Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2771–2780
Marseille, 20-25 June 2022

Mutual Gaze and Linguistic Repetition in a Multimodal Corpus

Anaïs Murat, Maria Koutsombogera and Carl Vogel
Trinity College Dublin
College Green, Dublin 2, Ireland
{murata, koutsomm, vogel}@tcd.ie

Abstract

This paper investigates the correlation between mutual gaze and linguistic repetition, a form of alignment, which we take as evidence of mutual understanding. We focus on a multimodal corpus made of three-party conversations and explore the question of whether mutual gaze events correspond to moments of repetition or non-repetition. Our results, although mainly significant on word unigrams and bigrams, suggest positive correlations between the presence of mutual gaze and the repetitions of tokens, lemmas, or parts-of-speech, but negative correlations when it comes to paired levels of representation (tokens or lemmas associated with their part-of-speech). No compelling correlation is found with duration of mutual gaze. Results are strongest when ignoring punctuation as representations of pauses, intonation, etc. in counting aligned tokens.

Keywords: Linguistic repetition, alignment, mutual understanding, gaze, interaction, multimodal corpus

1. Introduction

We provide an observational analysis of the interactions between gaze and evidence of mutual understanding in multiparty, multimodal communication. The analysis is based upon the Multisimo corpus (Koutsombogera and Vogel, 2018), a multimodal corpus which provides authentic task-based interactions among three dialogue participants. Our starting point hypotheses are that mutual gaze is in greater evidence at times of mutual understanding than times without mutual understanding, and that asymmetric gaze of one participant at another is in greater evidence at times without mutual understanding. We think of significant linguistic repetition as an index of mutual understanding.

Building on the works of Vogel (2013) and Reverdy et al. (2020), we investigate the relation between linguistic repetitions and mutual gaze. Motivated by the relationship between linguistic repetition and mutual understanding, our assumptions are consistent with the alignment theory developed by Pickering and Garrod (2004; 2006; 2007); repetition of linguistic forms provides evidence of a shared situational model, and therefore mutual understanding. We test imitative alignment by counting the number of repeated and non-repeated items per turn and analyze the interaction with both the occurrence and duration of mutual gaze. This method is a simplification of the approaches of Vogel (2013) and Reverdy et al. (2020) in that we explore a ratio between repeated forms and forms that could have been repeated, without computing whether repetition levels are significantly different from what might have been expected by chance. The null hypothesis is that there is no interaction between presence or duration of mutual gaze and linguistic repetition.

The structure of this paper is as follows. The next section addresses past works that focus on the role of gaze and alignment in communication. The third section explains how data were retrieved. The fourth section details the methods that were used to synchronise the gaze annotations, the turns as well as their speech tags. It also describes how the Jaccard Index was implemented to assess linguistic repetition, and how significance was calculated between repetitions and gaze features. Eventually, section 5 provides the raw results and section 6 discusses them. Section 7 concludes this paper by summarising its mains objectives, findings and limitations.

2. Background research

2.1. Gaze

Gaze study has made its path to the field of technology. For example, Fan et al. (2018) illustrates how “the study of shared-attention helps a computer vision system to better understand and interpret human activities in images or videos” (p.3). Yet, eye movements may be analysed at different scales of granularity. Taking into account smaller units of gaze such as fixations and saccades (Yarbus, 1967) requires specific recording material to make possible the measurement of such a high frame rate and detect such small variations in space. An alternative is to calculate gaze using facial landmarks and knowledge about the scenarios in which face-to-face dialogue is recorded (McLaren et al., 2020). Such a coarse-grained approach corresponds to what conversation partners may consciously perceive. The latter is all the more important that gaze is essential when it comes to the theory of mind and non-verbal communication. The theory of mind can be defined as “the ability to infer the psychological states; intentions, beliefs, desires, etc, of other individuals from non-verbal cues” (Emery, 2000, p.582). Gaze is also characterised by its double function: (1) the ability to perceive, as already discussed, and (2) the ability to express. This, therefore, means that when we study gaze, we need to know which function we want to investigate (Gobel et al., 2015). In our case, since the negotiation that leads to understanding is a back-and-forth movement between these two functions, we will have to focus on both: we can expect people to look at each other to both perceive any facial expression that could support the meaning that they are constructing and, almost at the same time, produce cues that would give hints about their understanding of the situation. Furthermore, mutual gaze not only signals communicative intents and initiates social interaction (Cary, 1978, in Pfeiffer et al. (2013)), but also indicates the willingness to pursue the conversation (Jokinen et al., 2010) by showing what the attention of the speaker and addressee are targeting. On the other hand, averted gaze is a way to reduce cognitive load and help focus on inner thoughts (Jording et al., 2018). Looking away might also provide cues that the person does not want to continue the interaction in the same terms (Jokinen et al., 2010). Like many other variables in human studies, gaze characteristics might greatly depend on the individuals and set-ups in which the interaction takes place. Gobel et al. (2015) as well as McLaren et al. (2020), for instance, showed how personality and hierarchy can impact gaze behaviours.

2.2. Understanding & Alignment

Conversation frequently involves negotiation and seeking agreement. Even debating or arguing requires people to, at least, agree on the terms of their disagreement. Pickering and Garrod (2004; 2006; 2007) theorized about dialogue by developing the idea of alignment. To understand a dialogue, not only is it needed to encode and decode messages, but also to align through interaction. Schober and Clark (1989) showed that, when asked to reproduce a certain shape, overhearers who could not interact with the instructor did not perform as well as actual addressees (who could communicate with the instructors and thus check on their understanding of the situation). It is a question of aligning their mental state – or situation models (Pickering and Garrod, 2004). Both the speaker and the addressee can reach the mutual belief that they have built a common ground. However, the belief that interlocutors share the same common ground can never be perfectly verified, and mutual understanding is generally assumed to be reached when both people lack evidence of misunderstanding (Taylor, 1992; Vogel, 2013). We focus on moments when people try to make sure that nothing has been misunderstood and see how they proceed. Pickering and Garrod (2004), showed that people can rely on several linguistic and conceptual levels to discuss and eventually align their situation models. Pickering and Garrod (2007) describe modalities of alignment: alignment via beliefs about one's interlocutor, via an agreement between interlocutors, via feedback, via physical co-presence, or via imitation. Imitation is of interest here: through the development (even temporary) of routines (Pickering and Garrod, 2006), people tend to reuse the same (normally) ambiguous terms through conversation, but with the precise meaning established during the exchange. They claim that people tend to reuse the same grammatical structures, copying the structure of their interlocutor's answers. To study these different levels of linguistic alignment (lexical, syntactic, and even phonetic), they detailed the widely used Interactive Alignment Model (IAM) (Pickering and Garrod, 2004). We base our experimental design on the work of Reverdy et al. (Reverdy and Vogel, 2017a; Reverdy and Vogel, 2017b; Reverdy et al., 2020) in which they proved that significant repetitions are witnessed in five levels of linguistic representation: three in-isolation levels: (1) token, (2) part-of-speech, and (3) lemma; and two paired levels: (1) token + part-of-speech, and (2) lemma + part-of-speech. Using lemmas and not only tokens allows to focus on conventional forms, for more qualitative results (Reverdy and Vogel, 2017b). Studying parts-of-speech, on the other hand, allows, for n-grams with $n \geq 2$, to highlight the grammatical structures that are being repeated. In comparison, unigram repetitions might possibly just reflect lexical repetitions (Doyle and Frank, 2016). Reitter and Moore (2007) consider that studying syntactic alignment is often more relevant than lexical one: lexical priming is difficult to distinguish from the vocabulary that is simply required by the topic. A challenge is that no method of calculating alignment is universally accepted (Doyle and Frank, 2016). Prior works mentioned (Reverdy and Vogel, 2017a; Reverdy and Vogel, 2017b; Reverdy et al., 2020) classify observed repetition as significant or not by comparing the forms of repetition observed in actual dialogues with that found in turn-based randomizations of the dialogues. This is meant to adjust for the fact that because words in closed-class part of speech categories are high-frequency, chance repetition is expected.

In this paper, we examine the ratio between the items that were repeated in relation to the items that could have been repeated but were not, as a fraction of total items (a Jaccard index), without reference to randomized counterpart dialogues. Our motivation for this simplification is that the former model is appropriate only for complete dialogues; however, the measure explored here could be used in live dialogues as they unfold (assuming accurate transcription and gaze information can be made in real time). Thus, there is importance to determining its efficacy. First, it is important to note that assessing the overall similarities of all interlocutor's productions is not enough: it does not allow to draw any conclusion on the impact of what one previously said on the current production (Mehler et al., 2011). Temporal cues are therefore crucial and we do so by comparing successive turns (Reverdy et al., 2020), as opposed to reaching arbitrarily far back within conversations (Reitter and Moore, 2007). Also, methods may differ in their counters: some consider utterances as "bags of words" and take into account all the n-grams, as in a multiset, whereas some prefer to only see them as sets and therefore rely on types of n-grams. Moreover, calculation models differ (Mekhaldi, 2006) in that some research only relies on counts while others take into account the length of the utterance and approach it as a frequency. Finally, there seem to be "contagion effects" which show that linguistic alignment can be accompanied by non-linguistic features such as the imitation of facial expressions (e.g. Bavelas et al. (1986)) and gesture alignment (Pickering and Garrod, 2006; Oben, 2018), providing prior evidence that our hypotheses regarding relations between mutual gaze and linguistic repetition might be verified.

3. Data Collection

3.1. The Multisimo Corpus

The Multisimo corpus (Koutsombogera and Vogel, 2018) is a multimodal corpus that investigates collaboration in three-party conversations. It is made of 18 dialogues of about 2,000 – 3,000 words each. Each dialogue includes two players and one facilitator. They all play a game in which the two participants have to guess and rank the most popular answers to three questions. The facilitator introduces the game and the tasks and provides feedback. Dialogues are recorded with a set of frontal view cameras and microphones. Although there were only three facilitators for the entire corpus, each dialogue involved new players. These collaborative tasks were carried out in English, but only 12 of the 36 were native speakers. A variety of nationalities was represented among the rest of the players and the facilitators. There were 20 male and 16 female participants divided into 6 mixed groups, 5 groups with only female participants, and 7 with only male participants.

3.2. Annotation of Gaze Tags

All dialogues were manually annotated with gaze tags within ELAN (Brugman and Russel, 2004). One tier was created for each of the three participants and four labels were used to describe the different gaze behavior of the participants:

1. GAZE_PLAYER_1 when the participant was looking at player 1,
2. GAZE_PLAYER_2 when the participant was looking at player 2,
3. GAZE_FACILITATOR when one of the players and was looking at the facilitator, or
4. GAZE_AWAY, when the participant was not looking at any of the two other persons around him.

Three MUTUAL_GAZE tiers were automatically generated, comparing two tiers at a time, corresponding to a pair of participants: (1) Player 1 & Player 2, (2) Player 1 & Facilitator, and (3) Player 2 & Facilitator.

3.3. Retrieving Utterances

The Multisimo corpus already contained manually made speech transcriptions created using Transcriber (Barras et al., 1998) and then synchronised with the other time-labelled annotations in ELAN. For this paper, we trusted the annotators' choices in turn individuation and punctuation. For the latter, they used common sense: full stops were placed when a coherent contribution was being concluded and coincided with adequate voice pitch. Commas and question marks relied on the audio and on semantic and syntactic cues such as clear enumerations, interrogative forms, etc. Analyses we describe below take punctuation into account as tokens. The use of square brackets indicates lengthened vowels, disfluencies [eh, ah, etc.], and laughter [laugh].

3.4. Annotation of the Speech Tags

To measure the same linguistic levels of representation as Reverdy et al. (2020), part-of-speech and lemma tags were required. We used TreeTagger (Schmid, 1994), whose reliability makes it a very frequently used tool in the literature (Moreau et al., 2019). It is a probabilistic part-of-speech and lemma tagger that makes use of decision trees relying on the context and a lexicon to determine the appropriate tags. By default, it is trained on a tagged training sample of the Penn Treebank corpus. Tested on another excerpt of the same corpus, it can reach up to 96% of accuracy (Schmid, 1994). However, as we run it on speech data, we cannot expect such high accuracy. Dialogues differ quite a lot in that they may include words that are not part of the lexicon (e.g. contractions such as "gonna"), sentences and words truncated or split by the words of another speaker, hesitations, and transcription special characters not belonging to standard written texts. Thus, there are reasons to worry about the use of a context-based tagger trained on a different genre of data. To evaluate it, the arbitrarily selected dialogue S02 (2696 lexical tokens), as well as other samples stratified to satisfy similar proportions as the ones observed in the overall corpus in terms of sex, familiarity, and age (1000 tokens in total, i.e. 2% of the overall data set) were manually annotated and compared to the tags obtained from the tagger. This showed that overall part-of-speech accuracy can be satisfied with minimal revision of the tags by defining a table of substitutions (table 1). To improve lemmatization, the original tokens were used when the word was not part of the tagger's vocabulary, and "+" signs marking elongated vowels were removed. All these changes allowed our part-of-speech tags and our lemmas to be respectively 93.80% and 97.01% accurate.

3.5. Punctuation in Speech Transcription

Speech transcriptions often integrate conventions of written language into the spoken one. Some can be misleading, for instance, capital letters. Every new sentence was introduced by a capital letter, but this creates a distinction between sentence initial and sentence medial tokens that is spurious for our analysis. We thus lowercased all tokens. Punctuation use was mainly left to the appreciation of the transcriber without prior measurements. Full stops were inserted when a coherent

Token	Tag	Meaning of the Tag
laugh	laugh	laugh
ok	UH	Interjection
eh	UH	Interjection
did	VDD	Verb do past tense
do	VDP	Verb do present tense (1,2,3P)
does	VDZ	Verb do present tense (3S)
done	VDN	Verb do past participle
thanks	UH	Interjection
,	,	,
!	!	!
?	?	?
hm[...]	UH	Interjection
ı[...]	UH	Interjection

Table 1: Substitution table

contribution was concluded, same common sense for commas and question marks. On one hand, it still reflects how an external listener perceived the speech, on the other hand, it cannot be said with certainty to be expressing exactly what the speakers wanted to express and might be influenced by the written reflex of the transcriber. To test the impact of punctuation on our method to measure alignment, we, therefore, created two sets of data, one which included punctuation, and another one which did not. All the upcoming steps were thus executed on both.

4. Methods

4.1. Synchronisation of Gaze Tags & Utterances

Once all our data were gathered, one difficulty was left: since both gaze annotations and utterances are temporally delimited, how can they be synchronised and clearly represented to allow the best possible analyses? A matrix was created on the basis of one line equals one turn. The starting and end time of both the turns and mutual gazes were compared, and the utterance line was duplicated as many times as a mutual gaze occurred, even partially, during the same time interval. In the case where the utterance did not match any mutual gaze, the columns about gaze were completed with predefined values such as: mutual gaze number="0". An example matrix can be found in table 2

4.2. Calculation of Self-Shared and Other-Shared Repetitions

As pointed out in Section 2, there is no consensus on methods to measure alignment. We clarify here the method we have used. Since we are working on a three-party conversation, one person can be responding to two people at the same time: therefore, it is not just the previous utterance that should be considered, but the last one of every single speaker. A register keeping track of the last enunciation of each participant was created. We then computed a Jaccard Index, counting both the number of n-grams ($1 \leq n \leq 3$) from the previous turns that were repeated and those that could have been repeated but that were not. As they appear functionally distinct in conversations, we both measured self-repetitions and other-repetitions [1] separately. Indeed, the main distinction being that while other-repetition is obviously signalling grounding and involvement, self-repetition seems to be rather indicating discourse plan perseverance and allows the person to keep the floor (Koutsombogera and Vogel, 2019).
The process is repeated for each level of linguistic representation (token, lemma, part-of-speech, token + part-of-speech, and lemma + part-of-speech), and each dialogue. This extended set of repetitions being motivated by the analysis of the extent to which repetition across levels of linguistic representation can be related to mutual understanding (Reverdy and Vogel, 2017b), this paper contributes to the study of this relation.

4.3. Comparing Repetitions & Gaze Annotations

We consider mutual gaze data through two intuitive variables: one is the presence or absence of mutual gaze, the other one is the duration of such a gaze. The binary scale expliciting the presence or absence of mutual gaze was paired with the repetition index of every single instance, and the significance between these variables was assessed using a Chi-Squared test. Each gaze duration for each turn was individually considered. This variable was compared with the repetition index per n-gram and type of repetition and evaluated using Spearman correlation test (as our data do not conform to a normal distribution).

5. Results

5.1. Presence of Mutual Gaze and Repetition

The first step in our analysis is to assess whether we can reject the null hypothesis stating that there is no relation between mutual gaze and repetition. The counts of repetition paired with the complement category of items that could have been repeated but were not, along with the presence or absence of mutual gaze are thus compared according to two dimensions, in four different situations. The two dimensions correspond to (1) the five levels of representation (token, part-of-speech, token, lemma, token + part-of-speech, and lemma + part-of-speech), and (2) the length of the n-grams (from $n = 1$ to $n = 3$). The four situations are (1) the case of other-shared repetitions when punctuation is counted, (2) the case of self-shared repetitions when punctuation is counted, (3) the case of other-shared repetitions when punctuation is not taken into account, (4) the case of self-shared repetitions when punctuation is ignored. The contingency test results are reported in Tables 3, 4, 5 and 6. Two meta-analyses were then run. Table 7 lists the counts of χ^2 tests that resulted in significance

[1] We refer to these also as self-sharing and other-sharing.

		Speech Tags		
Turn	Mutual Gaze	Token	Lemma	Part-of-Speech
Hey	None	hey	hey	UH
Hi	MG1	hi	hi	UH
Hello Leah	MG1	hello, leah	hello, leah	UH, NP
Hello Leah	MG2	hello, leah	hello, leah	UH, NP

Table 2: Made-up example of the matrix that aligns turns, mutual gazes and speech tags. Mutual Gaze MG1 occurs both during turn "Hi" and turn "Hello Leah". Two gazes (MG1 and MG2) occur while "Hello Leah" is being uttered. UH=Interjection, NP=Proper Noun

	n-grams,$n = 1$		n-grams,$n = 2$		n-grams,$n = 3$	
Level	χ^2	p	χ^2	p	χ^2	p
Token	36.272	1.72E-09	5.4979	0.01904	2.4851	0.1149
POS	249.9	<2.2e-16	59.463	1.25E-14	12.664	0.0003728
Lemma	46.374	9.77E-12	5.2342	2.22E-02	1.1553	0.2824
Token+POS	12.801	0.0003464	0.028789	0.8653	0.20761	0.6486
Lemma+POS	10.375	0.001277	0.0041934	0.9484	0.17917	0.6721

Table 3: χ^2 statistics and p-values resulting from the interaction of Mutual Gaze being present or not and items being **other-repetition or not; punctuation counted**; each is a 2×2 contingency table, therefore with 1 degree of freedom. POS = Part-of-Speech

($p < 0.05$) or not in relation to the length categories associated with n-grams considered. Table 8 shows the counts corresponding to the relationship between significance and level of linguistic representation.

The analysis shows significant results in all contexts: there are significant relations between the presence of mutual gazes and the self-shared and other-shared repetitions no matter whether punctuation is counted or not. Yet, what is striking is that out of the 34 significant instances, 18 are unigrams. Indeed, unigrams are not proved significant only for the paired levels of representation (token and lemma + part-of-speech) when punctuation is counted. When it comes to bigrams, only those in isolation (token, part-of-speech, and lemma) are significant. For trigrams, parts-of-speech are the only level for which significance appears. The two meta-analyses which led to table 7 and table 8 can confirm the relevance of our two dimensions (levels and lengths of repetitions). The first meta-analysis asks the question of whether there is an interaction between sequence length and the determination of significance in the underlying test of interaction with mutual gaze, and the second meta analysis tests whether there is an interaction between the linguistic levels and the determination of significance in the underlying tests. Both meta-analyses were conducted using χ^2 tests on the contingency tables of test outcomes (Table 7 and Table 8).

Both meta-analyses reveal significant interactions (P-values ≪ 0.05)[2] entailing that both the linguistic levels and the length of repetitions interact with the significance or non-significance of the underlying tests. Inspection of residuals in χ^2 tests helps identify the locus of interaction.[3] In meta-analysis of the interaction with sequence length, significantly fewer ($R = -2.26$) underlying tests were not significant for unigrams than would be expected if there were no interaction with sequence length; significantly fewer ($R = -2.18$) underlying tests were significant for trigrams than would be expected; significantly more ($R = 2.49$) underlying tests were not significant for trigrams than would be expected under the null hypothesis. For the interaction between linguistic levels and significant outcomes in the underlying tests, inspection of residuals show that there were significantly fewer ($R = -2.28$) tests involving POS tokens than would have been expected if there were no interaction between linguistic levels and outcomes in the underlying tests.

5.2. Mutual Gaze Duration and Repetition

Another part of our study investigated the correlation between mutual gaze duration and repetition counts relativized to the counts of items that could have been repeated. The null hypothesis, in that case, was H0: there is no relation between mutual gaze duration and repetitions measured using the Jaccard index. These data did not follow a normal distribution; thus, a Spearman correlation test was conducted. Again there are three

[2] Interaction between sequence length and significance: $\chi^2 = 20.09$, df= 2, $p < 4.339e - 05$; Interaction between linguistic level and significance: $\chi^2 = 19.955$, df= 4, $p = 0.0005098$.

[3] Residuals with magnitude between 2 and 4 are significant ($p < 0.05$); residuals with magnitude greater than 4 are highly significant ($p < 0.001$). The sign indicates the direction of divergence between expected and observed values.

	n-grams,$n = 1$		n-grams,$n = 2$		n-grams,$n = 3$	
Level	χ^2	p	χ^2	p	χ^2	p
Token	70.724	<2.2e-16	6.595	0.01023	2.99E-28	1
POS	317.54	<2.2e-16	80.628	<2.2e-16	14.335	0.000153
Lemma	97.116	<2.2e-16	12.707	0.0003643	0.18828	0.6643
Token+POS	1.4719	0.2251	0.18024	0.6712	1.1845	0.2764
Lemma+POS	0.3333	0.3333	0.92186	0.337	0.18594	0.6663

Table 4: χ^2 statistics and p-values resulting from the interaction of Mutual Gaze being present or not and items being **self-repetition or not; punctuation counted**; each is a 2 × 2 contingency table, therefore with 1 degree of freedom. POS = Part-of-Speech

	n-grams,$n = 1$		n-grams,$n = 2$		n-grams,$n = 3$	
Level	χ^2	p	χ^2	p	χ^2	p
Token	16.138	5.89E-05	8.809	0.002996	2.11	0.1463
POS	243.87	<2.2e-16	61.479	4.48E-15	13.248	0.0002728
Lemma	25.58	4.23E-07	7.5885	0.005874	1.1872	0.2759
Token+POS	13.092	0.0002965	0.11381	0.7358	0.030156	0.8621
Lemma+POS	11.213	0.0008121	0.50578	0.477	0.065683	0.7977

Table 5: χ^2 statistics and p-values resulting from the interaction of Mutual Gaze being present or not and items being **other-repetition or not; punctuation not counted**; each is a 2 × 2 contingency table, therefore with 1 degree of freedom. POS = Part-of-Speech

(for each length of n-gram between one and three) times five (for each level of linguistic tokenizations) times two (for other-sharing and self-sharing) times two (for counting punctuation or not), that is, 60 tests conducted. Even though some p-values were significant, the correlation coefficient was always less than 0.01. Therefore, the null hypotheses saying that there is no existing correlation between the duration of mutual gaze and repetitions could not be rejected. Because of the overwhelming null effect, in the interest of brevity we do not report here the complete table of results.

6. Discussion

6.1. Mutual Gaze and Linguistic Repetition

6.1.1. Levels of Representation in Isolation

Similarly to Reverdy et al. (2020) who worked on the same corpus and used a related alignment method, one of the questions that naturally arises from our analysis is whether there are some differences between the different linguistic levels of representation used when tokenizing and determining repetition. The analysis of the linguistic levels of repetition in interaction with the presence of mutual gaze suggests a number of remarks to answer this question. One of them is perceptible as soon as the p-values of all the individual contingency tables are calculated: is part-of-speech level a better predictor of the presence of mutual gaze? This question is based on two simple observations: first, part-of-speech is the only level to still be significant when considering trigrams; second, its residuals also usually show the greatest magnitudes ($R > 10$). The conclusion thus drawn that part-of-speech sequences involve the linguistic level where the interaction with presence of mutual gaze is most stark (and therefore possibly the only one to consider in an optimised practical exploitation of our results) is all the more tempting that it echoes the work of Reitter and Moore (2007) showing that syntactic alignment might be the most relevant types of alignment to address.

However, that a single level can be identified as having strongest positive interactions does not dismiss the fact that found significant interactions between tokens, parts-of-speech, lemmas, and the presence of mutual gaze, both for other-shared and self-shared repetitions. Oben (2018) showed similar results, even though the chronology was not exactly the same: if the person had been looking at the face of the speakers while they were talking, there were more chances for them to reuse the same vocabulary. This positive correlation in both types of repetitions therefore indicates more collaboration when they are looking at each other. They might want to check either that they have well understood what the other has just said, or that they are well understood themselves.

6.1.2. Paired Linguistic Levels and Repetition

Paired linguistic levels of repetitions (token + part-of-speech, and lemma + part-of-speech) were introduced in Reverdy and Vogel (2017a) and required more research to back up their relevance. When taken in isolation, unigrams of token, lemma, and part-of-speech, all proved to be significant under all conditions. On the contrary, out of the eight possible instances in which unigrams of token + part-of-speech, and lemma + part-of-speech were tested, only six showed that their repe-

	n-grams,$n = 1$		n-grams,$n = 2$		n-grams,$n = 3$	
Level	χ^2	p	χ^2	p	χ^2	p
Token	53.419	2.69E-13	4.7997	0.02847	0.24214	0.6227
POS	306.65	<2.2e-16	83.497	<2.2e-16	17.761	2.50E-05
Lemma	79.783	<2.2e-16	10.92	0.0009475	1.3245	0.2498
Token+POS	6.3182	0.01195	0.97688	0.323	1.0491	0.3057
Lemma+POS	5.0635	0.02444	0.23492	0.6279	8.45E-27	1

Table 6: χ^2 statistics and p-values resulting from the interaction of Mutual Gaze being present or not and items being **self-repetition or not; punctuation not counted**; each is a 2×2 contingency table, therefore with 1 degree of freedom. POS = Part-of-Speech

Significance	$n = 1$	$n = 2$	$n = 3$
$p < 0.05$	18	12	4
$p \geq 0.05$	2	8	16

Table 7: Counts underlying meta-analysis of χ^2 test outcomes: the interaction between significance and the length of n-grams, each $n, 1 \leq n \leq 3$ as a distinct category.

titions were significantly related to the presence of mutual gaze, excluding the exact same context for both levels: the case in which we were investigating self-shared repetitions with punctuation being counted as possible repetitions. Since Reverdy and Vogel (2017a) showed significant repetitions when all the different levels were simultaneously taken into account, but not when each level was taken separately, it was deemed pertinent to exploit the outcome of the meta-analysis presented in which significance was assessed throughout all levels of repetition. Indeed, it showed overall significance, and confirmed that these levels of representation are worthwhile categories to look at. Furthermore, the absence of one or two heavy constituents (most Pearson residuals being comprised between -2 and +2 for every level), as well as the change of pattern (both for the meta-analysis, and for the in-depth investigations) pointing out real distinctions between paired and in-isolation levels should motivate the consideration of all these five levels. Results for the paired levels of repetition (i.e token + part-of-speech, and lemma + part-of-speech) display totally different patterns since they are negatively correlated to the presence of mutual gaze, which might indicate two things: first, a certain independence between the levels (i.e paired levels cannot just be interpreted as the intersection of parts-of-speech, and tokens or lemma); second, a difference in the kinds of information conveyed. Further analysis, albeit rather hypothetical, might suggest that clear paraphrasing entailed by paired levels leads to the absence of mutual gaze. Could that mean that people tend to look at each other when they talk about the same thing, when they try to adapt to the other's discourse, or see that they need to repeat themselves to help their interlocutor? But that, when the repetition is too obvious and mutual understanding is deemed so certain, the attention's scope of the speaker and listeners switches to something else? To look deeper into these hypotheses, it would have been interesting to know exactly who the participants involved in the mutual gaze were, and what was the attitude of participants when no mutual gaze was occurring. Finally, the meta-analysis meets with the results obtained in Reverdy and Vogel (2017b): both token + part-of-speech, and lemma + part-of-speech levels behave likewise. As they suggested, this could be explained by 2 factors: either the complexity of the task only led participants to use basic vocabulary with not too many inflexions, especially since most of the participants were non-native speakers; or maybe inflexions would be more relevant for other languages than English (Reverdy and Vogel, 2017b).

6.2. Contribution to the Methods

6.2.1. The Evaluation of Part-of-Speech and Lemma Tags

Our data and methods were similar to Reverdy et al. (2020). Yet, some improvements were tested in this paper. The first one being that we conducted an in-depth evaluation of the tags, which had not been done in previous works on this corpus. Doing so, we minimised as much as possible tagging errors and made as relevant as possible the use of these tags for the planned analyses.

6.2.2. Contribution of Punctuation

We tested two ways to count repetitions. The first one included the repetitions of punctuation when counting the number of repeated n-grams, the second did not. The goal of this manipulation was to establish whether punctuation should or should not be taken into account for this kind of repetition measurement method. Our results showed overall more significance without punctuation. Indeed, not only two more instances were deemed significant for unigrams (token + part-of-speech, and lemma + part-of-speech for self-

Significance	Token	POS	Lemma	Token+POS	Lemma+POS
$p < 0.05$	8	12	8	3	3
$p \geq 0.05$	4	0	4	9	9

Table 8: Counts underlying meta-analysis of χ^2 test outcomes: the interaction between significance and the level of linguistic tokenization in repetition counts.

shared repetitions) but it also led to greater significance in the case of token and part-of-speech other-shared bigrams (p $\ll$ 0.01, as opposed to p = 0.02 when punctuation is included). Removing punctuation from the count of repetitions gives more weight to lexical items. As punctuation conventions adopted by annotators are influenced by written works they do not always perfectly apply to speech data and might sometimes fail to reflect the speaker's intentions, potentially confounding the interpretation of punctuation symbols.

6.2.3. Lack of mutual gaze duration effects

We hypothesized a relation between gaze durations and the magnitude of repetitions. However, our way to consider duration did not lead to significant results. One could argue that there must not be any link between mutual gaze duration and repetitions; however, the null result may also be an artifact of the manner in which we individuated gazes and their durations in relation to turns. Indeed, our experiment took the whole length of any mutual gaze occurring simultaneously with the utterance, but did not record the time without mutual gaze. Thus, further research could investigate duration differently. This should find a way to deal with the same gazes spanning several utterances, and, consequently, only representing a very small proportion of the actual utterance. Such study might also take into account the moments when no mutual gaze occurs, and the duration of other types of gaze, etc. This could be done by reversing the organisation of the data. Instead of being individuated by utterance, it could be individuated by gaze. In such a case, moments without mutual gazes could be analysed to help describe what the gaze patterns are when people are not looking at each other. Also, a weakness of our study is that we have not taken into account whether the speaker was involved in the mutual gazes co-temporal with utterances. This could also be given a more in-depth analysis in future research. A Wilcoxon rank sum test showed a very significant interaction (W = 40975016, p = 5.941e-10) between mutual gaze duration and familiarity: participants did share longer looks when they did not know each other (mean = 2,146.83 ms) than when they did (mean = 1,841 ms). Further investigations of gaze duration might therefore want to control these variables.

7. Conclusion

Our main aim here was to investigate a possible relation between linguistic repetition and gaze. We focused on two features of mutual gaze: its presence (or absence), and its duration. Repetition was assessed in real time through different levels of linguistic representation and different lengths of n-grams. Our results, albeit mainly significant for unigrams and bigrams showed that there exists a positive correlation between the presence of mutual gaze and the number of repetitions for in-isolation levels (i.e. when participants look at each other, it is more likely that they will repeat the tokens, lemmas or parts-of-speech that were previously said). The correlation turns out negative when relating mutual gaze to paired linguistic levels of repetition (tokens and parts-of-speech, and lemmas and parts-of-speech). This highlights that there must be something behind paired levels of repetition. Not only are they relevant, but they must reflect different kinds of information.

This paper also adds to the methodology of alignment. First, it details an instantaneous – as opposed to post-hoc – way of measuring alignment. Second, it investigates different levels of linguistic repetitions and different lengths of segments. Finally, it highlights the optional character of punctuation in transcripts. Yet, as for our experimental design, it could certainly be improved, and our results should be qualified by pointing out limitations. We individuated our data per utterances and mutual gaze, glossing over what other gaze patterns could have been, and giving excessive weight to utterances containing mutual gazes, without identifying which participants were actually involved in them. Furthermore, all mutual gazes included in a sentence were not necessarily initiated by the given utterance, but could have just had some milliseconds of overlap. Although this might have confounded effects associated with the duration of mutual gaze, it also leaves room for further follow-up studies.

8. Acknowledgement

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223, and was supported by the GEHM research network (Independent Research Fund Denmark grant 9055-00004B).

9. Bibliographical References

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (1998). Transcriber: a free tool for segmenting, labeling and transcribing speech. In *First International Conference on Language Resources and Evaluation (LREC)*, pages 1373–1376.

Bavelas, J., Black, A., Lemery, C., and Mullett, J. (1986). "I show how you feel": motor mimicry as a communicative act. *Journal of Personality and Social Psychology*, 50:322–329.

Brugman, H. and Russel, A. (2004). Annotating multi-media/multi-modal resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 2065–2068, Lisbon, Portugal. European Language Resources Association (ELRA).

Cary, M. S. (1978). The role of gaze in the initiation of conversation. *Social Psychology*, 41(3):269–271.

Doyle, G. and Frank, M. C. (2016). Investigating the Sources of Linguistic Alignment in Conversation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 526–536, Berlin, Germany. Association for Computational Linguistics.

Emery, N. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & Biobehavioral Reviews*, 24(6):581–604, August.

Fan, L., Chen, Y., Wei, P., Wang, W., and Zhu, S.-C. (2018). Inferring shared attention in social scene videos. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6460–6468.

Gobel, M. S., Kim, H. S., and Richardson, D. C. (2015). The dual function of social gaze. *Cognition*, 136:359–364, March.

Jokinen, K., Harada, K., Nishida, M., and Yamamoto, S. (2010). Turn-Alignment Using Eye-Gaze and Speech in Conversational Interaction. *Interspeech*, pages 2018–2021.

Jording, M., Hartz, A., Bente, G., Schulte-Rüther, M., and Vogeley, K. (2018). The "Social Gaze Space": A Taxonomy for Gaze-Based Communication in Triadic Interactions. *Frontiers in Psychology*, 9:226, February.

Koutsombogera, M. and Vogel, C. (2018). Modeling collaborative multimodal behavior in group dialogues: The multisimo corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2945–2951.

Koutsombogera, M. and Vogel, C. (2019). Observing Collaboration in Small-Group Interaction. *Multimodal Technologies and Interaction*, 3(3):45, June.

McLaren, L., Koutsombogera, M., and Vogel, C. (2020). Gaze, Dominance and Dialogue Role in the MULTISIMO Corpus. In *2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 83–88, Mariehamn, Finland, September. IEEE.

Mehler, A., Lücking, A., and Menke, P. (2011). Assessing Lexical Alignment in Spontaneous Direction Dialogue Data by Means of a Lexicon Network Model. In Alexander F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 6608, pages 368–379. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.

Mekhaldi, D. (2006). *A study on multimodal document alignment: bridging the gap between textual documents and spoken language*. Ph.D. thesis, Faculty of Science, University of Fribourg (Switzerland).

Moreau, E., Vogel, C., and Barry, M. (2019). A paradigm for democratizing artificial intelligence. In Ana Esposito, et al., editors, *Innovations in Big Data Mining and Embedded Knowledge*, volume 159, pages 137–166. Springer.

Oben, B. (2018). Chapter 10. Gaze as a predictor for lexical and gestural alignment. In Geert Brône et al., editors, *Advances in Interaction Studies*, volume 10, pages 233–264. John Benjamins Publishing Company, Amsterdam, October.

Pfeiffer, U. J., Vogeley, K., and Schilbach, L. (2013). From gaze cueing to dual eye-tracking: Novel approaches to investigate the neural correlates of gaze in social interaction. *Neuroscience & Biobehavioral Reviews*, 37(10):2516–2528, December.

Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(02):169–190, April.

Pickering, M. J. and Garrod, S. (2006). Alignment as the Basis for Successful Communication. *Research on Language and Computation*, 4(2-3):203–228, October.

Pickering, M. J. and Garrod, S. (2007). Alignment in dialogue. In M. Gareth Gaskell, editor, *The Oxford Handbook of Psycholinguistics*, pages 442–452. Oxford University Press, August.

Reitter, D. and Moore, J. D. (2007). Predicting success in dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Comp utational Linguistics*, pages 808–815. Association for Computational Linguistics.

Reverdy, J. and Vogel, C. (2017a). Linguistic repetitions, task-based experience and a proxy measure of mutual understanding. In *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 395–400, Debrecen, Hungary, September. IEEE.

Reverdy, J. and Vogel, C. (2017b). Measuring Synchrony in Task-Based Dialogues. In *Interspeech 2017*, pages 1701–1705, Stockholm, Sweden, August. ISCA.

Reverdy, J., Koutsombogera, M., and Vogel, C. (2020). Linguistic Repetition in Three-Party Conversations. In Anna Esposito, et al., editors, *Neural Approaches to Dynamics of Signal Exchanges*, volume 151, pages 359–370. Springer Singapore, Singapore. Smart Innovation, Systems and Technologies.

Schmid, D. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of Interna-*

tional Conference on New Methods in Language Processing, pages 44–49.

Schober, M. F. and Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, 21(2):211–232, April.

Taylor, T. J. (1992). *Mutual Misunderstanding: Scepticism and the Theorizing of Lang uage and Interpretation*. Duke University Press.

Vogel, C. (2013). Attribution of mutual understanding. *Journal of Law & Policy*, pages 101–145.

Yarbus, A. (1967). *Eye Movement ad and Vision*. Springer US.