Does Ability Affect Alignment in Second Language Tutorial Dialogue?

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
19th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2018)

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
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 interlocutor’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.
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 have already been seen through the success of apps such as Duolingo 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 prediction (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 vocabulary to the student; in other words, scaffolding 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

---

1 Also know as Dialogue-based Computer Assisted Language Learning (CALL)
2 bots.duolingo.com

3 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).

4 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.
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.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Type</th>
<th>English</th>
<th>Dialogues</th>
</tr>
</thead>
<tbody>
<tr>
<td>BELC</td>
<td>L2 tutoring</td>
<td>non-native</td>
<td>118</td>
</tr>
<tr>
<td></td>
<td>(levels 1-4)</td>
<td>fluent</td>
<td></td>
</tr>
<tr>
<td>Switchboard</td>
<td>conversational</td>
<td>fluent</td>
<td>1155</td>
</tr>
<tr>
<td>Map Task</td>
<td>task-based</td>
<td></td>
<td>128</td>
</tr>
</tbody>
</table>

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 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 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 the words in a text, and counting lemmas used by the speaker as prime. 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.

---

5This also has the effect of removing function words from consideration.

6Using 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\(^7\) 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 \( p(\text{prime}|\text{target}, n) \), 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 \textit{time in seconds} as measured in Reitter and Moore (2006). We look for repetitions within windows of 85 words\(^8\).

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\(^9\). We model this as binomial, where the success probability depends on the following explanatory variables: \textit{Categorical: corpus choice, priming type from speaker role, ability level}; and \textit{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 \textit{prime} and \textit{target}, and the other explanatory variables we are interested in; we refer to this as the probability estimates in 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 \textit{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

---

\(^7\)The use of priming is not intended to imply that priming is the only explanation for lexical repetition

\(^8\)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.

\(^9\)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 \textit{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 \textit{beds} at distance 2. The number of trials = target words \( \times \) window size.
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
effect between low and high levels, lowest at the
highest ability. Per role, priming effect is gener-
ally smaller at higher ability levels than lower.

Figures 7 and 8 show the effects of $\text{WOF}$ 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, in-
dicative of less complex words, where the repeti-
tion probability is now lower for high ability stu-
dents compared to low ability students (see Fig-
ure 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 be-
tween prime and target in lexical repetition, pro-
viding 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 asymmet-
ric in the same manner as for the task-based di-
ologue MT. This is in contrast to the symmetric
effects in the conversational dialogue of SB (Fig-
ure 1). ST alignment also has the greatest prim-
ing effect compared to the other roles in BELC,
which supports our hypothesis that student-to-
tutor alignment is an artefact of both tutor scaf-
folding, 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 rela-
tionship between GG and FF in Map Task. We in-
terpret 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 di-
ologue is harder, the leading speakers use more
consistent language in order to reduce the cogni-
tive 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 ef-
effs (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 immedi-
ate encouraging backchanneling seen in the repe-
tition in Table 2. We note (Figure 4) that tutor-to-
tutor repetition is more probable at lower levels,
which supports the above hypothesis. Addition-
ally, 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-
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**: \( \text{lemma}_\text{occ} \sim \text{window} + \log(WOF) \times \text{categorical level} \)

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

**Formula**: \( \text{lemma}_\text{occ} \sim \text{window} + \log(WOF) \times \text{role} \)

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\(^{10}\) level, or measures such as counts of word sense per word, or whether a word is concrete or abstract\(^{11}\), exploiting existing readability features (Vajjala and Meurers, 2014).

Acknowledgements

We would like to thank the three anonymous reviewers for their useful comments and ideas in relation to this paper. A. Sinclair especially acknowledges the help and support of Jon Oberlander during the early development of this idea.

\(^{10}\)The Common European Framework of Reference (CEFR) defines the 6 levels of English proficiency in ascending order as: A1, A2, B1, B2, C1, C2.

\(^{11}\)Using WordNet or other word/lemma concreteness rating database.
References


