Lexical Level Alignment in Dialogue

Yikai Tseng **Takenobu Tokunaga**
Tokyo Institute of Technology / Tokyo, Japan
tseng.y.ab@m.titech.ac.jp take@c.titech.ac.jp

Hikaru Yokono
Meisei University / Tokyo, Japan
hikaru.yokono@meisei-u.ac.jp

Abstract

Alignment in dialogue is believed to make communication progress smoothly. Lexical alignment has been particularly well studied. However, we hypothesise that it is not just words that get aligned but also the overall difficulty level of the vocabulary used. For instance, when talking to children or non-native speakers, one chooses familiar words to ensure their partner can understand their utterances. We call this phenomenon "lexical level alignment (LLA)". This study investigates whether LLA occurs in natural dialogues and the factors influencing LLA by analysing an existing Japanese dialogue corpus. The analysis revealed that LLA occurs in dialogues between firstly-encountered native and non-native speakers.

Sys 1:	Is your *router* connected to the computer?
Usr 1:	Uh. What's a router?
Sys 2:	It's *the big black box*.
Usr 2:	Ok.. yes.
Sys 3:	Do you see a small white box connected to the router?
Usr 3:	Yes.
Sys 4:	Ok. Is there *a flashing monitor symbol at the bottom right of the screen*?
Usr 4:	*The network icon*?
Sys 5:	Yes. Is it flashing?
Usr 5:	Yes. It is flashing.
Sys 6:	Ok. Please open your *browser*.

Figure 1: Dialogue example (Janarthanam and Lemon, 2009)

1 Introduction

It is well known that alignment at various levels occurs between interlocutors in dialogue for successful communication (Pickering and Garrod, 2006). Pickering and Garrod (2004) proposed the interactive alignment account of dialogue, which assumes that the linguistic representations employed by the interlocutors become aligned at various levels as a result of a largely automatic process. However, a single-level alignment does not necessarily lead to a successful dialogue. The alignment at different levels depends on each other, i.e., alignment at one level leads to those of other levels, and the alignment in total leads to a successful dialogue (Pickering and Garrod, 2006).

Lexical alignment is a typical alignment phenomenon where linguistic descriptions by interlocutors converge during the course of dialogue, and they gradually use the same expression referring to an object (Garrod and Anderson, 1987). Lexical alignment is also attracting attention in the context of human-computer interaction, conversational agents and explainable artificial intelligence (Branigan et al., 2010; Srivastava et al., 2023).

Janarthanam and Lemon (2009) proposed a dialogue system for troubleshooting which can choose referring expressions depending on the lexical knowledge of the user. Figure 1 shows an example of their dialogue data. In Sys 1, the system uses the term "router", but the user does not understand the word and clarifies what it is in Usr 1. This clarification makes the system rephrase "router" with a simple expression "the big black box" in Sys 2, assuming that the user has little lexical knowledge in the network domain. The system continues to use simpler expressions like "a small white box" and "a flashing monitor symbol at the bottom right of the screen". However, once the user rephrases "a flashing monitor symbol" with "the network icon" in Usr 4, the system updates the user's lexical knowledge again. It starts to use technical terms like "browser" in Sys 6. Janarthanam and Lemon (2009) aimed to dynamically adapt the lexical choice to the user's lexical knowledge, as this example illustrates. Although they call this phenomenon lexical alignment as well, we claim it should be distinguished from conventional lexical alignment.

Besides troubleshooting dialogues, when adults talk to children or native speakers to non-native speakers, they try to avoid difficult words in the first place and use easier words if their partner cannot understand their utterances. Namely, the native speaker aligns the lexical level of words in their utterance to their partner's. This phenomenon is different from well-known lexical alignment, where the lexical choice of interlocutors converges to align during the progress of dialogue. We call this phenomenon "lexical level alignment (LLA)". We expect LLA occurs in natural dialogue as likely as lexical alignment does. The system by Janarthanam and Lemon (2009) above can be considered to aim at realising LLA.

In this study, we investigate the phenomenon of LLA by analysing an existing Japanese dialogue corpus. Our research question is twofold.

RQ1: Does LLA occur in natural dialogue?

RQ2: What factors affect LLA?

Examining RQ2, we consider the following two factors: firstly, the intimacy between two interlocutors, whether friends or first-encounters; secondly, the language proficiency level of the interlocutors, whether a pair of native speakers or a pair of a native speaker and a non-native speaker.

2 Related Work

Lexical alignment, the alignment of words, has been widely studied and confirmed in various dialogues. Campano et al. (2014) confirmed that lexical alignment occurs in human-human dialogues both in natural settings and in Wizard of Oz settings, where one of the interlocutors plays the role of the virtual agent using limited utterances. Sinclair et al. (2018) analysed dialogues between second language (L2) learners and tutors and confirmed lexical priming, which indicates lexical alignment. They observed that alignment increases according to the ability of the L2 learners and the word complexity, and student-to-tutor alignment has a stronger priming effect than tutor-to-student alignment. Misiek et al. (2020) analysed child-adult dialogues and confirmed that lexical alignment occurs in both directions. In addition, they observed that adults align with children more than vice versa, even if the factor of language production ability was controlled. Although both Sinclair et al. (2018) and Misiek et al. (2020) consider the difference in lexical knowledge between interlocutors, their interest remains in lexical alignment. Wang et al. (2014) analysed multi-party conversations in online health communities and observed a strong lexical alignment effect.

Xu and Reitter (2015) compared three metrics for measuring linguistic alignment: indiscriminate local linguistic alignment, repetition decay, and Spearman's correlation coefficient. The indiscriminate local linguistic alignment has the overall best performance; it is especially favourable concerning individuals' inherent propensity of alignment. The repetition decay is favourable for exploring the correlations between alignments at different linguistic levels. Spearman's correlation coefficient has poorer normality and consistency than the other two. These metrics are developed for lexical alignment. We need to develop metrics to measure LLA.

Buschmeier et al. (2009) presented an alignment-capable micro planner, SPUD prime, which uses a priming-based interactive alignment model to model human speakers' alignment behaviour. Hu et al. (2018) proposed the Dialog Adaptation Score (DAS) measure to evaluate the adaptation in generated dialogues.

While the past lexical alignment research focuses on individual words from a microscopic viewpoint, we look at the alignment of a macroscopic property, i.e., the lexical level of interlocutor utterances.

3 Data

3.1 Dialogue corpus

We use the BTSJ[1] Japanese 1000-person natural conversation corpus[2] (BTSJ-1000 corpus hereafter) (USAMI, 2023) for analysis. The BTSJ-1000 corpus contains 514 dialogues in various settings totalling 127 hours. The interlocutors have various demographic properties regarding gender, age, first language, and professions. Relations between interlocutors also vary. This demographic information is helpful for us to investigate the factors that affect LLA. The BTSJ-1000 corpus contains dialogues in various situations, such as paper writing, interview, role-play of apology dialogues, and so on. Since we want to analyse the phenomenon of LLA in natural dialogues, we consider only 368 chat dialogues of general topics such as travel and school life in this study. Most of the themes of these chat dialogues are left to the interlocutors.

[1] Basic Transcription System for Japanese

[2] https://isplad.jp/lab/btsj_corpus_2023/

Table 1: Number of dialogues in the BTSJ-1000 corpus

	N-N	N-L
friend	141	43
first-encounter	125	59

Furthermore, we categorise these dialogues regarding two factors: intimacy between interlocutors and the language proficiency level of interlocutors. We have two cases for each factor, i.e., "friend" vs. "first-encounter" for intimacy, and "N-N" and "N-L" for proficiency level, where N-N stands for a pair of native Japanese speakers, while N-L means a pair of a native speaker and a Japanese learner, i.e., a non-native speaker. Table 1 shows the number of dialogues in each category.

3.2 Metric of lexical level

We need a metric to measure lexical level in Japanese (Tellols et al., 2023) to assess LLA in Japanese dialogues. One common metric of Japanese lexical level is the JLPT[3] level, which classifies the vocabulary into five discrete levels from N5 to N1, with N5 being the easiest and N1 the hardest. However, there is no available official vocabulary list for the JLPT level, and the coverage is lower, with only less than 10,000 words in total in an unofficial vocabulary list[4]. Another common metric is the occurrence frequency of BCCWJ (Balanced Corpus of Contemporary Written Japanese) (Maekawa et al., 2014). However, the frequency-based metric heavily depends on the corpus. In addition, since BCCWJ collects written text, it is unsuitable for dialogue analysis.

To remedy these problems, we adopt the WLSP[5] familiarity rate as the metric of lexical level. WLSP is a popular Japanese thesaurus, including 96,557 words with four syntactic categories (nominal, verbal, modifier, and other) and hypernymy and synonymy relations among them (National Institute for Japanese Language and Linguistics, 2004). Asahara (2019) collected familiarity ratings of words in WLSP through the Yahoo! crowdsourcing platform with 3,392 participants. The participants were asked to answer the familiarity of words regarding the five perspectives: KNOW, WRITE, READ, SPEAK, and LISTEN. To remove individual participant bias, a Bayesian linear mixed model was employed to estimate the familiarity rate for each word. Familiar words are assigned a higher value.

As we are focusing on dialogue, we use the LISTEN familiarity as the metric for lexical level, which represents how often one listens to the word. Low-familiarity words would be difficult to understand for the listener.

4 Analysis

4.1 Preprocessing

4.1.1 Dialogue data

The transcribed text of the BTSJ-1000 corpus contains annotation of paralinguistic information, such as filler, intonation and interruption. Since we are measuring lexical level, we remove this paralinguistic annotation in utterances and leave the content of the utterances.

4.1.2 Extracting WLSP words from utterances

To measure the lexical level in terms of WLSP familiarity, we need to extract WLSP words from the utterances. Since words are not separated by whitespaces in Japanese sentences, we first conduct the segmentation of utterances into tokens by a morphological analyser MeCab[6] with UniDic[7] v3.1.0 as the dictionary.

To convert the tokens in utterances into WLSP words, we use the WLSP2UniDic[8] list, which provides the association between WLSP words and UniDic tokens. However, this list only covers WLSP words corresponding to a single UniDic token. Covering multi-token words is essential since a token may occur more often with other tokens than occurs alone, resulting in a higher familiarity for the multi-token word than the token itself. For instance, the WLSP word "*gozaimasu* (an auxiliary verb for a polite form)" with the familiarity rate (LISTEN) of 1.48 is tokenised into two UniDic tokens "*gozaru*" and "*masu*". Although the token "*gozaru*" is also a WLSP word, it does not frequently occur and its familiarity rate (LISTEN) is -0.51, being less familiar than that of "*gozaimasu*".

To ensure the validity of word familiarity, we extend the WLSP2UniDic list to cover multi-token WLSP words as follows. First, we tokenise the

[3]Japanese Language Proficiency Test

[4]http://www7a.biglobe.ne.jp/nifongo/data/noryoku.html

[5]Word List by Semantic Principles

[6]https://taku910.github.io/mecab/

[7]https://clrd.ninjal.ac.jp/unidic/en/

[8]https://github.com/masayu-a/wlsp2unidic

WLSP words not in the original WLSP2UniDic list. However, we cannot simply apply MeCab to the unlisted WLSP words since MeCab cannot accurately tokenise them without a surrounding context. Therefore, we conduct a string-based search for utterances that include the unlisted WLSP words in the dialogue corpus and tokenise the utterances with MeCab. After confirming the consistency of the token boundary and readings between the unlisted word and the MeCab output, the corresponding token sequence for the unlisted WLSP word is added to the extended WLSP2UniDic list. In addition, we ignore the 19 unlisted words consisting of a single *Hiragana*[9] since they are not commonly used and cause many false matches.

We then construct a UniDic-to-WLSP list by inverting the extended WLSP2UniDic list, which maps UniDic token sequences to WLSP words. If multiple WLSP words in the extended WLSP2UniDic list correspond to a token sequence, we select the WLSP word with the highest familiarity, assuming that words with higher familiarity are more likely to occur.

Finally, with the tokenised utterances and the UniDic-to-WLSP list, we extract WLSP words from the utterances using a dynamic programming algorithm. Specifically, we compare the tokenised utterances with the UniDic-to-WLSP list to find a WLSP word sequence that minimises the number of unmatched tokens and the number of extracted words.

4.2 Method

As with lexical alignment, LLA is expected to occur as the dialogue progresses. When LLA occurs, the difference in the lexical level of the words used by the two interlocutors becomes smaller. To capture LLA, we divide each dialogue into two halves with the same length (in terms of the number of turns) and measure the lexical level of the utterances by each interlocutor in each dialogue segment. Consider a dialogue between A and B. Let $LL_p^{(j)} (p \in \{A, B\}, j \in \{1, 2\})$ be the lexical level of the utterances by p in the j-th half of the dialogue. When LLA occurs, we have

$$\Delta := |LL_A^{(2)} - LL_B^{(2)}| - |LL_A^{(1)} - LL_B^{(1)}| < 0.$$

That is, the difference in the lexical level of the interlocutors' utterances decreases in the later half of the dialogue. We calculate $LL_p^{(j)}$ based on the word types used in all p's utterances in the j-th dialogue segment.

We classify the dialogues into four groups by the two factors described in the research questions and perform a hypothesis test to check whether LLA occurs in each group of dialogues.

4.2.1 Lexical level of utterances

We assume that each interlocutor has their lexical level, representing that they understand all words with familiarity at least this level. Under this assumption, we define the lexical level of a word set as follows. After arranging the words in ascending order of familiarity, i.e., less familiar to more familiar, we assume that interlocutors can communicate even though they do not know the first $q\%$ of the difficult words in the list. Then, we define the lexical level of utterances $LL_p^{(j)}$ as the lexical level of the word set used in the utterances by p in the j-th dialogue segment. In this study, we consider 25 and 50 (the first and second quartiles) for q. That is, we assume that the interlocutors understand 75% and 50% of the words used by their partners.

Since the lexical alignment implies the interlocutors use the same words, it automatically induces LLA. To ensure that LLA is not just a by-product of lexical alignment, we exclude those words used by both interlocutors when calculating $LL_p^{(j)}$.

4.2.2 Hypothesis test

We conduct a hypothesis test to show that LLA occurs. The null hypothesis (H_0) assumes that LLA does not occur; in this case, there is no change in the lexical-level difference of the interlocutors' utterances between the first and second half of the dialogues. The alternative hypothesis (H_1) assumes that LLA occurs; in this case, the difference change Δ is negative, meaning the difference of the lexical level becomes smaller as the dialogue progresses.

$$\begin{aligned} H_0 &: \quad \Delta = 0 \\ H_1 &: \quad \Delta < 0 \end{aligned}$$

Since the distribution of the lexical level of utterances is unknown, we test the hypothesis with a one-sided permutation test with the resample count set to 100,000.

4.3 Result

Table 2 shows the result of the permutation test. The # column shows the number of dialogues in each group, and the "$q = N$" columns show the mean values of the lexical level difference change

[9] One of the Japanese writing scripts, a phonogram.

Table 2: Result of the permutation test. The numbers outside and inside the parentheses are the mean values of Δ and the P-values, respectively. The asterisk ($*$) indicates statistical significance at $p < .05$.

Dialogue group	#	$q = 25$	$q = 50$
N-L first-encounter	59	$-.040^{*}$ (.026)	.003 (.569)
N-L friend	43	.005 (.601)	$-.002$ (.441)
N-N first-encounter	125	$-.013$ (.103)	$-.005$ (.247)
N-N friend	141	.020 (.932)	.015 (.974)

Δ and their P-values in parentheses. From this result, we can see that while there is no significant change in lexical level in the N-L friend dialogues and N-N dialogues, there is a significant decrease in lexical level difference in the N-L first-encounter dialogues. This result suggests that LLA occurs in the N-L first-encounter dialogues even if we eliminate the effect of lexical alignment.

5 Discussion

5.1 Factors that affect LLA

LLA occurs in the first-encounter dialogues but not in the friend dialogues. In first-encounter dialogues, the interlocutors initially do not know their partners' lexical level but can estimate the level as the dialogue progresses. Therefore, they try to align their lexical level later in the dialogue. On the other hand, in the friend dialogues, the interlocutors already know their partners' lexical level before the dialogue. Therefore, their lexical levels can be aligned from the beginning. We cannot observe LLA during dialogue in this case.

LLA occurs in the N-L dialogues but not in the N-N dialogues. In the N-L dialogues, the native speaker might consider that their partner may not be able to understand difficult words and try to estimate and align the lexical level to their partner. In addition, the non-native speaker might also try to align their lexical level to their partner, which might be considered a language learning process. On the other hand, in the N-N dialogues, the interlocutors might assume that their partners would understand most of the words. It is unnecessary to consider the lexical level, so LLA does not occur.

LLA is not observed when measuring the lexical level of utterances calculated with $q = 50$. We do not observe LLA when calculating the lexical level of utterances with $q = 50$. Knowing more than 50% of the words might be needed to understand their partners' utterances.

Table 3: Number of dialogues with and without LLA when using the entire and the difference word sets. "O" and "X" indicate whether LLA is observed ($\Delta < 0$) or not ($\Delta \geq 0$), respectively.

Entire	Diff.	#	(%)
O	O	135	(34.5)
X	X	127	(36.7)
O	X	48	(13.0)
X	O	58	(15.8)

5.2 The validity of using difference word set

To ensure that LLA is not just a by-product of lexical alignment, we excluded the words used by both interlocutors, i.e., we excluded the intersection of the interlocutor word sets and considered the difference word sets when calculating the lexical level of utterances. To assess the validity of using the difference word sets, we compare Δ of each dialogue when using the entire and difference word sets in the calculation of lexical level.

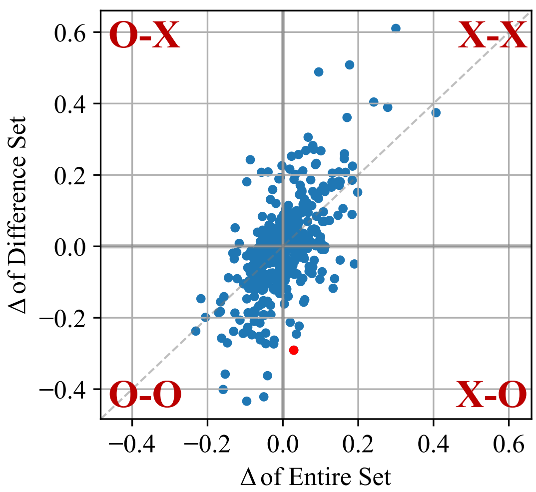

Figure 2: Δ pairs when using the entire and the difference word sets

Table 3 shows the number of dialogues which

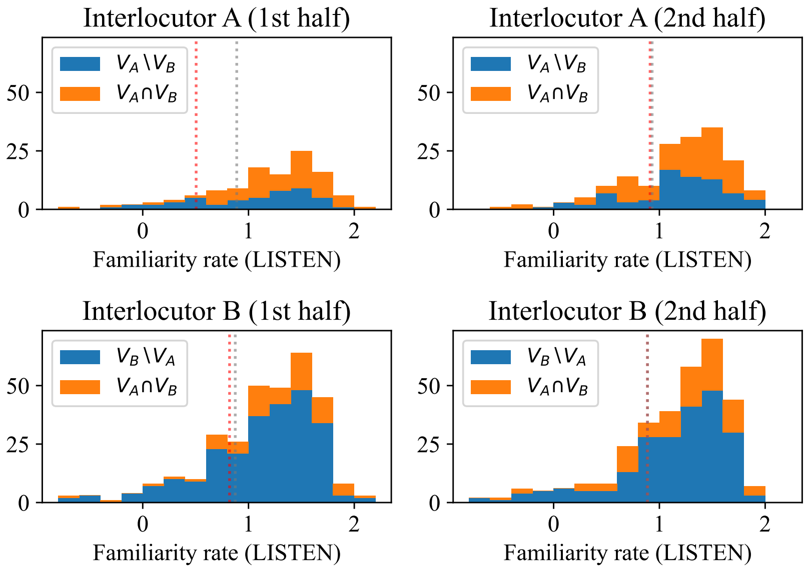

Figure 3: Familiarity rate distribution of word sets in the dialogue denoted by the red data point in Figure 2. V_A and V_B represent the word set of utterances by interlocutors A and B, respectively. The blue and orange bars represent the distribution of word familiarity rate for the difference set and the intersection of the word sets of utterances by both interlocutors, respectively. The orange bars are stacked on the blue bars; therefore, the sum of both bars represents the distribution of word familiarity rate for the entire set. The grey and red dashed lines represent the lexical level of the entire and the difference word sets of utterances, respectively.

are divided into categories according to whether LLA is observed ("O") using the entire and the difference word sets. For instance, we have 135 dialogues where we observe LLA using both the entire and the difference word sets (the O-O column). These dialogues indicate that LLA is not a by-product of lexical alignment. On the other hand, in the dialogues of the O-X column, LLA observed using the entire word set can be considered a spurious one induced by lexical alignment.

However, 15.8% of the dialogues (the X-O column) are unexpected since it indicates that LLA is not observed when considering the entire set while observed after the intersection is excluded. To investigate the reason for these unexpected cases, we analyse the familiarity rate distribution of word sets in these dialogues. Figure 2 shows the scatter diagram plotting Δ using the entire word set on the x-axis and Δ for the difference word set on the y-axis. Each quadrant corresponds to each column in Table 3. We pick up the red data point, which has the largest difference in Δ between the entire and the difference word set, and calculate the familiarity rate distribution of the words in the dialogue.

Figure 3 shows the distribution of the familiarity rate for the dialogue indicated by the red point in Figure 2. Comparing the distribution of the interlocutor A and B, while B's distributions are similar between the entire and the difference sets, A's distributions are less similar. As the distribution determines the lexical levels (grey and red dashed lines), its shape directly affects the lexical level value. In the second half of the dialogue, since the proportion of words with a familiarity rate higher than 1.0 is almost the same for the entire and the difference sets, the lexical levels of the two sets are similar for A, even though their distributions are quite different. On the other hand, in the first half of the dialogue, there is a non-negligible peak at $0.4 \sim 0.6$ for the difference set, which deviates the lexical level of the difference set from that of the entire set. Considering that the lexical level of the entire set tends to be lower than that of the difference set, we observe LLA only for the difference set. This example reveals the limitation of using a fixed q value for measuring the lexical level of utterances regardless of the familiarity distribution.

5.3 LLA patterns

In 4.2.2, we analysed LLA from a macroscopic viewpoint with a hypothesis test. We can also analyse it from a microscopic viewpoint by investigating the lexical level change of the interlocutors between the first and the second half of individual dialogues.

First, we consider whether both interlocutors contribute to LLA (two-way alignment) or only one of them does (one-way alignment). For dialogues between interlocutor A and B, let $\Delta_p := LL_p^{(2)} - LL_p^{(1)} (p \in \{A, B\})$. When LLA occurs ($\Delta < 0$), the two-way and one-way alignments are formulated as follows.

Two-way alignment: $\Delta_A \Delta_B < 0$

One-way alignment: $\Delta_A \Delta_B > 0$

Figure 4 illustrates the corresponding alignment patterns[10].

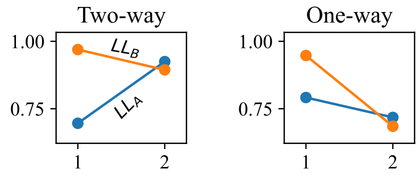

Figure 4: Example of the two LLA patterns regarding contributors. The “1” and “2” on the horizontal axis represent the first and the second half of the dialogue. The vertical axis represents the lexical level of utterances used by each interlocutor in each half.

Table 4 shows the number of each pattern. We can see that regardless of the intimacy of interlocutors, the two-way alignment occurs more than the one-way alignment in the N-L dialogues, but the opposite happens in the N-N dialogues. This suggests that native and non-native speaker pairs jointly tend to align their lexical level with each other, but it is not the case for native speaker pairs.

Regarding LLA in the N-L dialogues, we also consider the direction of alignment. Specifically, we focus on the absolute lexical level change of the interlocutor utterances from the first half to the second half of the dialogue $|\Delta_p|$, and assume that the interlocutor with larger $|\Delta_p|$ aligns to their partner. We have the following two patterns of lexical level change (Figure 5).

- N-to-L alignment: $|\Delta_N| > |\Delta_L|$
- L-to-N alignment: $|\Delta_N| < |\Delta_L|$

[10]The one-way alignment example in figure 4) shows the case where both interlocutors use more difficult words in the second half (i.e., $\Delta_A, \Delta_B < 0$). However, it is also possible that both of them use easier words (i.e., $\Delta_A, \Delta_B > 0$).

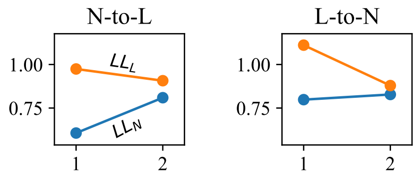

Figure 5: Example of the two LLA patterns regarding alignment direction. The “1” and “2” on the horizontal axis represent the first and the second half of the dialogue. The vertical axis represents the lexical level of utterances used by each interlocutor in each half.

Table 5 shows the number of each pattern. We can see that the L-to-N alignment occurs more than the N-to-L alignment, especially in the first-encounter dialogues. This result indicates that non-native speakers try to align their lexical level to native speakers as a part of the language learning process.

5.4 Alignment and dialogue quality

Alignment contributes to a successful dialogue. Here, we investigate the relationship between LLA and dialogue quality. It is, however, difficult to define dialogue quality in general. We focus on an aspect of to what extent both interlocutors speak equally to assess dialogue quality. Specifically, we consider the ratio of the WLSP word count per turn[11] between two interlocutors as the metric for dialogue quality. We take the larger word count as the denominator to make the metric range between 0 and 1. Therefore, a larger value means both interlocutors speak equally, and hence the dialogue has higher quality.

Figure 6 shows the relation between Δ and dialogue quality. While there is no correlation between Δ and dialogue quality for all dialogues, there is a weak tendency that smaller Δ, i.e. high LLA, leads to higher dialogue quality for the N-L first-encounter dialogues only, with Pearson correlation coefficient being -0.343 and P-value being 0.008. Our metric for dialogue quality is a rough approximation and captures only one of many other aspects. We need to investigate further the relationship between LLA and other aspects of dialogue quality more precisely.

[11]We also tried “ratio of UniDic token count per turn” and “ratio of vocabulary set size” and obtained similar results.

Table 4: Distribution of LLA patterns regarding contributors

Dialogue group	#	Two-way	One-way	No alignment
N-L first-encounter	59	20 (34%)	14 (24%)	25 (42%)
N-L friend	43	13 (30%)	7 (16%)	23 (54%)
N-N first-encounter	125	30 (24%)	47 (38%)	48 (38%)
N-N friend	141	30 (21%)	32 (23%)	79 (56%)
Total	368	93 (25%)	100 (27%)	175 (48%)

Table 5: Distribution of LLA patterns regarding alignment direction

Dialogue group	#	N-to-L	L-to-N	No alignment
N-L first-encounter	59	6 (10%)	28 (47%)	25 (42%)
N-L friend	43	9 (21%)	11 (26%)	23 (53%)

6 Conclusion

This study discussed lexical level alignment (LLA) in dialogue, which has not received explicit attention in past research. Analysing a Japanese dialogue corpus, we showed that LLA is observed (RQ1) when the interlocutors' lexical levels differ, and they do not know their partner's lexical level (RQ2).

We used WLSP familiarity rate (LISTEN) as the metric of lexical level and defined the lexical level of utterances as the required lexical level for the interlocutor to communicate without knowing the most difficult $q\%$ (we used 25 and 50 for q in this study) of the words used in the utterances. Specifically, we excluded those words used by both interlocutors when calculating their lexical level to ensure that LLA is not just a by-product of lexical alignment.

We classified the dialogues into four groups by the familiarity between interlocutors (friend or first-encounter) and their language proficiency level (N-N or N-L). We performed a permutation test to see if LLA occurs in each group. Specifically, we considered the change of lexical level difference between the utterances by the two interlocutors from the first half to the second half of the dialogue, and verified whether the difference decreased. As a result, we confirmed that LLA occurs in first-encounter dialogues between a native speaker and a non-native speaker when q is set to 25.

In addition, we checked the validity of using the difference word set when calculating lexical level and confirmed that 71.2% of the dialogues have the same result after excluding the words used by both interlocutors; 13.0% of the dialogues have spurious LLA; 15.8% of the dialogues are unexpected, which suggests the limitation of using the fixed q value for measuring the lexical level of utterances.

We also analysed the LLA patterns. We first analysed whether both interlocutors contribute to LLA or only one of them does. We found that the two-way alignment occurs more than the one-way alignment in the N-L dialogues but not in the N-N dialogues. This tendency indicates that native and non-native speaker pairs jointly try to align with each other, but it is not the case for native speaker pairs. We then analysed the direction of alignment in the N-L dialogues. We found that the L-to-N alignment occurs more than the N-to-L alignment, especially in the first-encounter dialogues. This indicates that non-native speakers try to align with native speakers, which might be considered a part of the language-learning process.

Finally, we investigated the relationship between LLA and dialogue quality. We considered the word count ratio per turn between interlocutors as the metric of dialogue quality, assuming that interlocutors speak equally in successful dialogues. We observed a weak tendency that LLA leads to higher dialogue quality in the N-L first-encounter dialogues.

7 Future Work

As we discussed in 5.2, we had unexpected dialogues where LLA was observed only using the difference word sets of interlocutors. The detailed analysis suggests that such an anomaly is caused by calculating the lexical level of utterances without considering the word familiarity distribution in utterances. More sophistication is needed in

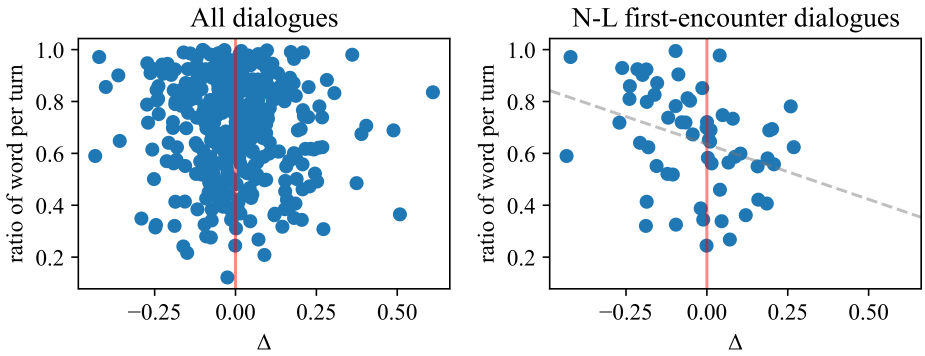

Figure 6: The relation between Δ (the metric of LLA) and "ratio of word count per turn" (the metric of dialogue quality). The grey dashed line represents the linear regression line.

measuring the lexical level of utterances.

We considered to what extent both interlocutors speak equally as a metric of dialogue quality in analysing the relation between LLA and dialogue quality. As we already mentioned, there are many other aspects of dialogue quality. For instance, lexical level gaps might cause frequent clarification, misunderstanding, or even dialogue breakdown. We would like to shed light on other aspects of dialogue quality and investigate their relation to LLA in future.

In addition, word difficulty is likely affected by the topic in dialogues. Therefore, the distribution of the lexical level of utterances can be unstable when the topic changes in the dialogue. We need to investigate the influence of the topic and how the lexical level aligns as it changes. In this study, we adopted a popular Japanese thesaurus WLSP, which assigns semantic categories to each word. We would also look at the relationship between the dialogue topic (change) and the distribution of the word categories in the utterances for investigating the LLA process.

Limitations

This study uses the WLSP familiarity rate for lexical level measurement, which might not be available for other languages. Besides, since we capture LLA from a macroscopic viewpoint, even though we confirmed that LLA occurs in the N-L first-encounter dialogues, the alignment process is still an open question. Further study is necessary for the dynamic nature of LLA.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP21K18358.

References

Masayuki Asahara. 2019. Word familiarity rate estimation using a Bayesian linear mixed model. In *Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP*, pages 6–14, Hong Kong. Association for Computational Linguistics.

Holly P. Branigan, Martin J. Pickering, Jamie Pearson, and Janet F. McLean. 2010. Linguistic alignment between people and computers. *Journal of Pragmatics*, 42(9):2355–2368.

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 (ENLG 2009)*, pages 82–89, Athens, Greece. Association for Computational Linguistics.

Sabrina Campano, Jessica Durand, and Chloé Clavel. 2014. Comparative analysis of verbal alignment in human-human and human-agent interactions. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4415–4422, Reykjavik, Iceland. European Language Resources Association (ELRA).

Simon Garrod and Anthony Anderson. 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218.

Zhichao Hu, Jean Fox Tree, and Marilyn Walker. 2018. Modeling linguistic and personality adaptation for natural language generation. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 20–31, Melbourne, Australia. Association for Computational Linguistics.

Srinivasan Janarthanam and Oliver Lemon. 2009. Learning lexical alignment policies for generating referring expressions for spoken dialogue systems. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 74–81, Athens, Greece. Association for Computational Linguistics.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, 48(2):345–371.

Thomas Misiek, Benoit Favre, and Abdellah Fourtassi. 2020. Development of multi-level linguistic alignment in child-adult conversations. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 54–58, Online. Association for Computational Linguistics.

National Institute for Japanese Language and Linguistics. 2004. *Bunrui goihyo zôho kaitei-ban* (Word List by Semantic Principles Revised and Enlarged Edition). Dainippon Tosyo, Tokyo.

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*, 4:203–238.

Arabella Sinclair, Adam Lopez, C. G. Lucas, and Dragan Gasevic. 2018. Does ability affect alignment in second language tutorial dialogue? In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 41–50, Melbourne, Australia. Association for Computational Linguistics.

Sumit Srivastava, Mariět Theune, and Alejandro Catala. 2023. The role of lexical alignment in human understanding of explanations by conversational agents. In *The 28th International Conference on Intelligent User Interfaces (IUI '23),*, pages 423–435.

Dolça Tellols, Takenobu Tokunaga, Hilofumi Yamamoto, and Hikaru Yokono. 2023. Pic2PLex: A test for assessing free productive vocabulary. *Transactions of the Japanese Society for Artificial Intelligence*, 38(1):B–M52 1–11.

Mayumi (ed.) USAMI. 2023. Building of a Japanese 1000 person natural conversation corpus for pragmatic analyses and its multilateral studies, and ninjal institute-based projects: Multiple approaches to analyzing the communication of japanese language learners.

Yafei Wang, David Reitter, and John Yen. 2014. A model to qualify the linguistic adaptation phenomenon in online conversation threads: Analyzing priming effect in online health community. In *Proceedings of the Fifth Workshop on Cognitive Modeling and Computational Linguistics*, pages 55–62, Baltimore, Maryland, USA. Association for Computational Linguistics.

Yang Xu and David Reitter. 2015. An evaluation and comparison of linguistic alignment measures. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*, pages 58–67, Denver, Colorado. Association for Computational Linguistics.