b>
2.1	Second Language Learning . . . . .	11
2.1.1	Computer Assisted Language Learning (CALL) Dialogue . . . . .	12
2.1.2	Tutoring Strategies . . . . .	14
2.2	Adaptation . . . . .	16
2.2.1	Measuring Linguistic Complexity . . . . .	16
2.2.2	Alignment & Grounding . . . . .	19
2.2.3	Word Frequency and Alignment . . . . .	20
2.2.4	Alignment in L2 learning . . . . .	20
2.2.5	Measuring Alignment . . . . .	22
2.2.6	Dialogue Acts . . . . .	22
2.3	Corpora . . . . .	24
2.3.1	L2 Learner . . . . .	25
2.3.2	Conversational . . . . .	26
2.3.3	Task based . . . . .	28
2.3.4	Student-Agent . . . . .	28
<b>3</b>	<b>Linguistic Complexity Adaptation</b>	<b>31</b>

3.1	Finding the Zone of Proximal Development: Student-Tutor L2 Dialogue Interactions . . . . .	32
3.2	Further Discussion . . . . .	42
3.2.1	Automatic Assessment . . . . .	42
3.2.2	Complexity Prediction . . . . .	44
3.2.3	Dialogue Act Annotation . . . . .	50
3.3	Contributions . . . . .	52
<b>4</b>	<b>Alignment</b>	<b>55</b>
4.1	Does Ability Affect Alignment in Second Language Tutorial Dialogue? . . . . .	56
4.2	Further Discussion . . . . .	67
4.2.1	Capturing Vocabulary - Concept Introduction . . . . .	67
4.2.2	Initiative & Alignment . . . . .	69
4.3	Contributions . . . . .	70
<b>5</b>	<b>Dialogue Style Adaptation</b>	<b>73</b>
5.1	I wanna talk like you: Comparing Speaker Adaptation in L2 Practice Conversation to Fluent Speakers . . . . .	74
5.2	Contributions & Discussion . . . . .	101
<b>6</b>	<b>Human-Agent Alignment</b>	<b>103</b>
6.1	Tutorbot Corpus: Evidence of Human-Agent Verbal Alignment in Second Language Learner Dialogues . . . . .	104
6.2	Further Discussion . . . . .	111
6.3	Contributions . . . . .	112
<b>7</b>	<b>Conclusion</b>	<b>115</b>
7.1	Summary of Contributions . . . . .	115
7.2	Implications . . . . .	117
7.3	Future Directions . . . . .	118
	<b>Bibliography</b>	<b>121</b>

# Chapter 1

## Introduction

Total immersion is one of the best ways to learn a second language language, since students learn more quickly when they are given extended opportunities to use the language interactively (Genesee 1985, 2004). A large part of learning a language while living in a country where the target language is spoken is that the learner is forced to interact in the target language on a regular basis as best as they can. This constant dialogue practice is often missing in a classroom setting due to syllabus, class size and time constraints (Rossiter et al. 2010).

Second Language (L2) tutoring dialogue agents can help make up for the lack of dialogue practice in a classroom setting, and have been trialled by online learning platforms such as Duolingo<sup>1</sup> and Babbel<sup>2</sup>. Autonomy and independence in L2 language learning is important, and providing learner-centric technology which helps learners find what works for them plays an increasingly important role in language learning (Benson & Voller 2014). Tutorial dialogues are very effective teaching methods (Vanlehn 2006), and a better understanding of a tutor's approach in one-to-one settings could help improve upon automated tutoring.

The goal of dialogue practice for a second language learner is to facilitate their production of dialogue similar to that between native speakers. Dialogue allows a learner to practice constructing newly learnt concepts, the paraphrasing and repetition of their tutor in a collaborative context leading to better retention (Ahmadian et al. 2014).

Tutors adapting their strategy to student ability is very intuitive, and has been well studied at a theoretical level (Costa et al. 2008, de la Colina & Mayo 2009, Lantolf

---

<sup>1</sup>duolingo.com

<sup>2</sup>babbel.com

2000a). There is less empirical work to understand how this works in practice. This thesis takes some of these theoretical ideas and tests them empirically.

**Thesis Statement:** *In L2 practice dialogues, both student and tutor adapt to each other, and this adaptation depends on student ability. Tutors adapt to push students of higher ability, and to encourage students of lower ability.*

This thesis demonstrates that tutors adapt in terms of linguistic complexity, lexical alignment and dialogue act usage, challenging the students who are more capable, and simplifying their language for the less capable. These effects can be seen over the course of a single interaction. Student adaptation is also demonstrated, with alignment correlating with ability level, and with ability corresponding to greater levels of both convergence of and change in dialogue act usage.

### Example

A tutor's role in conversational second language practice is to adapt to the needs of the learner in order to foster their learning. This can include interacting *within* the learners' capabilities such as translating their language to the language of the learner (such as in the Beginner dialogue in Table 1); and *outwith* it, to introduce new linguistic concepts or vocabulary within a space where they are comfortable and can learn through context, such as adding more natural conversational interaction to a routine practice dialogue. For example, at line 10 of the intermediate dialogue in Table 1, the tutor comments "*quite early*" in response to the student, instead of a simple "*ok good*". There is less empirical work to understand how this works in practice.

	Beginner	Intermediate	
	1 do you like the school ?	do you like this school ?	1
	2 [-spa] m-entens ?	yes .	2
	3 <i>0 [= says nothing]</i>	yes ?	3
	4 "do you like" ?	what are you planning to do next year ?	4
	5 do you like the school ?	<i>I would like to study zoology .</i>	5
	6 <i>hmm</i>	what time did you arrive here this morning ?	6
[htp]	7 no si t-agrada l-escola ?	<i>this morning ?</i>	7
	8 do you like the school ?	yes .	8
	9 <i>yes .</i>	<i>I ... I am here since eight o'clock .</i>	9
	10 yes ok .	uhhuh right quite early .	10
	11 now what time do you begin in the morning ?	and when will you leave ?	11
	12 <i>0 [= says nothing] .</i>	<i>I ... I finish my time-Table in half-past-two .</i>	12
	13 [-spa] m-entens ?		13
	14 <i>[- spa] no .</i>		14

A human tutor makes constant adjustments throughout a lesson, making choices about how to deliver feedback, corrections, content, new concepts, and reassurances. They must also consider adapting the level of the interaction to be an appropriate balance between being within reach of the learner yet still stretching them. Table 1 shows snippets from two tutor interactions, one with a Beginner student and one with an Intermediate student. When working with beginners, tutors provide learners with much more support in the form of repetition (for example in Table 1 lines 5 and 8) and translation (lines 7 and 8). Even acknowledgements are short: for example, at line 10, a simple confirmation is given to the beginner student, since they would in this case probably not understand anything more. With students of higher ability however, since tutors do not need to provide as much support, they can push the student to converse in a more fluent manner, asking longer-form questions, follow up questions, and giving longer-form response acknowledgement. This can be seen in lines 10 and 11 in the Intermediate column of Table 1, where the student is given a commentary on their response, as would be the case in a real conversational setting.

Table 1 shows that tutor *adaptation* to the student is key to ensure that any dialogue takes place at all. What these examples illustrate is that good tutors, subconsciously or not, adapt in a manner in-keeping with Vygotsky's Zone of Proximal Development (ZPD; Vygotsky 1987). According to the ZPD model, a 'good' dialogue tutor should adapt their interactions to remain within reach of the learner's capabilities, yet provide sufficient challenge to push the learner to the farthest extent of their abilities. This adaptation can present itself in the form of a number of different tutoring strategies, such as: *simplifying statements*, gradually increasing their complexity over time; *providing example answers*; *repeating* or *providing key vocabulary*; *grounding* new concepts within the context of what the student already knows; or *describing more clearly the task*.

For automatic L2 tutoring agents, contextual adaptation to the needs of an individual in dialogue is not yet possible. Before we can build tutoring agents to adapt to learners, we need to understand what human tutors do. To better model contextual adaptation, this thesis analyses aspects of interlocutor adaptation at different levels of student ability to better understand how tutors leverage their adaptation and that of their students within their teaching. We present our findings with respect to their informing the design of better automatic tutoring dialogue agents, and their potential for use in existing pedagogic applications.

## 1.1 Thesis Outline

A main aim of this thesis is to model aspects of dialogue to extend our understanding of adaptation within L2 learner dialogue. We therefore model how each speaker in a one-to-one L2 conversational practice dialogue *adapts* to the other over the course of a dialogue. There is a large difference between how tutors interact with high and low ability students, as can be seen in Table 1. Firstly, it is noticeable that when interacting with a Beginner as opposed to an Intermediate student, the tutor uses simple language and short phrases, which is a sign of reduced **Linguistic Complexity**. Secondly, the tutor repeats both their own words and phrases and that of their students' (and vice versa for Intermediate students), which is an indicator of **Lexical Alignment**. Finally, the interaction styles between Beginner and Intermediate is considerably different. One way to model interaction is with **Dialogue Acts**, which can capture how the Intermediate student makes more *statements* and *questions* in comparison to the Beginner, who makes more simple *yes-no* responses.

The following subsections outline the main content chapters of this thesis, exploring *linguistic complexity* (Chapter 3), *lexical alignment* (Chapter 4), and *dialogue act usage* (Chapter 5) in the context of L2 dialogues; comparing the effects found to those within fluent conversational dialogues and fluent task-based dialogues. Lexical alignment is also used to explore the contrast between the behaviour of students interacting with a human tutor, to that with an L2 tutoring agent (Chapter 6).

### 1.1.1 Linguistic Complexity

When modelling linguistic complexity adaptation, it helps to think of how we ourselves alter our conversation when confronted with someone who has less of a grasp on the language being used to communicate than we do. In practice, this may mean we use fewer complicated words, simpler tense and grammatical structure, and shorter sentences as seen when simplifying sentences in Heilman et al. (2007) and Aluisio et al. (2010). This adaptation that we automatically make is in order to interact at a level where our interlocutor is able to understand us, but not such a low level where our interaction is artificially impeded by communicating at their exact level; since often comprehension is at a higher level than production in L2 speakers. In a context where the goal of interaction with an L2 interlocutor is to foster their learning of the target language, adapting language in order to maintain this ideal distance between what they are capable of producing and what they can comprehend from context is key; therefore

being aware of the competencies of the learner is a key aspect of oral L2 tutoring.

At line 4 in Table 1, lack of a response from the learner prompts the tutor to simplify the question and highlight a phrase that the learner may know without the distraction of the topic word. It is the low-complexity of the tutor's response and the clear indication of non-understanding which prompts them to translate the sentence before repeating the question in English. This informs the tutor's learner-model that the learner is a beginner. When working with the intermediate student, the simple yes response on line 8, although linguistically non-complex, informs the tutor that the learner has understood the question and can respond. This response prompts a follow-up question which is more open and therefore more complex.

In Chapter 3, we show that tutors adapt their linguistic complexity to maintain a certain distance from that of their students, either through reducing it for lower ability students, or increasing it for those of higher ability. We interpret that this could be evidence of their adherence to the ZPD. In this thesis, we show evidence of and demonstrate that tutor adaptation of their linguistic complexity can be measured via a set of surface features in the text within the utterances. We show that utterances with different functions exhibit different complexity traits. This work is published in Sinclair et al. (2017), and, along with an extended description of the experiments involved, is presented in Chapter 3.2.1.

### 1.1.2 Lexical Alignment

Lexical alignment, in layman's terms, is the phenomenon whereby if we interact over time with an interlocutor, we will begin to use the same words and phrases as they do (Pickering & Garrod 2004a). This can also be seen in lines 6 and 7 of the Intermediate dialogue in Table 1, where the learner repeats the final noun phrase of the tutor's question. The *Interactive Alignment Model* (Pickering & Garrod 2004b) suggests that successful dialogue arises from an alignment of representations (including phonological, lexical, syntactic and semantic), and therefore of speakers' situation models. This model assumes that these aspects of the speakers' language will align automatically as the dialogue progresses, and will greatly simplify both production and comprehension in dialogue. Alignment in L2 dialogue has been less well studied, and the factors leading to alignment within L2 learning may differ from how this occurs between native speakers (Costa et al. 2008), and can be due to both the goals of the tutor and the ability of the learner. The tutor may repeat themselves during an interaction to repeat and



reinforce new or unfamiliar vocabulary to the learner (such as in lines 1, 5 and 8 of the beginner column in Table 1), or through repeating what the learner says to acknowledge or encourage the correct usage of words which stretch them. A learner will be constrained by their production abilities in terms of what words they can repeat from the tutor's language. However, the repetition of these less familiar words is part of the learner's process of picking up new vocabulary from the context of the interaction (this may be what is happening in line 7 of the Intermediate dialogue in Table 1).

Consider what happens in Table 1 on line 5. The intermediate learner response to the question could have been framed as '*I want to...*' or '*I will*' or simply '*to study zoology*' each of which would have resulted in a linguistically less complex sentence (due to the use of the conditional tense, and the higher frequency of the phrases *want to* and *will*). It could simply be the case that '*I would like to...*' is a phrase that the learner has learnt corresponds to '*quiero*', i.e., *want* in Spanish. It could also be the case that the recent usage of '*like*' by the tutor in line 1 has primed the learner to use this vocabulary item which arguable is more complex to them than '*want*' due to its less frequent use in the English language<sup>3</sup>.

In this thesis, we expand on previous work measuring alignment effects within conversational and task based dialogue, measuring alignment within L2 dialogue. We show that alignment correlates with student ability, and that between speaker alignment is asymmetric in L2 dialogues, with students aligning more to tutors than vice versa. Finally, we find that students at higher levels align to 'harder' words more so than students at lower levels, which we hypothesise may be evidence of them using alignment as a vocabulary learning strategy. This work is published in Sinclair et al. (2018), and it is presented in Chapter 4 along with additional, unpublished experiments.

### 1.1.3 Dialogue Act Usage

Dialogue Acts (DAs) are often used to model discourse structure within dialogue; a DA describes the meaning of an utterance (Stolcke et al. 2000). Examples of DAs include *statement* (lines 5, 9, 10, 12 Intermediate Table 1) *response-acknowledgement* (line 10 Beginner Table 1), *question* (lines 1, 2, 4, 5, 7, 8, 11, 13 Beginner and lines 1, 4, 6, 11 Intermediate Table 1), and *yes-answer* (line 9 Beginner and lines 2, 8 Intermediate Table 1), among others. The types of interactions within the dialogue at a high level as captured by DA usage is another aspect of interaction which will, in L2 dialogue,

<sup>3</sup>*want* has a ranking of 83rd and *like* 208th in terms of most common word according to a corpus of contemporary American English (<https://www.wordfrequency.info>)

differ from conversational dialogues between fluent speakers. The difference is due to tutoring strategy, which can be seen at the level of DAs (Rus et al. 2017, Boyer et al. 2011). In conversational dialogue between native speakers, which is the goal for L2 practice dialogue, the speakers' interactions will be largely symmetrical (Sinclair et al. 2017), with an even give and take of statements, questions, and acknowledgements. The hierarchical nature of the tutor-student relationship will influence this symmetry, as the tutor has the lead both in terms of their language proficiency, and in their more influential status within the dialogue, regardless of their goal to facilitate relaxed conversational dialogue practice. Over the course of a dialogue, both speakers may adapt to the other in terms of the style of dialogue they produce; the tutor to set the learner at ease, and the learner as they use the example of the tutors interaction in context.

For example, in Table 1, the beginner student is unable to respond to the question on line 5, however the intermediate one displays another important aspect indicative of learner proficiency - the ability to ask a clarification question. This change in interaction from the intermediate student from passively answering questions to actively taking part in the dialogue is a marker of proficiency. The tutor adapts to this by changing their role of questioning and support, to response (line 8) and commentary (line 10).

Chapter 5 shows that the usage of dialogue acts in L2 dialogue changes for both speakers as a function of student ability. We find that interlocutors tend to converge to similar DA usage patterns, contrasting our findings with fluent conversational and task based dialogues. We find that speaker DA usage across corpora has some distinctive differences, and show convergence over the course of a dialogue to be greater in L2 than in the non-educational dialogues compared. We propose using dialogue acts as an additional feature when observing alignment at a higher level; that is, not at a level of sequential utterances, rather at a less fine-grained level to measure conversational symmetry, and whether it is present within the dialogue, as a sign of student competence. This work is published in Sinclair, Ferreira, Lopez, Lucas & Gasevic (2019), and an extended journal version under submission is presented in Chapter 5.

#### **1.1.4 Student to Agent**

Chapters 3, 4 and 5 explore adaptation between humans, comparing the relationships between linguistic complexity, lexical alignment and dialogue act usage of students and tutors to those between speakers in fluent, non-educational interactions. Since one

of the goals of this thesis is applying our better understanding of human interactions to the task of *automatic* tutoring, in Chapter 6, we compare alignment within the human L2 corpus to that between student and agent in an automatic tutoring setting.

In an interaction where a learner knows they are interacting with a dialogue agent, the adaptation of learner to tutor may be very different to how they would act with a human. Interaction with a tutor-agent could have positive aspects for the student such as reducing social pressure e.g. students will not be embarrassed about their accent, such as could be the case for the hesitation in the Beginner dialogue in Table 1. However, for less engaged students, an automated tutor may not have the ability to keep them on track; which the tutor of the Beginner, in the example in Table 1, has to do. We know, that humans do align, thereby adapting, to computers more so than to other humans, in the context of task-based (Branigan et al. 2010) and negotiation dialogues (Duplessis et al. 2017). In educational settings, by contrast, alignment has been found to predict both student learning and engagement (Ward & Litman 2007a). In the case of user interaction with a virtual agent, alignment increases perceived interaction naturalness and maintains user engagement (Yu et al. 2016). Although alignment may be different in L2 dialogues with an agent, using alignment as a predictor of engagement could prove useful in the analysis of effective agent interactions. Whether L2 students will be able to align to an agent, which does not provide them with the same level of support as a tutor would, or whether this would be done in a similar manner to which they align to a human tutor is a question we address in Chapter 6 of this thesis.

In this thesis, we expand on previous work comparing alignment in human-human vs human-agent in negotiation dialogues, contrasting human-human vs human-agent in L2 learner corpora. We use alignment, at the level of expression repetition, to find different patterns of learner engagement behaviour within the human-agent corpus in comparison to the more uniform alignment distribution present in human corpora. We offer a discussion on the qualitative differences between these corpora. This work is published in Sinclair, McCurdy, Lopez, Lucas & Gasevic (2019), and is presented in Chapter 6.

## 1.2 Contributions

This thesis contributes to our understanding of adaptation in L2 dialogue. We find that within the L2 practice dialogues we explore, tutors and students adapt to each other in different ways as a function of learner ability. We find evidence of tutor adaptation

to the linguistic complexity of a student (Sinclair et al. 2017); that alignment effects correlate with student ability (Sinclair et al. 2018); that speakers' Dialogue Act usage converges over the course of an interaction (Sinclair, Ferreira, Lopez, Lucas & Gasevic 2019); and finally, that L2 learners will align to a conversational agent, albeit in a manner different to that of a human tutor (Sinclair, McCurdy, Lopez, Lucas & Gasevic 2019). This understanding of adaptation opens up new questions about how adaptation may change on more levels than those explored in this thesis, and how learning and tutoring strategy may influence these. Our work has implications for the design of automatic tutoring agents, and as pedagogic tools in L2 language learning. It also serves as a starting point for further investigation into the role of alignment and adaptation within L2 dialogue.



# Chapter 2

## Background

This chapter introduces and motivates the importance of conversational practice in L2 learning, the need for better tools, and the significance of dialogue practice in L2 education (Section 2.1). This chapter also provides background on the aspects of speaker adaptation modelled in this thesis and why they are relevant to L2 learning (Section 2.2). Finally, the corpora used for our analyses are described in Section 2.3.

### 2.1 Second Language Learning

Learning a second language is a useful and desirable skill, since this opens doors for communication, also allowing the learner to see the world through the eyes of the culture their target language expresses, leading to an expanded view of concepts, as expressed in both their mother tongue and their L2 (Cook et al. 2006). Immersion<sup>1</sup> has been described as the “mother’s method” (Lenneberg 1960) for learning an L2. Immersion is a communicative approach that reflects the essential conditions of first language learning and at the same time responds to the special needs of L2 learners (Genesee 1985).

Learning a language via immersion is not always practical or possible, however true it is that L2 acquisition is enhanced when students are given extended opportunities to use the language interactively (Genesee 2004). Immersion gives learners the context to constantly practice expressing themselves in their target language: the best environment in which to learn if viewing L2 acquisition through the lens of Socio-Cultural Theory (SCT) (Lantolf 2000*b*), where the development of human cognitive and higher

---

<sup>1</sup>Immersion can both mean the educational technique of teaching bilingual learners in both languages, or immersing a learner in their target language (Genesee 2004)

mental functions comes from social interactions. SCT emphasises the central role of dialogic interaction in all learning, where, as a result of dialogic inter-psychological activity, new knowledge is appropriated. In other words, students learn through talking (Hawkes 2012).

One-to-one spontaneous dialogue practice represents an important aspect of L2 learning in both a classroom setting and online learning platforms. This form of dialogue practice has been shown to provide greater opportunities for L2 learning (Hawkes 2012, Bailey 2001, Samana 2013, Birjandi & Jazebi 2014, Lantolf 2000a), as learners can both take advantage of the example of their interlocutor, and learn through practice.

However, student speaking difficulties are common many L2 learning classrooms, with reasons cited being lack of student support, student shyness or simply lack of exposure to practice (Al Hosni 2014). Similarly, Rossiter et al. (2010) argue that current L2 resources for oral practice must be supplemented, as a lack of oral exercise in classroom materials contributes to a general lower oral proficiency in the L2 learning.

Dialogue agents have high potential for serving as an inexhaustible source of L2 practice. An ideal agent should be able to mimic a tutor in its ability to adapt its linguistic complexity to suit the learner (Morales-Jones 2011), keeping the dialogue in the Zone of Proximal Development (Vygotsky 1987), and scaffolding linguistic knowledge (Abbott 2014). If we view dialogue as a mediated or collaborative learning process, we can expect to see both speakers trying to arrive at a shared understanding at the utterance exchange level (Lantolf 2000a). While we expect speakers to arrive at a communicative symmetry (Van Lier 1996), we do not expect them to be able to do so in all cases, as the nature of a tutor-student relationship has an expert-novice asymmetry. That said, at higher levels of student ability, the tutor should begin to alter their role to that of conversational peer, to better encourage student independence and autonomy; thus slowly removing some support (Birjandi & Jazebi 2014). An ideal L2 dialogue tutoring system would take all of the above into account.

### **2.1.1 Computer Assisted Language Learning (CALL) Dialogue**

The automation of L2 teaching is extremely desirable as it allows learners better access to education through their smartphone or computer, and addresses some of the constraints of the L2 classroom mentioned above. Examples of large scale app-based

CALL systems include Duolingo<sup>2</sup>, Memrise<sup>3</sup>, and Babbel<sup>4</sup>. These applications have a large user base and already serve to make language learning more accessible. CALL for adaptive feedback can take the form of guided self assessment, for example, in an interface which “grades” input text on language proficiency level, using a colour scale to highlight irregularities or errors (Yannakoudakis & Briscoe 2012). This “grading” analyses the coherence of user-input discourse, having learnt from both large collections of correctly formed texts, and error-annotated English assessment scripts from the Cambridge Learner Corpus (Briscoe et al. 2010). These applications are adaptive in so far as they have a model tailored to individual user progression, rather than forcing a user to follow a static script, or allow the user to regulate their own learning.

Conversational dialogue agents have been explored for their potential use in e-learning (Kerry et al. 2009, Graesser et al. 2005, Dzikovska et al. 2014) and as a partner for L2 practice (Levy 2009, Stewart & File 2007). Examples of dialogue agents for one on one L2 learning include CLIVE, an agent which allows learners to practice basic conversation and fall back on their native language for clarification (Zakos & Capper 2008), and more form-based teaching, which varies the explicitness of corrective feedback (Wilske & Wolska 2011). Most L2 dialogue agents are heavily constrained and do not adapt their *language complexity* to the student, rather they focus on the troubleshooting of the content communication (Wilske & Wolska 2011), or other, less open-ended conversational goals. An example of this constrained style of interaction was Duolingo’s chat-bot, which allowed the user to participate in a heavily scripted dialogue with constrained text entry<sup>5</sup>. Immersive games-based dialogue tutoring has also been proven an effective environment for language learning (Johnson & Valente 2009).

*Adaption* of agent to learner, however, is an ongoing research task, although outside L2 tutoring, is a well-explored area from which we can learn much about how different strategies in automatic tutoring affect student learning (Graesser et al. 2005). Alignment, or “*more lexical similarity between student and tutor*” has been shown to be more predictive of increased student motivation (Ward et al. 2011, Forbes-Riley & Litman 2012). Adaptation of feedback strategies, for example, are important to best tailor agent interaction to student needs, such as adapting to the common errors a stu-

---

<sup>2</sup>Duolingo is an automated language teacher with syllabus automatically adapting to the learner (Settles & Meeder 2016) (<https://www.duolingo.com/>)

<sup>3</sup>Memrise is an online platform for L2 tuition (<https://www.memrise.com/>)

<sup>4</sup>Babbel is an online platform for L2 tuition, ([learn.languages.babbel.com/](http://learn.languages.babbel.com/))

<sup>5</sup>[www.duolingo.com](http://www.duolingo.com)



dent makes (Ferreira et al. 2007).

### 2.1.2 Tutoring Strategies

**The Zone of Proximal Development (ZPD)** is defined as:

*“the distance between the 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 (1980).*

In other words, between the space of what is known and what is not known, lies the ZPD, where skills which are too difficult for the learner to solve alone can be achieved with guidance and encouragement from another knowledgeable person. The construct of ZPD specifies that development cannot occur if too much assistance is provided or if the task is too easy (Lantolf 2013). Studies finding support for the ZPD can be found comparing the success of children working alone or with supportive instructions and encouragement (Freund 1990, Wood & Middleton 1975), which indicated scaffolding works best when the support is matched to the needs of the learner.

The ZPD is often synonymous with the term scaffolding (McLeod 2012). *Reciprocal teaching* is a more modern application of Vygotsky’s theories, in which students and teachers collaborate in learning, practising summarising, questioning, clarifying and predicting to improve a student’s ability to learn from text. *“From a Vygotskian perspective, the teacher’s role is mediating the child’s learning activity as they share knowledge through social interaction”* (Dixon-Krauss 1996).

Second language acquisition techniques include *alignment*, *recast*<sup>6</sup>, *explicit correction*, *scaffolding*, *key vocabulary identification* and crucially *discussion* whereby the teacher will discuss the elements known to the learner, or rephrasing what has been said/read to mirror the student level. The complexity of automatically delivering such scaffolds varies. Exploring recast techniques in the context of comparing explicit and implicit feedback on form-based language teaching has been tested in a L2 tutoring dialogue system, finding that meaning oriented, implicit adaptive instruction results in superior long term learning effects to more accuracy-focused drill-like activities (Wilske & Wolska 2011). The scope of the dialogue however, was constrained to specific grammatical exercises to explore the effects of these different scaffolding

---

<sup>6</sup>rephrasing a student’s incorrect attempt correctly

styles and not for a more general L2 conversational tutoring system due to implementation constraints.

*Scaffolding* in the context of education refers to a variety of instructional techniques used to move students progressively towards a stronger understanding, and ultimately, a greater independence in the learning process (Abbott 2014). Introduced by Wood et al. (1976), the term itself provides the metaphor to the kind of temporary support at successive levels of development needed to construct, in this case, knowledge, or to support learning. For example, if a student is not at the reading level required to understand a text of a certain complexity, a teacher can use “instructional scaffolding to incrementally improve their reading ability” (Abbott 2014). Common scaffolding techniques for this can be seen in Table 2.1.

Table 2.1: Common scaffolding techniques

1	Presenting a simplified version of the lesson or reading, and gradually increase the complexity or sophistication over time.
2	Describing or presenting a concept in a different way or simpler context to help understanding
3	Providing the student with an example answer
4	Giving key vocabulary before the reading of a difficult text
5	Clear description of the task and motivation for its benefits
6	Referencing how the current task builds on their existing knowledge

Line 6 in Table 2.1 can also be referred to as *Grounding* in dialogue. This consists of the participants establishing a common basis, or ground, on which their communication takes place. For example, if a student has used some vocabulary on a topic, and the tutor introduces a new word they don’t understand, the tutor can contextualise this within the language they know the student understands. Grounding can be viewed as a strategy for managing uncertainty and therefore error handling in dialogue (Skantze 2007).

Automatic scaffolding tries to attain the positive effects of scaffolding through the identification of particular traits associated with scaffolding in the field of learning which the automated tutoring is taking place. Scaffolding delivered by an automated tutor has been successful in some domains. For example, in basic science, agent scaffolds can help students learn how to create hypotheses, and thus what a hypothesis is (Sao Pedro et al. 2014). This work, while centred around the learning of a skill, is still very applicable to the learning of a new language, as the structure of the learning

is laid out to the student, and they are left to try to understand carefully selected elements. Better understanding the patterns of L2 learner dialogues at different levels of expertise can inform work in the field of CALL, specifically adaptive dialogue agents for L2 tutoring.

## 2.2 Adaptation

This thesis measures adaptation in terms of linguistic complexity, lexical alignment and DA usage. The thesis then compares lexical alignment of students in Human-Agent dialogues. In the following subsections, we therefore present some of the most relevant literature to our research in the areas of linguistic complexity, lexical alignment, DAs, and human-agent alignment.

### 2.2.1 Measuring Linguistic Complexity

Linguistic complexity of text can also be referred to as its *readability*. An example of linguistically less complex text is WikiSimple, where each article has been simplified in order to be understood by readers who either have a lower level of English, or of education in the subject matter of the article. Authors of these articles are instructed to “use basic English vocabulary and shorter sentences”<sup>7</sup>. Linguistic complexity describes qualitatively how hard it is for a reader or listener to understand a sentence or a text. There are many long-standing measures to compare how hard sentences are such as those proposed by Flesch (1948), Senter & Smith (1967) and Gunning (1952), which all use some weighted combination of words per sentence combined with either ratings of how many ‘hard’ or ‘easy’ words per sentence, or the addition of syllables per word to arrive at a score to indicate how ‘readable’ a sentence is.

Automatic readability measures are used to judge a text’s suitability or lack thereof for readers of lower ages, literacy, or, in the case of L2 learners, fluency. Readability prediction can take many forms, and is often applied to simplify or classify discourse at a document level. Examples of readability analysis include: the fine-grained task of predicting which words within a sentence make that sentence complex (Siddharthan 2006); the ranking of sentences by their readability (Vajjala & Meurers 2016), the grouping of documents into similarly appropriate reading levels for use in examinations (Heilman et al. 2007) and classrooms (Crossley et al. 2008); or the identifi-

---

<sup>7</sup>[https://simple.wikipedia.org/wiki/Main\\_Page](https://simple.wikipedia.org/wiki/Main_Page)

<i>Lexical</i>	Lexically complex words are those for which a “simpler” synonym exists; “simpler” can be defined for example as higher frequency, shorter length, or through using lexical semantic properties from WordNet (Miller 1995) such as average number of senses of a word, with a lower average indicating a more simple word. Lexical diversity and density, which can be measured by type-token and POS ratios, are also indicators of text complexity drawn from SLA research.
<i>Syntactic</i>	Features include patterns extracted from parse trees, as well as measures of syntactic complexity from SLA research - which includes measures such as sentence length.
<i>Morphological</i>	These encode the morpho-syntactic properties of lemmas, estimated from the Celex (Baayen et al. 1993) database.
<i>Psycholinguistic</i>	Word-level features such as concreteness, meaningfulness and image-ability can be extracted from the MRC psycholinguistic database (Coltheart 1980) and other Age of Acquisition (AoA) measures (Kuperman et al. 2012) can be used
<i>Simple Counts</i>	Features can include: n-grams, ”difficult” words from frequency lists, word length, syllables per word and other counts are commonly the first step in common readability measures such as the Flesch-Kincaid Grade Level Index calculation (Farr et al. 1951) based on a combination of average word, sentence, and syllable lengths.
Word Frequency	The higher frequency of a word, the more exposure a student has had to it, the more likely they are to learn it faster (Vermeer 2001). Word Frequency has also been shown to act as a reasonable indication of word ‘difficulty’ (Chen & Meurers 2017).

Table 2.2: Feature Sets successful in Readability Analysis

*This list was partially taken from this summary of features Vajjala & Meurers (2016)*

cation of complex aspects above a certain readability threshold in order to simplify them (Aluisio et al. 2010). Specific methods to target L2 speakers have also more

recently been developed by Xia et al. (2019).

Measuring the complexity of L2 dialogue has the additional aspect of how difficult the context of the exchange is. For example a learner may respond to a complex tutor question with a simple answer, but if the answer shows understanding of the question, then it should indicate a higher level of ability than the same answer to a very simple question. This is hard to achieve automatically as it requires modelling dialogue context. However, measures of linguistic complexity which work for discourse have been found to generalise with some success to dialogue (Vajjala & Meurers 2016) where subtitles from television programs targeted at young children, teens and adults are correctly assigned to their audience level category - this is a similar task to automatically differentiating between the linguistic complexity of the Beginner and Intermediate dialogue samples in Table 1. The linguistic measures used to achieve this are summarised in Table 2.2. Specific work in dialogue has found that the ranking of age appropriate television subtitles by their target audience age (and therefore linguistic complexity) can be effectively performed (Vajjala & Meurers 2014a) with support from features compiled in the SUBTLEX-UK corpus as part of the British Lexicon Project (Keuleers et al. 2012). The SUBTLEX-UK corpus consists of frequencies of words in Cbeebies, Cbbc, and BBC news programs: targeted at young children, children, and adults respectively. The English Vocabulary Profile (EVP) (Capel 2012a) can also act as a good measurement of word difficulty for L2 learners. It consists of common learner vocabulary grouped by the ability level in which this word is likely acquired. The grading used in the EVP is the Common European Framework of Reference (CEFR), which defines 6 levels of English proficiency in ascending order as: A1, A2, B1, B2, C1, C2. with A being beginner, B intermediate, and C advanced. The EVP dataset has been employed in recent work on predicting readability levels for L2 speakers (Xia et al. 2019). Cambridge automatic assessment of *English as a Foreign Language* (EFL) texts make up a large corpus of annotated and error corrected exam scripts from L1 and L2 writing exams known as the Cambridge Learner Corpus (Briscoe et al. 2010). The Cambridge Learner Corpus is used as a resource for the task of automatic assessment of exam scripts (Yannakoudakis et al. 2011), assigning grades to essays based off features present in the human-graded texts. Features are derived from the fields of Readability analysis and *Second Language Acquisition* (SLA) research, including syntactic and lexical features, and achieve impressive accuracy.

## 2.2.2 Alignment & Grounding

The Interactive Alignment Model (IAM; Pickering & Garrod 2004c) defines alignment with respect to underlying speaker *representations*, not behaviour. Within an interaction, interlocutors' situation models – a representation which corresponds to understanding and thus successful communication – are said to align. Alignment can occur at every linguistic level. For example, *lexical alignment* means that both interlocutors will strongly activate a concept with the same term as each other, such as the concept of *couch* in the following interaction: *speaker\_A*: “I love lying sprawled on the sofa”, *speaker\_B*: “me too, your sofa is the best”, without alignment of the word *sofa* for the concept *couch*, the reply may have been *speaker\_B*: “me too, your settee is the best”. Crucially, this alignment on the state of linguistic representations can potentially give rise to patterns of behaviour which can be measured. When interlocutors use the same words, this is referred to as *lexical entrainment* (Brennan & Clark 1996). Alignment, as Pickering & Garrod (2004c) define it refers to the underlying representations which give rise to this behaviour. Entrainment<sup>8</sup> is often used as an indicator of alignment. Alignment can be brought about via *priming*, an automatic mechanism whereby an interlocutor's use of a word activates a representation of that word, increasing its likelihood of subsequent use. In other words, priming can be defined as an interlocutor's use of a syntactic construction or lexical phrase more frequently than would occur by chance (Reitter et al. 2011). It has also been shown that speakers are sensitive to priming within their own speech and from that of their interlocutor (Branigan et al. 2000). In other words, priming brings about alignment of representations and hence linguistic entrainment (Costa et al. 2008).

A key assumption of the IAM is that in dialogue, production and comprehension are tightly coupled, leading to automatic alignment at many measurable levels of linguistic representation. One outcome of viewing dialogue as an interactive alignment process leads us to expect to see interlocutors develop and use routine expressions, developing a shared common ground. *Shared common ground* involves modelling the interlocutor's mental state, while *common ground* refers to the background knowledge that the interlocutors share. Clark & Wilkes-Gibbs (1986) show that the establishment of common ground is critical to the success of communication. Pickering & Garrod (2004c) argue that interlocutors align on an *implicit* common ground, drawing upon it in order to repair misalignment. Implicit common ground, and interlocutors' model

---

<sup>8</sup>Entrainment the process of a speaker adopting the reference terms of their interlocutor (Garrod & Anderson 1987)

of it is arrived at as a result of an automatic mechanism, rather than a more explicit constant check on the knowledge and understanding of an interlocutor that common ground would require. Implicit common ground, or the information shared between interlocutors, is extensive when in alignment, each interlocutors' situation model contains or foregrounds information that both speakers have either produced or comprehended.

The process of alignment means each speaker draws upon representations which have been set up over the course of the dialogue on the fly between two speakers for the purpose of that dialogue alone (Pickering & Garrod 2004c). The use of routinisation therefore contributes to the fluency of the dialogue in comparison to most monologue (Tavakoli 2016). That is, the interlocutors' space of alternatives to consider is smaller, and they have access to the others' words, grammatical constructions and concepts. Pickering & Garrod (2004c) suggest that language production can be greatly enhanced, even though the aim of dialogue is not the repetition of an interlocutors' sentences. The increase in production is due to the fact that the previous utterances will activate syntactic and lexical representations, leading to their re-use in the dialogue, and therefore, creating interlocutor alignment.

### **2.2.3 Word Frequency and Alignment**

One of the most universally accepted phenomena in experimental psychology is the word frequency effect (Taft 1979): more frequent words are understood and produced faster than less frequent words. However, Pickering & Garrod (2004c) argue that local context is so central that the frequency of an expression should become far less important. They also argue that frequency is replaced by accessibility of the expression within the dialogue context, and therefore, predicting that frequency effects will be dramatically reduced in dialogue. From the IAM, Pickering & Garrod (2004c) predict that children will tend to repeat a phrase that is novel to them to a greater extent when repeating lexical items.

### **2.2.4 Alignment in L2 learning**

Costa et al. (2008) consider the IAM in the context of L2 learning, and how the fluency of the L2 speakers may impact alignment. Costa et al. (2008) sketch a range of experimental predictions about how alignment may take place within L2 dialogue. Alignment is particularly of interest within L2 because the language itself is the learn-

ing outcome, and the repetition and backchanneling<sup>9</sup> along with shared context of the dialogue impact learner language acquisition. Costa et al. (2008) discuss the various factors which may affect the success of communication in L2 dialogue, and derive some hypotheses for future studies. In a L2 learning setting, a learner will have a more limited scope for alignment, and their proficiency will dictate to what extent they are capable of aligning lexically, syntactically and semantically (Pickering & Garrod 2006). Alignment within L2 learner dialogue will differ from alignment in fluent dialogues due to the different constraints mentioned above (Costa et al. 2008). This is both because of the difficulty of the task leading to a greater need for alignment (Pickering & Garrod 2006) and that aligning with their interlocutor allows learners to bootstrap their knowledge from the more competent linguistic example being given to them (Robinson 2011). L2 learners' lexical complexity have been found to increase in a dialogue setting due to the shared context words within that dialogue, compared to the level at which they are capable of expressing themselves in monologue (Robinson 2011), more so when their interlocutor is of a higher level of proficiency than they are (Khodamoradi et al. 2013).

Even once a situational alignment is reached (i.e. the learner understands the context of their interlocutor's interaction with them), there remains the question of the learner's *receptive* vs. *productive* vocabulary knowledge (words they understand when others use them vs. words they can use themselves), both of which are active in L2 dialogues (Takač 2008) and constrain their scope for alignment. Learner alignment therefore will also be influenced by the tutor's strategy; or by how much of the learner's receptive language the tutor produces which facilitates the learner productive ability in this context. L2 learners have been shown to learn vocabulary through taking part in dialogue (Hawkes 2012), suggesting this process of alignment and repetition of their interlocutor's speech produces learning gains. It has been hypothesised that learners may leverage alignment to improve achievement of pedagogic goals (Michel 2011). In an L2 setting, learners have been shown to imitate tutors as part of their learning process (Holley & King 1971). In educational settings, alignment has been found to predict both student learning and engagement (Ward & Litman 2007a), and in the case of user interaction with a virtual agent, increases perceived interaction naturalness and maintain user engagement (Yu et al. 2016).

---

<sup>9</sup>In linguistics, a backchannel during a conversation occurs when one participant is speaking and another participant interjects responses to the speaker (Yngve 1970). Examples include such expressions as "yeah", "uh-huh", "hmm", and "right"



### 2.2.5 Measuring Alignment

While *lexical* alignment consists of speakers beginning to use the same words (Ward & Litman 2007b, Sinclair et al. 2018) or phrases (Duplessis et al. 2017) as each other, *syntactic* alignment consists of the same parts of speech patterns, such as similar noun-phrase constructions (e.g. “*the basketball game*”), or similar adjuncts (e.g. “*in the morning*”) (Reitter et al. 2006, Reitter & Moore 2014) as the conversation progresses. Finally, *semantic* alignment can range from adaptation to individual differences in personality (Isard et al. 2006) to convergence at a higher level of representation such as DAs (Sinclair, Ferreira, Lopez, Lucas & Gasevic 2019) or in the alignment of terminology when speakers in a task based dialogue converge on a semantic representation of the problem (Reitter & Moore 2014, Mills & Healey 2008).

Methods for measuring alignment can range from simple count statistics (Duplessis et al. 2017) to linear regression on prime target distance<sup>10</sup> (Ward & Litman 2007a) or using a generalised linear mixed model to take into account the random effects present in dialogue (Reitter & Moore 2014) for a similar sliding window of prime and target occurrence. Alignment measures have potential to augment existing measures of linguistic sophistication prediction (Vajjala & Meurers 2016) to better deal with individual speakers within a dialogue, using alignment as a predictor of learner ability as has been suggested by Ward & Litman (2007a). Dialogue is inherently sparse, particularly when considering the lexical contribution of a single speaker. Accordingly, alignment could be a useful predictor of learner receptive and productive knowledge when in combination with lexical complexity of the shared vocabulary.

### 2.2.6 Dialogue Acts

Dialogue Acts (DAs) describe the meaning of an utterance at a certain level of granularity depending on the coding scheme used (Stolcke et al. 2000). DAs are often used to infer discourse structure, and are an important aspect in the automatic understanding of spontaneous dialogue (Stolcke et al. 2000). DAs are similar to *Speech Acts* (Searle & Searle 1969), but are often more specific, and are commonly used in natural language processing settings for the annotation of single utterances (Sinclair et al. 2017). DA sequences have been used to examine differences in tutor adaptation to introverted vs. extroverted students, as measured by the big-five personality metric (Vail & Boyer

---

<sup>10</sup>The item being aligned to in this context is known as the *prime*, and the subsequent usage of this prime by the other speaker is known as the *target*, or sign of alignment

2014). DAs have been used to try to detect tutoring strategy in human tutoring dialogues such as in Rus et al. (2017), Chen et al. (2011), Boyer et al. (2011) and are very useful in better automatic understanding of these strategies at an abstract level.

The 42 dialogue act labels. DA frequencies are given as percentages of the total number of utterances in the overall corpus.

Tag	Example	%
STATEMENT	<i>Me, I'm in the legal department.</i>	36%
BACKCHANNEL/ ACKNOWLEDGE	<i>Uh-huh.</i>	19%
OPINION	<i>I think it's great</i>	13%
ABANDONED/ UNINTERPRETABLE	<i>So, -/</i>	6%
AGREEMENT/ ACCEPT	<i>That's exactly it.</i>	5%
APPRECIATION	<i>I can imagine</i>	2%
YES-NO-QUESTION	<i>Do you have to have any special training?</i>	2%
NON-VERBAL	<i>&lt;Laughter&gt;, &lt;Throat_clearing&gt;</i>	2%
YES ANSWERS	<i>Yes.</i>	1%
CONVENTIONAL-CLOSING	<i>Well, it's been nice talking to you.</i>	1%
WH-QUESTION	<i>What did you wear to work today?</i>	1%
NO ANSWERS	<i>No.</i>	1%
RESPONSE ACKNOWLEDGMENT	<i>Oh, okay.</i>	1%
HEDGE	<i>I don't know if I'm making any sense or not.</i>	1%
DECLARATIVE YES-NO-QUESTION	<i>So you can afford to get a house?</i>	1%
OTHER	<i>Well give me a break, you know.</i>	1%
BACKCHANNEL-QUESTION	<i>Is that right?</i>	1%
QUOTATION	<i>You can't be pregnant and have cats</i>	.5%
SUMMARIZE/ REFORMULATE	<i>Oh, you mean you switched schools for the kids.</i>	.5%
AFFIRMATIVE NON-YES ANSWERS	<i>It is.</i>	.4%
ACTION-DIRECTIVE	<i>Why don't you go first</i>	.4%
COLLABORATIVE COMPLETION	<i>Who aren't contributing.</i>	.4%
REPEAT-PHRASE	<i>Oh, fajitas</i>	.3%
OPEN-QUESTION	<i>How about you?</i>	.3%
RHETORICAL-QUESTIONS	<i>Who would steal a newspaper?</i>	.2%
HOLD BEFORE ANSWER/ AGREEMENT	<i>I'm drawing a blank.</i>	.3%
REJECT	<i>Well, no</i>	.2%
NEGATIVE NON-NO ANSWERS	<i>Uh, not a whole lot.</i>	.1%
SIGNAL-NON-UNDERSTANDING	<i>Excuse me?</i>	.1%
OTHER ANSWERS	<i>I don't know</i>	.1%
CONVENTIONAL-OPENING	<i>How are you?</i>	.1%
OR-CLAUSE	<i>or is it more of a company?</i>	.1%
DISPREFERRED ANSWERS	<i>Well, not so much that.</i>	.1%
3RD-PARTY-TALK	<i>My goodness, Diane, get down from there.</i>	.1%
OFFERS, OPTIONS & COMMITS	<i>I'll have to check that out</i>	.1%
SELF-TALK	<i>What's the word I'm looking for</i>	.1%
DOWNPLAYER	<i>That's all right.</i>	.1%
MAYBE/ ACCEPT-PART	<i>Something like that</i>	<.1%
TAG-QUESTION	<i>Right?</i>	<.1%
DECLARATIVE WH-QUESTION	<i>You are what kind of buff?</i>	<.1%
APOLOGY	<i>I'm sorry.</i>	<.1%
THANKING	<i>Hey thanks a lot</i>	<.1%

Figure 2.1: Switchboard Dialogue Acts from the DAMSL (Core & Allen 1997) coding scheme.

There are many different types of DA coding scheme, and this is often dependent on the type of corpus being analysed and the task (Walker & Passonneau 2001, Bunt et al. 2017, Chen et al. 2011, Anderson et al. 1991). The DAMSL coding scheme<sup>11</sup>

<sup>11</sup>An example of which applied to the Switchboard corpus (Godfrey et al. 1992) can be seen in Figure 2.1

Dialog Act	Description
instruction	Commands partner to carry out action
explanation	States information that partner did not elicit
align	Checks partner attention & agreement, or readiness for next DA
check	Requests partner to confirm information that checker is partially sure of
query-yn	Yes/No question other than a check or align
query-w	Any other question
acknowledge	Minimal verbal response showing that speaker has heard the preceding DA
clarify	repetition of information already stated by speaker, often in response to a check DA
reply-y	affirmative reply to any query
reply-n	negative reply to any query
reply-w	any other reply
ready	DA that occurs after end of a dialogue game and prepares conversation for a new game

Table 2.3: Dialogue Acts used to annotate the Map-Task Corpus.

was developed to provide a general purpose coding scheme, which works for multiple corpora and downstream uses and which allows for a consistent baseline set of interactions to be captured (Core & Allen 1997), although the level of granularity of the DAs will not capture all the nuances present in very specific corpora (Bunt 2006). An example of DAs which are more corpus-specific can be seen in Table 2.3, for a task-based dialogue corpus, where some of the granularity of the DAMSL tag-set is unnecessary. In Chapter 3 we choose a subset of the DAMSL tagset which are most relevant to the L2 dialogue corpora used.

## 2.3 Corpora

In this thesis, we compare aspects of adaptation present within three *Human-Human* dialogue corpora: L2 student, conversational and task based, and one of *Human-Agent*. Table 2.4 presents some basic details of the four corpora. To understand the behaviour of tutors and learners in L2 dialogue, we need a corpus of such dialogues. To understand which behaviours are specific to L2 learning rather than common to two person dialogue, we also need to compare them to other types of dialogue between native speakers. In order to minimise the differences between the L2 and fluent dialogue as

well as to investigate interactions in as naturalistic a setting as possible, the Human-Human corpora chosen consisted of transcripts from spoken dialogue. We choose a corpus of *spontaneous conversation* which allows us to compare the dialogue to the goal of the L2 practice: that the learner is able to achieve a natural interaction, with more equal participation between interlocutors. We also choose a *task based* dialogue corpus which allows us to explore whether the effects of the task of teaching, and the hierarchical nature of the tutor student relationship has more influence over the conversational dynamics than the conversational goal.

	<i>L2 learner</i>		<i>Conversational</i>	<i>Task-Based</i>
	<b>Tutorbot</b>	<b>BELC</b>	<b>Switchboard</b>	<b>Map-Task</b>
number of dialogues	3689	118	1155	128
average Num. utterances	20.41	130.69	193.60	207.98
average Num. tokens	128.99	634.28	1239.59	1193.00
average tokens/utterance	6.32	4.85	6.40	5.74
communication medium	typed	spoken	spoken*	spoken
speakers	H-A	H-H	H-H	H-H
student L1	German	Spanish	–	–

Table 2.4: Corpora detail comparison. H-H: Human-Human, H-A: Human-Agent, Tutorbot is introduced in Sinclair, McCurdy, Lopez, Lucas & Gasevic (2019), Switchboard in Godfrey et al. (1992) and Map-Task in Anderson et al. (1991). \*telephone mediated

We use the understanding of how speakers in L2 dialogue interact to analyse how students respond in the context of a Human-Agent dialogue, where the tutor is an automated L2 ‘tutorbot’. Chapter 6 explores aspects of alignment present within the fourth corpus we analyse: computer mediated dialogues between L2 students and a chatbot in the context of an online learning platform for ESOL. The following subsections provide more details of the corpora outlined in Table 2.4.

### 2.3.1 L2 Learner

The dataset used is the Barcelona English Language Corpus (BELC) (Muñoz 2006).

BELC consists of 118 transcripts from conversational practice between students of English as a foreign language, and tutors. These vary from 60 to 140 utterances in length. The tutors’ instructions for the dialogue were to elicit as much conversation from the learner as possible, and to set them at ease while having as natural a conver-

sation as possible. The tutors follow a similar script of questions with each participant resulting in the dialogues covering similar topics. The corpus was gathered at four different timestamps over a period of three years, with the students involved receiving approximately one school year of weekly English tuition between sessions. Thus, the corpus can be divided into four general levels of student ability. The corpus has been annotated at an utterance level with a set of DAs (Sinclair et al. 2017), which were chosen from Stolcke et al. (2000) for their relevance to the corpus. The choice of DAs and the annotation procedure is described in Chapter 3. Examples of the lowest and highest level of learner ability, alongside the utterance-level DA annotations can be seen in Table 1.

Resources consisting of two person conversational dialogue practice between an L2 learner and tutor are scarce, and gathering a new dataset with a range of student first languages is outwith the scope of this thesis. BELC consists solely of L2 speakers with an L1 of Spanish/Catalan. To make our contributions as general as possible given corpus constraints, we limit our analysis to the adaptation between student and tutor and the differences between students at different levels. We also base our hypotheses on general theories of communication and L2 acquisition which are applicable to all language. We discuss how future studies can take into consideration differences in the L1 of the students in Section 7.3 where we also offer a discussion of potential effects of different L1-L2 pairings

### 2.3.2 Conversational

In order to contextualise our analysis of the L2 corpus, we compare the results with those from conversational dialogue between fluent English speakers, the target L2 of this thesis. The desired fluent conversation should have a symmetry of speaker interaction which is the goal of the L2 practice dialogue: that both speakers contribute more or less equally to the dialogue, and do not differ greatly in status. Corpora for creating conversational dialogue agents often make use of the large resource of movie subtitles such as Open Subtitles (Lison & Tiedemann 2016). Using film dialogue has the disadvantage of being *scripted* rather than naturalistic, where instead of the communication being spontaneous with the goal of achieving understanding between interlocutors, it is crafted to communicate most strongly with the viewer: a passive third dialogue participant. Therefore while Open Subtitles is a useful resource for dialogue examples, the smaller Switchboard corpus (Godfrey et al. 1992) is more suitable for our purposes.

Table 2.5: Switchboard dialogue example extract

DA: dialogue act, S: speaker, *resp\_ack*: response-acknowledgement, *stmt*: statement, *bkchnl*: backchannel, *yes\_A*: yes-answer, *gen\_Q*: general-other-question. [...] indicates truncated utterance

DA	S	Utterance
resp_ack	B	Okay .
stmt	A	Great . Um , currently , I 'm not doing a whole lot of exercise in any type of program .
bkchnl	B	Huh-uh .
stmt	A	I 'm mainly do a lot of walking . I have a son [...] be dedicated towards the ,
yes_A	B	Yeah .
stmt	A	exercise area , is covered in boxes .
gen_Q	B	Um , what did you do when you did exercise regularly ?
stmt	A	Well , I had , uh , a little routine that I did for warm ups .
bkchnl	B	Huh-uh .
stmt	A	And then I did some very mild [...] not trying to make big bulging muscles ,
bkchnl	B	Huh-uh .
stmt	A	just trying to try and stay as firm as I can stay in my old age .
stmt	B	Yeah . Um , right now , um , I try when it 's nice out [...] that 's pretty popular in Texas , I do n't know ,
bkchnl	A	Huh-uh .
stmt	B	if it 's up north , but every weekend [...] that 's a lot of fun .
bkchnl	A	Huh-uh .

Switchboard's consistent equal speaker roles, comparative lack of drama and diverse balanced range of *conversational* topics makes it a useful 'gold standard' with which to compare L2 English learners of varying ability.

The Switchboard corpus (Godfrey et al. 1992) is a large corpus collected from telephone conversations between English 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. The conversations range in length from one and a half to ten minutes, averaging six and a half minutes. The resulting transcripts are broken into speaker turns and an example of the type of dialogue can be seen in Table 2.5. From Table 2.5 it can be seen that the predominant DAs used are statements and backchannel, which is different to the questions

answer/statement dynamic within BELC. Switchboard provides an example of what naturalistic dialogue could look like for high ability L2 English speakers and therefore is a useful counterpoint to our comparisons of student and tutor interactions at different levels of student ability.

### 2.3.3 Task based

A striking aspect of the L2 dialogues between the lowest ability students and tutors is the asymmetry between the speakers. This is due both to the huge difference in fluency, and in the fact that the dialogue goals of the speakers are so very different. At this level, the tutor's role with respect to the student resembles very much a *task*, that of tutoring, or providing information in the form of easy to understand dialogue. We choose to compare this asymmetry in the L2 dialogues with asymmetry in fluent dialogue, and choose the Map-Task corpus (Anderson et al. 1991) for its properties as a particularly asymmetric example of a task based dialogue. While other task based dialogue corpora exist, they often focus on either very closed domains such as restaurant reservation, or very domain specific, such as the very technical Ubuntu dialogue corpus (Lowe et al. 2015). They are also typically written dialogue rather than spoken, gathered either from extracting chat logs (Ubuntu, NPS chat (Forsythand & Martell 2007)), or from Mechanical Turk ( MultiWoz (Budzianowski et al. 2018), Persona Chat Zhang et al. (2018)). While more contemporary corpora have the advantage of scale, the consistent asymmetry of knowledge between speakers in the Map Task corpus is more akin to the difference in language knowledge we wish to compare the dialogues to in the L2 corpora.

The Map-Task corpus (Anderson et al. 1991) 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.

### 2.3.4 Student-Agent

We are interested in the comparison between student alignment in human-human (H-H) and human-agent (H-A) dialogues. The H-A corpus analyzed in this study com-

Table 2.6: Map-Task dialogue example

DA	speaker	utterance
statement	giver	just slightly below it
decl_YNQ	follower	on the left-hand side?
yes_answers	giver	mmhmm
statement	follower	okay
decl_YNQ	follower	so i'm going underneath it?
statement	giver	above it
resp_ack	follower	right
yes_no_Q	giver	and then you'll be underneath the waterfall?
yes_answers	follower	that's right
statement	giver	and go up the left-hand side of the waterfall in a straight line
resp_ack	follower	mmhmm
statement	giver	and then turn to your right
resp_ack	follower	mmhmm
statement	giver	and go for about an inch
statement	giver	and then turn upwards again
yes_no_Q	follower	have you got public footpath ?
no_answers	giver	no
resp_ack	follower	oh right okay

prises dialogues drawn from a large-scale commercial platform for L2 learners<sup>12</sup>. 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 (Bibauw et al. 2019); 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 2.7

Key differences are shown in Table 2.4. However, it should also be noted that the Tutorbot corpus only consists of single utterance turns, whereas BELC has multiple. The single utterance turns in Tutorbot are a result of the design of the chatbot, and may constrain the verbosity of the learner. The topics are also more diverse in BELC, as the Tutorbot explicitly guided learners to review practised material rather than engage

<sup>12</sup>This data was kindly shared with us by Babbel, <https://www.babbel.com/>



Table 2.7: Tutorbot dialogue example. Italics indicate Expression Repetition

1.	bot:	What is your <i>favorite day of the week</i> ?
2.	user:	My <i>favorite day of the week</i> is Friday ...
3.	bot:	Do you play sports ?
4.	user:	yes
5.	bot:	What sport do you <i>play</i> ?
6.	user:	I <i>play</i> 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

in open-ended discussion. Nevertheless, certain main topics (*how are you, where are you from, tell me about your family, hobbies, and what time do you do that*) and the beginner/lower-intermediate range of learner ability are common to both, facilitating automated alignment comparison.

## Chapter 3

# Linguistic Complexity Adaptation

Interlocutors have the tendency to align in fluent conversation, creating a symmetry between speakers which learners of a language may be unable at first to achieve. Our hypothesis was that as a learner's competence increases, symmetry between learner and tutor language increases. We compare characteristics of student and tutor dialogue to speakers in fluent conversational and task-oriented corpora, finding more linguistic complexity adaptation within L2 dialogues.

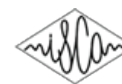
This chapter uses existing measures of readability to analyse L2 corpora in terms of the linguistic sophistication of the language used by each interlocutor. Section 3.2.1 outlines the measures used to compare the linguistic complexity of the dialogue, to allow us to make the distance comparisons between student and tutor complexity within different dialogues. We measure speaker linguistic complexity convergence within full dialogues and for individual Dialogue Acts. Section 3.2.3 is additional to the results reported in the paper, describing how we performed a human evaluation of the automatic DA labels used in our work. The data supports our hypothesis at the level of individual DA usage and complexity for statements and questions. We found at the level of the full dialogue that tutors adapt to learner ability: in the case of low ability students, we see evidence of tutor convergence when student and tutor exhibit very different levels of linguistic complexity. In the case of higher ability students, tutors diverge: tutor and student begin the dialogue with a more similar linguistic complexity and the tutor increases their linguistic complexity in the latter half of the dialogue. We interpret this as evidence of tutor adherence to the Zone of Proximal Development.

### 3.1 Finding the Zone of Proximal Development: Student-Tutor L2 Dialogue Interactions

This section includes the verbatim copy of the following publication:

Sinclair, A., Oberlander, J. and Gasevic, D., 2017. Finding the Zone of Proximal Development: Student-Tutor Second Language Dialogue Interactions. *Proceedings of SEMDIAL*, pp.107-115.

**Contributions:** The ideas and analysis in the paper were developed and discussed between all authors of the work. The original idea, the experiments and the bulk of the writing were the work of the first author.



## 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,<sup>1</sup> 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

<sup>1</sup>Here we use second language (L2) in the broad sense, to include any language additional to the speaker's native language.

### 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: uhuh good.  
INV: and what time will you finish?  
PAR: hmm three.  
INV: at three uhuh.

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

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.<sup>2</sup>

<sup>2</sup>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<sup>3</sup> 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)

<sup>3</sup>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).

### 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/fluient	128
Switchboard (SB)	conversational	native/fluient	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

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

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 as-

Rule based	Map Task	Switchboard
<i>yes-no-question</i>	query-yn	yes-no-Question
<i>declarative yes-no-question</i>	check	declarative yes no question
<i>backchannel-question</i>	–	backchannel question tag question
<i>wh-question</i>	–	wh-question
<i>general-other-question</i>	query-w (other q)	open question rhetorical question declarative wh question or-clause (or question)
<i>yes answers</i>	reply-y	yes answer
<i>no answers</i>	reply-n	no answer reject
<i>signal-non-understanding</i>	–	signal non understanding
<i>backchannel-acknowledge</i>	–	backchannel ack
<i>response acknowledge-ment</i>	acknowledge	response ack
<i>repeat-phrase</i>	–	repeat phrase
<i>statement</i>	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.

sign 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

the symmetry of Switchboard, asymmetry of Map-Task 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

Bigram	BELC				MT	SB
<b>Speaker</b>	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
<b>Statement</b>	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
<b>Question</b>	L1	L2	L3	L4		
TT/AA/GG	<b>0.573</b>	<b>0.289</b>	<b>0.190</b>	<b>0.138</b>	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.

their *statements* increase slightly along with *WH-questions*, and *signal\_non\_understanding*.

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’



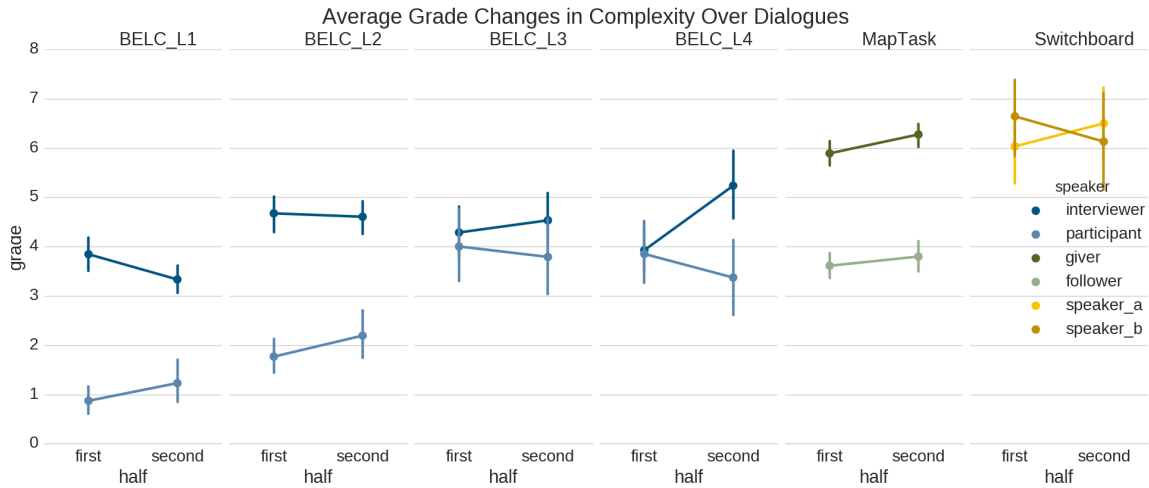


Figure 2: Average Student tutor complexities for first and second halves of dialogues by level. In the BELC results, the convergence and divergence of the *tutor's* complexity grade in relation to the student's in level (L)1 and 4 is significant ( $t = 6.25, p = 1.60e-08, t = -4.18e+00, p = 2.95e-04$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

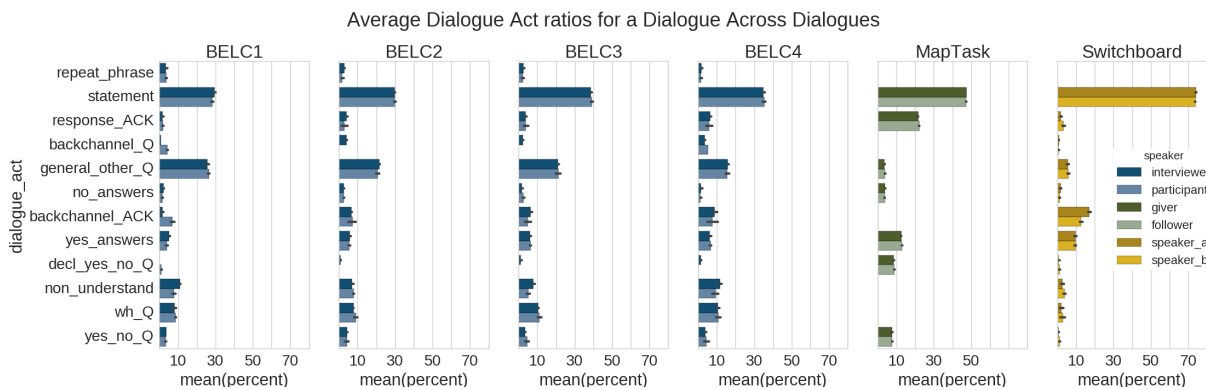


Figure 3: Dialogue Act percentages by corpus for the average dialogue.

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 struc-

ture 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

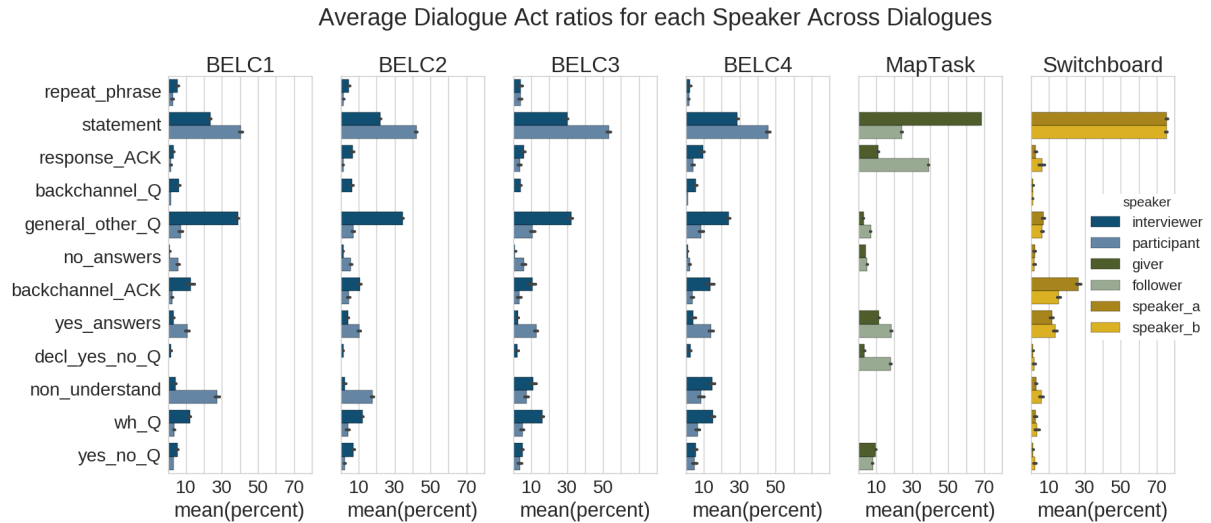


Figure 4: Average Dialogue Act percentages per dialogue by corpus: for an individual speaker’s average share of the dialogue.

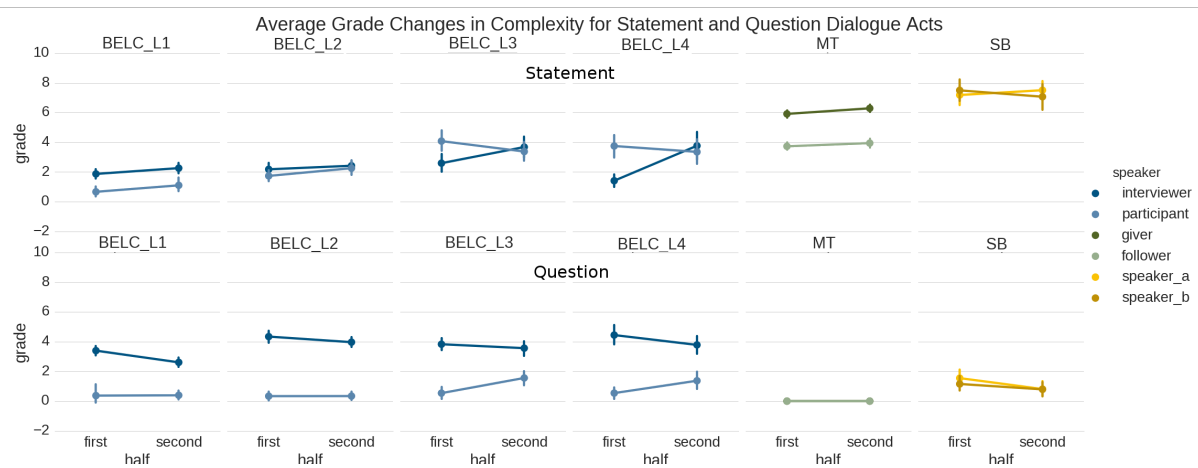


Figure 5: Complexities within Statement and Question dialogue acts in the three corpora. For the Statements (upper row), the interviewer’s statements between the first and second half of Levels 3 and 4 significantly ( $t = -2.28, p = 2.72e-02, t = -4.18, p = 2.95e-04$ ) increase in grade complexity. 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.

## 3.2 Further Discussion

Sinclair et al. (2017) presents an analysis of BELC in terms of predicted linguistic complexity of the dialogue both as a whole, and at the level of specific dialogue acts. Since neither the development of new complexity measures nor DA annotation techniques was the goal of this research, discussion of the techniques employed to achieve this analysis were limited. The following subsections justify the choice of training data, development and evaluation of the model used, describe in detail the annotation procedure, evaluation and accuracy of the DA labels.

### 3.2.1 Automatic Assessment

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

In the field of automated assessment, linguistic complexity is an important feature when training models to assign a grade to either essays or other student work via supervised methods. Since our corpus is small, using similar methods would not make sense. Our data does not have fine grained grading such as CEFR labels for training and our research goal is not grading the learners. Rather, we aim to select features which are common to both discourse and dialogue, using a model trained on a larger, more finely graded dataset of written English with varying levels of ‘complexity’, and use the predictions of this model on the L2 dialogues in order to compare the contributions of the interlocutors according to the same criteria.

Existing measures of text complexity were developed to predict the readability of discourse, but also have been applied to dialogue in the form of subtitles from television shows of varying age of audience (Vajjala & Meurers 2014a). The metrics suc-

cessfully differentiated 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 linear regression model as a way to ‘grade’ the transcribed dialogue text in order to compare the complexities of language used between the corpora.

### **Learner Fluency**

In order to better understand SLL dialogue, and to measure fluency as a dialogue takes place, a model should predict the complexity of the utterances with reasonable accuracy. The measurement of dialogue complexity for language learners should consist of a general assessment of correctness coupled with a model of vocabulary use and the potential knowledge it implies. A third aspect specific to fluency assessment in dialogue is the level of linguistic competence shown in a learner’s response.

Measuring the complexity of dialogue utterances is similar to measuring the readability of discourse, though with emphasis on features which are suitable for small quantities of text. Bootstrapping from labelled discourse using common features is a good first step, where the semi-supervised learning of dialogue grades can be achieved.

The nature of second language learner dialogue means that the resulting utterances are likely to be informal, not well formed, either grammatically, or in terms of structure, and there are likely to be errors. Since the goal of this complexity measurement is not assessment, but rather comparing the linguistic complexity of student and tutor utterances, the dialogue text should be cleaned to an extent, such that the fluency is not completely determined by the errors. The fluency or complexity of an utterance will be measured by the complexity of the language it exhibits.

Any student utterance in an L2 setting indicates the knowledge of how to construct it, irrespective of its relation to the question. However two utterances of the same surface level complexity demonstrate very different levels of fluency if one expresses misunderstanding and the other a coherent grasp of the preceding dialogue. Another factor which can influence the demonstrated competence of the student is the direction of the tutor within the dialogue: that is, since the tutor ‘leads’ the dialogue, if they do not provide the student with the conversational opportunity to communicate at their best, whether in terms of topic choice, or in terms of social dynamics, the student can only partially demonstrate their knowledge.

Part of SLL dialogue consists of the task of listening and comprehension. Some parts of a more advanced learner dialogue may contain periods where the learner has more of a listening role in the dialogue, where their responses are brief, thus on surface

Table 3.1: Linguistic Complexity Features for Text in Dialogue

Age Of Acquisition	<i>average AoA over utterance</i>
Word Frequency	<i>average word frequency counts over utterance drawn from googleNews corpus or similar</i>
Context Complexity	<i>Average complexity of the context window of a particular word</i>
Lexical	<i>can be measured in type/token ratio, or counting the number of distinct words for a length of text</i>
Number of Senses	<i>WordNet synsets count for words</i>
POS tag depth	<i>when applicable; what sorts of constructions are used and how deep parse tree is</i>

level complexity display low fluency, yet if these are indications of understanding, communicate a high listening and comprehension fluency. Modelling the language of a students interlocutor is therefore equally important when discerning their fluency.

Oral tutoring allows a lower pressure environment where tutor input can actively scaffold knowledge and contribute to learner fluency, giving the learner the potential to achieve a greater level of fluency by the end of the interaction.

To achieve a general *model of language knowledge* which can be used to model the progression of complexity of language in a dialogue at varying levels of linguistic competence, the mining of dialogues and bootstrapping from labelled graded discourse is necessary as a first step. Features learnt will initially be those typically used in the fields of automatic readability analysis and text simplification systems, with an emphasis on lexical complexity.

### 3.2.2 Complexity Prediction

#### Approach

Language fluency and complexity can be assessed over many dimensions, and changes depending on the context of that language, be that discourse of different types, or dialogue. While dialogue is known to be more simple in terms of lexical complexity than discourse (Robinson 2011), large-scale corpus resources for second language *dialogue* are less readily available than those for discourse. Since BELC is neither particularly large, nor finely graded, an unsupervised approach to assign ‘grades’ to these dialogues

seemed appropriate under the assumption that a regression model trained on discourse can be used to compare linguistic complexity in dialogue. Our approach therefore was to train a readability classification model on a large corpus of graded discourse using features which are common to both styles of communication, and to only use the complexity predictions of this model to *compare* rather than *grade* speakers' contributions.

Table 3.2: Newsela dataset example sentences written at multiple levels of text complexity. the bold font highlights the parts of the sentence that are different from the adjacent version. The grades correspond to the whole article in which the sentence occurs, so the same sentences may receive different scores such as the sentences for level 6 and 7 here.

Level	Text
12	<b>Slightly more</b> fourth-graders <b>nationwide are reading proficiently compared with</b> a decade ago, <b>but</b> only a third of them <b>are now reading well</b> , according to a new report.
7	Fourth-graders <b>in most states are better readers than they were a decade ago. But</b> only a third of them <b>actually are able to read well</b> , according to a new report.
6	Fourth-graders <b>in most states</b> are better readers than they were a decade ago. But <b>only a third of them</b> actually <b>are able to read well, according to a new report.</b>
4	<b>Most</b> fourth-graders are better readers than <b>they were</b> 10 years ago. But <b>few of them can actually</b> read well.
3	Fourth-graders are better readers than 10 years ago. But few of them read well.

To train our model, we use the graded hand-simplified collection of simple discursive articles provided in the Newsela corpus (Xu et al. 2015). The Newsela dataset consists of text from news articles professionally simplified 5 times, then labelled with reading level grades applicable to the American school system; examples of the grades and sentences from the corpus can be seen in Table 3.2. This corpus was developed as a resource for sentence simplification however we use it as training data for a readability ranking model such as in Vajjala & Meurers (2014b) and Vajjala & Meurers (2016). Newsela is a valuable resource due to its size and variety of articles, which have been human graded and simplified. We chose this corpus for two main reasons. Firstly the corpus is written for learners (not by learners) at a known level of competence, which is ideal for our work as the dialogue has been transcribed and therefore



the language is in ‘correct’ English, at least at the level of spelling. 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. The data is not balanced according to grade label, with some grades containing very few example documents as is shown in Figure 3.1. This imbalance is due to the corpus creation method: a set of complex articles (grade 12) were used as source material for the writing of 5 successively more simple versions. The resulting article set were then graded. This explains to some degree the very few documents graded 10 and 11, and the large number of documents graded 12 (the highest grade) suggesting that the subsequent simplifications typically resulted in a version two or more grades easier than the source document. However, since the goal of using this data is not to use the predicted grades themselves, rather to use them as a method of ranking and comparing graded texts, any skewed prediction effects will not affect our results.

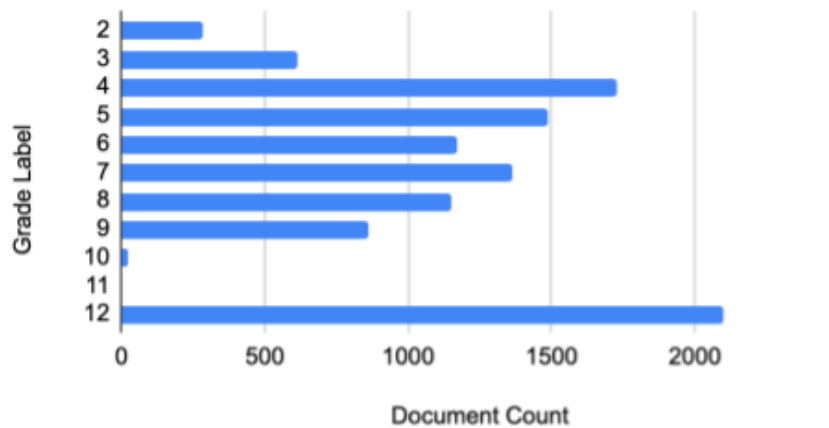


Figure 3.1: Distribution of document labels in the Newsela corpus.

### Feature Selection

The features used in this initial model were features chosen from work on readability ranking (Vajjala & Meurers (2014b)) which fall under the categories of lexical and psycholinguistic features. Since the goal was exploration of SLL dialogue and testing using discourse as training data, feature choice was driven by their relevance to dialogue text, and the general nature of the feature. As such, syntactic features were avoided at this stage due to the dissimilarity between dialogue and discourse in these respects. The main feature types used by Vajjala & Meurers (2016) to measure readability are described below:

**Lexical** Lexically complex words are those for which a simpler synonym exists. Lexical diversity and or lexical density are measured by type-token (TTR) and part-of-speech (POS) ratios. e.e. Higher TTR is an indication of lexical diversity, and more even POS ratios of lexical density.

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

**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 following Farr et al. (1951)

<b>Feature Selection - Ablation</b>				
i	Feature	F1	P	R
0	Baseline - TFIDF	7.45%	9.11%	6.67%
1	Average Word/Sent. lengths	33.89%	33.58%	34.22%
2	FKGL scores	39.15%	38.04%	40.44%
3	Average Age of Acquisition	35.98%	36.89%	35.15%
4	Lexical diversity	42.18%	40.50%	44.44%
5	Average Occurrence Freqs	63.67%	63.27%	64.44%
6	Smog index	61.63%	61.31%	62.22%
7	Dale Chall score	66.29%	66.04%	66.67%
8	Linsear Write Formula	70.16%	70.11%	70.22%
9	Automatic Readability Index	74.69%	74.73%	74.67%
10	Hard word-count	76.87%	76.85%	76.89%
11	Coleman-Liau score	79.08%	79.07%	79.11%
<b>12</b>	<b>All Features</b>	<b>79.08%</b>	<b>79.07%</b>	<b>79.11%</b>

Table 3.3:

Ablation table of selected model features trained with a logistic regression classifier. The results recorded are the average of results from cross validation. Age of acquisition was looked up for each word from the compilation of Kuperman et al. (2012).

An additional feature which was not present in the work of Vajjala & Meurers (2014b) but we included was *Average Occurrence Frequency (AOF)* of a word. AOF

has been shown to correlate highly with the age of acquisition of the same word and the more frequently that an L2 learner is exposed to a word will influence their rate of learning it (Brysbaert et al. 2000). This motivated our inclusion of AOA as a feature in the model. The word frequency was obtained from counts from the Google News corpus which is commonly used in training Word2Vec (Mikolov et al. 2013).

To explore the effects of different features, we train and test the model using One Stop English (OSE) corpus introduced by Vajjala & Meurers (2014b) who report 90% accuracy for document level ranking with a full set of features using an SVM classifier. The accuracy of the linear regression model at prediction when trained and tested on subsets of the OSE corpus is shown in Table 3.3. This model is a promising start, suggesting that reasonable results can be obtained with a simple regression model and partial features, without any model regularization or smoothing.

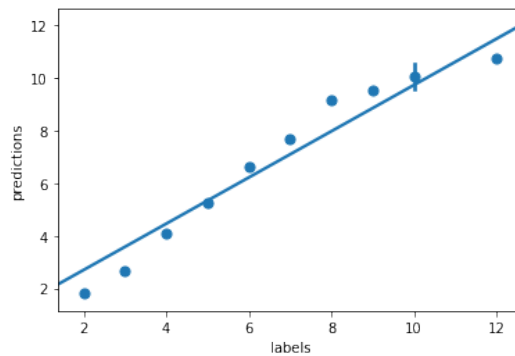


Figure 3.2: Results of the Linear regression model on Newsela dataset: strong correlation between predicted grade and true label ( $R^2 = 0.81$ )

### Model Evaluation

The final model was trained on the full Newsela corpus. In order to evaluate whether the features selected work well with this dataset, we performed K-fold cross validation on the dataset with  $k = 6$ . The results of the predictions of the linear regression model can be seen in Figure 3.2. The mean absolute error was 0.96, with an  $R^2$  value of 0.81. Since the purpose of the model is the exploratory analysis of the BELC corpus with the goal of *comparing* complexity between selected dialogue segments, we evaluate its performance at predicting the *relative* complexity of the Newsela documents. We therefore compare the ranking of the predicted grades with the observed grades in the data using Spearman's rank correlation coefficient. We find that the correlation coefficient is 0.94 with  $p < 0.0001$  suggesting that the model is suitable for comparing

document complexity successfully. While we think Spearman's coefficient is most relevant to the ordinal nature of the labels, we also report Pearson's R, which, with a coefficient of  $0.90$  and  $p < 0.0001$ , additionally indicates a strong linear relationship between labels and predictions.

### Effects of Document Size

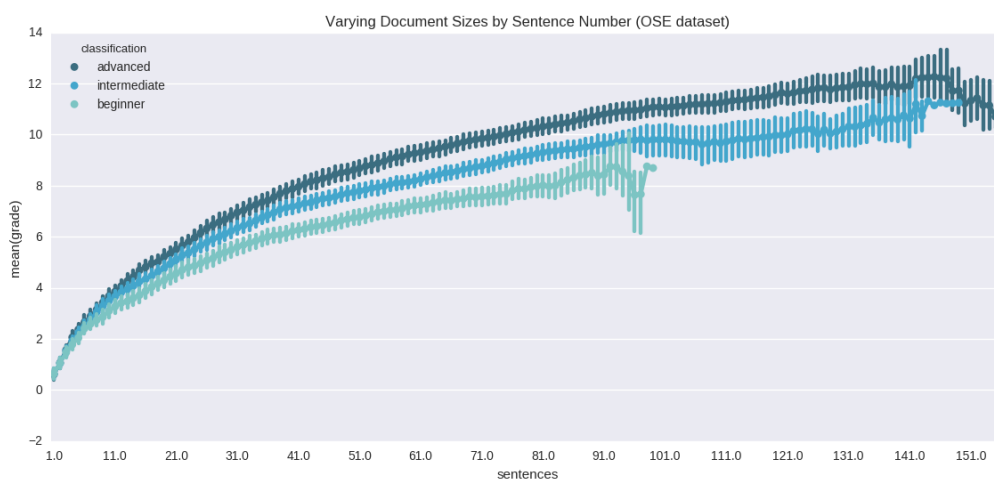


Figure 3.3: For each article, sub documents containing an increasing number of its sentences were given predicted grades. The classifier was trained on the Newsela dataset. As can be seen from the shape of the data, text size has a large impact on the model, although after a certain length of text, the model is very effective at correctly ranking the documents' complexity.

Sampling sentences of varying length and the discovery of how much is identifiable from what quantity of text is essential to deciding how much belief to place on any measure of dialogue complexity. Dialogue is inherently less formal than discourse, and the dynamic and conversational aspect means that there is less cognitive time to construct longer range dependencies and more complex clauses. Conversely, in dialogue, there is also the aspect of shared context, whereby referring expressions and omissions are common and refer to objects or concepts not explicitly expressed in the dialogue at all since doing so would be unnecessary to the understanding of both dialogue participants at a given time. All this being said, it is logical to assume that simply transferring assessment measures and complexity features from discourse analysis to that of dialogue should not be done without some checking of their application and relevance to dialogue.

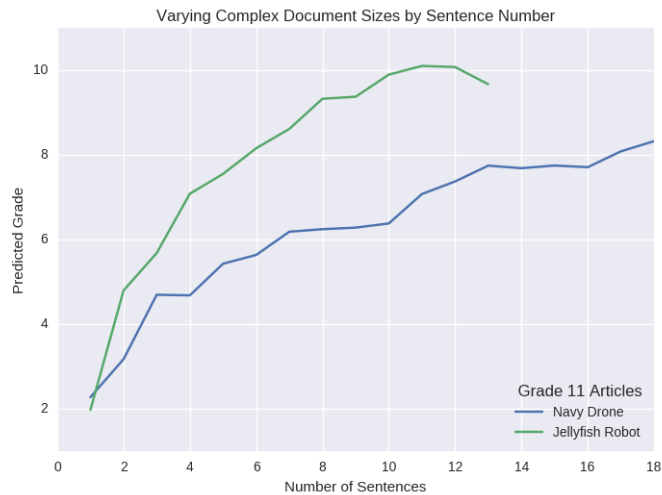


Figure 3.4: The same technique as in Figure 3.3 was applied, but with only two documents to examine the individual relationship between document length and grade

In order to explore current techniques in the context of dialogue, and assess their relevance to the classification of shorter texts, we compared the accuracy of models each trained with varying selections of features on the classification of sub-texts; excerpts of varying length from the corpus held back for evaluation, and with a separate graded dataset, OSE. The results of testing varying size texts can be seen in Figure 3.3 and 3.4

In Sinclair et al. (2017), we compare the “grades” of the utterances in the first and second halves of the dialogues to one another, making the same comparison for the sets of statement and question DAs. This is due to the fact that with only small quantities of text, our model will struggle to predict reliably accurate “grades”. We therefore compare only the linguistic complexity differences between the first and second halves of the dialogue, and for the two majority class DAs, where there will be enough text to ensure a more reliable prediction and thus comparison.

### 3.2.3 Dialogue Act Annotation

In Sinclair et al. (2017), we use regular expressions to automatically label the Dialogue Acts (DAs) used. Since BELC had no human-annotated labels, we tested the accuracy of our labelling on its ability to predict the DAs in Switchboard, and examined by hand the utterances in BELC for each DA label. To better validate our DA labels against how a human would label the same utterances, we gathered some gold-standard labels via

DA Accuracy				
DA	human labels	correct	percent	
no_answers	10	10	100.0	
statement	478	465	97.3	
signal_non_understanding	172	167	97.1	
wh_question	110	105	95.5	
general_other_question	224	209	93.3	
yes_answers	59	55	93.2	
yes_no_question	40	27	67.5	
backchannel_question	56	37	66.1	
backchannel_acknowledgement	80	48	60.0	
declarative_yes_no_question	5	2	40.0	
response_acknowledgement	108	37	34.3	
repeat_phrase	17	1	5.9	
Total	1359	1163	85.6%	

Table 3.4: DA labelling accuracy

human annotation. Two annotators were asked to independently label the utterances of 10% of the corpus with the set of 12 dialogue acts in table 3.4. The selected dialogues were equivalently balanced between each of the 4 student ability levels. The coders were given definitions of the DAs used, and were instructed to independently annotate the full validation set, which consists of 1359 utterances. The coders initially achieved 71.2% agreement, with a total of 392 disagreements on utterance labels. The coders were then asked to review the utterances upon which they disagreed and come to an agreement on what the final DA label should be. This final set of labels was then used as a gold standard set with which to test the accuracy of our automatic labelling method. The resulting accuracy of our labels when compared to the human labels was 85.6%.

When individual DA accuracy was inspected (shown in Table 3.4), over 90% accuracy was achieved for *no-answers*, *statements*, *signal-non-understanding*, *wh-questions*, *general-other-questions*, and *yes-answers*. Over 60% accuracy was achieved for *yes-no-questions*, *backchannel-questions*, and *backchannel-acknowledgements*. The three DAs which achieved under 50% accuracy were *response-acknowledgement*, *declarative-yes-no-question*, and *repeat-phrase*. These last three are clearly much harder to label with a regular-expression based approach. The poor accuracy of *repeat-phrase* can in

part be explained by the fact that the rule to label an utterance as a repeat only matches if it contains a sub-string within the *previous* utterance only. Some of the human labels for repeat phrase can be further back than 5 utterances previous to the utterance being considered. A secondary reason for this particularly poor accuracy is that if a speaker gives a yes-answer, or a backchannel, and their interlocutor repeats this as part of a response acknowledgement, or as an introduction to a statement (e.g. Student: “yes”, Tutor: “yes, ok that must have been interesting”) then this can be incorrectly labelled as a repeat phrase by our rules.

A full breakdown of incorrectly labelled utterances can be seen in Table 3.5. The first two columns show the automatically labelled DAs which were incorrect according to the human labels. The right three columns break down these incorrect labels, to show what they should have been according to our annotators. This can help to better understand the lower accuracy DAs in Table 3.4. For example, *response-acknowledgement* (respAck) has only one instance where it predicts the label and is incorrect, however, there are 55 instances where *response acknowledgement* is labelled as *statement*, showing that we fail to recognise many of the true instances of that label.

### 3.3 Contributions

The contributions of this chapter are threefold:

- We find evidence of and show that tutor’s adaptation of their linguistic complexity can be measured via a set of surface features in the text within the utterances. We see tutors converge to low ability learners and diverge from high ability learners in terms of the linguistic complexity of their language (Figure 2 in Sinclair et al. (2017))
- We show that utterances with different functions (*statements vs. questions*) exhibit different complexity traits, per speaker and between halves of the dialogue (Figure 5 in Sinclair et al. (2017)).
- We show that DA usage becomes more symmetric between student and tutor in L2 dialogues as student ability increases. We contrast this to the symmetric DA contributions between fluent native speakers in conversational, and to the asymmetry present in task-based dialogues (Figure 4 in Sinclair et al. (2017))

These findings have some implications for the design of an automatic L2 tutor: we see that human tutors adapt complexity to remain within a certain accessible range

DA Error Analysis					
Incorrectly		% total	Human		
Predicted Label	count	error	Label	count	% of error
decYNQ	3	1.5	genQ	3	100.0
genQ	52	26.1	backQ	18	34.6
			decYNQ	3	5.8
			repeat	9	17.3
			sigNA	3	5.8
			stmt	2	3.8
			whQ	5	9.6
			YNQ	12	23.1
repeat	21	10.6	genQ	6	28.6
			respAck	4	19.0
			stmt	11	52.4
respAck	1	0.5	backAck	1	100.0
sigNA	32	16.1	backAck	24	75.0
			respAck	8	25.0
stmt	76	38.2	backAck	7	9.2
			backQ	1	1.3
			genQ	2	2.6
			noA	2	2.6
			repeat	3	3.9
			respAck	55	72.4
			yesA	5	6.6
			YNQ	1	1.3
whQ	6	3.0	sigNA	2	33.3
			repeat	4	66.7
yesA	4	2.0	respAck	4	100.0
YNQ	4	2.0	genQ	4	100.0

Table 3.5: Label errors in the Dialogue Acts. Of the incorrect automatic labels, we present what the human labels were for each DA category.

of the learner’s ability. Since we can model this adaptation, the same features could be used as criteria to optimise dialogue generation. These features are lightweight and surface level; therefore, given a certain dialogue history, the generation of the next tutor utterance could be a function of learner level prediction, or to maximise the probability



that the learner will respond given previous similarly complex prompts in the past.

Modelling the differences in complexity within certain DAs suggests that any automated assessment of learner ability should incorporate interlocutor expressions as well as learner expressions. Likewise in an automated tutoring scenario the complexity adaptation shouldn't necessarily be uniform; but rather influenced by the type of dialogue move the automated tutor plans to make.

# Chapter 4

## Alignment

In order to have a successful conversation, second language learners must rely on both their *productive* and *receptive* knowledge of vocabulary (Takač 2008). Words that a learner can say in conversation show their *productive* vocabulary, and words they can understand from their tutor's language demonstrates their *receptive* vocabulary. A student's *productive* knowledge of language can be signal a much larger latent range of vocabulary at a *receptive* level, *alignment* of the student to the tutor could indicate language that a student only has partial productive knowledge of, depending on how much the tutor pushes the student or remains within the student's known vocabulary. This is discussed further in Section 4.2.1 due to space constraints in (Sinclair et al. 2018). L2 learners have greater production ability in dialogue than in monologue (Robinson 2011). Robinson (2011) reason this is due to recent examples from their interlocutor of vocabulary from the learner's receptive knowledge. With the priming effects of dialogue, if this vocabulary in a monologue setting is only just below a learner's productive knowledge, then in the context of conversational dialogue, they can leverage alignment to move from receptive to productive knowledge of the primed words.

In Chapter 3, we show that tutors adapt their language complexity to the student, and that over the course of a dialogue, speakers' contributions become more symmetric (Sinclair et al. 2017). In this chapter, we investigate alignment differences between speakers at different levels of learner ability and for different word complexities. We find that there is an effect of word frequency, which we use as an indication of word complexity<sup>1</sup>, on strength of alignment at different levels of learner ability.

In an L2 context, the potential for alignment will be shaped by the different goals

---

<sup>1</sup>The higher frequency of a word, the more exposure a student has had to it, the more likely they are to learn it faster (Vermeer 2001). Word Frequency has also been shown to act as a reasonable indication of word 'difficulty' (Chen & Meurers 2017).

of the speakers (Costa et al. 2008). L2 learners can benefit through alignment in terms of the vocabulary and grammatical example of their interlocutor, but their ability will affect to what extent they do. The tutor's goals will also affect how they align, potentially using alignment as a ZPD strategy. It has been hypothesised that learners leverage alignment to achieve pedagogic goals (Michel 2011). This chapter explores whether student ability affects alignment.

We measured lexical alignment in L2, fluent conversation and task-based corpora, finding that alignment effects were greater at higher levels of student ability, similar to the fluent corpora. We also found that the more complex the word, the greater the likelihood of alignment within L2 dialogue. We hypothesise that this is evidence of the learners leveraging alignment to learn through repetition of this near-production-level vocabulary in context.

## 4.1 Does Ability Affect Alignment in Second Language Tutorial Dialogue?

This section includes the verbatim copy of the following publication:

Sinclair, A., Lopez, A., Lucas, C.G. and Gasevic, D., 2018, July. Does Ability Affect Alignment in Second Language Tutorial Dialogue?. *In Proceedings of the 19th Annual SIGDIAL Meeting on Discourse and Dialogue* (pp. 41-50).

**Contributions:** The ideas and analysis in the paper were developed and discussed between all authors of the work. The original idea, the experiments and the bulk of the writing were the work of the first author.

# Does Ability Affect Alignment in Second Language Tutorial Dialogue?

Arabella Sinclair Adam Lopez

Christopher G. Lucas

University of Edinburgh

s0934062@sms.ed.ac.uk

{alopez, clucas2}@inf.ed.ac.uk

Dragan Gasevic

Monash University

dragan.gasevic@monash.edu

## Abstract

The role of alignment between interlocutors in second language learning is different to that in fluent conversational dialogue. Learners gain linguistic skill through increased alignment, yet the extent to which they can align will be constrained by their ability. Tutors may use alignment to teach and encourage the student, yet still must push the student and correct their errors, decreasing alignment. To understand how learner ability interacts with alignment, we measure the influence of ability on lexical priming, an indicator of alignment. We find that lexical priming in learner-tutor dialogues differs from that in conversational and task-based dialogues, and we find evidence that alignment increases with ability and with word complexity.

## 1 Introduction

The *Interactive Alignment Model* (Pickering and Garrod, 2004) suggests that successful dialogue arises from an alignment of representations (including phonological, lexical, syntactic and semantic), and therefore of speakers' situation models. This model assumes that these aspects of the speakers' language will align automatically as the dialogue progresses and will greatly simplify both production and comprehension in dialogue.

In a Second Language (L2) learning setting, a learner will have a more limited scope for alignment due to their situational understanding, and their proficiency will dictate to what extent they are capable of aligning lexically, syntactically and semantically (Pickering and Garrod, 2006). Even once a situational alignment is reached (i.e. the learner understands the context of their in-

terlocutor's interaction with them) there remains the question of the learners *receptive* vs. *productive* vocabulary knowledge (words they understand when others use them vs. words they can use themselves), both of which are active in L2 dialogues (Takač, 2008) and constrain their scope for alignment. Student alignment therefore will also be influenced by the tutor's strategy; or by how much of the student's receptive language the tutor produces which facilitates the student productive ability in this context.

We expect that alignment within L2 learner dialogue will differ from alignment in fluent dialogues due to the different constraints mentioned above (Costa et al., 2008). We also expect learners to align to their interlocutor to a comparatively greater degree than found in native dialogue. This is both because of the difficulty of the task leading to a greater need for alignment (Pickering and Garrod, 2006), and because we know that an L2 learner's lexical complexity increases in a dialogue setting due to the shared context words within that dialogue, compared to the level at which they are capable of expressing themselves in monologue (Robinson, 2011).

In order to find out whether ability affects alignment in L2 dialogue, we investigate *lexical priming* effects between L2 learner and tutor. *Priming* is a mechanism which brings about alignment and entrainment, and when interlocutors use the same words, we say they are *lexically entrained* (Brennan and Clark, 1996). We compare the effects against two different corpora: task-based (Anderson et al., 1991) and conversational (Godfrey et al., 1992), and between different levels of L2 student competency. We expect that alignment of tutor to student and vice versa will be different, and that the degree of alignment at a higher level of L2 learner competence will be more similar to that of conversational dialogue than that at a lower level

(Sinclair et al., 2017). We are interested in the difference between tutor-to-student (TS) and student-to-tutor (ST) alignment, as there are various factors which could contribute to both increased and decreased alignment to that existing between two fluent interlocutors (Costa et al., 2008).

## 1.1 Motivation

By examining alignment differences, we aim to better understand the relationship between tutor adaptation and L2 learner production. This understanding can inform analysis of “good” tutoring moves, leading to the creation of either an L2 tutoring language model or more informed L2 dialogue agent design, which can exploit this knowledge of effective tutor alignment strategy to contribute to improved automated L2 tutoring. The potential benefits of automated tutoring for L2 dialogue<sup>1</sup> have already been seen through the success of apps such as Duolingo<sup>2</sup> bots which allow the user to engage in instant-messaging style chats with an agent to learn another language. *Adaptation* of agent to learner however is an ongoing research task, although outside L2 tutoring, is a well-explored area (Graesser et al., 2005). Alignment, or “*more lexical similarity between student and tutor*” has been shown to be more predictive of increased student motivation (Ward et al., 2011), and agent alignment to students’ goals can improve student learning (Ai et al., 2010). We build on previous research by investigating lexical priming effects for *each interlocutor* in dialogue both within- and between-speaker, and at *different ability levels in L2 dialogue*. This adds the dimension of lexical priming and individual speaker interactions to the work of Reitter and Moore (2006) and the inspection of student to tutor, and within-speaker priming to that of Ward and Litman (2007b). By also making comparisons across L2 ability levels, we can now analyse priming effects in terms of L2 acquisition. Similar work in this area outside the scope of this paper includes work analysing alignment of *expressions* in a task-based dialogue setting (Duplessis et al., 2017) and the analysis of alignment-capable dialogue generation (Buschmeier et al., 2009).

In addition to informing dialogue tutoring agent design, this work has potential to augment existing measures of linguistic sophistication predic-

<sup>1</sup>Also known as *Dialogue-based Computer Assisted Language Learning (CALL)*

<sup>2</sup>bots.duolingo.com

tion (Vajjala and Meurers, 2016) to better deal with individual speakers within a dialogue, using alignment as a predictor of learner ability as has been suggested by Ward and Litman (2007a). Dialogue is inherently sparse, particularly when considering the lexical contribution of a single speaker. Accordingly, alignment could be a useful predictor of student receptive and productive knowledge when in combination with lexical complexity of the shared vocabulary.

## 1.2 Research Questions

We present evidence which strengthens our hypothesis that tutors take advantage of the natural alignment found in language, in order to better introduce, or *ground*<sup>3</sup> vocabulary to the student; in other words, *scaffolding*<sup>4</sup> vocabulary from receptive to productive practice in these dialogues.

Our work investigates the following research questions:

**RQ1** *How does L2 dialogue differ from task-based and conversational in terms of alignment?*

**We find ST alignment has the strongest effect within L2 dialogue.**

**RQ2** *Does alignment correlate with ability in L2 dialogue?*

**We find priming effects are greater at higher levels of student ability.**

**RQ3** *Does linguistic sophistication of the language used influence alignment of speakers at different ability levels in L2 dialogue?*

**We find the more complex the word, the greater the likelihood of alignment within L2 dialogue.**

## 2 Corpora

We compare the alignment present within three dialogue corpora: *L2-tutoring, conversational and task-based*. A summary of the corpora is presented in Table 1. The Barcelona English Language Corpus (BELC) (Muñoz, 2006) was gathered at four different periods over the course of

<sup>3</sup>Grounding in dialogue consists of the participants establishing a common basis, or ground, on which their communication takes place. This can be viewed as a strategy for managing uncertainty and therefore error handling in dialogue (Skantze, 2007).

<sup>4</sup>Scaffolding (Wood et al., 1976) provides a metaphor to the kind of temporary support at successive levels of development needed to construct knowledge, or to support learning.

Corpus	Type	English	Dialogues
BELC	L2 tutoring	non-native (levels 1-4)	118
Switchboard	conversational	fluent	1155
Map Task	task-based	fluent	128

Table 1: Corpora types and details. *Map Task* is referred to in later diagrams as MT, *Switchboard* as SB. The levels in BELC indicate increasing learner ability, with 1 indicating the lowest ability level and 4 the highest.

three years, with the students involved receiving approximately one school year of weekly English tuition between sessions. Table 2 shows a short 20-utterance long extract from a dialogue. The Switchboard Corpus is conversational dialogue over telephone between two fluent English speakers (*A* and *B*), and MapTask is a task-based dialogue where the *instruction-Giver* (*G*) directs the *instruction-Follower* (*F*) from a shared start point to an end point marked on *G*'s map but which is unknown to *F*, who also has access to a similar map, although some features may only be present on one of the interlocutors' copies.

### 3 Methods

To address *RQ1* and *RQ2*, section 3.1 discusses how we measure lexical priming so that we can compare priming effects in different situations. Section 3.2 discusses the measure we use for word complexity in order to address *RQ3*, so that we can use this as an additional parameter in our model.

#### 3.1 Lexical Convergence

Lexical priming predicts that a given word (*target*) occurs more often closely after a potential *prime* of the same word than further away. In order to measure lexical convergence, we count each word used by the speaker being considered as a potential prime. Following Ward and Litman (2007b), who measure the lexical convergence of student to tutor in physics tutorial dialogues, we only count words as primes if in WordNet (Miller, 1995), the word has a non-empty synset<sup>5</sup> e.g. if there was a choice of potential words and the speaker used the same word as their interlocutor, this can be counted as a prime, since it was not simply used because it was the only choice.

Since the learning content of L2 dialogues is the

<sup>5</sup>This also has the effect of removing function words from consideration.

Tutor	Student
<i>do you have a bedroom for just you ?</i>	<i>yes .</i>
<i>ok .</i>	<i>two .</i>
<i>how many beds are there in your room ?</i>	<i>two beds .</i>
<i>two beds .</i>	<i>two beds .</i>
<i>ok one for you...</i>	<i>... and his friend algúns amigos .</i>
<i>and a friend that's good .</i>	<i>and a friend that's good .</i>
<i>hmm what is the room like ?</i>	<i>hmm...</i>
<i>tell me about your room .</i>	<i>my room ?</i>
<i>uhhuh .</i>	<i>my room is...</i>
<i>describe it .</i>	<i>there's two beds...</i>
<i>there's two beds...</i>	<i>...very big...</i>
<i>uhhuh .</i>	

Table 2: Example of lexical alignment in BELC dialogue. *room*, *beds* and *friend* are examples of lexical alignment from student to tutor and from tutor to student respectively. Underlined text indicates within-speaker (*TT* or *SS*) alignment, and **bold** text indicates between-speaker (*TS* or *ST*) alignment (*algúns amigos* means *some friends*).

language itself, we group the words into *word families*, which is a common method used to measure L2 student vocabulary (Graves et al., 2012). We do this by lemmatizing<sup>6</sup> the words in a text, and counting *lemmas* used by the speaker as prime. Thus, we count the forms *want*, *wants*, *wanted* & *wanting* as a single word.

We also distinguish between the speakers when looking at between-speaker, or *comprehension-production* (CP) priming where the speaker first comprehends the prime (uttered by their interlocutor) and then produces the target, and within-speaker or *production-production* (PP) priming, where both the prime and the target are produced by the same speaker. Since we are also interested in tutor *T* behaviour vs. student *S* in these interactions we map PP priming to *TT* and *SS* respectively and CP to *TS* and *ST*.

<sup>6</sup>Using NLTK (Loper and Bird, 2002)

## Lexical Repetition

In our data, each repetition of an occurrence of a word  $W$  at distance  $n$  is counted as *priming*<sup>7</sup> where  $W$  has a non-empty synset, and is of the same *word-family* as its prime (section 3.1). Each case where  $W$  occurs but is not primed  $n$  units beforehand in the dialogue, is counted as *non-priming*. Our goal is to model  $\hat{p}(\text{prime}|\text{target}, n)$ , that is the sampling probability that a *prime* is present in the  $n$ -th word before *target* occurs. Without lexical priming’s effect on the dialogue, we would assume that

$$\hat{p}(\text{prime}|\text{target}, n) = \hat{p}(\text{prime}|\text{target}).$$

The distance  $n$  between stimulus and target is counted in words, as this has the advantage over utterances for capturing within-utterance priming and is less sensitive to differences in average utterance length between corpora when comparing priming effects. *Words* were chosen as the closest approximate available to *time in seconds* as measured in Reitter and Moore (2006). We look for repetitions within windows of 85 words<sup>8</sup>.

## Generalized Linear Mixed Effects Regression

For the purposes of this study, following Reitter and Moore (2006), we use a Generalized Linear Mixed Effects Regression Model (GLMM). In all cases, a word instance  $t$  is counted as a repetition at distance  $d$  if at  $d$  there is a token in the same word-family as  $t$ . To measure speaker-speaker priming effects, we record both the prime and target producers at  $d$ . GLMMs with a binary response variable such as ours can be considered a form of logistic regression. We model the number of occurrences  $\text{prime} = \text{target} | d \leq n$  (where  $n$  is window size) of priming being detected<sup>9</sup>. We model this as binomial, where the success proba-

<sup>7</sup>The use of *priming* is not intended to imply that priming is the only explanation for lexical repetition

<sup>8</sup>We chose this window size based on Reitter and Moore (2006) using an utterance window of 25 and a time window of 15 seconds. We calculated the average number of words to occur in the utterance window chosen, and the average number of words which are spoken in the 15 second window and chose the average of the two as our window.

<sup>9</sup>For example, if we were only interested in priming within a window size of 3 words, In table 2, for the student’s first use of the word *beds* we would record 3 data points: (window:1, target:bed, role:SS, prime=target:0), (window:2, target:bed, role:ST, prime=target:1), (window:3, target:bed, role:ST, prime=target:0) indicating there is a prime for our target *beds* at distance 2. The number of trials = target words  $\times$  window size.

bility depends on the following explanatory variables: *Categorical: corpus choice, priming type from speaker role, ability level*; and *Ordinal: word frequency*, as explained in Section 3.2. The model will produce coefficients  $\beta_i$ , one for each explanatory variable  $i$ .  $\beta_i$  expresses the contribution of  $i$  to the probability of the outcome event, in our case, successful priming, referred to as *priming effect size* in the following sections. For example, the  $\beta_i$  estimates allow us to predict the decline of repetition probability with increasing distance between *prime* and *target*, and the other explanatory variables we are interested in; we refer to this as the *probability estimates* in subsequent sections. The model outputs a statistical significance score for each coefficient, these are reported under each figure where relevant.

## 3.2 Complexity Convergence

To capture *linguistic complexity* within the priming words, we use Word Occurrence Frequency (WOF) as a predictor of the relative difficulty of the words used. We use  $\log(WOF)$  to normalise the data before using it as a factor in our model. *WOF* has been found to predict L2 vocabulary acquisition rates - the higher frequency of a word, the more exposure a student has had to it, the more likely they are to learn it faster (Vermeer, 2001). Word Frequency has also been shown to act as a reasonable indication of word ‘difficulty’ (Chen and Meurers, 2017). We therefore expect a negative correlation between learner level and frequency of vocabulary used, given a certain prime window. We gathered frequency counts from the Google News Corpus introduced by Mikolov et al. (2013), for its size and diverse language.

## 4 Results

### 4.1 Lexical Convergence Cross Corpora

To find how L2 dialogue differs from task-based and conversational in terms of alignment (*RQ1*), we investigate the priming effects present across corpora of different speaker roles. Figure 1 shows that the BELC corpus has a similar asymmetry in speaker alignment to MT, and that the alignment of speakers in SB is more symmetrical, mirroring the speakers’ equal role in the dialogue. This can be seen in the different priming effects between speakers in BELC and MT, and the same effects between speakers in SB. Figure 2 shows the different decay of repetition probability with window



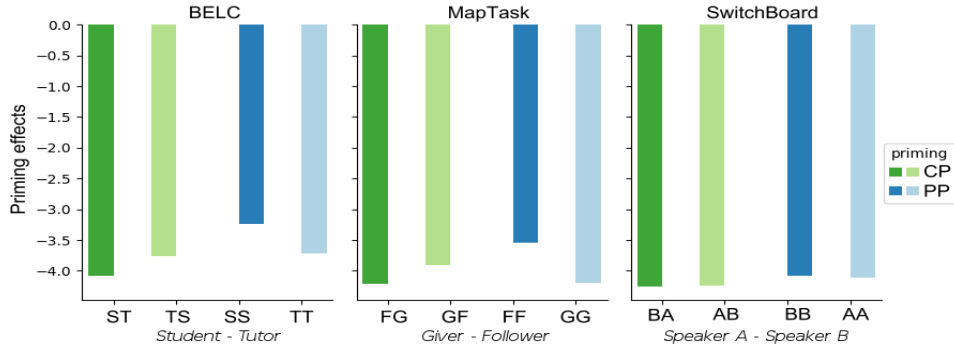


Figure 1: Priming effects of distance across Corpora for different speaker roles. *S:Student, T:Tutor, F:Follower, G:Giver, A& B:Speaker A& B*. AB indicates alignment of A to B. CP: comprehension-production, or between-speaker priming, PP: production-production, or within-speaker priming. The results are all significant with ( $p < 0.0001$ ) except *BB* within Switchboard, with ( $p < 0.01$ ).

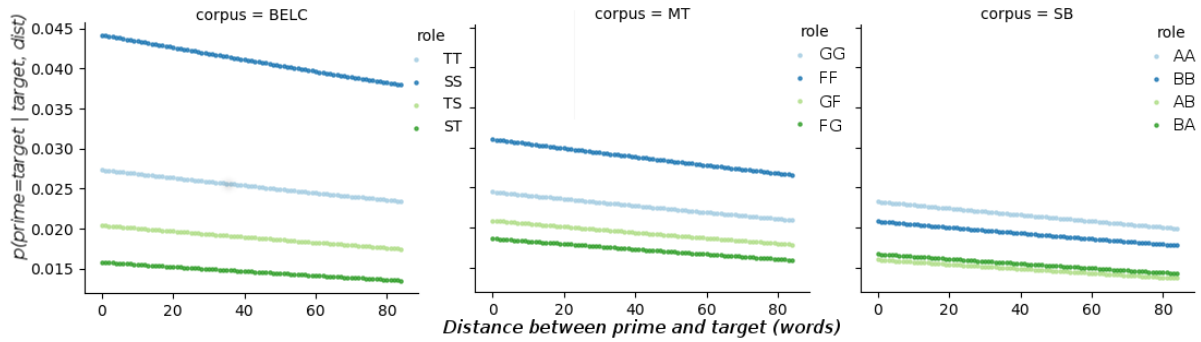


Figure 2: Decaying probability estimates for window lengths for different speaker roles across corpora. Formula :  $lemma\_occ \sim window + role * corpus$

size for the different roles for all three corpora. This shows the same symmetry and asymmetry of between- and within-speaker repetition decay probability as Figure 1.

#### 4.2 Lexical Convergence by Level

We investigate priming effects within BELC between levels to find whether alignment correlates with ability in L2 dialogue (*RQ2*). Figure 3 shows the strong student-tutor priming occurring at each ability level, and the general increase in priming effect size as ability level increases for all priming types. When comparing both Figure 1 and 3, we see that as ability level increases, BELC priming effect sizes tend towards those seen in SwitchBoard, particularly those of ST and TS, the effect size of which also becomes more symmetrical with ability level, although the imbalance between SS and TT priming remains similar to that of MapTask.

We also examine the model predictions for different window sizes for different conditions. Figures 4 and 5 describe the relationship between role

and ability level on the probability of seeing a prime word at different window sizes. Figure 4 shows a sharper decay in the probability of tutor to student (TS) priming than in student to tutor (ST) priming. Figure 5 shows that tutor self-priming is more probable at lower ability levels, and that ST alignment at lower levels is less likely than at higher levels of ability.

#### 4.3 Linguistic Complexity Convergence

Exploring the question of whether linguistic sophistication of the language used influences alignment of speakers at different ability levels in L2 dialogue (*RQ3*); we find  $\log(WOF)$  to have a significant negative correlation ( $p < 0.0001$ ) with priming effects. Thus the more complex the word (as measured by a lower *WOF*), the greater the likelihood of alignment. Figure 6 shows the priming effects of *WOF*. It shows that priming effects of *WOF* are stronger for *ST* and *TT*, than for the other roles, but this difference is less pronounced at higher levels than it is for lower levels of ability. The *ST* shows the most marked difference in



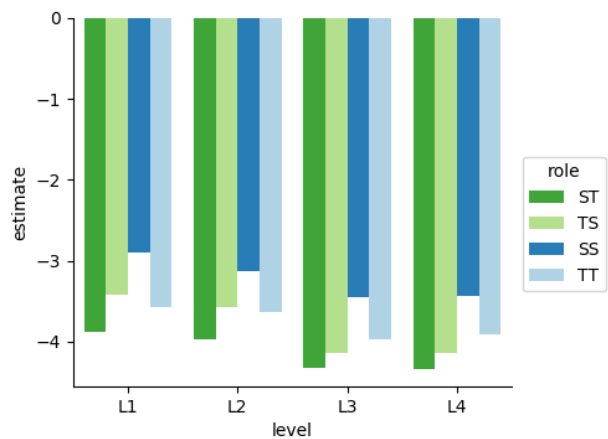


Figure 3: Priming effect sizes under different speaker role situations, across levels in BELC. Effects estimated from separately fitted nested regression models for each subset of BELC split by level(1-4). The results are all significant ( $p < 0.0001$ ).

effect between low and high levels, lowest at the highest ability. Per role, priming effect is generally smaller at higher ability levels than lower.

Figures 7 and 8 show the effects of *WOF* on level and role respectively. In Figure 7, lower  $\log(WOF)$  values are indicative of more complex words. In such cases (see Figure 7, column 1), the repetition probability is higher for high ability students, compared to low ability students. This stands in contrast to higher  $\log(WOF)$  values, indicative of less complex words, where the repetition probability is now lower for high ability students compared to low ability students (see Figure 7, column 6). Figure 8 shows differences in self-priming and within speaker priming, in that for both TS and ST, the probability of repetition is greater for higher frequency words, while for TT and SS, the probability of repetition is higher for lower frequency words.

## 5 Discussion

The three spoken dialogue corpora we investigated demonstrate a significant effect of distance between prime and target in lexical repetition, providing evidence of a lexical priming effect on *word family* use. We also found evidence of priming for each interlocutor in both between-speaker and within-speaker roles.

**ST alignment has the strongest effect within L2 dialogue.** To find how L2 dialogue differs

from our other two corpora in terms of role (*RQ1*), we measured the priming effects for Tutors (TT, TS) and Students (SS, ST) and find it asymmetric in the same manner as for the task-based dialogue MT. This is in contrast to the symmetric effects in the conversational dialogue of SB (Figure 1). ST alignment also has the greatest priming effect compared to the other roles in BELC, which supports our hypothesis that *student-to-tutor* alignment is an artefact of both tutor scaffolding, and students' productive range benefiting from the shared dialogue context.

When considering within-speaker priming, it is also interesting to note that TT priming has a more marked effect than SS priming, similar to the relationship between GG and FF in Map Task. We interpret this similarly to Reitter and Moore's (2006) comparison of Map Task and Switchboard, in that since the task-based or tutoring nature of the dialogue is harder, the leading speakers use more consistent language in order to reduce the cognitive load of the task (tutoring/instruction-giving).

**Priming effects are greater at higher levels of student ability.** In order to investigate our main hypothesis, that ability *does* affect alignment (*RQ2*), we measured priming effects in different ability levels of L2 tutorial dialogue (Figure 3), and found that priming effects are greater at higher levels of student ability, which provides evidence that as ability increases, dialogues have more in common with conversational dialogue. We also measured how role influences these priming effects (Figures 4 and 5) and hypothesise that the faster decay of TS repetition probability (Figure 5) is an indication that the tutor is using the immediate encouraging backchanneling seen in the repetition in Table 2. We note (Figure 4) that tutor-to-tutor repetition is more probable at lower levels, which supports the above hypothesis. Additionally, student-to-tutor repetition probability is more likely at higher levels which is a good indication that student ability is higher, since we argue that they are now *able* to align to their interlocutor.

**The more complex the word, the greater the likelihood of alignment within L2 dialogue.**

Lastly, to find whether linguistic sophistication of language aligned to is affected by ability (*RQ3*), we investigated the influence of word frequency on alignment within BELC. Figure 7 shows that at lower  $\log(WOF)$  values (which we use to in-

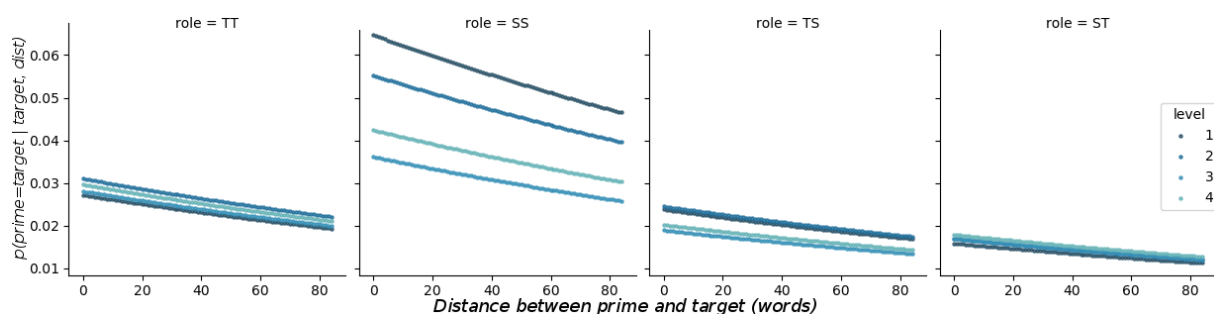


Figure 4: Decaying repetition probability estimates depending on the increasing distance between prime and target, contrasting different speaker roles at different levels.

Formula :  $lemma\_occ \sim window + role * categorical\_level$

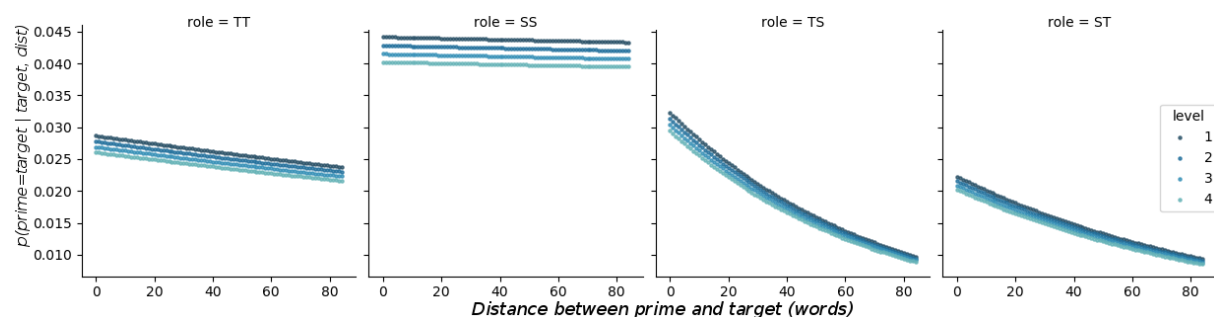


Figure 5: Decaying repetition probability estimates depending on the increasing distance between prime and target, contrasting different speaker roles at different levels.

Formula :  $lemma\_occ \sim window * role + categorical\_level$

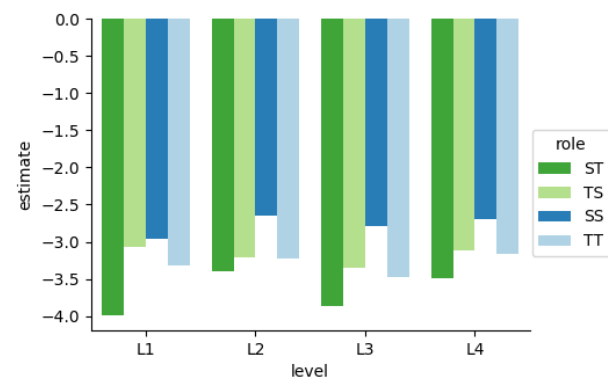


Figure 6: Word Occurrence Frequency Priming effects under different selections of role and level situations in BELC. Each model was separately fitted on the relevant subset of data to show the priming effect sizes for *Word Occurrence Frequency*. (L1:SS, L2:TS and L3:ST are insignificant, all other results are significant with at least  $p < 0.001$  and most with  $p < 0.0001$ .)

icate more complex words), repetition probability is higher in the higher ability levels compared to the lower levels, and at higher  $\log(WOF)$ , the repetition probability of the higher ability levels

is now *lower* than at the lower levels. This has interesting implications for using these results as features for student alignment ability prediction. This fits with the Interactive Alignment Model (Pickering and Garrod, 2004), which suggest that alignment will happen more with greater cognitive load, and (Reitter and Moore, 2006), who find stronger priming for less frequent syntactic rules which supports the cognitive-load explanation. The stronger priming effect identified for less frequent vocabulary also supports this hypothesis. Figure 6 shows the priming effects are slightly smaller at higher ability levels.  $\log(WOF)$  has a negative correlation, meaning there is more likely to be alignment the lower the  $WOF$ . The results at each level have a similar priming effect distribution over role, with the most marked difference in priming effect being for ST (Student to Tutor alignment), which shows a decrease in priming effect for harder words at higher ability levels. This provides an interesting first indication that there is a measurable effect of student leveraging contextual vocabulary to augment their productive reach in L2 dialogue.

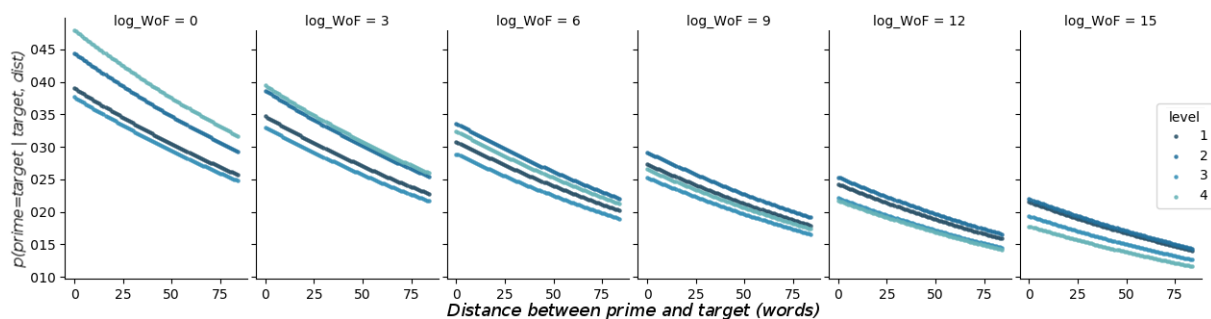


Figure 7: Decaying repetition probabilities of different  $\log(WoF)$  values on probability of word occurrence by level. Lower  $\log(WoF)$  values correspond to *lower* frequency, an indication of *more* complex words, and *higher* frequency as *less* complex words.

Formula :  $lemma\_occ \sim window + \log(WoF) * categorical\_level$

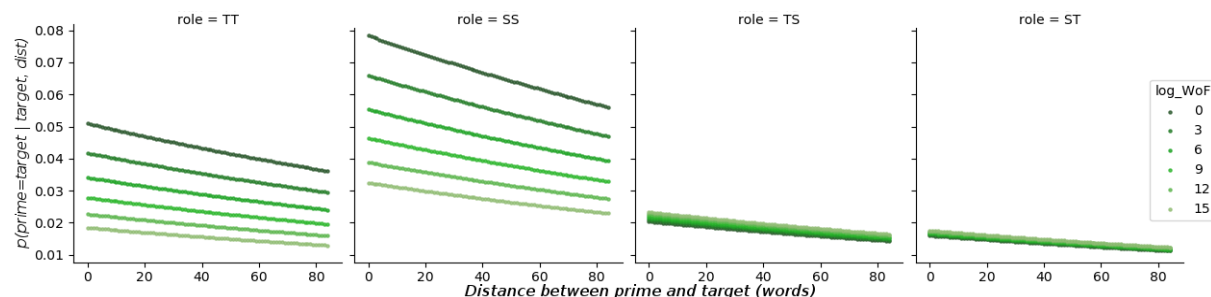


Figure 8: Decaying repetition probabilities of different  $\log(WoF)$  values on probability of word occurrence by role. *Higher*  $\log(WoF)$  indicates *easier* words.

Formula :  $lemma\_occ \sim window + \log(WoF) * role$

## 6 Conclusions and Future Work

We see these results as an indication that measuring lexical alignment combined with lexical sophistication of vocabulary has potential as a predictor of student competency. We also hypothesise that measurements of ‘good tutoring’ actions could consist of how and to what extent tutors adapt interactively to individual students’ needs in terms of their conversational ability. Tutor self-priming seems to be an interesting possible feature for measuring this adaption. We want to further investigate different measures of alignment and both lexical and syntactic complexity to inform systems that aim to automate L2 tutoring. We plan to consider which speaker *introduces* the word being aligned to, in order to better understand the relationship between productive and receptive vocabulary of the student in dialogue settings. It is also important to separate the effects of priming per se from other factors that can influence lexical convergence, such as differences in vocabulary and topic specificity. As a first step toward that goal, we plan to compare lexical convergence in the original corpus with convergence in matched

baselines of randomly ordered utterances (Duplessis et al., 2017), which will account for vocabulary effects and corpus-specific factors. To explore more measures of word complexity in addition to simple  $WoF$ , we will further investigate measures specific to L2 dialogue, such as the English Vocabulary Profile (EVP) (Capel, 2012), with word lists per CEFR<sup>10</sup> level, or measures such as counts of word sense per word, or whether a word is *concrete* or *abstract*<sup>11</sup>, exploiting existing readability features (Vajjala and Meurers, 2014).

## Acknowledgements

Thanks to Amy Isard, Maria Gorinova, Maria Wolters, Federico Fancellu, Sorcha Gilroy, Clara Vania and Marco Damonte as well as the three anonymous reviewers for their useful comments in relation to this paper. A. Sinclair especially acknowledges the help and support of Jon Oberlander during the early development of this idea.

<sup>10</sup>The Common European Framework of Reference (CEFR) defines the 6 levels of english proficiency in ascending order as: A1, A2, B1, B2, C1, C2.

<sup>11</sup>Using WordNet or other word/lemma concreteness rating database.

## References

- Hua Ai, Rohit Kumar, Dong Nguyen, Amrut Nagasunder, and Carolyn P Rosé. 2010. Exploring the effectiveness of social capabilities and goal alignment in computer supported collaborative learning. In *International Conference on Intelligent Tutoring Systems*, pages 134–143. Springer.
- 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 hrc map task corpus. *Language and speech*, 34(4):351–366.
- Susan E Brennan and Herbert H Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482.
- Hendrik Buschmeier, Kirsten Bergmann, and Stefan Kopp. 2009. An alignment-capable microplanner for natural language generation. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 82–89. Association for Computational Linguistics.
- Annette Capel. 2012. [Completing the english vocabulary profile: C1 and c2 vocabulary](#). *English Profile Journal*, 3:e1.
- Xiaobin Chen and Detmar Meurers. 2017. [Word frequency and readability: Predicting the text-level readability with a lexical-level attribute](#). *Journal of Research in Reading*, pages n/a–n/a. JRIR-2017-01-0006.R1.
- Albert Costa, Martin J Pickering, and Antonella Sorace. 2008. Alignment in second language dialogue. *Language and cognitive processes*, 23(4):528–556.
- Guillaume Dubuisson Duplessis, Chloé Clavel, and Frédéric Landragin. 2017. 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.
- 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, Patrick Chipman, Brian C Haynes, and Andrew Olney. 2005. Autotutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, 48(4):612–618.
- Michael F. Graves, Diane August, and Jeannette Mancilla-Martinez. 2012. *Teaching Vocabulary to English Language Learners*. TESOL Press/Teachers College Press.
- Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, pages 63–70. Association for Computational Linguistics.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.
- Carmen Muñoz. 2006. *Age and the rate of foreign language learning*, volume 19. Multilingual Matters.
- Martin J Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.
- Martin J Pickering and Simon Garrod. 2006. Alignment as the basis for successful communication. *Research on Language and Computation*, 4(2-3):203–228.
- David Reitter and Johanna D Moore. 2006. Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 28.
- P. Robinson. 2011. *Second Language Task Complexity: Researching the Cognitive Hypothesis of Language Learning and Performance*. Task-based language teaching : issues, research and practice. John Benjamins Publishing Company.
- Arabella Sinclair, Jon Oberlander, and Dragan Gasevic. 2017. Finding the zone of proximal development: Student-tutor second language dialogue interactions. *SEMDIAL 2017 SaarDial*, page 134.
- Gabriel Skantze. 2007. *Error Handling in Spoken Dialogue Systems-Managing Uncertainty, Grounding and Miscommunication*. Gabriel Skantze.
- Višnja Pavičić Takač. 2008. *Vocabulary learning strategies and foreign language acquisition*. Multilingual Matters.
- 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*.
- Anne Vermeer. 2001. Breadth and depth of vocabulary in relation to 11/12 acquisition and frequency of input. *Applied Psycholinguistics*, 22(2):217234.

- Arthur Ward and Diane Litman. 2007a. Dialog convergence and learning. In *Proceedings of the 2007 conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*, pages 262–269. IOS Press.
- Arthur Ward and Diane Litman. 2007b. Measuring convergence and priming in tutorial dialog. *University of Pittsburgh*.
- Arthur Ward, Diane Litman, and Maxine Eskenazi. 2011. Predicting change in student motivation by measuring cohesion between tutor and student. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 136–141. Association for Computational Linguistics.
- 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.

## 4.2 Further Discussion

While Sinclair et al. (2018) investigate lexical alignment, the idea behind exploring this came about due to considering the differences in a learner's production vs. comprehension ability. A possible method to measure this from dialogue interaction alone could start from analysing the differences in vocabulary between speakers, and specifically which speaker introduced that vocabulary to the dialogue. If a learner *introduces* a vocabulary item to the dialogue, it can be said that the learner has a good grasp on that word. The relationship between vocabulary used by the tutor and subsequently either picked up by the student or left unused can also tell us about the student's vocabulary knowledge: either receptive, productive, both or neither. Where the speakers adopt each other's language, they align, and there are many possible reasons leading to this phenomenon, including the productive vs receptive nature of the learner's vocabulary, and whether they take initiative to introduce rather than simply repeat it.

A possible additional explanation for the differences in alignment levels between student and tutor and the effects of alignment correlating with ability could be due to the role of speaker initiative. While we don't explore this empirically, in the subsequent sections we discuss our hypotheses about how Initiative, and productive vs. receptive student vocabulary may impact alignment.

### 4.2.1 Capturing Vocabulary - Concept Introduction

In Chapter 3, we discussed how some questions are harder than others for a student to answer, and answers can be an indication of full or partial understanding. This understanding may not be able to be detected if measuring surface linguistic features of the utterance as we do in Sinclair et al. (2017): "yes" in answer to a hard question will be classified as just as complex a response as "yes in answer to an easy question. Measuring alignment is a good first step in examining whether a learner is able to produce vocabulary independent of a tutor. When they exhibit alignment, which may indicate partial receptive knowledge: they may be leveraging alignment to learn (Michel 2011), taking advantage of the language of their interlocutor (Robinson 2011).

Taking the example of different questions showing different levels of learner ability we explore possible learner responses in the following paragraph. A sign of learner

Vocabulary	Student Reasons
Unused	Cannot reproduce tutor vocabulary- <i>the reason could be not useful to re-use vocabulary</i> or <i>showing non-understanding</i>
Used	Able to reproduce shared vocabulary - <i>shows understanding</i>
Introduced Unshared	Able to introduce new words to the dialogue - <i>but if unused by tutor they may be incorrect and implicitly corrected</i>
Introduced Shared	Able to introduce new and related words to the dialogue, and to have the tutor adopt these words

Table 4.1: Description of different vocabulary types and their meanings for student usage. *Used/Unused* is used to indicate whether the shared vocabulary of the whole dialogue is present in the speakers vocabulary. *Introduced* is used to indicate that this speaker was the first to use a word in the context of this dialogue.

proficiency is how they are able to respond to direct questions. In the case of student *understanding*, there are four main response type that we have found qualitatively in BELC: short/single word answering (not reproducing tutor-introduced vocabulary); repeating the correct object/subject under discussion (using that vocabulary); paraphrasing/repeating the question re-formulated to contain the answer (minimal introduction of new vocabulary); or extended answering, where a student both uses the contents of the question and brings new concepts/ideas/vocabulary to the answer (Introducing vocabulary). These response types have different implications for student knowledge. We tried to capture this at a high level through specifying the criteria of Table 4.1. Measuring the vocabulary according to these criteria within dialogues of different levels of learner ability allows us to test our hypothesis that there will be a relationship between learner ability and their produced vs used vocabulary ratio in dialogue compared to their tutor.

In order to keep track of vocabulary introduced, we traverse each dialogue by utterance to establish a set of *introduced* vocabulary per speaker and *used* dialogue per speaker. This means both independently productive and contextually productive (used, interlocutor-introduced) vocabulary<sup>2</sup> can be examined for the students, and the quantity of vocabulary introduction on the part of the tutor can be compared to its complexity and to student level.

<sup>2</sup>Learners who speak in a dialogue setting have increased productive vocabulary than when speaking individually due to the shared context/alignment/positive examples in *context* that they can bootstrap (Robinson 2011)



Level	Percentage of vocabulary used by speaker per level (%)										V_size	
	used		i_shared		i_unshared		intro (i)		vused_intro			v_dial
	T	S	T	S	T	S	T	S	T	S		
1	92%	27%	7%	7%	54%	6%	61%	13%	66%	48%	1403	
2	87%	37%	9%	8%	46%	10%	56%	18%	64%	47%	1740	
3	77%	52%	10.5%	10%	36%	17%	46%	27%	60%	53%	1946	
4	78%	51%	11.5%	9%	36%	16%	48%	26%	61%	50%	2129	

Table 4.2: Vocabulary usage statistics across levels

**used** = percentage of the dialogue vocabulary used by that speaker

**i\_shared** = percentage of the dialogue vocabulary that this speaker introduced which is shared by both speakers

**i\_unshared** = percentage of the dialogue vocabulary that this speaker introduced, which is only used by this speaker

**intro (i)** = percentage of the dialogue vocabulary which consists of introduced words by that speaker

**v\_used\_intro** = percentage of the vocabulary used by this speaker which is introduced (new) vocabulary

**v\_dial** = average dialogue vocabulary size (for the whole dialogue)

The results of comparing introduced and used vocabulary at different student ability levels of BELC are shown in Table 4.2, where it can be seen that students use a greater share of the words in a dialogue as ability increases. It can also be seen that as student ability increases, the proportion of vocabulary they introduce increases. Finally in a tutors language, introduced language that is shared by the student increases with student ability, showing that students at higher ability levels do indeed adopt more of the language of their interlocutor.

## 4.2.2 Initiative & Alignment

When viewing the L2 dialogues through the lens of speaker initiative (Walker & Whitaker 1990), tutors typically will take the role of *initiator*, thereby having more *control* within the dialogue than the student. However, this is not to say that we think of the student as a passive listener (Cohen 1987). Rather we hypothesise that a large



part of the tutor's role is to provide opportunity for student control, or to support the student's ability to take initiative as much as possible in the practice dialogue, while maintaining the direction and structure of the learning experience. With *backchannel* DAs being equivalent to the *prompts* of Walker & Whittaker (1990), their usage could be interpreted as indication of the tutor passing control to the student. Tutor usage of backchannels increases with student ability (Figure 3, Section 3.1), a sign that perhaps the relationship between initiative and L2 ability requires more exploration due to the multiple roles of backchanneling in these dialogues: whether prompt, passing control, encouragement, or reassurance.

We hypothesise that students at higher ability levels will show more initiative, or volunteer/contribute more information to the dialogue than those at lower ability levels, due to their increased fluency in the language allowing them to do so. We see possible evidence of this in that incidence of questions is higher at higher learner ability levels (Figure 4, Section 3.1).

It may be difficult to separate alignment from initiative when exploring initiative further as a feature of alignment. The concept introduction discussed in the previous section may also have a relationship with alignment and initiative in dialogue. If alignment is both a subconscious mechanism, and used as a tool by the speakers for both teaching and learning, then it would seem reasonable to hypothesise that there will be higher alignment on the part of the student when they are taking initiative.

Future work exploring the relationship between alignment and initiative in learner dialogues could aim to compare the alignment results we report with dialogues where the tutor is actively instructed to not take as much initiative. A first step could be in the annotation of BELC for markers of initiative in order to discover whether the DA labels used are enough to detect this speaker control dynamic.

### 4.3 Contributions

The main results reported in Sinclair et al. (2018) are the following:

- We find that student-to-tutor alignment has the strongest effect within L2 dialogue (Figure 1 in Sinclair et al. (2017))
- Priming effects are greater at higher levels of student ability (Figure 3 in Sinclair et al. (2017))

- Priming effects are greater for low frequency words (Figure 7 in Sinclair et al. (2017))
- The more complex a word, the stronger the priming effect for high over low ability students, and the less complex a word, the stronger for low over high ability students (Figure 7 in Sinclair et al. (2017)).

We see these results as an indication that measuring lexical alignment combined with lexical sophistication of vocabulary has potential as a predictor of student competency, which is important when a goal of a tutor is to interact within a learner's ZPD. We also hypothesise that measurements of '*good tutoring*' actions could consist of how and to what extent tutors adapt interactively to individual students' needs in terms of their conversational ability. Tutor self-priming seems to be an interesting possible feature for measuring this adaptation. In future work, we plan to investigate different measures of alignment and both lexical and syntactic complexity to inform systems that aim to automate L2 tutoring. We plan to consider which speaker *introduces* the word being aligned to, in order to better understand the relationship between productive and receptive vocabulary of the student in dialogue settings. It is also important to separate the effects of priming per se from other factors that can influence lexical convergence, such as differences in vocabulary and topic specificity. As a first step toward that goal, we plan to compare lexical convergence in the original corpus with convergence in matched baselines of randomly ordered utterances (Duplessis et al. 2017), which will account for vocabulary effects and corpus-specific factors. To explore more measures of word complexity in addition to simple *Word Occurrence Frequency* (WOF), we will further investigate measures specific to L2 dialogue, such as the English Vocabulary Profile (EVP) (Capel 2012b), with word lists per the Common European Framework of Reference (A1, A2, B1, B2, C1, C2) level, or measures such as counts of word sense per word, or whether a word is *concrete* or *abstract*. Measuring alignment can also be useful in the prediction of L2 dialogue sophistication, or in measuring how engaged the student is.



# Chapter 5

## Dialogue Style Adaptation

Interaction style can be measured in terms of Dialogue Act (DA) sequences and common usage per speaker Stolcke et al. (2000). Dialogue act sequences can reveal patterns in language, and have been used to explore effective tutoring DA sequences, common tutor DA sequences correlate with student learning (Chen et al. 2011). DAs give us a high level view of the dialogue which can generalise across languages and subjects. In Chapter 3, we reported on the complexity of text within certain dialogue acts, but did not explore how the full set of DAs used co-occur and change over the course of an interaction. In this chapter, we examine L2 interactions at the level of DAs, contrasting the symmetry between L2 speakers in the dialogue as a whole, with speakers in fluent dialogues. We find that at higher levels of ability, students exhibit more symmetry and closer overlap of DA use than they do at lower levels, similar to the symmetrical speaker role in fluent conversation. At lower levels of student ability, we see more pronounced asymmetry of DA use, similar to the difference in speaker roles we show for task based dialogue. We also compare the convergence of speaker DA usage over the course of a dialogue between corpora, finding that both L2 and fluent conversational speakers converge, although to a different extent and in a different manner, possibly indicating student competence and confidence improve over the interaction.

## 5.1 I wanna talk like you: Comparing Speaker Adaptation in L2 Practice Conversation to Fluent Speakers

This section includes the verbatim copy of the following journal submission which builds on the work of the following publication:

Sinclair, A.J., Ferreira, R., Gašević, D., Lucas, C.G. and Lopez, A., 2019, June. I Wanna Talk Like You: Speaker Adaptation to Dialogue Style in L2 Practice Conversation. *In International Conference on Artificial Intelligence in Education* (pp. 257-262). Springer, Cham.

**Contributions:** The ENA experiments and statistics reported were run by the second author, who also contributed to the writing of the methodology section of the paper. All remaining authors contributed to the development of the ideas and the refinement of the analysis and the paper as a whole. The original idea, and bulk of the analysis and writing is the work of the first author.

# I wanna talk like you: Comparing Speaker Adaptation in L2 Practice Conversation to Fluent Speakers

## Speaker Dialogue Style Adaptation

Arabella J. Sinclair · Rafael Ferreira · Dragan Gašević · Christopher G. Lucas · Adam Lopez

Received: date / Accepted: date

**Abstract** This paper presents a novel method for analysing speaker adaptation in second language conversational practice dialogue, comparing it to fluent conversational dialogues. In particular, we make use of Dialogue Acts (DAs) specific to one-to-one tutoring dialogues, and use Epistemic Network Analysis to show their co-occurrence patterns for each speaker. Drawing on the foundation of Socio-Cultural Theory, Vygotsky's Zone of Proximal Development, and the Interactive Alignment Model, we hypothesised that there would be convergence between speakers. However, the degree to which they converged would be influenced by learner ability and tutor strategy. We found that as student ability increased, the distribution of student and tutor DAs was both initially closer, and converged more over the course of the dialogues. We also found that speaker DA use changed both with ability and during the interactions. In comparison to fluent conversational dialogues, we found speaker role influenced DA patterns, and noticed higher ability students interacted in a more similar manner to the fluent speaker dialogues. These results contribute to our knowledge of speaker adaptation and convergence in terms of dialogue style as measured by DAs. We also demonstrate a novel method for the automatic analysis of both learner ability and tutor strategy, which can inform the development of personalised automatic tutoring tools, and can be used in formative assessment and feedback in an educational setting.

**Keywords** Socio-Cultural Theory · Epistemic Network Analysis · Scaffolding · Dialogue · Natural Language Processing · Alignment

---

Arabella J. Sinclair, Christopher G. Lucas, Adam Lopez  
University of Edinburgh, Informatics Forum, 10 Crichton St, Edinburgh, EH8 9AB, United Kingdom  
E-mail: s0934062@sms.ed.ac.uk, cglucas2@ed.ac.uk, alopez@ed.ac.uk,

Rafael Ferreira  
Federal Rural University of Pernambuco, Rua Dom Manuel de Medeiros, Recife, Brazil  
E-mail: rafael.mello@ufrpe.br

Dragan Gašević  
Monash University, 25 Exhibition Way, Clayton, VIC, 3800, Australia  
E-mail: dragan.gasevic@monash.edu

## 1 Introduction

One to one spontaneous dialogue practice represents an important aspect of Second Language (L2) learning in both classroom settings and online learning platforms. This form of dialogue practice has been shown to provide better opportunity for L2 learning [13, 2, 24, 3, 15] as learners can both take advantage of the example of their interlocutor, and learn through practice. According to the Zone of Proximal Development (ZPD) model, first proposed by Vygotsky [34], a ‘good’ dialogue tutor should adapt their interactions to remain within reach of the learner’s capabilities, yet provide sufficient challenge to push the learner to the farthest extent of their abilities. The improved analysis of student and tutor interactions is important as this can not only provide a tool for tutors to inform their practice, but also a method of formative assessment of students.

Within dialogue, alignment consists of interlocutors adapting their interaction to one another, resulting in convergence, or in their sharing of the same concept space [20]. In conversational and task-based dialogue, speakers have been found to align at both lexical and syntactic levels [22, 30]. Lexical alignment has also been found in L2 dialogues, with greatest alignment effects reported for student to tutor alignment [30]. Alignment between learners and teachers has been linked to both student engagement and learning [35]. In an L2 context, the potential for alignment will be shaped by the different goals of the speakers [9]: L2 learners can benefit from vocabulary and grammatical examples provided by their interlocutors, but the learners’ abilities will affect how useful those examples are. The tutor’s goals will also affect how they align, potentially using alignment as a ZPD strategy. It has been hypothesised that learners may leverage alignment to achieve pedagogic goals [18].

In this paper, we are particularly interested in conversational practice dialogue where there is no explicit form-based teaching and the goal is to encourage spontaneous dialogue, similar to those held between fluent speakers. We expand on our previous work which demonstrates that as a student’s ability improves, the contribution of both student and tutor becomes increasingly symmetric [31], as is the case between fluent speakers in conversation [29]. We compare *dialogue style* within L2 conversations to spontaneous fluent conversation, with the hypothesis that at higher student ability levels, student dialogue style will more resemble that of spontaneous fluent conversation. We also contrast L2 dialogue with task-based conversations, since there is a similarly asymmetrical relationship between speaker role (the tutor in some sense performing the *task* of teaching which is similar to *instruction giving* within task-based dialogue). We expect however, that alignment of dialogue style will adapt as a function of learner ability level, since the tutor will adhere to the ZPD.

Our study has implications for automatic tutoring systems, which remove some of the social barriers to learner conversational practice. Dialogue practice is beginning to be offered by language learning apps such as Duolingo and Babbel<sup>1</sup>, although the experience in them is not personalised to adapt constantly to the student’s level. The analysis of the resulting dialogues necessitates automatic measures of student engagement and ability such as analysing alignment of learners, something which has

---

<sup>1</sup> bots.duolingo.com, babbel.com

also been shown between humans and computers [6]. Additionally, the development of better dialogue agents cannot be achieved without further analysis of dialogues with human tutors in order to better understand their adaptation strategies to cater to the needs of individual learners.

We examine aspects of tutors' adaptation to students at a higher level than purely lexical, in order to identify scaffolding actions associated with this behaviour. We contrast this analysis of dialogue style with both spontaneous conversation and task based dialogue in order to contextualise our findings within a range of conversation environments. Better understanding of and automatic ability to identify these actions can lead to more personalised, student-centric tools and models for automatic L2 tuition. Personalisation of learning experience in the form of one to one tutoring is important to learner progression, resulting in significantly higher cognitive learning gains than group education [4]. Therefore, our work proposes a language agnostic method for recognising the needs of an individual in terms of their interaction patterns, and analysing how tutors adapt to these needs, so as to provide insights when developing automated conversational aides.

We are interested in two main aspects of alignment: the symmetry between speakers within the dialogue as a whole, and the convergence of speakers over the course of a given interaction. In particular, we study how speaker role influences dialogue style, and the effect of learner ability on adaptation. We choose to examine alignment at the utterance level; that is, what types of utterances are more prominent within each ability level and within the dialogue of each speaker? We use Dialogue Acts (DAs) [14] as labels to describe the role of each utterance in the dialogue. Alignment is analysed in terms of DA usage since this allows a more high-level view of the *types* of interaction present in the dialogue, and the resulting *dialogue style*. DAs allow for a topic agnostic method of comparison of student and tutor contribution in different learning contexts, and they have also been used to identify common tutor interactions in other educational settings [7]. It is also particularly suited to the analysis of L2 learning, where the language itself is the educational content. We therefore explore the following three research questions, building on our previous work investigating dialogue style and convergence between student and tutor [31], adding further analysis of the impact of learner ability on dialogue style adaptation, and contrasting this to alignment found within other conversational dialogue.

**RESEARCH QUESTION 1:**

*What is the relationship between students' and tutors' DA usage and student ability?*

**Hypothesis:** *DA usage will be more similar as student ability improves, because speaker contributions become increasingly symmetric [9].*

**RESEARCH QUESTION 2:**

*How does the distribution of DAs used change over the course of a dialogue?*

**Hypothesis:** *speakers will converge within the course of an interaction [29].*

**RESEARCH QUESTION 3:**

*What is the relationship between DA usage and speaker role in different conversational settings?*



**Hypothesis:** *DA usage will be different depending on both the type of conversation and speaker role [29].*

To answer our three research questions, we apply Epistemic Network Analysis (ENA) [27] to study and visualise how DAs co-occur within student-tutor one to one dialogues. ENA allows us to analyse the DA co-occurrence in a multi-dimensional space, showing the strength of co-occurrence between different DAs for different speakers in different dialogue settings and, within L2 dialogue, at different levels of student ability. This allows us to quantify an interlocutor’s dialogic contribution, which enables us to measure (i) each speaker’s dialogic changes over time and (ii) alignment between speakers.

Our main contributions are twofold: Firstly, we contribute to the existing literature on speaker adaptation within L2 dialogue, providing evidence to support our hypothesis that increased DA alignment can be seen both with increasing ability level and across dialogues within L2 corpora, building on our initial analysis reported in [31]. We also analyse dialogue style adaptation within spontaneous fluent conversation and task-based dialogue, and compare them to L2 dialogue adaptation. The contrast between fluent dialogues with clearly different speaker roles allows us to contextualise our analysis of the L2 dialogue adaptation relative to conversational dialogue in general. We use this understanding to analyse tutor strategy, and learner progression. Second, we apply a novel method for modelling speaker contribution and dialogue style to L2, spontaneous and task based conversational dialogue, combining the descriptive powers of ENA with DAs. This has implications for both formative assessment in an instructional setting, and continuous feedback for tutors and students, providing a data-informed reflection on their practice. Our work also has implications for (i) the design of learning analytic tools, (ii) informing tutoring strategy, and (iii) the design of automatic tutoring systems.

## 2 Background

Alignment in dialogue is a well studied phenomenon [5, 11]. The Interactive Alignment Model (IAM) [20] describes the process of speakers agreeing on a shared conceptual space. In educational settings, alignment has been used as a predictor of student learning and engagement [35]. Typically, alignment is measured either at a lexical or a syntactic level. While *lexical* alignment consists of speakers beginning to use the same words [36, 30] or phrases [10] as each other, *syntactic* alignment consists of the same parts of speech patterns, such as similar noun-phrase constructions, or similar adjunct phrases [22, 21] as the conversation progresses. Methods for measuring alignment can range from simple count statistics [10] to linear regression on prime target distance<sup>2</sup> [35] or using generalised linear mixed models to take into account the random speaker effects present in dialogue [21] for a similar sliding window of prime and target occurrence. We use ENA because it allows us to view the multidimensional space of DAs in two dimensions and therefore compare alignment within DA space,

<sup>2</sup> The item being aligned to in this context is known as the *prime*, and the subsequent usage of this prime by the other speaker is known as the *target*, or sign of alignment

rather than as more simple pairwise-co-occurrence. ENA was developed to quantify qualitative analysis of interaction sequences in learning [27]. It has been used to analyse student interactions within diverse learning environments [26, 28], modelling individuals' interactions that are characteristic of a particular group or context.

Within an L2 practice setting, alignment will have slightly different properties compared to a fluent conversational setting where speakers tend to have a symmetric contribution and equal status within the dialogue [9]. The tutoring context will also have an impact upon the speakers' interaction, with the tutor being more likely to take the lead role in moving the dialogue forward, therefore making it less symmetric. L2 learners have been found to perform at a higher level when speaking in dialogue with a peer than in a monologue context, which suggests they draw from the example language of their interlocutor leading us to expect evidence of alignment [23]. However, the ability of the learner will dictate how much of their interlocutor's dialogue they are able to understand and therefore align to. In the case of the tutor, their need to adhere to 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 [9].

The theoretical framework underlying our research is Socio-Cultural Theory (SCT) [16]. This emphasises the central role of dialogic interaction in all learning, and the concept of internalisation: as a result of dialogic inter-psychological activity, new knowledge is appropriated. In other words, students learn through talking [13]. Vygotsky's ZPD [34] states students will learn best when addressed at the correct level, therefore we also expect to see tutors adapt to student ability.

If we view dialogue as a mediated or collaborative learning process, we can expect to see the speakers trying to arrive at a shared understanding at the utterance exchange level [15]. While we expect speakers to arrive at a communicative symmetry [33], as speakers do in spontaneous conversational dialogue [29], we do not expect them to be able to do so in all cases. The nature of a tutor-student relationship has an expert-novice asymmetry, similar to the different roles of instruction *giver*, and instruction *follower* in task-based speaker dialogues [1]. At higher levels of student ability, the tutor should begin to alter their role to that of conversational peer to better encourage student independence and autonomy, thus slowly removing some support [3]. This change in tutor role as a learner gains proficiency is one aspect of ZPD which we examine at the level of interaction sequences, comparing to the more fixed roles in other conversational dialogue, modelled through the use of DA labels.

We choose to examine dialogue at the level of *Dialogue Acts* (DAs), which are labels given to utterances in dialogue to describe their function, such as *question*, *statement* or *backchannel*<sup>3</sup>. DAs are often used to infer discourse structure, and are an important aspect in the automatic understanding of spontaneous dialogue [32]. DAs are similar to *Speech Acts* [25], but are often more specific, and are commonly used in natural language processing settings for the annotation of single utterances [29]. DAs aim to capture the discourse structure of a dialogue, and allow us to understand better the dynamics between speakers and speaker communication style or strategy.

---

<sup>3</sup> a form of feedback a speaker gives to their interlocutor

DAs have been used in the analysis of tutoring action sequences to better understand effective teaching strategies in dialogue [7].

### 3 Corpora

**L2 Learner Corpus:** The dataset used is the Barcelona English Language Corpus (BELC) [19]. It consists of 118 transcripts from conversational practice between English language learners and tutors. These vary from 60 to 140 utterances in length. The tutors' instructions for the dialogue were to elicit as much conversation from the learner as possible, and to set them at ease while having as natural a conversation as possible. The tutors follow a similar script of questions with each participant resulting in the dialogues covering similar topics. The corpus was gathered at four different times over a period of three years, with the students receiving approximately one school year of weekly English tuition between sessions. Thus, the corpus can be divided into four general levels of student ability. The corpus has been annotated at an utterance level with a set of DAs [29], which were chosen from [32] for their relevance to the corpus. These can be seen, along with our abbreviations for the codes in Table 2.

**Table 1:** DA annotated dialogue examples at Levels 1 (Highest) and 4 (Lowest) in BELC

**P** = Participant **DA** = Dialogue Act

<b>P</b>	<b>Level 1</b>	<b>DA</b>	<b>P</b>	<b>Level 4</b>	<b>DA</b>
T	do you like the school ?	YNQ	T	do you like this school ?	YNQ
T	[- spa] m-entens ?	SPA	S	yes .	YesA
S	0 [= says nothing] .	SNA	T	yes ?	RAck
T	"do you like" ?	YNQ	T	what are you planning to do next year ?	WhQ
T	do you like the school ?	YNQ	S	I would like to study zoology .	Smt
S	xxx .	SNA	T	what time did you arrive here this morning ?	WhQ
T	no si t-agrada l-escola ?	SPA	S	this morning ?	GenQ
T	do you like the school ?	YNQ	T	yes .	YesA
S	yes .	YesA	S	I ... I am here since eight o'clock .	Smt
T	yes ok .	RAck	T	uhhuh right quite early .	Smt
T	now what time do you begin in the morning ?	WhQ	T	and when will you leave ?	WhQ
S	0 [= says nothing] .	SNA	S	I ... I finish my time-table in half-past-two .	Smt
T	[- spa] m-entens ?	SPA			
S	[- spa] no .	SPA			

The DAs occurring within BELC vary with ability level and speaker. Table 1 demonstrates some of the differences in the sorts of DA patterns present. The first column shows a section of a dialogue at level 1, where it can be seen that the student uses mainly *Yes-Answers* (*YesA*), and *Signal-non-understanding* (*SNA*) when they are not replying in *Spanish* (*SPA*), their native language. Their tutor meanwhile repeats *Yes-No-Questions* (*YNQ*) for most of the dialogue, interspersed with *acknowledgements* (*RAck*). This is clearly more asymmetrical than the diverse interaction seen in the second column, where both participants ask at least one *question* and *statement*, showing even via the dialogue acts the greater competence of the student at level 4.

**Table 2:** Dialogue Acts, Labels and Examples

% is the percentage of utterances in the corpus labeled with a specific Dialogue Act

Code	Tag	Example	%BELC	%MT	%SB
<b>YesA</b>	YES ANSWERS	<i>yes .</i>	5.2	11.3	1
<b>NoA</b>	NO ANSWERS	<i>no / nope / uh no</i>	1.7	4.8	1
<b>BAck</b>	BACKCHANNEL-ACKNOWLEDGE	<i>uhhuh</i>	3.3	↓	19
<b>RAck</b>	RESPONSE ACKNOWLEDGEMENT	<i>ok. / good. / right ok</i>	2.3	24.2	1
<b>NA</b>	SIGNAL-NON-UNDERSTANDING	<i>hmm. / ah. / [-spa] no se/ silence</i>	8.0	0	0.1
<b>repeat</b>	REPEAT-PHRASE	<i>XX ok/ ah XX*</i>	1.9	-	0.3
<b>YNQ</b>	YES-NO-QUESTION	<i>do you XX, are you XX</i>	3.5	6.5	2
<b>DYNQ</b>	DECLARATIVE YES-NO-QUESTION	<i>so XX ?</i>	6.8	5.2	1
<b>BackQ</b>	BACKCHANNEL-QUESTION	<i>yes? / oh yeah? / no? / really?</i>	2.7	↓	1.1
<b>whQ</b>	WH-QUESTION	<i>ok and wh*... / wh* .. / uhhuh ok wh*</i>	9.3	↓	1
<b>genQ</b>	GENERAL-OTHER-QUESTION	<i>Any other question</i>	25.0	11.6	0.8
<b>Smt</b>	STATEMENT	<i>Any other utterance</i>	36.4	32.3	68

\*when XX is in previous utterance

We compare BELC against two different fluent English corpora: task-based [1] and conversational [12].

**Task Based Corpus:** The MapTask corpus consists of 128 dialogues between two participants, the *Giver* and the *Follower*, with an average of 207 utterances per dialogue. The speakers were tasked with describing or marking a route on a map that was marked on only the giver’s map, while the follower had 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 was much less fluent than their interlocutor. The Map-Task corpus has a greater proportion of *response acknowledgements* (RAck) and *yes answers* (YesA) than BELC, and a similar proportion of *statements* (stmt). This can be seen in the right hand column of Table 3 This reflects the fact that the speakers have to constantly confirm their shared understanding of the task.

**Conversational Corpus:** The Switchboard corpus is a large corpus consisting of 1,155 dialogues of an average of 193 utterances in length. The dialogues were collected from telephone conversations between English speakers on a random topic selected from a set of pre-defined conversational topics such as sports, television or politics. The speakers did not necessarily know each other, had equal status, and the aim was to produce largely unconstrained conversation. The Switchboard corpus contains a much larger proportion of *statements* (stmt) and *backchannel acknowledge* (BAck) than either of the other corpora, shown in Table 3. This reflects the fact that in spontaneous informal dialogue, there is less need for clarification questions, and more exchange of opinion.

#### 4 Epistemic Network Analysis

We use Epistemic Network Analysis (ENA) [27] to derive the DA space in order to examine the relationship between speakers DA distribution and student ability. ENA is a graph-based analysis method for examining the association between different concepts (called *codes*) in textual datasets. Two codes are considered related if they appear in the same *stanza*, which in our case are either the full dialogue (research question 1), or quarters of the dialogue divided by number of utterances (research

**Table 3:** DA annotated dialogue examples of fluent conversational dialogue  
*S: Speaker, DA: Dialogue Act, G: instruction giver, F: instruction follower*

<i>Conversational- Switchboard</i>			<i>Task based - Map-Task</i>		
DA		S	DA		S
B	Okay .	RAck	G	just slightly below it	Stmt
A	Great . Um , currently , I 'm not doing a whole lot of exercise in any type of program .	Stmt	F	on the left-hand side?	DYNQ
B	Huh-uh .	BAck	G	mmhmm	YesA
A	I 'm mainly do a lot of walking . I have a son [...] be dedicated towards the ,	Stmt	F	okay	Stmt
B	Yeah .	YesA	F	so i 'm going underneath it?	DYNQ
A	exercise area , is covered in boxes .	Stmt	G	above it	Stmt
B	Um , what did you do when you did exercise regularly ?	genQ	F	right	RAck
A	Well , I had , uh , a little routine that I did for warm ups .	Stmt	G	and then you 'll be underneath the waterfall?	YNQ
B	Huh-uh .	BAck	F	that 's right	YesA
A	And then I did some very mild [...] not trying to make big bulging muscles ,	Stmt	G	and go up the left-hand side of the waterfall in a straight line	Stmt
B	Huh-uh .	BAck	F	mmhmm	RAck
A	just trying to try and stay as firm as I can stay in my old age .	Stmt	G	and then turn to your right	Stmt
B	Yeah . Um , right now , um , I try when it 's nice out [...] I don 't know ,	Stmt	F	mmhmm	RAck
A	Huh-uh .	BAck	G	and go for about an inch	Stmt
B	if it 's up north , but every weekend [...] that 's a lot of fun .	Stmt	G	and then turn upwards again	stmt
A	Huh-uh .	BAck	F	have you got public footpath ?	YNQ

question 2). ENA provides several networks and graphs to analyze the relationships between different codes called an *analysis unit*. As an example, if we consider Table 1 as a stanza, Table 4 is the resulting input file for ENA. Each utterance is represented as a one-hot encoding of the DA labels<sup>4</sup>. A co-occurrence matrix is then generated based on the summation of these codes. Dimensionality reduction is then performed using Singular Value Decomposition (SVD) [17]. Typically, a two dimensional representation (with axes [svd1, svd2]) of the analytic space called the projection graph (Figure 1a) is used in analysis. This graph shows the units of analysis (tutors and students), represented by the nodes and the mean network of these groups as squares. The units of analysis are presented by their centroids, whereby a centroid is calculated as an arithmetic mean of edge weights for a given unit of analysis. In other words, a centroid is a point representing the average position of a speaker in DA space.

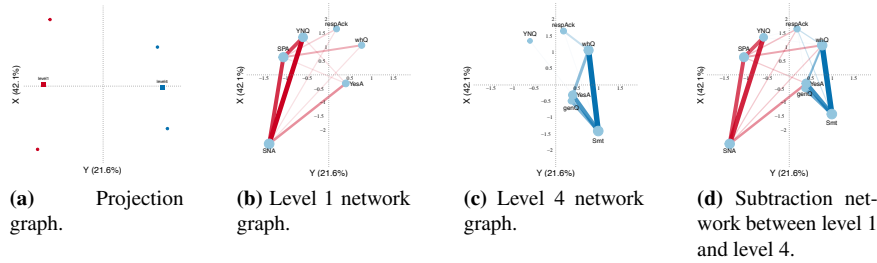
ENA also produces a network diagram that shows the code relationships for an individual unit of analysis as undirected graphs. Figures 1b and 1c show the network diagrams of the level 1 and level 4 groups, respectively. The size of the nodes represents their frequency, while the strength of the code relationship (line thickness between nodes) represents the frequency of their co-occurrence within a given dialogue (RQ1), or quartile (RQ2). To compare the network graphs, *subtraction graphs* (Figure 1d) can be used, which show the edges of the unit of analysis which have the stronger connection.

To answer the first research question, we used *speaker* (students and tutors) and *ability level* as the unit of analysis and individual dialogues as stanzas. The projection network was used to extract the projection points in the two dimensional space (i.e., svd1 and svd2) of each speaker, which we refer to as DA space. The differences

<sup>4</sup> vector of the presence (1) or absence (0) of each code

**Table 4:** ENA initial file example; **P** = Participant **DA** = Dialogue Act

P	Level 1	DA	SPA	SNA	whQ	RAck	YNQ	YesA
T	do you like the school ?	YNQ	0	0	0	0	1	0
T	[- spa] m-entens ?	SPA	1	0	0	0	0	0
S	0 [= says nothing] .	SNA	0	1	0	0	0	0
T	“do you like” ?	YNQ	0	0	0	0	1	0
T	do you like the school ?	YNQ	0	0	0	0	1	0
S	xxx .	SNA	0	1	0	0	0	0
T	no si t-agrada l-escola ?	SPA	1	0	0	0	0	0
T	do you like the school ?	YNQ	0	0	0	0	1	0
S	yes .	YesA	0	0	0	0	0	1
T	yes ok .	RAck	0	0	0	1	0	0
T	now what time do you begin in the morning ?	WhQ	0	0	1	0	0	0
S	0 [= says nothing] .	SNA	0	1	0	0	0	0
T	[- spa] m-entens ?	SPA	1	0	0	0	0	0
S	[- spa] no .	SPA	1	0	0	0	0	0
Stanza	Summation		4	3	1	1	4	1

**Fig. 1:** ENA examples derived from DA counts from the Dialogue sample in Table 1.

between the speaker groups on both svd1 and svd2 values were then compared by using a series of Mann-Whitney tests where the threshold for statistical difference was set initially at 0.05 and Bonferroni correction was then applied to avoid type I errors. The subtraction network was used to explain qualitative differences between the groups.

To answer our second research question, we split each dialogue in the dataset into four quartiles by number of utterances. The unit of analysis used for this network was *speaker*, *ability level* and *quartile*. The interpretation of the SVD vectors differs from those produced under the previous configuration. From the network graphs, we produced a *trajectory* graph in order to compare projection points at different stages in the dialogue and therefore see how each speaker’s DA use changed from one quartile to the next. The points represent the mean positions of students and tutors in DA space for each of the four quartiles of their dialogue.

To answer our third research question, firstly we used speaker as the unit of analysis and separate dialogues as the stanza and performed separate analysis on the Map-Task and Switchboard corpora. Thus allowing us to discover the differences in DA usage between speakers, and compare these to the BELC speakers at different levels of student ability. Secondly, we combined all the corpora, using corpus, speaker and student level as the unit of analysis to inspect whether tutor and student at different levels have DA patterns more similar to either of the speakers in either fluent conversational corpora. Finally, we split each dialogue into four quartiles in the same manner as done to address research question 2, and analysed the trajectories of the speakers, both in the individual corpora comparing speakers (with *speaker* and *quar-*

*tile* as units of analysis), and in the combined space to contrast all roles together (with *corpus*, *speaker*, *quartile*, and *ability level* as units of analysis).

## 5 Results

### 5.1 L2 corpus analysis: BELC

In order to answer our first research question, we used the full dialogue as a stanza. Figure 2(a) shows the projection of individual students' and tutors' mean networks (i.e., centroids) at different levels of student ability in DA space. The main difference between the students and tutors in DA distribution was along the X-axis, and a higher variance between groups over different levels (from 1 to 4) can be seen across the Y-axis. This shows that the main differences between the speakers was between the use of *statement*, *yes-answer*, *no-answer* and *signal-non-understanding* (student) and the rest of the DAs (tutor). Figure 2(b) shows the position of DA centroids in relation to these axes and helps interpret the type of DA change present in Figure 2(a). Figure 2(b) shows a subtraction network of student and tutor projections, which shows students had more connections between *statements*, *signal-non-understanding* and *yes-answers* than tutors, who had more connections in general, specifically between *questions*, *back-channeling* and *repetition*.

To further explore the differences in DA connections across levels of student ability, Figure 3 shows the differences between levels 1 and 4 for both student and tutor. In general, speakers at higher student ability levels showed more DA co-occurrence between *Statement*, *Wh-Question* and *Response-Acknowledgment*, while at lower levels, more between *Signal-non-understanding*, and *general-other-questions*. Tables 5 and 6 present the Mann-Whitney tests results over the X and Y axis for tutors and students in different levels, which shows the differences in co-occurrence between DAs across these axes. For both students and tutors, the differences between level 1 and the other levels were significant. The differences between tutor and student at level 1 and level 4 have large<sup>5</sup> ( $r = 0.55$ ,  $r = 0.82$ ) effect sizes, respectively. Moderate effect sizes can also be seen between students at level 2 compared to level 3 and 4. Differences between level 3 and 4 however were not significant.

**Table 5:** Tutor matrix of Mann-Whitney test results over the Y-Axis

	Tutor 1		Tutor 2		Tutor 3		Tutor 4	
	U	r	U	r	U	r	U	r
Tutor 1	-	-	439**	0.45	137**	0.68	113**	0.55
Tutor 2	-	-	-	-	316*	0.40	220	0.29
Tutor 3	-	-	-	-	-	-	164	0.02
Tutor 4	-	-	-	-	-	-	-	-

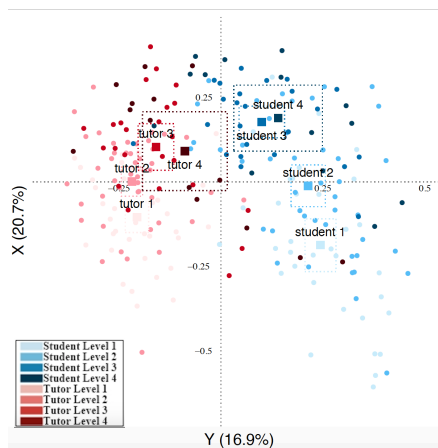
**Table 6:** Student matrix of Mann-Whitney test results over the Y-Axis

	Student 1		Student 2		Student 3		Student 4	
	U	r	U	r	U	r	U	r
Student 1	-	-	426**	0.43	59**	0.86	42**	0.82
Student 2	-	-	-	-	230**	0.56	136**	0.56
Student 3	-	-	-	-	-	-	146	0.13
Student 4	-	-	-	-	-	-	-	-

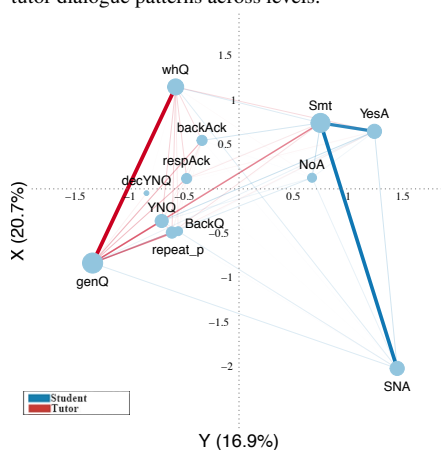
Note: \* indicates  $p < .05$ , \*\* indicates  $p < .001$ . Results for projections from Figure 3

To further answer RQ1, we also explored the differences in connection between student and tutor at the lowest and highest level (figure 4). The low ability dialogues

<sup>5</sup> according to Cohen's proposal [8]



(a) ENA scatter plot of the average student and tutor dialogue patterns across levels.

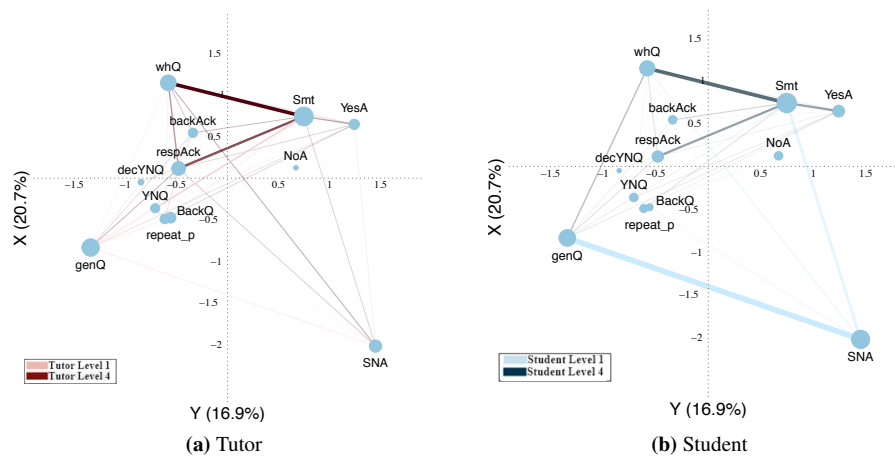


(b) Subtraction network for student and tutor networks. This shows which connections were stronger for each speaker compared to the other

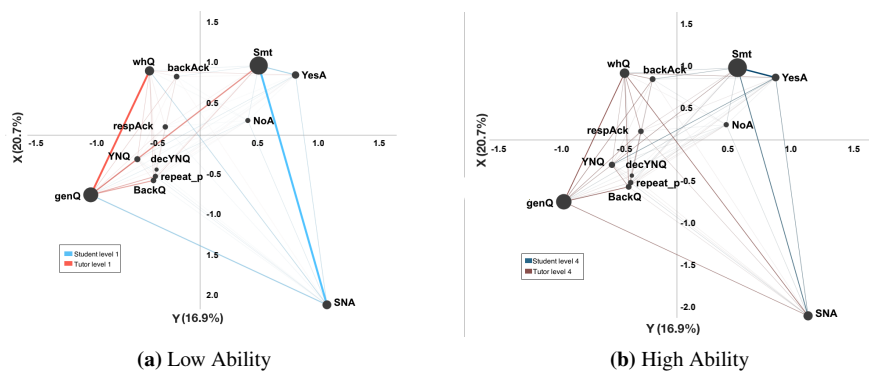
**Fig. 2:** ENA plots of DA space: Students and tutors at Higher levels have a position in the space closer to DAs such as *WH-questions*, and *Statements*, whereas at lower levels, they are closer to DAs such as *general questions*, and *Signal-non-understanding(SNA)* The subtraction network shows which speakers’ DA co-occurrence connection was stronger than their interlocutor for each pair, i.e. tutors whQ and genQ have a stronger connection than students, and students have a stronger connection between smt and SNA than tutors..

showed more *signal non understanding* produced with *wh-questions* and *general-questions* most often by students, whereas the tutors produced more *general-question* with *wh-question* and *statements*, creating the strong connections between the DAs in Figure 4. In comparison, the high ability students produced *statements* more often with *yes-answers*, *yes-no-questions* and *wh-questions*, with tutors producing *response*





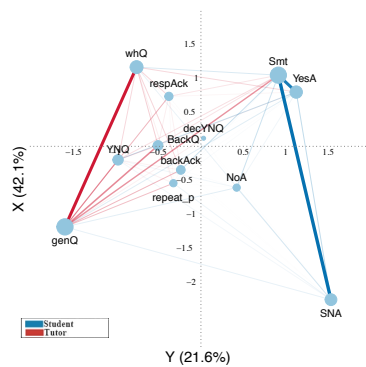
**Fig. 3:** ENA projections comparing the effects of student ability level on each speaker.



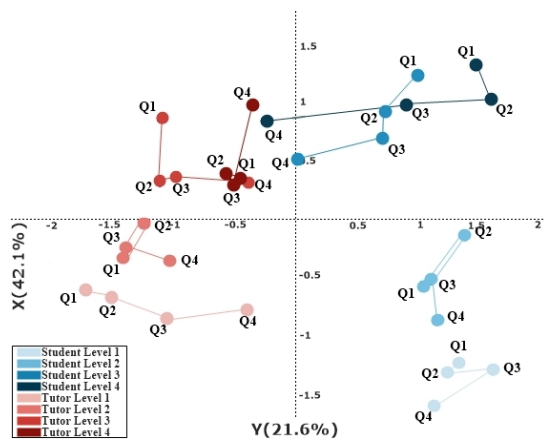
**Fig. 4:** Subtraction Graphs comparing speaker DA connections at low and high levels

*acknowledgment* more often with *statement* and *wh-questions*, as seen by the stronger connection in Figure 4. This difference shows both the student managing to participate more (more *wh-questions*), and the tutor needing to acknowledge this and make more *statements* and ask more specific *questions*. Less general questions could also be as a result of the students' ability to respond: in the lower ability dialogues, often when the tutor asked a question that the student was unable to reply to, they would paraphrase themselves in a simplified, follow-up question, which may have been a more general version of the original *wh-question*. e.g. "what time did you arrive at school this morning?" (*wh-question*) followed by "was it 7 o'clock? 8?" (*general question*).

To answer our *second research question* (RQ2), we used dialogue quartiles as stanzas. Figure 5 shows the subtraction network between students and tutors in the new projection space. Again, the visualisation was done using 1 and 2, which accounted for 42.1 and 21.6 percent of variability, respectively.



**Fig. 5:** Mean Subtraction between tutor and student at all levels across all quartiles. This shows the new DA space when stanzas are *dialogue quartiles*. This projection allows us to analyse the mean trajectories in Figure 6 better.



**Fig. 6:** ENA Trajectories over the course of the dialogue interactions. Each point shows the mean of students and tutors for each of the four quartiles of the dialogues at each student level. Euclidean distances between points are in Table 7. t-tests show significant differences between Q1 & Q4 for each trajectory except Students at Level 1 ( $x(D = 0.28 p = 0.26)$ ,  $y(D = 0.08 p = 0.74)$ ). The highest effect sizes were for Tutor Level 1 ( $D = 1.49 p = 0.001$ ) and Student Level 4 ( $D = 1.46 p = 0.001$ ).

Figure 6 shows the trajectory of different groups over the four quartiles, which is the same DA space as Figure 5. This movement of speakers in the DA space (indicating their change in DA co-occurrence and thus dialogue style) shows speaker convergence to a more similar DA distribution in quartile 4 than in quartile 1 for each level. Table 7 shows the Euclidean distances between the coordinates of Figure 6. It can be seen that, at higher levels of student ability, the difference between student and tutor DA co-occurrence was much smaller than it was at lower levels; that is, student and tutor contributions were much more similar by the end of the dialogue than they were at the beginning. It can also be seen that the students at higher levels moved

much further in the space in the direction of the tutor, than students at lower levels, who remained within a smaller area. The tutors on the other hand, travelled less in the space in general, except for students at level 1, where they moved in the direction of the students position.

**Table 7:** Distances between points on the BELC trajectories of Figure 6

Level	1	2	3	4
Student Distance Travelled $\ Q1-Q4\ $	0.44	0.33	1.28	1.88
Tutor Distance Travelled $\ Q1-Q4\ $	1.42	0.39	0.94	0.66
Q4 Student-Tutor dist $\ Q4\_S-Q4\_T\ $	1.84	2.40	0.48	0.17
Q1 Student-Tutor dist $\ Q1\_S-Q1\_T\ $	3.32	2.63	2.26	2.28
Difference in start-end distances $\ dist\_Q1-dist\_Q4\ $	1.48	0.23	1.79	2.11

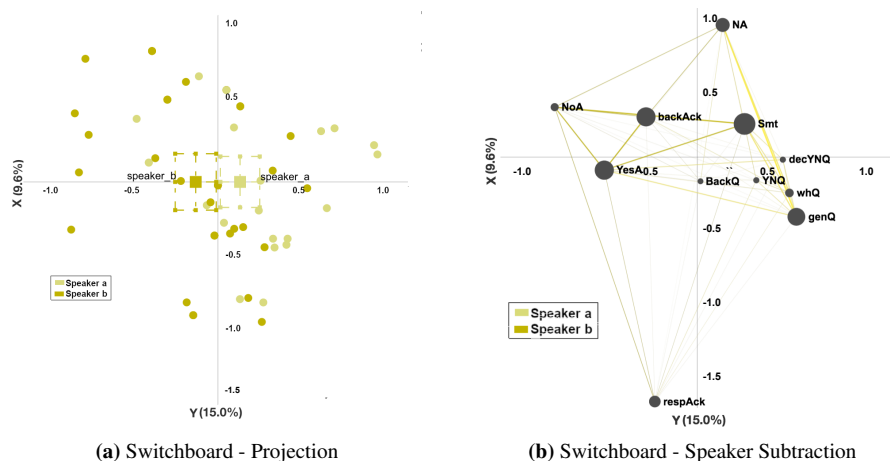
## 5.2 Conversational and Task-Based Dialogue Analysis

In order to address our *third research question* (RQ3) and explore the BELC analysis in the wider context of conversational dialogue, we compared the BELC DA distribution conversational dialogues where participants have equal status (switchboard); and to a task-oriented dialogue where one participant gives instructions to their interlocutor in order to achieve a shared goal (Map-Task). The other two corpora consisted of conversational interactions by fluent speakers of English.

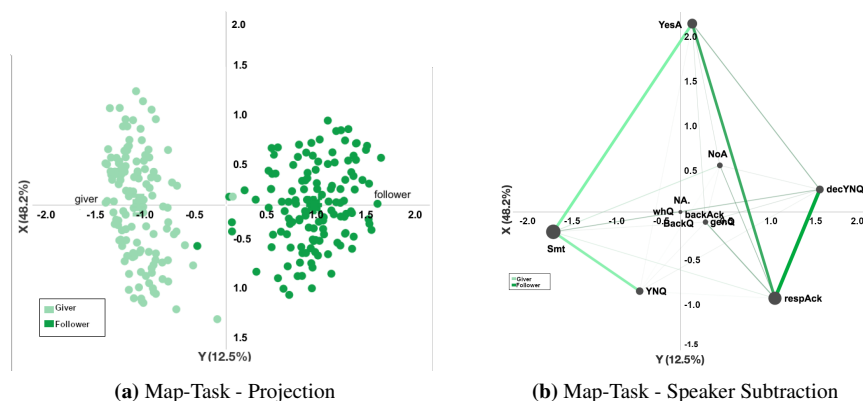
Initially, in order to explore how the DA space differs across corpora, we applied ENA to each corpus individually. Figures 7a and 8a show the projection graphs for Switchboard and Map-Task, respectively. Figures 7b and 8b show the speakers' DA distribution in the context of a subtraction network, showing the natural respective symmetry and asymmetry of these dialogues. Figure 7b shows that there was little to no difference between the DA contribution or context of the speakers in these dialogues. It also indicates the dominance of statements (Smt) within these dialogues, followed by back-channel acknowledgement and yes answers. Figure 8b however, shows a clear difference between the speaker roles, with the instruction giver showing stronger connections to response acknowledgement (RAck), and the follower to statements (Smt).

Comparing Figures 2 to Figures 7 and 8, we can begin to describe the differences in speaker DA space between student ability levels in terms of how they relate to the conversational (figure 7) and task-based dialogue (figure 8).

Something immediately different about the spaces was the position of '*signal non understanding*' (SNA) in relation to the other DAs in each space. For BELC, SNA is a clear outlier in terms of how it was used in the dialogue. In both Switchboard and Map-Task however, it is in a more similar position to DAs such as Wh-Questions (WhQ), and No-Answers (NoA), showing its usage in more similar contexts to these. Additionally, student means are further from SNA at higher ability levels (figure 2(a)), and there is a much stronger connection to SNA for students at lower than higher ability levels in terms of their subtraction graphs (figure 3(b)). We



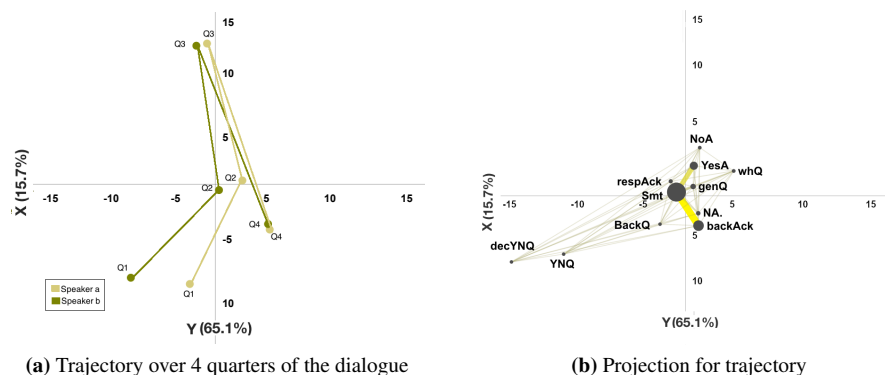
**Fig. 7:** ENA applied to the Switchboard corpus. The means of speaker\_a and speaker\_b in (a) show no significant difference, showing high similarity of DA usage, and thus conversational symmetry. The weight of the DA labels, and thus the distribution of DAs within switchboard show that Statement(Smt) is the dominant DA used between speakers. The means of the speakers are *Speaker a (0.21,0)* *Speaker b(-0.21,0)*, which are not significantly different ( $U= 882.00, p= 0.01, r=0.36$ )



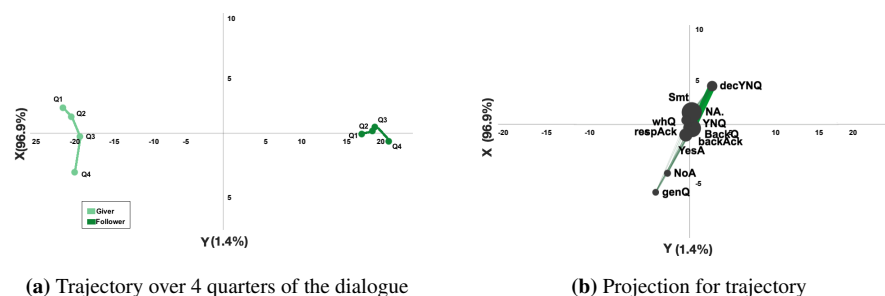
**Fig. 8:** ENA applied to the Map-Task corpus. The means of giver and follower are significantly different, with (a) showing a clear difference in DA usage between speaker roles. The subtraction network for the speakers (b) shows that instruction givers have many more connections between statements, yes-no-questions and yes answers than the followers, and that the follower has more connections between response acknowledgement, declarative yes no questions and yes answers than the giver. The means of the speakers are *giver (-1.09,0)* *follower(1.09,0)*, Which are significantly different ( $U= 5.00, p < 0.001, r=1.00$ )

interpret these differences as indicative of the very different purpose of this DA within L2 language learning, which becomes less important as student ability improves.

Another key difference is how the projection plots are distributed for the corpora. Switchboard has a clear overlap of speakers, and Map-Task a clear separation in terms of where they fall in dialogue space. BELC shows a less pronounced separation than Map-Task, yet still less overlap than Switchboard. However, when looked at separated by levels, speakers in BELC at lower levels of learner ability had a greater separation than those at higher levels. This mirrors the findings of Sinclair et al. [29], in that BELC speakers showed greater symmetry, at higher levels of learner ability than at lower levels.



**Fig. 9:** Switchboard Trajectory analysis



**Fig. 10:** Map-Task Trajectory analysis

In order to discover whether speakers in the two corpora with fluent speakers converge in terms of their DA usage over time, we performed the same trajectory analysis for both Switchboard (figure 9) and Map-Task (figure 10). For Switchboard, the movement through the space was the same for both speakers; clearly, there was a symmetry of conversation. Their movement is reflective of the different stages of the

**Table 8:** Distances between points on the trajectories for Switchboard (Figure 9) and Map-Task (Figure 10)

Corpus Speaker	Switchboard		Map-Task	
	Speaker A	Speaker B	Giver	Follower
Individual Speaker Distance Travelled $\ Q1-Q4\ $	9.07	13.82	8.15	2.86
Speaker Distances Q1		5.57		41.18
Speaker Distances Q2		2.37		41.41
Speaker Distances Q3		1.03		39.49
Speaker Distances Q4		0.53		43.25
Difference in start-end distances $\ dist\_Q1-dist\_Q4\ $		5.04		2.07

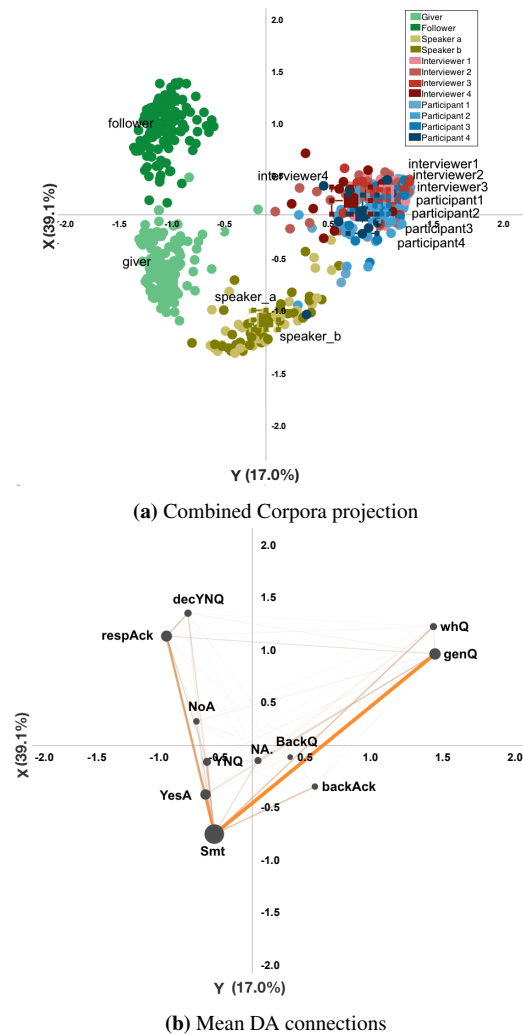
dialogue – the beginning comprising more *questions* and *backchannels* as the speakers get to know one another, and then movement between *yes answers* and *statements* as the conversation progressed. Figure 8 shows convergence within this change in DA usage, with speaker\_a and b being closer in quartile 4 (the end of the dialogue) than they were in quartile 1. This is consistent with what we know about alignment between speakers in conversation [20, 21, 30]. The convergence between speakers in Switchboard is more pronounced than in BELC, with speakers starting off further apart, and converging to a similar distance to the BELC speakers in higher student ability dialogues (Comparing Table 7 and 8). This DA convergence is an interesting finding for switchboard as the overall DA usage per speaker is not significantly different (Figure 7).

For Map-Task, in comparison, there was both no convergence, but movement through the space. The change in DA usage was very different for each speaker role, with Figure 8 showing that the giver travels more than double the distance in DA space than the follower. The instruction changed their interaction the most, moving to a position closer to asking more declarative yes-no questions and making more statements in the space. This might be reflect them becoming closer to their goal, and repeating confirmation style interactions such as *question* serving as an example or *statement* serving as confirmation. The follower, on the other hand, moved very little in the space, remaining closest in space to response acknowledgement. This is in-keeping with their dialogue role, as they must only answer the instruction giver and follow their direction.

### 5.3 Comparison of DA Space for All Corpora

To further address our *third research question* and to discover whether the DA usage within each corpus was significantly different, we combine all three corpora and analyse the means of the speakers within this new combined space of DAs. Figure 11 shows the projection plot and the means of the speakers within each corpus.

Ignoring the underlying DA positions in Figure 11b, it is clear that the usage of the DAs were very different in each dialogue type. The overlap of speakers within Switchboard is very clear, showing the conversational symmetry expected from the DA count statistics reported in Sinclair et al. [29]. The speakers of Map-Task showed a clear asymmetry, also in-keeping with those results. The *giver* in Map-Task showed a more similar position to Switchboard than the *follower*. This possibly indicated

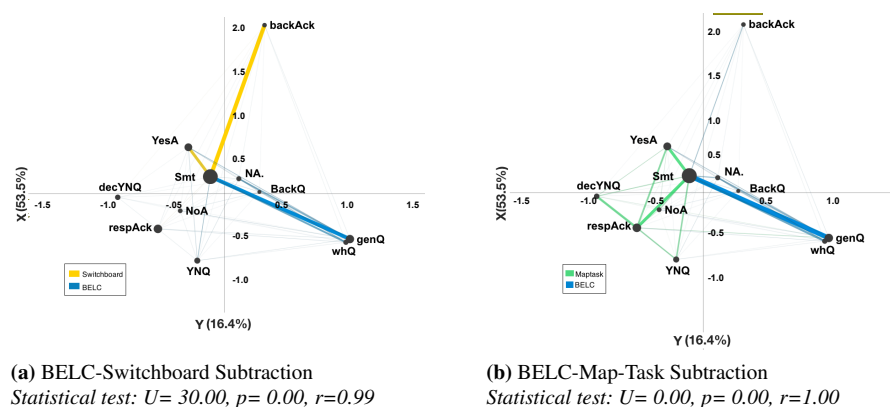


**Fig. 11:** ENA projections showing the three corpora within the same DA space, with BELC dialogues split by student ability level.

the giver's contribution to the instructional dialogue was more conversational and inquiring than that of the instruction follower, who has the more dominant role in the dialogue. The distribution of the BELC speakers in space shows both the overlap of speaker DA usage seen in Switchboard, but also some of the asymmetry seen in Map-Task, although to a lesser degree. Where the overlap is less pronounced, it is the student who is closer in space to the Switchboard speakers, and the tutor who is closer to the speakers in Map-Task. This may reflect the task based nature of tutoring: the tutor has a certain non-conversational agenda to ensure that the student is both being pushed to the boundary of their abilities, but is still supported (adhering to the ZPD). When we look at the projection with BELC split by student ability level (figure 11),

it can be seen that the means of the high student ability dialogues were closer to the position of the fluent dialogues than those of the lower abilities.

To compare in more detail the differences between the DA connections present within BELC to the other fluent corpora, Figures 12a and 12b show the subtraction graphs comparing these. Figure 12a shows a much stronger connection between *statement* and *backchannel-acknowledgement*, and between *statement* and *yes-answer* in Switchboard than there was in BELC. In contrast, the connection between *statement* and *general-other-question* was much stronger for BELC than Switchboard. The comparison of this result to figure 3 shows that for both student and tutor connections, higher student ability dialogues had a stronger connection between *statement* and both *backchannel-acknowledgement* and *yes-answers*, mirroring the stronger connections present in the conversational dialogues. We can also see from the tutor subtraction graph (figure 3(a)) that the connection between *statement* and *general-other-question* was stronger in lower ability student interactions than for high ability ones, suggesting that this dissimilarity between BELC and Switchboard diminishes as student ability increases, leading to better, more fluent conversational interactions.

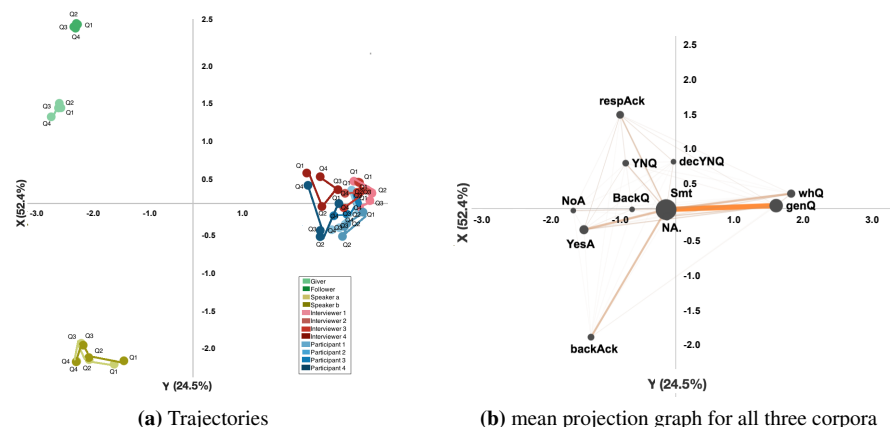


**Fig. 12:** Subtraction graphs comparing the connections between BELC and the other conversational corpora.

In contrast, figure 12b shows that for Map-Task, stronger connections were present between *statement* and response-acknowledgement, and declarative-yes-no-questions, while BELC had stronger connections with *statement* - general-other-questions and *statement* - wh-questions. At a high level, this is a similar difference to that between Switchboard and BELC, in that there seems to be more general *acknowledgement* present in the fluent corpora than in the learner corpus. In BELC, there was a greater emphasis on *general questions*, and *wh-questions*, in comparison to Map-Task where there were more *declarative-yes-no* and *yes-no-questions*. We hypothesise this difference indicates a greater degree of open-ended questioning in the learner scenario, and the succinct minimal response needed questioning of the task based dialogues. The comparison of figure 12b to the connections between BELC at different levels



(figure 3) shows no discernible difference between the connections to *declarative-yes-no*. This finding suggests that *declarative-yes-no* was simply not a DA of particular importance in L2 learner dialogue. However, there was a slight increase in YNQ connections at higher learner abilities for both speakers. There was a clearly stronger connection between statement and response-acknowledgement in higher learner ability dialogues for both speakers, suggesting again that at higher student ability levels, connection strengths become more similar to fluent dialogues.



**Fig. 13:** All Corpora same space Trajectory analysis. Please, can we make sure that the two figures are centred so that the X axes from the parts of the two figure (a and b) follow the same line

Finally, the trajectories of the three corpora in the shared DA space (figure 13) showed that neither Switchboard or Map-Task had much movement in the space, although there was more movement for Switchboard than for Map-Task. The BELC trajectories on the other hand, moved in general towards the centre of the graph, and therefore in the direction of the other corpora as a function of student ability. Showing the trajectories in the shared DA space allowed us to see how the speakers moved in relation both to each other, and compared to the speakers in the different corpora. The BELC speakers at a high level moved in different directions: the student moved towards and the tutor away from the Switchboard corpus.

## 6 Discussion

### 6.1 Convergence and Alignment

This study firstly investigated the different DA usage between student and tutor at different levels of learner ability (RQ1). We found interlocutors' means were closer to one another at higher levels of student ability than at lower ones (Figure 2). This speaker movement within the DA space at different levels of learner ability can in

the case of the tutor, be interpreted as adapting their strategy to meet the needs of the learner ability. Students showed *more* movement in the space across ability levels than tutors, indicating that learner ability influences the sorts of DAs produced. Figure 3 shows how each speaker’s DA usage adapted depending on student ability. As noted in Section 5, lower levels consisted of more *signal-non-understanding (SNA)* and *General-Questions (GENQ)*. By contrast, at higher levels there are more *Wh-Questions (whQ)*, *Response-Acknowledgements (RespAck)* and *Statements (Smt)* which shows a much more active role being taken in the dialogue. Overall, this evidence supports our hypothesis that speakers would have similar DA patterns at higher levels of student ability.

To address our second research question (RQ2), we investigated how student and tutor DA usage changed over the course of a dialogue. When we compared speakers’ position across the four quarters of the dialogue, we found evidence to support our hypothesis that we would see DA convergence over the course of an interaction. Table 7 shows that the student and tutor are closer to each other at the end of the dialogue than at the beginning across levels. There is clear student convergence towards the tutor’s DA distribution as the conversation moves forward at higher levels of student ability (Figure 6) which can be seen by the much higher *student distance travelled* reported in the first row of Table 7. This could indicate that their greater ability allowed them to align more, or that some DAs were simply used less. Movement in the DA space also occurred within tutor dialogue: by the end of the dialogue, the position of the tutor and the student (quartile 4) were very close in the space, although we see the most tutor movement in the case of lower ability students (Table 7, second row). We interpret this as evidence of the tutor’s ZPD strategy: converging when the student cannot, and adapting less when they are more capable.

Socio-Cultural Theory allows us to interpret the change in student movement as the students taking advantage of the context of this interaction to improve their contribution. Sinclair et al. [29] argue that the dialogues at higher levels of ability become more symmetric, mirroring the symmetric contributions of same-level speakers in conversation. We were able to see at a finer grained depth that this was the case for interlocutors’ use of DAs. While Sinclair et al. [30] found some evidence of alignment between speakers at a lexical level, our work demonstrates a technique that allows us to see this at a more abstract level (i.e. DAs) in terms of the conversational dynamics.

Finally, to address our third research question (RQ3), we explored the effects of speaker role on DA usage across fluent corpora, and compared both *DA usage* and *speaker convergence* to that found in our L2 analysis. We also compare all corpora in the same space in order to contrast the speaker DA usage in each corpus to one another. In terms of speaker role influencing DA usage, when we look at the fluent speaker dialogues, there is no interesting difference between speakers in Switchboard (Figure 7), and a clear difference in speaker role for Map-Task (Figure 8). When we contrast this to the differences between speakers across ability levels in BELC, we see that speaker means are closer at higher ability levels, in a manner more similar to Switchboard, and further apart at lower levels, in a manner more similar to Map-Task.

When analysing speaker convergence, speakers show convergence within Switchboard to a minor degree, although their interaction is very similar over the course of

their dialogues (Figure 9). Speakers in Map-Task do not show convergence, although the instruction giver does show movement over the course of a dialogue (Figure 10). The speaker roles in Map-Task may dictate the DA patterns, meaning speakers cannot align at the level of DAs due to this conversational role constraint. In comparison to the fluent corpora, the convergence seen in BELC (Figure 6) is greater than in Switchboard. For the Level 1 and 2 students in BELC, there is little movement, similar to the instruction follower in Map-Task, suggesting a similarly passive role in the conversation.

Our analysis of the DA space of the *combined* corpora (Figure 11) shows that the speakers have a clearly different position in DA space, across corpora, for speakers in Map-Task, and to some degree at higher student ability levels for speakers in BELC. When analysing the trajectories within the combined space, it can be seen that at *higher* levels of student ability, tutors are closer to Map-task speakers in DA space, and students closer to Switchboard speakers. This may indicate that tutors push more able students and ask harder, more involved questions, or, since higher level students can respond more to these questions, tutors make more response acknowledgements (RAck). Students at a higher ability level make more statements, and ask more questions, thus moving towards a more similar interaction style to both of the fluent corpora, across the x axis of Figure 13(a).

## 6.2 Dialogue Act Usage

An additional finding through the analysis of DA space is that the movement of speakers through the space indicates different underlying DA usage patterns. Based on this change, we can see in figure 3 that some DAs are ‘easier’ than others. From Figures 3 and 4, we can see that the use of more *Statements* by both speakers was a sign of a greater ability level, and can interpret this as the student taking a more active role in the dialogue, with the tutor taking a more conversational rather than supportive role. We can also see a change in the types of *Questions* being asked with increased student ability: some questions require more effort to respond to on the part of the student than others, and clearly some take more effort to form. This has implications for teaching and feedback methods in L2 oral instruction: part of the challenge of dialogue is in being able to interact in a contextual and timely manner.

Across corpora, there is a significant difference in DA usage (Figure 11), with BELC speakers closer to *general-other-question* and *wh-question*; Switchboard speakers and the instruction giver from Map-Task closer to *backchannel-acknowledge*, and the instruction follower from Map-Task closer to *response acknowledgement*. At higher levels of student ability both speakers make more acknowledgements: tutors make more *response acknowledgement* which is more common in Map-Task, and both student and tutor make more *backchannel acknowledgement* which is most common to Switchboard (Figure 12). This can be interpreted as follows: at higher levels of student ability, dialogues become more similar to conversational dialogues in general. Both speakers in BELC become more similar to one another as student ability increases, but tutors use DAs more similar to task based dialogue, and stu-

dents, while moving to be more similar to both, become more similar to spontaneous conversation.

## 7 Study contributions and conclusions

The present study contributes a novel method for the analysis of L2 dialogue transcripts, expanding on our previous work to contextualise our L2 corpus findings relative to conversational dialogue between fluent speakers. Our proposed method, combines the use of Dialogue Act labels with ENA. Our findings support the hypothesis that speakers in L2 dialogue practice exhibit convergence, both as ability level increases and over the course of a single dialogue. We also find that DA across different conversational settings is very different, and that with greater student ability, DA usage becomes more similar to other fluent dialogues. We find more significant DA convergence within L2 corpora than within either spontaneous fluent conversation or task-based dialogue. The implications of these results are as follows: firstly, a better understanding of tutor adaptation to learners of different ability levels can inform the design of automated tutoring dialogue systems; secondly, the proposed method can be used to offer formative assessment of learning progression not only to tutors in training, or as a tool for self reflection, but also as a resource for students; and finally, this method can be used by practitioners in learning analytics for the design of new tools for dialogue across different dialogue modalities.

While our proposed method provides evidence of tutor-student convergence, there are some accuracy limitations to the DA labels which were derived automatically [29], constraining our analysis to a certain degree of granularity. Our L2 corpus is also not large or diverse enough for us to make generalisations about particular dialogue characteristics at certain levels of student ability; accordingly, we limit our interpretation to higher level convergence and adaptation phenomena. We also limit our analysis of the combined corpora to high level differences, as there are differences in the length of the dialogues and the proportion of the dialogue acts which will influence speakers positions relative to each other for each corpus.

An important avenue of future research is to explore the functions that certain DAs perform within dialogue, and the associated difficulty of such acts. The shift in L2 speaker position in the DA space suggests a move to using different sorts of DA patterns to better suit the ability of the student, both moving closer to fluent speaker positions with increasing student ability. This leads us to hypothesise that certain DA sequences may be more indicative of scaffolding, others of the conversational symmetry seen in Switchboard and others of the conversational asymmetry in Map-Task. Identifying these sequences is therefore of great interest. We also intend to investigate other aspects of alignment such as the use of code switching<sup>6</sup>. We also intend to make further use of the dialogue within the code-switch utterances: currently the work focuses only on the English L2 quotient of the data, but expanding our analysis to the use of the L1 could bring greater understanding as to how the L1 is used as a cognitive tool in this setting.

---

<sup>6</sup> Where speakers switch to the L1 of the learner

## References

1. Anderson AH, Bader M, Bard EG, Boyle E, Doherty G, Garrod S, Isard S, Kowtko J, McAllister J, Miller J, et al. (1991) The hcr map task corpus. *Language and speech* 34(4):351–366
2. Bailey KM (2001) What my efl students taught me. *The PAC Journal* 1(1):7–31
3. Birjandi P, Jazebi S (2014) A comparative analysis of teachers' scaffolding practices
4. Bloom BS (1984) The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher* 13(6):4–16
5. Branigan HP, Pickering MJ, McLean JF, Cleland AA (2007) Syntactic alignment and participant role in dialogue. *Cognition* 104(2):163–197
6. Branigan HP, Pickering MJ, Pearson J, McLean JF (2010) Linguistic alignment between people and computers. *Journal of Pragmatics* 42(9):2355–2368
7. Chen L, Di Eugenio B, Fossati D, Ohlsson S, Cosejo D (2011) Exploring effective dialogue act sequences in one-on-one computer science tutoring dialogues. In: *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, Association for Computational Linguistics, Portland, Oregon, pp 65–75, URL <https://www.aclweb.org/anthology/W11-1408>
8. Cohen J (1992) A power primer. *Psychological bulletin* 112(1):155
9. Costa A, Pickering MJ, Sorace A (2008) Alignment in second language dialogue. *Language and cognitive processes* 23(4):528–556
10. Duplessis GD, Clavel C, Landragin F (2017) Automatic measures to characterise verbal alignment in human-agent interaction. In: *18th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pp 71–81
11. Garrod S, Pickering MJ (2007) Alignment in dialogue. *The Oxford handbook of psycholinguistics* pp 443–451
12. Godfrey JJ, Holliman EC, McDaniel J (1992) Switchboard: Telephone speech corpus for research and development. In: *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on, IEEE*, vol 1, pp 517–520
13. Hawkes R (2012) 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
14. Jurafsky D, Shriberg E, Biasca D (1997) Switchboard dialog act corpus. *International Computer Science Inst Berkeley CA, Tech Rep*
15. Lantolf JP (2000) Second language learning as a mediated process. *Language teaching* 33(2):79–96
16. Lantolf JP (2000) *Sociocultural theory and second language learning*, vol 78. Oxford University Press
17. Mandel J (1982) Use of the singular value decomposition in regression analysis. *The American Statistician* 36(1):15–24
18. Michel MC (2011) Effects of task complexity and interaction on l2 performance. *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* 2:141–173

19. Muñoz C (2006) Age and the rate of foreign language learning, vol 19. *Multilingual Matters*
20. Pickering MJ, Garrod S (2004) Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences* 27(2):169–190
21. Reitter D, Moore JD (2014) Alignment and task success in spoken dialogue. *Journal of Memory and Language* 76:29–46
22. Reitter D, Keller F, Moore JD (2006) Computational modelling of structural priming in dialogue. In: *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, Association for Computational Linguistics, Stroudsburg, PA, USA, NAACL-Short '06*, pp 121–124, URL <http://dl.acm.org/citation.cfm?id=1614049.1614080>
23. Robinson P, Gilabert R (2007) Task complexity, the cognition hypothesis and second language learning and performance. *IRAL-International Review of Applied Linguistics in Language Teaching* 45(3):161–176
24. Samana W (2013) Teacher's and students' scaffolding in an efl classroom. *Academic Journal of Interdisciplinary Studies* 2(8):338
25. Searle JR, Searle JR (1969) *Speech acts: An essay in the philosophy of language*, vol 626. Cambridge university press
26. Shaffer DW, Graesser A (2010) Using a quantitative model of participation in a community of practice to direct automated mentoring in an ill-formed domain. In: *Intelligent Tutoring Systems Conference, Pittsburgh, PA*
27. Shaffer DW, Hatfield D, Svarovsky GN, Nash P, Nulty A, Bagley E, Frank K, Rupp AA, Mislevy R (2009) Epistemic Network Analysis: A Prototype for 21st-Century Assessment of Learning. *International Journal of Learning and Media* 1(2):33–53, DOI 10.1162/ijlm.2009.0013
28. Shaffer DW, Hatfield D, Svarovsky GN, Nash P, Nulty A, Bagley E, Frank K, Rupp AA, Mislevy R (2009) Epistemic network analysis: A prototype for 21st-century assessment of learning. *International Journal of Learning and Media* 1(2)
29. Sinclair A, Oberlander J, Gasevic D (2017) Finding the zone of proximal development: Student-tutor second language dialogue interactions. In: *Proc. SEM-DIAL 2017 (SaarDial) Workshop on the Semantics and Pragmatics of Dialogue*, pp 107–115
30. Sinclair A, Lopez A, Lucas C, Gasevic D (2018) Does ability affect alignment in second language tutorial dialogue? In: *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pp 41–50
31. Sinclair A, Ferreira R, Lopez A, Lucas C, Gasevic D (2019) I wanna talk like you: Speaker adaptation to dialogue style in l2 practice conversation. In: *Proceedings of Artificial Intelligence in Education - 20th International Conference*
32. Stolcke A, Ries K, Coccaro N, Shriberg E, Bates R, Jurafsky D, Taylor P, Martin R, Ess-Dykema CV, Meteer M (2000) Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics* 26(3):339–373
33. Van Lier L (1996) *Interaction in the language classroom: Awareness, autonomy and authenticity*. London: Longman
34. Vygotsky L (1987) Zone of proximal development. *Mind in society: The development of higher psychological processes* 5291:157

- 
35. Ward A, Litman D (2007) Dialog convergence and learning. *Frontiers in Artificial Intelligence and Applications* 158:262
  36. Ward A, Litman D (2007) *Measuring convergence and priming in tutorial dialog*. University of Pittsburgh

## 5.2 Contributions & Discussion

In Sinclair, Ferreira, Lopez, Lucas & Gasevic (2019) and our extended journal version in the previous section, we showed evidence of the following:

- We find that DA usage was more similar between tutor and student at higher levels of student ability than at lower levels (Figure 1 in Sinclair, Ferreira, Lopez, Lucas & Gasevic (2019) and Figure 2 in our journal version in the previous section)
- We find convergence over the course of a dialogue in the usage of DAs between speakers in L2 tutorial dialogue, with greater adaptation of the tutor at lower student ability levels, and greater adaptation from the student at higher ability levels (Figure 2 in Sinclair, Ferreira, Lopez, Lucas & Gasevic (2019), Figure 6 in the previous section)
- Comparing to fluent dialogue, L2 speakers at higher levels of student ability, have more similar DA usage to fluent speakers in both conversational corpora (Figure 11 in the previous section)
- Convergence over the course of a dialogue can be seen for fluent conversational dialogue, but is not present in task based (Figures 9 and 10 in the previous section)
- The convergence in L2 is less pronounced than in conversational dialogue (Tables 7 and 8 in the previous section) but exhibits a greater degree of change in terms of the DAs used (Figures 5, 6 and 9 in the previous section)

We hypothesise that the findings above may be evidence of tutor ZPD strategy: converging when the student cannot, and adapting less when they are more capable.

Limitations of our study are firstly that the DA labels used (Sinclair et al. 2017) have some accuracy limitations, therefore we limit our trajectory analysis to comparing speaker behaviour across quartiles, rather than at a lower level of granularity. Secondly, the L2 corpus used is not large or diverse enough for us to make generalisations about particular dialogue characteristics at certain levels, rather we limit our interpretation to higher level convergence and adaptation phenomena.

The present study contributes a novel method for the analysis of L2 dialogue transcripts. Our proposed method, from which we provide preliminary results, combines



the use of Dialogue Act labels with ENA. Our findings support the hypothesis that speakers in L2 dialogue practice exhibit a degree of convergence, both as ability level increases and over the course of a single dialogue. The implications of these results are as follows: firstly, a better understanding of tutor adaptation to learners of different ability levels can inform the design of automated tutoring dialogue systems; secondly, the proposed method can be used to offer formative assessment of learning progression not only to tutors in training, or as a tool for self reflection, but also as a resource for students; and finally, this method can be used by practitioners in learning analytics for the design of new tools for dialogue across different dialogue modalities.

An important avenue of future research is to explore the functions that certain DAs perform within dialogue, and the associated difficulty of such acts. The shift in speaker position in DA space suggests a move to using different sorts of DA patterns to better suit the ability of the student. This leads us to hypothesise that certain DA sequences may be more indicative of scaffolding, and others of conversational symmetry. Identifying these sequences is therefore of great interest. We also intend to investigate other aspects of alignment such as the use of code switching, where speakers switch to the L1 of the learner. We also intend to make further use of the dialogue within the code-switch utterances: currently the work focuses only on the English L2 quotient of the data, and expanding our analysis to the use of the L1 could bring greater understanding as to how the L1 is used as a cognitive tool in this setting.

# Chapter 6

## Human-Agent Alignment

The previous three chapters (3, 4 and 5) compare adaptation between human speakers in L2 dialogue practice to that in fluent conversational and task based dialogues. In this chapter we compare adaptation of an L2 student to their tutor in *Human-Human* and *Human-Agent L2* corpora. As discussed in Chapter 4, alignment in an L2 context can be indicative of vocabulary learning through the student repeating unfamiliar keywords to learn them. We hypothesise that this may also be true in Human-Agent dialogues, where the tutor is a dialogue agent.

We apply measures of alignment at the level of expressions to L2 tutoring corpora, comparing the effects found in dialogues between human tutor and student to those found between an automated tutor and student. Section 6.1 contains extra discussion of the types of expressions aligned to. Understanding how alignment occurs within Human-Human dialogue allowed us to contextualise patterns of alignment within Human-Agent dialogue, and explore outliers in a large corpus collected from students using Babbel, an online platform for second language learning.

We found students, who may be limited by their constrained linguistic ability, would align to the agent in the H-A dialogue, more than they would by chance. We also revealed that there was evidence of greater variability of student alignment within the H-A corpus than in H-H. We hypothesise that this is due to different levels of student engagement.

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

This section includes the verbatim copy of the following publication:

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: The 12th International Conference on Educational Data Mining*, 2019, pp. 414 - 419

**Contributions:** The ideas and analysis in the paper were developed and discussed between all authors of the work, with special emphasis on the contribution of the second author to both development and refinement of the idea and the writing. The original idea, the experiments and the bulk of the writing were the work of the first author.

# 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  
lopez@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 an 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: *The 12th International Conference on Educational Data Mining*, Michel Desmarais, Collin F. Lynch, Agathe Merceron, & 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

**Table 1: Tutorbot dialogue example. Italics indicate Expression Repetition**

1.	bot:	What is your <i>favorite day of the week</i> ?
2.	user:	My <i>favorite day of the week</i> is Friday ...
3.	bot:	Do you play sports ?
4.	user:	yes
5.	bot:	What sport do you <i>play</i> ?
6.	user:	I <i>play</i> 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

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.

## 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<sup>1</sup>. 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-

<sup>1</sup>This data was kindly shared with us by Babbel, <https://www.babbel.com/>

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

	<b>Tutorbot</b>	<b>BELC</b>
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.

$$EV = \frac{\text{length}(EL)}{\text{numberoftokens}}$$

**Expression Repetition – speaker** (ER<sub>S</sub>) 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<sub>S</sub>) are the established expressions initiated by S

**Vocabulary Overlap** (VO) is the ratio of shared tokens between interlocutors S<sub>1</sub> and S<sub>2</sub>. The higher the VO, the more vocabulary is shared between speakers.

$$VO = \frac{(\text{Tokens}_{S_1} \cap \text{Tokens}_{S_2})}{(\text{Tokens}_{S_1} \cup \text{Tokens}_{S_2})}$$

## 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<sub>S</sub> 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<sub>S</sub>. 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<sup>2</sup>, 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**

	<b>Tutorbot</b>	<b>BELC</b>
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

<sup>2</sup><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.

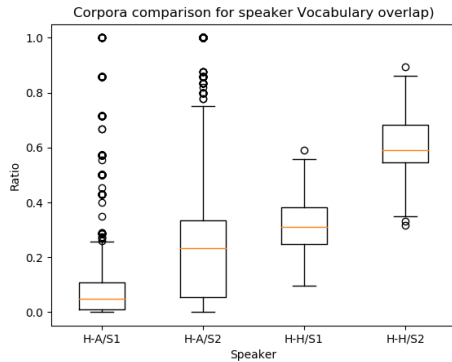


Figure 1: H-H/A corpora Vocabulary Overlap. Speaker difference was significant for H-A ( $p < 0.0001$ ) ( $statistic = 6.42, pvalue = num1.4e - 10$ ) and H-H ( $p < 0.001$ ) ( $statistic = -2.11, pvalue = 0.00036$ ) S1 = Tutor/Agent, S2 = Student

## 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<sub>R</sub> and H-A<sub>R</sub> 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<sub>R</sub> and H-A<sub>R</sub> 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

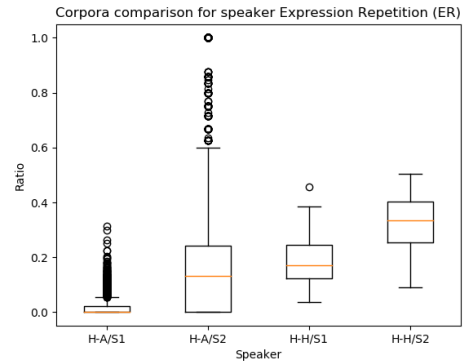
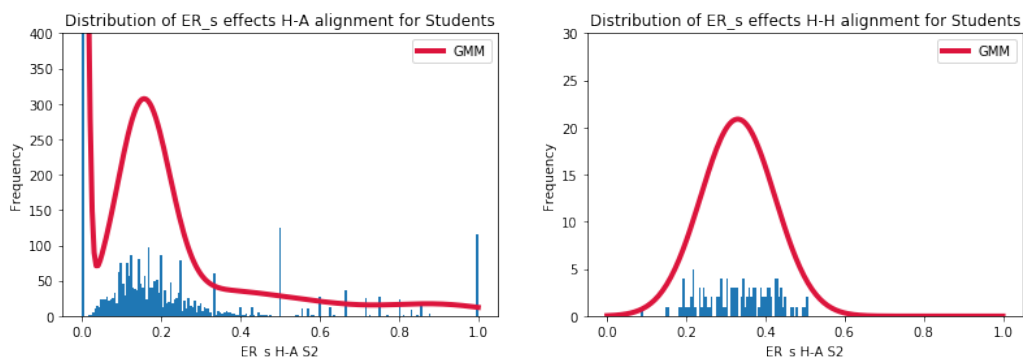


Figure 2: H-H/A corpora ER.s. Speaker difference is significant for H-A ( $p < 0.0001$ ) ( $statistic = -44.91, pvalue = 0.0$ ) and H-H ( $p < 0.0001$ ) ( $statistic = -12.71, pvalue = 1.77 \times 10^{-28}$ ) S1 = Tutor/Agent, S2 = Student

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 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	<b>no-response or request for help:</b> students either do not engage with the agent, or demonstrate inability to engage
0.1-0.15	<b>minimal response:</b> students respond curtly, appear less engaged <i>bot: What is your favorite day of the week ?</i> <i>user: That 's Sunday .</i>
0.25-0.4	<b>high engagement:</b> 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 <i>bot: Do [you] _have a_ boyfriend or _a_ girlfriend ? Or _a_ husband_ or _a_ wife ?</i> <i>user: I [have [a] husband] .</i>
0.5-0.55	<b>minimal response:</b> low rate of student production, typical response one high-frequency word, low engagement despite high alignment <i>'hi', 'bye.'</i>
0.85-0.9	<b>high repetition:</b> all student responses are rephrases, dialogues very short <i>bot: _Hello_ , _nice to see you_ !</i> <i>user: [Hello] [nice to see you] too</i>

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 Babel and specifically Zach Sporn and Joel Kieseay 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. FranÁgois, 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. ChrÁltien. 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.

## 6.2 Further Discussion

Sinclair, McCurdy, Lopez, Lucas & Gasevic (2019) present a comparison of the alignment present in Human-Human vs Human-Agent learner dialogues at the level of Expressions. Expressions allow for the analysis of the sorts of multi-word phrases which are being re-used, something that can be of particular interest for L2 tutoring if used to interpret a tutor's goals or strategy. Expressions can also be learnt by students in the same manner as vocabulary and could indicate students more explicitly repeating phrases to help themselves learn.

We provide examples and qualitative groupings of the main types of multi-token expressions present in the student language which due to space constraints was not included in Sinclair, McCurdy, Lopez, Lucas & Gasevic (2019). We hypothesise that the repetition of these types of expressions may indicate learning strategy on the part of the student. Knowing what the student is likely to repeat could have implications for the choice of tutoring language in agent design.

### Expression Lexicon

The Expression lexicon is the set of expressions which are shared between speakers. Table 6.1 shows a selection of the most common of the longer (length  $\geq 2$ ) expressions within Tutorbot since these showed more routinisation, and were less likely to be as a result of shared vocabulary (e.g. 'yes', 'ok' 'and' and other stopwords). 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 **favourite day of the week** ? user—5: My [favourite 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. There may be specific aspects of the tutor interaction in the BELC which we attribute to tutor alignment to student such as their confirmation/acknowledgement actions of repeating what the student says, 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 6.1: two main types of expressions being aligned to by students in the Tutorbot corpus

Reform question in response	'have a', 'are their names', 'a wife', 'a girlfriend', 'a husband', 'a boyfriend or a', 'favourite day of the week', 'of the week', 'brothers and sisters'
Reflect question back in same form	'you ! How are you ?', 'Do you have a', 'How are you ?', 'How old are you ?', 'Where are', 'How are you', 'you ! How are you', 'How old', do you live'

### 6.3 Contributions

In Sinclair, McCurdy, Lopez, Lucas & Gasevic (2019) we find evidence of the following:

- We find significant H-A alignment (Figure 2 in Sinclair, McCurdy, Lopez, Lucas & Gasevic (2019)). 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 find asymmetrical alignment within H-H L2 dialogues (Figure 2 in Sinclair, McCurdy, Lopez, Lucas & Gasevic (2019)), which was in keeping with results reported on lexical priming for the same corpus in which we found the strongest priming effects are those from student to tutor (Sinclair et al. 2018).
- 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, shown in Figure 3 in Sinclair, McCurdy, Lopez, Lucas & Gasevic (2019))
- We see differences in alignment across dialogues (Figure 3 in Sinclair, McCurdy, Lopez, Lucas & Gasevic (2019)), suggesting different patterns of student engagement, could be detected via their alignment levels. Outliers showed differ-

ent signs of student engagement which are summarised in Table 4 in Sinclair, McCurdy, Lopez, Lucas & Gasevic (2019).

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 intervention 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.

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 (Duplessis et al. 2017) and task-based dialogues (Braniġan et al. 2010). 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. There is a large difference between spoken and typed dialogue in terms of non-verbal cues, repetition, listening/speaking vs. reading/spelling to name a few, which may have an effect on alignment due to contextually different language styles. It should also be noted that the dialogue length is typically shorter in H-A dialogues, with greater variability. 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 and to explore any effects of dialogue length on the alignment metric reported. 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 ER<sub>S</sub> 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 ER<sub>S</sub> 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.



# Chapter 7

## Conclusion

Automatic adaption to an individual's zone of proximal development is not yet possible for L2 tutoring agents. This thesis analyses tutor adaption to L2 learners' linguistic ability, alignment, and dialogue style in order to better model how tutors leverage their adaption and that of their students within their teaching. This analysis is performed both on free form face to face spoken dialogue, and on naturalistic computer-mediated instant message dialogues between tutor or L1 speaker and L2 learner. This thesis also analyses student adaptation at different levels of ability, to understand common patterns of student interaction and engagement. These patterns of student adaptation to a human tutor are compared with how students adapt to an L2 tutoring dialogue agent. We learned that both tutors and students in L2 practice dialogue do adapt their language to one another, and adaptation changes with learner ability. We found that L2 students align to a tutoring agent, although in a more varied way than with a human tutor.

### 7.1 Summary of Contributions

The goal of this thesis was to explore how speakers adapt in an L2 teaching and learning context to arrive at a better understanding of how adaptation changes with learner ability. We found empirical evidence of adaptation, providing some concrete examples which corroborate many theories of L2 dialogue interaction. We found that tutors and students adapted to one another over the course of an interaction in terms of linguistic complexity, lexical alignment and dialogue act usage. We also found evidence of student lexical alignment to an L2 tutor agent. We hypothesised tutor adaptation was a sign of teaching strategy, and student adaptation an indication of learning. These

findings show that personalised adaptation is important and have implications both for the design of automated L2 tutoring agents, and for other pedagogical applications, as a learning analytic tool.

The contributions of this thesis are:

- We found evidence of and showed that tutor adaptation of their **linguistic complexity** could be measured via a set of surface features in the text within the utterances. We showed that utterances with different functions exhibited different complexity traits. This work is published in (Sinclair et al. 2017), and presented in Chapter 3.
- We expanded previous work on measuring **lexical alignment** effects within conversational and task based dialogue to include that of second language. We showed that alignment correlates with student ability, and that alignment between speakers was asymmetric in L2 dialogues, with students aligning more to tutors than vice versa. Finally, we found that students at lower ability levels aligned to ‘harder’ words more so than students at higher ability levels, which we hypothesised may be evidence of their leverage of alignment for vocabulary assimilation. This work is published in (Sinclair et al. 2018), and presented in Chapter 4.
- We showed that the usage of **Dialogue Acts** in L2 dialogue changes for both speakers as a function of student ability. We also demonstrated interlocutor convergence over the course of a dialogue in terms of dialogue act use. We proposed using dialogue acts as an additional feature when observing alignment at a higher level. This work is published in (Sinclair, Ferreira, Lopez, Lucas & Gasevic 2019), and presented in Chapter 5.
- We extend (Sinclair, Ferreira, Lopez, Lucas & Gasevic 2019) to include analysis of fluent conversational dialogues, contrasting convergence and Dialogue act use to that found in the L2 learner dialogues. This new work is under submission to the *International Journal of Artificial Intelligence in Education (IJAIED)* and presented in Chapter 5.
- Finally, we compared alignment in **human-human vs human-agent** second language learner corpora. We used alignment as a means to examine different user engagement behaviour within the human-agent corpus, and offer a discussion

on the qualitative differences between these corpora. This work is published in (Sinclair, McCurdy, Lopez, Lucas & Gasevic 2019), is presented in Chapter 6.

Beyond the concrete contributions of the adaptation we have examined, the more general contribution lies in the comparison of L2 dialogue to other conversational dialogue in order to contextualise our findings within how dialogue interactions happen in the wild. This approach allows us to be more confident in our L2 specific findings and better understand them through contrasting them to fluent speakers.

## 7.2 Implications

These findings have some implications for the design of an automatic L2 tutor: we see that human tutors adapt complexity to remain within a certain accessible range of the learner's ability. Since we can model this adaptation, the same features could be used as criteria to optimise dialogue generation. These features are lightweight and surface level; therefore, given a certain dialogue history, the generation of the next tutor utterance could be a function of learner level prediction, or to maximise the probability that the learner will respond given previous similarly complex prompts in the past.

Better understanding of tutor adaptation to learners of different ability levels can inform the design of automated tutoring dialogue systems; the method in Sinclair, Ferreira, Lopez, Lucas & Gasevic (2019) can be used to offer formative assessment of learning progression not only to tutors in training, or as a tool for self reflection, but also as a resource for students; finally, this method can be used by practitioners in learning analytics for the design of new tools for dialogue across different dialogue modalities.

Our work in Sinclair, McCurdy, Lopez, Lucas & Gasevic (2019) 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 ER<sub>S</sub> 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.



### 7.3 Future Directions

The potential future directions to our work fall under two main themes. Firstly, modelling speaker adaptation and alignment in dialogue can be taken further. As measuring linguistic complexity is a field of research within itself, it would be interesting to measure adaptation of different features of linguistic complexity at a discourse level, taking into account the utterance sequence and other markers of student ability than at the level of text that we explored in this thesis. Modelling alignment at different levels of interaction in this context is another aspect of our work which can be taken further. We explored linguistic complexity, lexical, and dialogue style alignment, but combining these, or measuring semantic alignment is the next step to better understand the dynamics of learner-tutor interaction. For example, some DAs may be more complex than others, and this should be taken into account when measuring the linguistic complexity of an utterance; if a learner has not aligned to some of the tutor's vocabulary this could be a good indication that they do not understand; and if the learner is able to introduce vocabulary to the dialogue independent of the tutor, this can be an indication of more competence than their re-use of the tutor's language. Measuring semantic alignment could be useful in an online setting in order to check that a student is not gaming the system and that their interaction is relevant to the context. All these aspects of adaptation combined could create a more general model of L2 adaptation, which has the scope to improve personalised L2 tutoring agents.

#### L1 Effects

This thesis uses a Human-Human corpus where the learner L1 is Spanish/Catalan. In future experiments, exploring the effects of L1 on the differences between linguistic complexity exhibited between tutor and student, alignment, and dialogue style could yield greater variation in results. We hypothesise that the main trends found, i.e. that tutors push higher ability students, and simplify for low ability students would remain consistent, but that they would manifest themselves to a different degree depending on the linguistic distance between the L1 and L2 used. Linguistic distance (difference between the L1 and L2) shows that learning an L2 is easier for some L1s than for others (Chiswick & Miller 2005). For example, as Spanish (L1) and English (L2) are both Indo-European languages, some lexical features are shared and may therefore be less indicative of concepts learnt. This may mean that alignment to these 'easier' items be-

tween students and tutors is different than for the same items when the L1 is Arabic or Mandarin, where there is less similarity between the L1 and L2. Exploring hypotheses about how adaptation occurs within different language pairings experimentally could lead to a broader understanding of the diverse range of alignment and adaptation phenomena present in L2 learning.

### **Linguistic Complexity**

In Sinclair et al. (2017), we explore linguistic complexity differences between the first and second halves of the dialogue, it would be interesting to compare complexity adaptation at the level of utterances. Our model would not allow us to investigate this at such a low level of granularity, therefore developing better automatic complexity prediction tools designed for L2 dialogue is a useful future avenue of research. Knowing how complex utterances are in comparison to each other could be useful in automatically identifying when a tutor rephrases a question to make it easier for a student to understand, or in the case of the student, identifying self-correction behaviours.

### **Lexical Alignment**

In Sinclair et al. (2018), we do not explore self-priming as a possible tutoring strategy: a tutor may repeat a word several times to reinforce a concept to a learner. This is an interesting possible feature for using alignment to understand dialogue structure. In future work, we plan to investigate different measures of alignment and both lexical and syntactic complexity to inform systems that aim to automate L2 tutoring. We plan to consider which speaker *introduces* the word being aligned to, in order to better understand the relationship between productive and receptive vocabulary of the student in dialogue settings.

### **Dialogue Style**

From Sinclair, Ferreira, Lopez, Lucas & Gasevic (2019) and our paper in Chapter 5, an important avenue of future research is to explore the functions that certain DAs perform within dialogue, and the associated difficulty of such acts. The shift in speaker position in DA space suggests a move to using different sorts of DA patterns to better suit the ability of the student. This leads us to hypothesise that certain DA sequences may be more indicative of scaffolding, and others of conversational symmetry. Identi-

fyng these sequences is therefore of great interest. We also intend to investigate other aspects of alignment such as the use of code switching, where speakers switch to the L1 of the learner. We also intend to make further use of the dialogue within the code-switch utterances: currently the work focuses only on the English L2 quotient of the data, and expanding our analysis to the use of the L1 could bring greater understanding as to how the L1 is used as a cognitive tool in this setting.

## **Wider Context**

There is scope for using our results in applied learning analytics, in the manner of Chapter 6. Improving computer aided language learning is an active field, and with massive open online courses (MOOCs) becoming ubiquitous, it is vital that we develop automatic methods for understanding learners' needs and interactions. The methods we demonstrated in chapters 3, 4, and 5 could be incorporated in learning analytic dashboards as teaching tools for educational dialogue, allowing tutors to reflect on their lessons after the fact, and plan their teaching to address specific student DA patterns. These techniques can also be used in the analysis and evaluation of more automatic tutoring tools such as in Chapter 6. The better understanding of tutoring adaptation that we provide can also be used in the design of more personalised learner experiences with L2 tutoring agents. This ideal tutoring agent would: tailor the linguistic complexity of their interactions to fall within the learner's zone of proximal development; leverage alignment within their interactions to make the learner feel more at ease, and to scaffold vocabulary; and adapt their dialogue style to encourage the learner to take a more active role in the dialogue, encouraging the student to ask their own questions, or simply to interact via a more diverse range of dialogue acts. An agent which can react to learner needs in context gives L2 students the chance to take control of their own learning, and learn at their own pace, making this technology less one size fits all, and therefore more inclusive.

# Bibliography

- Abbott, S. (2014), 'Hidden curriculum', *The glossary of education reform*. .  
**URL:** <http://edglossary.org/hidden-curriculum>
- Ahmadian, M., Amerian, M. & Tajabadi, A. (2014), 'The effect of collaborative dialogue on efl learner's vocabulary acquisition and retention', *International Journal of Applied Linguistics and English Literature* **3**(4), 38–45.
- Al Hosni, S. (2014), 'Speaking difficulties encountered by young efl learners', *International Journal on Studies in English Language and Literature (IJSELL)* **2**(6), 22–30.
- Aluisio, S., Specia, L., Gasperin, C. & Scarton, C. (2010), Readability assessment for text simplification, in 'Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications', IUNLPBEA '10, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1–9.  
**URL:** <http://dl.acm.org/citation.cfm?id=1866795.1866796>
- Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J. et al. (1991), 'The hrc map task corpus', *Language and speech* **34**(4), 351–366.
- Baayen, R. H., Piepenbrock, R. & van H, R. (1993), 'The {CELEX} lexical data base on {CD-ROM}'.
- Bailey, K. M. (2001), 'What my efl students taught me', *The PAC Journal* **1**(1), 7–31.
- Benson, P. & Voller, P. (2014), *Autonomy and independence in language learning*, Vol. 1, Routledge.
- Bibauw, S., François, T. & Desmet, P. (2019), '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.  
**URL:** <https://doi.org/10.1080/09588221.2018.1535508>

- Birjandi, P. & Jazebi, S. (2014), 'A comparative analysis of teachers' scaffolding practices', *International Journal of Language and Linguistics* **2**(3), 154–164.
- Boyer, K. E., Phillips, R., Ingram, A., Ha, E. Y., Wallis, M., Vouk, M. & Lester, J. (2011), 'Investigating the relationship between dialogue structure and tutoring effectiveness: a hidden markov modeling approach', *International Journal of Artificial Intelligence in Education* **21**(1-2), 65–81.
- Branigan, H. P., Pickering, M. J. & Cleland, A. A. (2000), 'Syntactic co-ordination in dialogue', *Cognition* **75**(2), B13–B25.
- Branigan, H. P., Pickering, M. J., Pearson, J. & McLean, J. F. (2010), 'Linguistic alignment between people and computers', *Journal of Pragmatics* **42**(9), 2355–2368.
- Briscoe, T., Medlock, B. & Andersen, Ø. (2010), 'Automated assessment of esol free text examinations', *University of Cambridge Computer Laboratory Technical Reports* **790**.
- Brysbaert, M., Lange, M. & Wijnendaele, I. V. (2000), 'The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition: Further evidence from the dutch language', *European Journal of Cognitive Psychology* **12**(1), 65–85.
- Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O. & Gašić, M. (2018), 'Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling', *arXiv preprint arXiv:1810.00278*.
- Bunt, H. (2006), Dimensions in dialogue act annotation., in 'LREC', pp. 919–924.
- Bunt, H., Petukhova, V. & Fang, A. C. (2017), Revisiting the iso standard for dialogue act annotation, in 'Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)'.
- Capel, A. (2012a), 'Completing the english vocabulary profile: C1 and c2 vocabulary', *English Profile Journal* **3**.
- Capel, A. (2012b), 'Completing the english vocabulary profile: C1 and c2 vocabulary', *English Profile Journal* **3**, e1.
- Chen, L., Di Eugenio, B., Fossati, D., Ohlsson, S. & Cosejo, D. (2011), Exploring effective dialogue act sequences in one-on-one computer science tutoring dialogues, in

'Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications', Association for Computational Linguistics, Portland, Oregon, pp. 65–75.

**URL:** <https://www.aclweb.org/anthology/W11-1408>

Chen, X. & Meurers, D. (2017), 'Word frequency and readability: Predicting the text-level readability with a lexical-level attribute', *Journal of Research in Reading* pp. n/a–n/a. JRIR-2017-01-0006.R1.

**URL:** <http://dx.doi.org/10.1111/1467-9817.12121>

Chiswick, B. R. & Miller, P. W. (2005), 'Linguistic distance: A quantitative measure of the distance between english and other languages', *Journal of Multilingual and Multicultural Development* **26**(1), 1–11.

**URL:** <https://doi.org/10.1080/14790710508668395>

Clark, H. H. & Wilkes-Gibbs, D. (1986), 'Referring as a collaborative process', *Cognition* **22**(1), 1–39.

Cohen, R. (1987), 'Analyzing the structure of argumentative discourse', *Computational linguistics* **13**(1-2), 11–24.

Coltheart, M. (1980), 'Mrc psycholinguistic database user manual: Version 1', *Birkbeck College* .

Cook, V., Bassetti, B., Kasai, C., Sasaki, M. & Takahashi, J. A. (2006), 'Do bilinguals have different concepts? the case of shape and material in japanese l2 users of english', *International journal of bilingualism* **10**(2), 137–152.

Core, M. G. & Allen, J. (1997), Coding dialogs with the damsl annotation scheme, Vol. 56.

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

Crossley, S. A., Greenfield, J. & McNamara, D. S. (2008), 'Assessing text readability using cognitively based indices', *Tesol Quarterly* **42**(3), 475–493.

de la Colina, A. A. & Mayo, M. d. P. G. (2009), 'Oral interaction in task-based efl learning: The use of the l1 as a cognitive tool', *IRAL-International Review of Applied Linguistics in Language Teaching* **47**(3-4), 325–345.

- Dixon-Krauss, L. (1996), *Vygotsky in the Classroom: Mediated Literacy Instruction and Assessment.*, ERIC.
- Duplessis, G. D., Clavel, C. & Landragin, F. (2017), Automatic measures to characterise verbal alignment in human-agent interaction, in '18th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)', pp. 71–81.
- Dzikovska, M., Steihauser, N., Farrow, E., Moore, J. & Campbell, G. (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.
- Farr, J. N., Jenkins, J. J. & Paterson, D. G. (1951), 'Simplification of flesch reading ease formula.', *Journal of applied psychology* **35**(5), 333.
- Ferreira, A., Moore, J. D. & Mellish, C. (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.
- Flesch, R. (1948), 'A new readability yardstick.', *Journal of applied psychology* **32**(3), 221.
- Forbes-Riley, K. & Litman, D. (2012), Adapting to multiple affective states in spoken dialogue, in 'Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue', Association for Computational Linguistics, pp. 217–226.
- Forsythand, E. N. & Martell, C. H. (2007), Lexical and discourse analysis of online chat dialog, in 'International Conference on Semantic Computing (ICSC 2007)', IEEE, pp. 19–26.
- Freund, L. S. (1990), 'Maternal regulation of children's problem-solving behavior and its impact on children's performance', *Child development* **61**(1), 113–126.
- Garrod, S. & Anderson, A. (1987), 'Saying what you mean in dialogue: A study in conceptual and semantic co-ordination', *Cognition* **27**(2), 181 – 218.  
**URL:** <http://www.sciencedirect.com/science/article/pii/0010027787900187>

- Genesee, F. (1985), 'Second language learning through immersion: A review of u.s. programs', *Review of Educational Research* **55**(4), 541–561.  
**URL:** <https://doi.org/10.3102/00346543055004541>
- Genesee, F. (2004), '21 what do we know about bilingual education for majority-language students?', *The handbook of bilingualism* p. 547.
- Godfrey, J. J., Holliman, E. C. & McDaniel, J. (1992), Switchboard: Telephone speech corpus for research and development, in 'Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on', Vol. 1, IEEE, pp. 517–520.
- Graesser, A. C., Chipman, P., Haynes, B. C. & Olney, A. (2005), 'Autotutor: An intelligent tutoring system with mixed-initiative dialogue', *Education, IEEE Transactions on* **48**(4), 612–618.
- Graesser, A. C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Tutoring Research Group, T. R. G. & Person, N. (2000), 'Using latent semantic analysis to evaluate the contributions of students in autotutor', *Interactive learning environments* **8**(2), 129–147.
- Gunning, R. (1952), *The Technique of Clear Writing*, Toronto: McGraw-Hill.
- Hawkes, R. (2012), 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.
- Heilman, M., Collins-Thompson, K., Callan, J. & Eskenazi, M. (2007), Combining lexical and grammatical features to improve readability measures for first and second language texts, in 'Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference', pp. 460–467.
- Holley, F. M. & King, J. K. (1971), 'Imitation and correction in foreign language learning', *The Modern Language Journal* **55**(8), 494–498.  
**URL:** <http://www.jstor.org/stable/323789>
- Isard, A., Brockmann, C. & Oberlander, J. (2006), Individuality and alignment in generated dialogues, in 'Proceedings of the Fourth International Natural Language



Generation Conference', INLG '06, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 25–32.

**URL:** <http://dl.acm.org/citation.cfm?id=1706269.1706277>

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

Kerry, A., Ellis, R. & Bull, S. (2009), Conversational agents in e-learning, in 'Applications and Innovations in Intelligent Systems XVI', Springer, pp. 169–182.

Keuleers, E., Lacey, P., Rastle, K. & Brysbaert, M. (2012), 'The british lexicon project: Lexical decision data for 28,730 monosyllabic and disyllabic english words', *Behavior Research Methods* **44**(1), 287–304.

Khodamoradi, A., Iravani, H. & Jafarigohar, M. (2013), 'The effect of teacher's scaffolding and peers' collaborative dialogue on the acquisition of english tenses in the zone of proximal development: A sociocultural perspective', *European Online Journal of Natural and Social Sciences* **2**(2s), pp–336.

Kuperman, V., Stadthagen-Gonzalez, H. & Brysbaert, M. (2012), 'Age-of-acquisition ratings for 30,000 english words', *Behavior Research Methods* **44**(4), 978–990.

Lantolf, J. (2013), *Sociocultural Theory and Second Language Learning*., OXFORD APPLIED LINGUISTICS, Oxford University Press.

**URL:** <https://books.google.co.uk/books?id=z-dBgAAQBAJ>

Lantolf, J. P. (2000a), 'Second language learning as a mediated process', *Language teaching* **33**(2), 79–96.

Lantolf, J. P. (2000b), *Sociocultural theory and second language learning*, Vol. 78, Oxford University Press.

Lenneberg, E. H. (1960), *Language* **36**(1), 97–112.

**URL:** <http://www.jstor.org/stable/410626>

Levy, M. (2009), 'Technologies in use for second language learning', *The Modern Language Journal* **93**(s1), 769–782.

Lison, P. & Tiedemann, J. (2016), 'Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles'.

- Lowe, R., Pow, N., Serban, I. & Pineau, J. (2015), 'The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems', *arXiv preprint arXiv:1506.08909*.
- McLeod, S. A. (2012), 'Zone of proximal development'.  
**URL:** [www.simplypsychology.org/Zone-of-Proximal-Development.html](http://www.simplypsychology.org/Zone-of-Proximal-Development.html)
- Michel, M. C. (2011), 'Effects of task complexity and interaction on L2 performance', *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* **2**, 141–173.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013), Distributed representations of words and phrases and their compositionality, in 'Advances in neural information processing systems', pp. 3111–3119.
- Miller, G. A. (1995), 'Wordnet: a lexical database for english', *Communications of the ACM* **38**(11), 39–41.
- Mills, G. J. & Healey, P. G. T. (2008), Semantic negotiation in dialogue: The mechanisms of alignment, in 'Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue', SIGdial '08, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 46–53.  
**URL:** <http://dl.acm.org/citation.cfm?id=1622064.1622072>
- Morales-Jones, C. (2011), 'Methods/approaches of teaching esol: A historical overview', *Fundamentals of Teaching English to Speakers of Other Languages in K-12 Mainstream Classrooms* pp. 63–74.
- Muñoz, C. (2006), *Age and the rate of foreign language learning*, Vol. 19, Multilingual Matters.
- Pickering, M. J. & Garrod, S. (2004a), 'The interactive-alignment model: Developments and refinements', *Behavioral and Brain Sciences* **27**, 212–225.
- Pickering, M. J. & Garrod, S. (2004b), 'Toward a mechanistic psychology of dialogue', *Behavioral and brain sciences* **27**(2), 169–190.
- Pickering, M. J. & Garrod, S. (2004c), 'Toward a mechanistic psychology of dialogue', *Behavioral and Brain Sciences* **27**(2), 169–190.

- Pickering, M. J. & Garrod, S. (2006), 'Alignment as the basis for successful communication', *Research on Language and Computation* **4**(2-3), 203–228.
- Reitter, D., Keller, F. & Moore, J. D. (2006), Computational modelling of structural priming in dialogue, in 'Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers', NAACL-Short '06, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 121–124.  
**URL:** <http://dl.acm.org/citation.cfm?id=1614049.1614080>
- Reitter, D., Keller, F. & Moore, J. D. (2011), 'A computational cognitive model of syntactic priming', *Cognitive science* **35**(4), 587–637.
- Reitter, D. & Moore, J. D. (2014), 'Alignment and task success in spoken dialogue', *Journal of Memory and Language* **76**, 29–46.
- Robinson, P. (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.  
**URL:** <https://books.google.co.uk/books?id=aqXpMG5ZenwC>
- Rossiter, M. J., Derwing, T. M., Manimtim, L. G. & Thomson, R. I. (2010), 'Oral fluency: The neglected component in the communicative language classroom', *Canadian Modern Language Review* **66**(4), 583–606.
- Rus, V., Maharjan, N., Tamang, L. J., Yudelson, M., Berman, S., Fancsali, S. E. & Ritter, S. (2017), An analysis of human tutors' actions in tutorial dialogues, in 'The Thirtieth International Flairs Conference'.
- Samana, W. (2013), 'Teacher's and students' scaffolding in an efl classroom', *Academic Journal of Interdisciplinary Studies* **2**(8), 338.
- Sao Pedro, M. A., Gobert, J. D. & Baker, R. (2014), 'The impacts of automatic scaffolding on students' acquisition of data collection inquiry skills', *Roundtable presentation at American Educational Research Association* .
- Searle, J. R. & Searle, J. R. (1969), *Speech acts: An essay in the philosophy of language*, Vol. 626, Cambridge university press.
- Senter, R. & Smith, E. A. (1967), Automated readability index, Technical report, CINCINNATI UNIV OH.

- Settles, B. & Meeder, B. (2016), A trainable spaced repetition model for language learning, *in* 'Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', pp. 1848–1858.
- Siddharthan, A. (2006), 'Syntactic simplification and text cohesion', *Research on Language and Computation* **4**(1), 77–109.  
**URL:** <http://dx.doi.org/10.1007/s11168-006-9011-1>
- Sinclair, A., Ferreira, R., Lopez, A., Lucas, C. & Gasevic, D. (2019), I wanna talk like you: Speaker adaptation to dialogue style in l2 practice conversation, *in* 'Proceedings of Artificial Intelligence in Education - 20th International Conference'.
- Sinclair, A., Lopez, A., Lucas, C. & Gasevic, D. (2018), Does ability affect alignment in second language tutorial dialogue?, *in* 'Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue', pp. 41–50.
- Sinclair, A., McCurdy, K., Lopez, A., Lucas, C. & Gasevic, D. (2019), Tutorbot corpus: Evidence of human-agent verbal alignment in second language learner dialogues, *in* 'Proceedings of Educational Data Mining - 12th International Conference'.
- Sinclair, A., Oberlander, J. & Gasevic, D. (2017), 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', pp. 107–115.
- Skantze, G. (2007), *Error Handling in Spoken Dialogue Systems-Managing Uncertainty, Grounding and Miscommunication*, Doctoral Dissertation, KTH.
- Stewart, I. A. & File, P. (2007), 'Let's chat: A conversational dialogue system for second language practice', *Computer Assisted Language Learning* **20**(2), 97–116.
- Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C. V. & Meteer, M. (2000), 'Dialogue act modeling for automatic tagging and recognition of conversational speech', *Computational linguistics* **26**(3), 339–373.
- Taft, M. (1979), 'Recognition of affixed words and the word frequency effect', *Memory & Cognition* **7**(4), 263–272.

- Takač, V. P. (2008), *Vocabulary Learning Strategies and Foreign Language Acquisition*, Vol. 27, Multilingual Matters.
- Tavakoli, P. (2016), 'Fluency in monologic and dialogic task performance: Challenges in defining and measuring l2 fluency', *International Review of Applied Linguistics in Language Teaching* **54**(2), 133–150.
- Vail, A. & Boyer, K. (2014), Adapting to personality over time: examining the effectiveness of dialogue policy progressions in task-oriented interaction, in 'Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)', pp. 41–50.
- Vajjala, S. & Meurers, D. (2014a), 'Exploring measures of "readability" for spoken language: Analyzing linguistic features of subtitles to identify age-specific tv programs', *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)@ EACL* pp. 21–29.
- Vajjala, S. & Meurers, D. (2014b), 'Readability assessment for text simplification: From analysing documents to identifying sentential simplifications', *ITL-International Journal of Applied Linguistics* **165**(2), 194–222.
- Vajjala, S. & Meurers, D. (2016), 'Readability-based sentence ranking for evaluating text simplification', *arXiv preprint arXiv:1603.06009* .
- Van Lier, L. (1996), 'Interaction in the language classroom: Awareness, autonomy and authenticity', *London: Longman* .
- Vanlehn, K. (2006), 'The behavior of tutoring systems', *International journal of artificial intelligence in education* **16**(3), 227–265.
- Vermeer, A. (2001), 'Breadth and depth of vocabulary in relation to l1/l2 acquisition and frequency of input', *Applied psycholinguistics* **22**(2), 217–234.
- Vygotsky, L. (1987), 'Zone of proximal development', *Mind in society: The development of higher psychological processes* **5291**, 157.
- Vygotsky, L. S. (1980), *Mind in society: The development of higher psychological processes*, Harvard university press.
- Walker, M. & Passonneau, R. (2001), Date: a dialogue act tagging scheme for evaluation of spoken dialogue systems, in 'Proceedings of the first international conference

- on Human language technology research', Association for Computational Linguistics, pp. 1–8.
- Walker, M. & Whittaker, S. (1990), Mixed initiative in dialogue: An investigation into discourse segmentation, in 'Proceedings of the 28th annual meeting on Association for Computational Linguistics', Association for Computational Linguistics, pp. 70–78.
- Ward, A. & Litman, D. (2007a), 'Dialog convergence and learning', *Frontiers in Artificial Intelligence and Applications* **158**, 262.
- Ward, A. & Litman, D. (2007b), 'Measuring convergence and priming in tutorial dialog', *University of Pittsburgh* .
- Ward, A., Litman, D. & Eskenazi, M. (2011), Predicting change in student motivation by measuring cohesion between tutor and student, in 'Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications', pp. 136–141.
- Wilske, S. & Wolska, M. (2011), 'Meaning versus form in computer-assisted task-based language learning: A case study on the German dative.', *JLCL* *26*, no. 1 .
- Wood, D., Bruner, J. S. & Ross, G. (1976), 'The role of tutoring in problem solving', *Journal of child psychology and psychiatry* **17**(2), 89–100.
- Wood, D. & Middleton, D. (1975), 'A study of assisted problem-solving', *British Journal of Psychology* **66**(2), 181–191.
- Xia, M., Kochmar, E. & Briscoe, T. (2019), 'Text readability assessment for second language learners', *arXiv preprint arXiv:1906.07580* .
- Xu, W., Callison-Burch, C. & Napoles, C. (2015), 'Problems in current text simplification research: New data can help', *Transactions of the Association for Computational Linguistics* **3**, 283–297.
- Yannakoudakis, H. & Briscoe, T. (2012), Modeling coherence in ESOL learner texts, in 'Proceedings of the Seventh Workshop on Building Educational Applications Using NLP', Association for Computational Linguistics, pp. 33–43.
- Yannakoudakis, H., Briscoe, T. & Medlock, B. (2011), A new dataset and method for automatically grading ESOL texts, in 'Proceedings of the 49th Annual Meeting

of the Association for Computational Linguistics: Human Language Technologies-Volume 1', Association for Computational Linguistics, pp. 180–189.

Yngve, V. (1970), 'On getting a word in edgewise.', *Sixth Regional Meeting of the Chicago Linguistic Society* **6**, 567–577.

Yu, Z., Nicolich-Henkin, L., Black, A. W. & Rudnicky, A. (2016), A wizard-of-oz study on a non-task-oriented dialog systems that reacts to user engagement, in 'Proceedings of the 17th annual meeting of the Special Interest Group on Discourse and Dialogue', pp. 55–63.

Zakos, J. & Capper, L. (2008), Clive – an artificially intelligent chat robot for conversational language practice, in J. Darzentas, G. A. Vouros, S. Vosinakis & A. Arnellos, eds, 'Artificial Intelligence: Theories, Models and Applications', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 437–442.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D. & Weston, J. (2018), 'Personalizing dialogue agents: I have a dog, do you have pets too?', *arXiv preprint arXiv:1801.07243*.