https://dx.doi.org/10.1016/j.system.2022.102931

Multimodal alignment in telecollaboration: a methodological exploration

Marco Cappellini[1]

Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France

marco.cappellini@univ-amu.fr

Benjamin Holt

Université de Lille, CNRS, STL, Lille, France

benjamin.holt@univ-lille.fr

Yu-Yin Hsu

The Hong Kong Polytechnic University, Hong Kong, China

yyhsu@polyu.edu.hk

Abstract. This paper presents an analysis of interactive alignment (Pickering & Garrod, 2004) from a multimodal perspective (Guichon & Tellier, 2017) in two telecollaborative settings. We propose a framework to analyze alignment during desktop videoconferencing in its multimodality, including lexical and structural alignment (Michel & Cappellini, 2019) as well as multimodal alignment involving facial expressions. We analyze two datasets coming from two different models of telecollaboration. The first one is based on the *Français en (première) ligne* model (Develotte et al., 2007) which puts future foreign language teachers in contact with learners of that language. The second one is based on the teletandem model (Telles, 2009), where students of two different mother tongues interact to help each other use and learn the other's language. The paper makes explicit a semi-automatic procedure to study alignment multimodally. We tested our method on a dataset that is composed of two one-hour sessions. Results show that in desktop videoconferencing-based telecollaboration, facial expression alignment is a pivotal component of multimodal alignment.

Keywords. Telecollaboration, virtual exchange, interactive alignment, lexical alignment, structural alignment, facial expression alignment

Authors statement and acknowledgement.

The authors would like to thank Bryan Smith and Audrey Chery from Arizona State

University for their help in data collection in the AMU-ASU teletandem project, as well as

Brigitte Bigi for her assistance in using SPPAS. This work was supported by the French

[1] Corresponding author

Agence Nationale de la Recherche [ANR-18-CE28-0011-01] and by a grant provided by the Research Grants Council of Hong Kong [A-PB1C] and the Department of Chinese and Bilingual Studies at the Hong Kong Polytechnic University.

The following article is a collaborative endeavor. However, the writing of each part of the article was under the responsibility of one of the authors. Yu-Yin Hsu was responsible for sections about facial expression alignment; Benjamin Holt was responsible for sections about the data collection and lexical alignment; Marco Cappellini was responsible for the other sections.

1. Introduction

Research on second language acquisition has centered on interaction dynamics at least as early as the communicative approach (Van Ek & Trim, 1991). Such research has converged toward a focus on interactive punctual phenomena, conceptualized in different ways depending on the epistemological stance taken by authors. For instance, from a cognitive interactionist perspective, much attention has been given to phenomena such as negotiation of meaning (Varonis & Gass, 1985) or different forms of feedback (Long, 1996; Nassaji, 2016). From a sociocultural perspective, feedback has been conceptualized rather in relation to the concept of Zone of Proximal Development (e.g. Aljaafreh & Lantolf, 1994) or in terms of Language-Related Episodes (Swain, 2000).

In this paper, we have chosen Pickering and Garrod's (2004, 2021) interactive alignment framework as a basis since it provides a way to move the focus from punctual phenomena to more pervasive interactional dynamics at work in conversation. Besides, it favors a socio-cognitive approach that transcends the confrontations between cognitive and socio-cultural approaches (Michel & Cappellini, 2019). Indeed, the study of alignment allows us to formulate hypotheses on the cognitive level without limiting ourselves to case studies of punctual phenomena. Moreover, whereas Coste and Cavalli (2015) allow us to conceive of foreign language learning as a process of social mobility and socialization where the learner tends to adopt the norms—including linguistic ones—of the target language community, the framework of alignment (see below) provides insights into the learner's integration through adopting formulations employed by speakers of the target community.

Since the mid-90s, possibilities for learners' interaction in the target language have been extended both by increasing physical mobility and, above all, by the widespread adoption of Computer-Mediated Communication (CMC), which has been harnessed by pedagogical approaches such as those of telecollaboration and virtual exchange (O'Dowd & O'Rourke, 2019). In this paper, we focus on telecollaboration, defined as the interaction of groups of learners in different geographical locations, interacting through CMC to accomplish learning tasks (Belz, 2003). More precisely, we focus on interactions through desktop videoconferencing platforms such as Skype and Adobe Connect. Although desktop videoconference was already one of the main means of telecollaboration (O'Dowd, 2018), this type of CMC tools became most prominent because of the Covid-19 pandemic, which sparked a mass migration of teaching towards online synchronous settings (ECML, 2021). We therefore think that it is important to research the specific multimodality of these CMC settings in order to gain a better understanding of how the modalities at work are drawn upon to co-construct meaning.

The main purpose of this article is to explore innovative methodological procedures to study alignment in desktop videoconferencing-based telecollaboration in a multimodal way. We combine different tools to gain insight into the multimodality at work in interactive alignment through desktop videoconference in telecollaboration. More precisely, we draw upon NLP tools and facial recognition combined with manual annotation and analysis. In this sense, this study contributes to the necessary expansion of research on alignment taking a multimodal stance (Rasenberg et al., 2020).

2. Literature review

We will first consider existing research on second language learning through telecollaboration before focusing on research on interactive alignment in computer-mediated communication.

2.1 Telecollaboration and SLA

Telecollaboration emerged as a sub-field of Computer-Assisted Language Learning (CALL) when CMC enabled language learners to practice their L2 communication skills online. Since its inception in the mid-90s, a large number of studies in telecollaboration have dealt with language learning, combined with other objects of research such as the development of intercultural skills and of different forms of multimodal literacies (see Avgousti, 2018 for a recent review). Methodologically, the first studies on telecollaboration in the late 90s and early 2000s (see Kern, 2006 for a review) drew their units of analysis from existing theoretical frameworks, mainly cognitivist interaction analysis (Gass, 1997), in order to construct the research object of language learning. The main foci were on particular communicative events, such as negotiation of meaning (e.g. Kötter, 2003). With the advent of sociocultural approaches (Lantolf & Thorne, 2006), other conceptualizations and categories were introduced, such as Language-Related Episodes (e.g. Ware & O'Dowd, 2008). The confrontation between cognitivist and sociocultural approaches generated epistemological debates in the field of telecollaboration (O'Rourke, 2008). More recently, the categories of these broad approaches were complexified. For instance, Van der Zwaard and Bannik (2019) argued for the distinction between task-related and face-related dynamics in the study of negotiation of meaning. Another example is Akiyama (2019), who studied the links between Language-Related Episodes and lexical categories in a task-based exchange through Skype.

Moreover, a large number of studies began to take interest not only in the verbal level of interaction, but also in the multimodality allowed by more recent CMC settings such as desktop videoconferencing (e.g. Akiyama, 2019; Wigham, 2017).

Aside from some rare cases of corpus-based investigation, especially with a focus on pragmatic competence (e.g. Cunningham, 2017), the main focus of empirical studies on telecollaboration has been on punctual linguistic phenomena that are considered to have an impact or an influence on second language acquisition and development. We argue that the framework of interactive alignment provides a possibility to go beyond the focus on punctual phenomena and to study second language development with a wider perspective on this pervasive feature of communication (Michel & Cappellini, 2019). The present article is part of this endeavor as it proposes a methodological framework that we tested on a valid but small dataset.

2.2 Alignment: from F2F L1 interaction to CMC L2 multimodality

The term *alignment* designates the convergence of mental states during dialogue, or global alignment, which is often observed through the adoption of the same or similar communicative behaviors at particular moments of interaction through focal alignment (Pickering & Garrod, 2021) [2]. Rasenberg et al. (2020) argue that the study of alignment has generated a lot of research in psycholinguistics over the past 30 years. In their literature review, they observe a wide variety of theoretical approaches, which they group around two main perspectives: priming (e.g. Pickering & Garrod, 2004) vs. grounding (e.g. Brennan &

[2] Readers may have come across related concepts such as mirroring or orientation. Alignment is a psycholinguistic phenomenon, as we described in the paper. *Mirroring* is first of all a neurological phenomenon that has been widely observed and interpreted by Giacomo Rizzolatti and his colleagues. *Orientation* is a concept from a conversationalist perspective, which can provide a framework for interpretation of the behaviors in terms of co-construction of methods (in the ethnomethodological sense) for interlocutors to orient themselves and make sense of the social situation.

Clark, 1996). The latter perspective deals with high-level, often intentional processes, while the former usually focuses on lower-level automatic processes (see section *Analysis procedures*). The psycholinguistic framework of interactive alignment is arguably the best-known example of the priming perspective. It was first elaborated to describe L1 interaction by Pickering and Garrod (2004) and has resulted in a recent framework embedding alignment more widely into joint action and into a theory of a shared workspace for collaborative dialogue (Pickering & Garrod, 2021). According to these authors, during a conversation, understanding between interlocutors is achieved when their models of the situation (Zwaan & Radvansky, 1998) are as similar as possible, in other words, when these models are aligned. Alignment of situational models is achieved in conversation mainly through the alignment of representations at the semantic, syntactic, and phonological levels (Pickering & Garrod, 2021). Pickering and Garrod (2009) and Rasenberg et al. (2020) add to these levels other dimensions relevant to the multimodality of communication such as posture, gesture, and laughter. Alignment at all levels occurs through the mechanism of *priming* and *entrainment* (Pickering & Ferreira, 2008), which can be roughly defined as the appearance of a communicative element, for instance a word (*priming*), then repeated by the interlocutor (*entrainment*). For our purposes it is important to note that previous work has largely focused on linguistic features, often ignoring other modalities (see also the contributions of the present special issue by McDonough et al. and Tekin et al.). In our terms, the different linguistic and paralinguistic levels of analysis correspond to modalities for meaning making (Bezemer & Kress, 2016). In their recent review, Rasenberg et al. (2020) also note that studies on alignment are mainly concerned with alignment in one modality. However, theoretically, the framework posits that alignment on one level “percolates” towards the other levels and facilitates alignment on them, thus resulting in alignment of mental representations (Pickering & Garrod, 2004, 2009, 2021). Notably, they found only two studies which explored cross-modal alignment, and conclude that “more work is needed on the causal relations between alignment at various channels or (linguistic) levels of behavior” (Rasenberg et al., 2020, p. 22), or, in our terms, in different modalities. In this paper, we are interested in cross-modal alignment, that is, alignment of lexical, structural and facial expressions.

The framework of alignment has been developed for L1 dialogue, but it has also been introduced for the study of L2 interaction for instance by Costa et al. (2008). Alignment in L2 interaction has raised a number of questions which have been explored over the past 15 years (see Jackson, 2018 and Kim & Michel in this issue for reviews). For our purposes, it is interesting to note the difference between alignment by the NS to the NNS or the other way around. The literature on NS-NNS interaction has long highlighted that NSs adapt their language to make it more accessible to NNSs, a phenomenon also known as “foreigner talk” (Gass, 1997 p. 59). Recent research within the alignment framework showed that this goes as far as the NS adopting a disfavored name after this has been used by the NNS (Suffil et al., 2021). On the other hand, Michel and Stiefenhöfer (2019) argue that NNSs sometimes consciously decide not to align with their interlocutor, for instance when they consider the grammatical structure used by the latter to be too difficult. In other words, in contrast with alignment in L1 interaction, alignment in L2 might not always be an automatic process. Moreover, the “direction” of alignment deserves to be considered in L2 interactions.

A recent body of work has started to investigate L2 alignment in CMC settings, mainly written and synchronous (SCMC). Michel and Smith (2018) studied lexical alignment in written SCMC among learners of English, taking as units of analysis multi-word expressions or n-grams from 3 to 10 words. With the help of eye-tracking technology, they studied whether more focal attention to an n-gram resulted in a higher probability of reuse. This was indeed the case, since when an interlocutor looked at an n-gram more often, they were more likely to reuse it in their subsequent productions. However, many reuses of n-grams were produced without any longer previous focal attention. The authors concluded that some instances of L2 alignment are the fruit of strategic choice, while others are presumably more automatic in nature, similar to L1 alignment. In a similar setting, Michel and O'Rourke (2019) explored text chat between learners of German interacting with peers or with a tutor. The results were consistent with those of Michel and Smith. Moreover, thanks to subsequent interviews, they highlighted the fact that the learners' perceptions of their own proficiency level as well as the nature of their interlocutor (peer or tutor), influenced strategic behavior in choosing whether to lexically align or not.

Moving away from lexical alignment, Michel and Stiefenhöfer (2019) studied how learners of Spanish reacted to primed subjunctives, finding that the experimental group using CMC did produce more target structures in Spanish than the control group. CMC seems therefore to be a relevant communication environment to study alignment. Kim et al. (2020) compared face-to-face and written SCMC settings to investigate Korean learners' structural alignment of direct vs. indirect questions in English. Their results indicate that a primed direct or indirect question is more likely to produce structural alignment. Moreover, in their comparison of face-to-face vs. CMC settings, they found that the latter elicited more direct questions, but not indirect ones. Alignment may therefore vary depending on the structures considered. Consequently, we have chosen in our current exploratory study to focus on as many structures as possible.

Finally, all these studies investigated alignment in *written* SCMC. To our knowledge, the only study that dealt with audiovisual SCMC (i.e. desktop videoconferencing) is Michel and Cappellini (2019), which the present article expands on. In this study, the authors developed a comparison between written and audiovisual SCMC looking at both lexical and structural alignment. Their case study raised a certain number of methodological issues, of which we aim to address three in the present paper. The first is that the oral nature of videoconferencing results in the presence of conversational disfluencies (Pallaud et al., 2013), which need to be identified in order not to be coded as instances of alignment. The second issue is the need for a lemma-based approach in the study of lexical alignment, rather than an n-gram approach as in previous studies, in order to detect alignment subject to morphological inflections. The third issue is the need to take into account not only the verbal features of communication, but also the visual ones. The present study tackles these issues and proposes a methodological framework for the study of alignment in videoconferencing settings.

3. Research questions

The present paper aims to continue exploring lexical and structural alignment in CMC settings and more precisely in telecollaboration while developing methodological innovation. Moreover, since we focus on CMC through desktop videoconferencing, we are interested in the particular features of this setting. Research on desktop videoconferencing in pedagogical settings showed the importance of visual cues (Develotte et al., 2010; Guichon & Wigham, 2016). Since in this setting the main bodily element is the presence of the interlocutor's face, we are interested in studying if and how alignment is realized in the modality of facial expressions.

Our first and main research question is: how can interactive alignment be studied in telecollaborative settings through desktop videoconferencing? More precisely, we are interested in:

A. improving the methodological framework for lexical alignment suggested in Michel and Cappellini (2019);

B. applying the framework for structural alignment from the same authors;

C. building a framework to study interactive alignment in the modality of facial expressions.

Our second research question is: how do occurrences of alignment in different modalities relate to each other? This question aims to explore the issue of multimodality underscored by Rasenberg et al. (2020). Since the framework of interactive alignment posits that alignment at one level percolates to other levels, we are interested in studying if and to what extent this happens across the levels of lexis, grammatical structures and facial expression.

4. Methods

4.1 Context and participants

The interactions under analysis for the present article are part of the VAPVISIO project[3], which aims to compare two telecollaboration configurations (Cappellini & Azaoui, 2017) in order to understand what techno-pedagogic skills are necessary to teach online, and which of them require formal training to be developed. For the present study, we focus on two of the four datasets of the whole project.

The first dataset comes from *Le français en (première) ligne* (Develotte et al., 2007; hereafter F1L), in which future teachers of a foreign language are connected with learners of the same language, in our case French (see Cappellini & Hsu, 2020, for further details). In this setting, each trainee teacher from France interacted with two learners from a US university on the Adobe Connect platform. All the participants were in their twenties. The trainee teacher in this study was a male student pursuing a master’s degree in teaching French as a foreign language. The two American students were third-year undergraduates who were studying science and enrolled in a French course in order to attain the B2 level.

The second dataset comes from a French-American teletandem setting. In a teletandem setting (Telles, 2009; hereafter TT), two learners who want to learn each other’s mother tongue

[3] https://www.ortolang.fr/market/corpora/vapvisio

interact half the time in each language in order to help each other (cf. Michel, Appel & Cipitria, this issue). Learners interacted using Skype. The two learners examined in this study were also in their twenties and were enrolled in their third year of undergraduate studies, respectively in foreign languages and cultures for the French student and in international commerce for the American one.

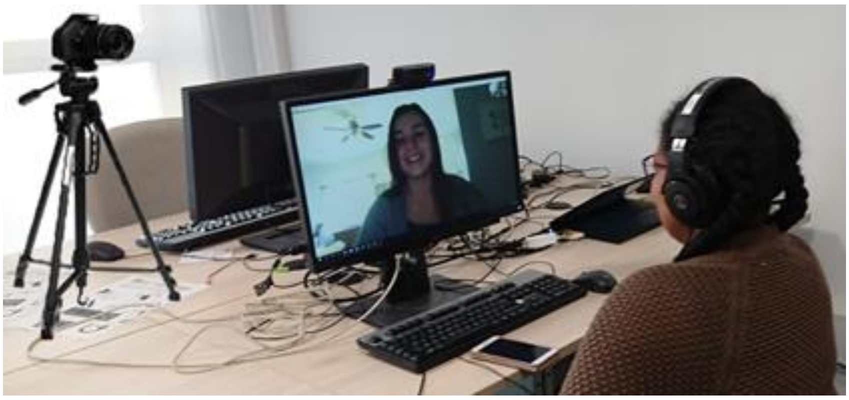
In each setting, the same learners met three (F1L) or five (TT) times over a semester. During each one-hour session, the participants completed learning tasks related to different topics, including students' physical mobility and sustainable development. In this paper, we selected the first session for an F1L group and the fifth session for a TT pair. We chose these sessions because learners completed the same task, consisting in an analysis and commentary of the UNESCO's infographic about the objectives of sustainable development and a meme about "green washing" by multinational corporations (Appendix 1). Choosing a session with the same learning task is a way for us to neutralize the possible influence of the learning task on alignment (see also Michel et al., this issue).

4.2 Data collection and corpus

The interactions took place during the spring semester of the 2018–2019 academic year. The interactions that are being analyzed were recorded in an isolated room at Aix-Marseille University's language center. Participants wore headphones and a microphone while sitting in front of an external monitor, webcam, mouse and keyboard that were hooked up to a laptop computer. The bottom of the monitor was equipped with a Tobii X3-120 infrared bar that captured participants' eye movements. In addition to the laptop computer, which was used for the videoconferencing and eye-tracking software, there was a second desktop computer that was used to record the sound via an external soundcard in order to have one audio channel per interlocutor and enable use of SPPAS (see below). Finally, an external camera was aimed at the interlocutor in order to capture contextual elements such as hand gestures that were produced outside the webcam's field of view (Guichon, 2009) [4]. The picture below presents the data collection setting.

Figure 1

Data Collection Setting



We recorded everything that was displayed on the external monitor using the Tobii Pro Studio software, and audio data were recorded using Audacity. The video files were compressed using the VSDC free video editor before being synchronized and mixed with the sound using Adobe Premiere Pro video editing software. Video files were exported in MP4 format at 1920x1080 resolution at 30 fps, and sound files (one per interlocutor) were exported in Waveform (WAV) format at 44.1 kHz 16-bit mono.

For audio data, SPPAS (Bigi, 2015) was used to automatically generate empty labels corresponding to Inter-Pausal Units (IPUs) of each interlocutor, meaning blocks of speech separated by pauses of 200 milliseconds or more. These "labels" were subsequently imported as tiers in ELAN (Sloetjes & Wittenburg, 2008) for manual transcription. The final corpus for the two sessions is represented as follows.

[4] We consider eye-tracking data and gestures outside the frame of the webcam in other publications. For instance Cappellini & Hsu (2022).

Table 1

Corpus of Study

Interaction	Time	Participant	Number of IPUs	Number of words	Total IPUs	Total words
F1L	38:16	Alain[5] (tutor)	933	2844	1153	4185
		Stephen	158	1056		
		Isa	62	285		
TT	1:04:35	Océlia	2750	5218	4074	8100
		Tabitha	1324	2882		

4.3 Analysis procedures

According to Rasenberg et al. (2020), alignment can be studied with respect to time, sequence, meaning, form and modality. These variables can be used as either grouping criteria or measurement variables. The relationship between the different modalities depends on whether one adopts a priming or a grounding perspective. For example, priming, which is automatic, non-intentional and low-level, prioritizes form and time. Grounding, which is controlled, intentional and higher-level, prioritizes sequence, form, and meaning. In our study, we adopt a priming perspective since the alignment framework is part of it and operationalize it in different ways for each modality as we will explain in the following.

4.3.1 Lexical alignment

To study lexical alignment, we looked for other-repetitions within the next five IPUs. For the two interactions selected to study lexical alignment, we used the automatic annotation tools

[5] All names are pseudonyms.

built into SPPAS (Bigi, 2015) to detect lexical repetitions. Before automatic annotation was possible, several steps were necessary to prepare the data.

After the speech transcription, each speaker tier in ELAN was exported as a TextGrid file to be edited using PRAAT. PRAAT was then used to duplicate the transcription file for each speaker in order to have one tier for English and one for French (in the case of TT interaction) and in order to discard any English words (in the case of the F1L interaction). We changed the name of each tier to "Transcription" so that it would be recognized as such by SPPAS, and exported one TextGrid transcription file per speaker per language. We then edited each TextGrid file with NotePad++ in order to fill in any empty intervals with the "#" symbol, which is recognized by SPPAS as a silence. We also ensured that each laugh was represented by just one "@" symbol and not multiple symbols. Then, we duplicated and renamed the sound file for each speaker (in Waveform format) so that each speaker had a corresponding sound file with an identical name. Once the WAV and TextGrid files were prepared, we imported them into SPPAS in order to perform automatic annotations.

We linked speakers, two at a time, in order to create "interactions" recognized by SPPAS. Four interactions were created: two for the F1L session (tutor-learner1 and tutor-learner2[6]), and two for the TT session (one interaction for French and one for English). For each interaction, SPPAS performed speaker annotations and interaction annotations. The following automatic annotations were performed on each individual speaker:

- Normalization: punctuation is removed, numbers are converted to letters, text is tokenized, etc.
- Phonetization: text is converted into its phonetic constituents.

[6] We did not consider learner-learner alignment since learners did not interact among themselves.

- Sound-transcription linking: phonemes and tokens are located in time and aligned with the sound file.

For the interaction annotations, SPPAS searched for “other repetitions” (Bigi et al., 2014) at the level of lemmas, within a window of five IPUs, meaning that it searched for words or sequences of words that were repeated either strictly or with a variation, a split or a reduction. An example of a “variation” is when the original utterance is “which is better” and the echo is “which is the better one.” A “split” is when the repeated utterance is divided across separate IPUs, and SPPAS lists the number of IPUs over which it is split, for example split:2. An example of a “reduced” echo is when the original speaker says “bonjour bonjour” and the interlocutor only says “bonjour.” This analysis was possible thanks to lemmatization of the words in the transcription. SPPAS also compared each word with a list of “stop words,” that is, words that are very common, in order to avoid counting them as repetitions. The English list of stop words contains 149 words, and the French list contains 65. Some examples for English are: more, not, on, then, and with.

After performing these steps, we combed through the SPPAS output log in order to correct any spelling errors, unrecognized words, or IPUs that had been assigned to the wrong language. After each correction, we recreated a new TextGrid file and repeated the steps above.

For each interaction, two sets of other repetition annotations were created, one for each direction (Speaker A repeats Speaker B, and speaker B repeats Speaker A). Therefore, for our two telecollaboration sessions, we had four “interactions” in SPPAS and eight sets of annotations in total. Each set of annotations contains the following eight elements:

Table 2

Set of Annotations for Lexical Alignment

Other repetition source	This is the name of the IPU that is automatically created containing a word or series of words that are repeated by the other speaker. They are named S1, S2, S*n* for *n* sources that are repeated.
Source strain	This is the first word that is repeated.
Source Len	This is the number of words in the series that are repeated.
Source type	This indicates whether the repetition is strict, a variation, a split or a reduction.
Echo	This is the name of the IPU that is created containing the word or series of words that are repeated. They are named R1, R2, R*n* for *n* repetitions.
Token Strain	These are the tokenized words that are repeated.
Stop word	This tier indicates whether the tokenized word is part of the list of stop words. If a word is considered a stop word, then it is not counted as a repetition.
Token strain	These are the tokenized words that are repeated, labelled “echo” in SPPAS.

We configured SPPAS to export the annotation files in the TextGrid format so that they could be easily imported into ELAN for viewing, or into an Excel file for annotation of multimodality (below).

After importing the other repetition annotations into ELAN for viewing, we manually checked 10% of them to make sure that they corresponded to real lexical repetitions. For instance, for the F1L interaction, we manually verified repetitions #1, 11, and 21. In the IPU marked by SPPAS as S1, meaning “Source 1,” Alain says “bonjour bonjour” (“hello hello”). In the IPU marked by SPPAS as R1, meaning “Repetition 1,” Isa says “bonjour.” SPPAS marked this as a “reduction” because she says “bonjour” only once. This manual check therefore counts as a success. Repetitions 11 and 21 were also successes, so “Isa repeats Alain in French” earns a success rate of 100%. The table below shows the accuracy rate for the manual annotation.

Table 3

Manual Verification of Automatic Annotations

Telecollaboration setting	Interaction	Accuracy rate of 10% manual verifications
F1L	Isa repeats Alain in French	100% for 3 verifications
	Alain repeats Isa in French	100% for 2 verifications
	Stephen repeats Alain in French	100% for 4 verifications
	Alain repeats Stephen in French	100% for 2 verifications
TT	Tabitha repeats Océlia in French	100% for 7 verifications
	Océlia repeats Tabitha in French	100% for 4 verifications
	Océlia repeats Tabitha in English	100% for 3 verifications
	Tabitha repeats Océlia in English	100% for 4 verifications

Occasionally, a repetition label did not align with the source word label. For example, SPPAS marked a source as S5, but there is no R5 to be found. As it turned out, S5 actually aligned with R4. When this happens, it is because a source word was said multiple times by an interlocutor (for example "document"), but was only repeated once. SPPAS therefore marked both utterances of "document" as S4 and S5 for the original speaker, but there was only one R4 for the interlocutor, which was the "echo" of both S4 and S5. In total, 29 manual verifications were carried out for 234 automatic annotations, with a success rate of 100%. Finally, we calculated the ratio of the number of lexical alignment instances per 100 seconds to allow comparison between the two sessions with different lengths.

4.3.2 Structural alignment

For structural alignment, we drew on the categories of analysis from Michel and Cappellini (2019), slightly adapting them for French. For the TT session, categories were coded for the English part as well. We coded the following categories:

Table 4

Categories for Structural Alignment

Aller + complément	To go + complement
Avoir + complément	To have + complement
Il y a existentiel	There is / there are
Construction infinitive	/
Participe présent	/
Passif	Passive voice
Futur proche	Going to
Subordonnée relative	Relative clause

Pouvoir	Modal verbs
Devoir	
Vouloir	
J'aime / j'aime pas	Like / dislike
Être + complément de lieu	To be + place
Être copule	Copular be
État mental	Mental state

We coded these categories manually within ELAN. To do so, we created a tier for each interlocutor with a controlled vocabulary composed of the categories above. For the TT session, we created separate tiers for French and English. After coding, we exported the verbal turns as well as the structural alignment annotations and imported them into an MS Excel file. In this file, manual annotations were built for primes and entrainments, distinguishing between those of the NS and those of the NNS. Structural alignment was annotated as far as one minute after the prime.

There were two differences from a previous case study (Michel & Cappellini, 2019). The first one is that alignment was considered only across participants, in order to focus only on the interactive dimension of it. The second one is that an occurrence representing an entrainment could serve as a prime for a subsequent alignment by the interlocutor, within a one-minute frame. This annotation scheme resulted from considering the tension between two phenomena. First, we decided to count an entrainment as a prime for a subsequent entrainment because each appearance of the structure in the workspace (Pickering & Garrod, 2021) implies the activation of the mental representation related to that structure. Second, we decided to arbitrarily set the frame size to one minute because some of the structures (e.g. *copular be*) are fairly frequent. Therefore, we thought that their appearance after a long time in the interaction was not due to alignment to a lasting activated mental representation of the structure, but simply the result of constructing a sentence to express an idea.

Finally, as for lexical alignment, for the F1L session we only included alignment between the tutor and one learner because of the characteristics of the interaction.

We only present descriptive statistics, given that our aim is mainly methodological and that our sample consists of only two telecollaboration sessions. After counting instances of each prime and each alignment, distinguishing between the NS and the NNS(s), we calculated the total number of occurrences for each structure, the alignment score and the non-alignment score. Finally, as for lexical alignment, we calculated the ratio of the number of cases of alignment per 100 seconds to ensure comparability.

4.3.3 Facial expression alignment

For emotional facial expression alignment, we used a commercial facial emotion recognition programming interface, BaiduAPI [7], to train and process images extracted from our video data. BaiduAPI is based on Ekman's (1992) classification and can identify seven basic emotion types: angry, disgusted, fearful, happy, sad, surprised and neutral. In our analysis, we were less interested in the automatic identification of emotion than in the alignment of facial expressions.

Before training and identifying the facial expression types, for each tutoring session that we analyzed, we used a customized script to split videos of both sides of the speakers into individual images, with an interval of one second between images. In cases when there was more than one speaker in the video, another customized script was used to locate and extract

[7] https://ai.baidu.com/ai-doc/FACE/yk37c1u4t

the face area of that speaker before extracting the images from the video. Images were extracted with specific time stamps of their source video.

We then used BaiduAPI to process the images and identify the facial expressions associated with emotions. The result of each image was one of the seven emotion types. Analysis was not possible throughout the entire interaction due to different sizes of the interlocutors' pictures, or possible moments when the French interlocutor covered the image of one or more interlocutors. In the TT session, only 1 minute and 32 seconds for the French learner and 6 seconds for the American learner were missing. In the F1L session, facial expression analysis was not possible for 2 minutes and 58 seconds for the tutor, 32 minutes and 27 seconds for the first learner and 8 minutes and 6 seconds for the second learner (21% of the interaction). As we will discuss, this had an impact on our results.

The final step of analysis was to enter the results of the automatic facial expression annotation in separate columns in an MS Excel sheet, then to automatically compare the annotations of facial expressions, therefore identifying when they matched. For the F1L session, analysis was run for tutor-learner1 and tutor-learner2. To minimize the impact of the missing analyses, we calculated the rate of alignment cases per 100 seconds, taking as the total number of seconds the number of seconds available for analysis. Moreover, we did not take into account alignment with the "neutral" facial expression, since this does not correspond to a behavior that one can align to.

4.3.4 Analysis of multimodality

To analyze how cases of alignment in different modalities relate to each other, we exported the transcriptions of speech, the automatic annotations of lexical alignment, the manual annotations of structural alignment, and the automatic annotations of facial expression alignment into a MS Excel file. All the annotations and the transcription were organized in different columns, aligned according to the time labels of each.

After this preparation, we searched for co-occurrences of alignment in two or three modalities. For each annotation of alignment, we looked one second before and one second after, to see if an annotation in another modality also occurred. We restricted it to one second as we considered multimodal alignment happening in more than one modality from a synchronous or near-synchronous point of view (Rasenberg et al., 2020). For lexical alignment including more than one lemma (for example when a multi-word utterance is repeated), we counted only one occurrence, since at this stage our focus is on multimodality and not on the number of words aligned. The possible categories for multimodal alignment were the following: facial-lexical (FL), facial-structural (FS), lexical-structural (LS), facial-lexical-structural (FLS). After manual annotation, we calculated the number of occurrences for each type of multimodal alignment, then the ratio with respect to the total number of annotated occurrences of alignment.

5. Results

In the following, we first present the descriptive statistics for each modality under investigation, with a short discussion. Second, we present a multimodal analysis, and third a comparison between the results in the two telecollaboration models.

5.1 Alignment at the different levels

5.1.1 Lexical alignment

The following table presents the total number of occurrences for each lexical alignment.

Table 5

Occurrences of Lexical Alignment per SPPAS' Interaction

Interaction	Number of lexical alignments
Isa repeats Alain in French	29
Alain repeats Isa in French	8
Stephen repeats Alain in French	38
Alain repeats Stephen in French	20
Tabitha repeats Océlia in French	68
Océlia repeats Tabitha in French	38
Océlia repeats Tabitha in English	23
Tabitha repeats Océlia in English	40

If we group results for telecollaboration setting, and we differentiate between NNS-NS alignment and NS-NNS alignment, this results in the following table.

Table 6

Lexical Alignment NNS-NS vs. NS-NNS

Telecollaboration setting	NNS-NS alignment	NS-NNS alignment
F1L	67	28
TT French	68	38
TT English	23	40

The results show similar figures for F1L and the French part of TT. In these cases, the NNSs align more to the NSs than the other way around. In English, we find the opposite, with the NS aligning more to the NNS. As for languages, for the F1L session, learners lexically align to the tutor 67 times, and he aligns to them 28 times. For the TT session, learners align much more often in French than they do in English (106 times vs. 63 times, ratio of 1.7). This could be partly due to the fact that they spend more time talking in French than in English (2309 IPUs vs. 1837 IPUs, ratio 1.2).

As for the ratio of number of occurrences of alignment per 100 seconds, in the F1L session there are 95 lexical alignment examples for 2296 seconds, that is 4.14 occurrences of alignment per 100 seconds. In the TT session, there are 169 examples of alignment in 3875 seconds, that is 4.36 occurrences of alignment per 100 seconds.

5.1.2 Structural alignment

Analysis of structural alignment resulted in the following tables.

Table 7

Structural Alignment in the F1L Session (French only)

	prime NS	prime NNS	NS aligns NNS	NNS aligns NS	Nonalignment	Total occurrences	Alignment score	Nonalignment score
aller + compl	0	0	0	0	0	0	0	0
avoir + compl	0	2	2	0	11	15	0.13	0.73

il y a existentiel	3	0	1	4	19	27	0.18	0.70
construction infinitive	0	1	1	1	4	7	0.28	0.57
participe présent	0	0	0	0	0	0	0	0
passif	0	0	0	0	0	0	0	0
futur proche	1	0	0	1	16	18	0.05	0.89
subordonnée relative	0	0	0	0	0	0	0	0
pouvoir	1	2	4	1	12	20	0.25	0.6
devoir	0	0	0	0	0	0	0	0
vouloir	0	0	0	0	1	1	0	1
j'aime / j'aime pas	0	0	0	0	0	0	0	0
ê + compl lieu	0	0	0	0	1	1	0	1
ê copule	1	2	5	1	18	27	0.22	0.67
état mental	1	0	0	3	17	21	0.14	0.81
	7	7	13	11	99	137	0.17	0.72

Table 8

Structural Alignment in the Teletandem Session – French

	fr_primeNS	fr_primeNNS	fr_NS aligns NNS	fr_NNS aligns NS	Nonalignment	Total occurrences	Alignment score	Nonalignment score
aller + compl	0	0	0	0	7	7	0	1
avoir + compl	2	1	1	2	19	25	0.12	0.76
il y a existentiel	2	0	0	5	25	32	0.16	0.78
construction infinitive	0	0	0	0	9	9	0	1
participe présent	0	0	0	0	0	0	0	0
passif	0	0	0	0	0	0	0	0
futur proche	1	0	0	2	12	15	0.13	0.8
subordonnée relative	0	0	0	0	0	0	0	0
pouvoir	0	0	0	0	12	12	0	1
devoir	0	0	0	0	2	2	0	1
vouloir	0	0	0	0	1	1	0	1
j'aime / j'aime pas	0	0	0	0	4	4	0	1

ê + compl lieu	0	0	0	0	4	4	0	1
ê copule	7	5	32	17	58	119	0.41	0.49
état mental	5	1	16	14	31	67	0.45	0.46
	17	7	49	40	184	297	0.30	0.62

Table 9

Structural Alignment in the Teletandem Session – English

	eng_primeNS	eng_primeNNS	eng_NS aligns NNS	eng_NNS aligns NS	Nonalignment	Total occurrences	Alignment score	Nonalignment score
go + compl	0	0	0	0	3	3	0	1
have + compl	0	1	4	1	15	21	0.29	0.71
there is/there are	1	1	8	2	12	24	0.42	0.5
passive voice	0	0	0	0	0	0	0	0
going to	0	0	0	0	0	0	0	0
relative clause	0	0	0	0	8	8	0	1

modal verb	3	1	7	3	24	38	0.26	0.63
like/dislike	0	0	0	0	0	0	0	0
be + place	0	0	0	0	4	4	0	1
copular *be*	5	2	17	15	38	77	0.41	0.49
mental state	3	1	11	10	20	45	0.47	0.44
	12	6	47	31	124	220	0.35	0.56

The analysis shows that the structures we considered are mostly present, with only 5 categories out of 26 never used in the interactions. The total number of occurrences, all categories combined, during the 2 hours of interaction is 654.

The alignment and nonalignment scores show that in general, interlocutors do not align within 1 minute on the structures considered. When they do, it is usually in both directions in French, with the NS speaker aligning to the NNS roughly as much as the NNS aligns to the NS: 13 times in the F1L session and 49 times in the TT session vs. 11 times in the F1L and 40 times in the TT sessions, respectively. The situation is different for the English part of the TT session, in which the NS aligns to the NNS 47 times, while the NNS aligns to the NS only 31 times. This result is consistent with those at the level of lexical alignment.

Overall, there is less alignment in the F1L session than in the TT session, both in terms of absolute number of occurrences (26 vs. 167; among which 89 in French and 78 in English) and in terms of alignment scores (0.17 vs. 0.30 in French and 0.35 in English). The rate of structural alignment occurrences per 100 seconds also confirm this trend, with 1.04 in the F1L session vs. 4.31 in the TT session. In other words, for this dataset, the TT setting seems more conducive to structural alignment than the F1L setting.

5.1.3 Facial expression alignment

The results of facial expression types and their occurrences of alignment are summarized in the following tables.

While the majority of emotion types identified in both videos fall under the categories Happy and Neutral, emotion types presented in each video show some differences.

Table 10 below focuses on the F1L session and shows that all three speakers expressed five types of emotion, with Neutral being the most common, followed by Happy. In terms of the third dominant facial expression type, the tutor showed more occurrences of Sad and Surprise, while the third most common category among the learners was Sad. Learner 1 showed few instances of Angry facial emotion, but this type was not observed in the other learner and the tutor. The numbers relate to the total number of seconds during which the facial expression was detected.

Table 10

Counts (and Percentages) of Emotion Types for each Speaker in the F1L Session

	Tutor	Learner 1	Learner 2
Happy	533 (25%)	22 (7%)	95 (5%)
Neutral	1357 (64%)	292 (90%)	1688 (94%)
Sad	106 (5%)	6 (2%)	10 (1%)

Fear	8 (0%)	1 (0%)	6 (0%)
Surprise	104 (5%)	1 (0%)	1 (0%)
Angry	0 (0%)	4 (1%)	0 (0%)
Total (4234)	2108	326	1800

As shown in Table 10 above, the analysis of the TT session shares the same major categories with the F1L session, except for one extra emotion type: Disgust. Unlike the F1L session, the most frequent type that we identified in this exchange was Happy for the French learner and Sad for the American learner, while the second was Neutral for the French learner and Happy for the American learner.

Table 11

Counts (and Percentages) of Emotion Types for each Speaker in the TT Session

	Tutor	Learner 1
Happy	1655 (44%)	864 (24%)
Neutral	1075 (29%)	462 (13%)
Sad	664 (18%)	2156 (59%)
Fear	2 (0%)	33 (0%)
Surprise	2 (0%)	34 (0%)
Angry	12 (0%)	75 (2%)

Disgust	316 (8%)	50 (1%)
Total (7400)	3726	3674

To observe possible facial expression alignment during tutor-learner conversations, we measured for each second during the conversation whether the same type of emotion was observed between the tutor and the learner. The results are summarized in the following tables.

Table 12

Counts (%) of Tutor-Learner Emotion Alignment in the F1L Session

	Tutor-Learner1	Tutor-Learner2	Teletandem Tutor-Learner
Happy	11 (5.5%)	64 (5.5%)	347 (51%)
Neutral	189 (94.5%)	1080 (94.5%)	55 (8%)
Sad	0	0	275 (41%)
Fear	0	0	0
Surprise	0	0	0
Angry	0	0	0
Disgust	0	0	0
Total	200	1144	677

In the F1L session, the most common type was Neutral, followed by Happy. The ratios that we observed between each tutor-learner interaction were very similar.

In the TT session, the tutor and the learner were aligned in expressing Happy and Sad emotion types, which were the most common. Neutral came in third, but was much less common than the first two types. Compared with the F1L session, we can see that speakers in the TT session seemed to be more expressive in their facial expressions during their conversions. This is also confirmed by the rate of total instances of alignment per 100 seconds. In the F1L session, there are 3.15 instances of alignment per 100 seconds between the tutor and learner 1 and 3.53 between the tutor and learner 2, while it is 16.5 in TT.

5.2 Multimodal alignment

The analysis of co-occurrences of alignment in different modalities within a one-second time frame resulted in the following numbers. As mentioned above, the possible categories for multimodal alignment were facial-lexical (FL), facial-structural (FS), lexical-structural (LS), facial-structural-lexical (FLS).

Table 13

Multimodal Alignment

	F1L		TT	Total
	Tutor-learner1	Tutor-learner2		
FL	0	0	41	41
FS	3	9	55	67
LS	0	1	22	23
FLS	0	0	6	6
Total	3	10	124	137

Results show a striking imbalance between the two telecollaboration settings, with the TT session presenting almost ten times more occurrences of multimodal alignment than the F1L session. This may be partly due to characteristics of the dataset. In fact, we noted that the TT session lasted longer than the F1L session, and that in the F1L session, large parts of facial expression annotations are missing. However, for the different lengths, the rate of multimodal alignment instances per 100 seconds confirms the trend, with 0.32 in F1L vs. 3.2 in TT. As for the impact of missing analysis for facial expression, we can note that even for the category where facial expression is not involved (LS), the difference cannot be explained solely on the lesser amount of interaction in the F1L session. Despite the biases induced by our dataset, the TT pair aligns more than the F1L learners do with their tutor.

As for the categories of multimodal alignment that are the most common, the combination of facial expression and structural alignment is the most present in both sessions, followed by the co-occurrence of facial expression and lexical alignment. The co-occurrence of alignment on three levels of multimodality is rare, and happens for only 6 seconds throughout the whole corpus.

6. Discussion

This paper proposes a methodological approach to study alignment in desktop videoconferencing-based telecollaboration from a multimodal perspective. Multimodal alignment is defined here as the co-occurrence, within one second, of instances of alignment in two or three modalities.

6.1 Methodological considerations

Our first research question was a methodological one. Above we have provided extensive details on the procedure that enabled us to study alignment. More precisely, we explained how the study of lexical alignment can be augmented by the use of an automatic annotation tool, SPPAS (Bigi, 2015), to analyze other-repetitions at the level of lemmas, and to take into consideration one-word repetitions as well as multi-word expressions (up to 6 words in the present dataset). For the present study, we arbitrarily set the search window to 5 IPUs after each word. In future research, it may be useful to vary the search window and to qualitatively compare the results from different parts of the datasets. This would allow us to understand if a smaller or larger limit can be more interesting and theoretically grounded to study lexical alignment.

As for structural alignment, this study confirmed the validity of the method we derived from previous ones (Dao et al., 2018; Michel & Cappellini, 2019). Automatic detection of facial expression alignment also proved to be a viable method. However, we experienced a large amount of data loss for the F1L session, in which interlocutors' faces were sometimes covered by other graphical elements on the screen, due to the manipulation of interfaces by the tutor. For future data collection aimed at studying (also) facial expression alignment, it will be important to ensure optimal conditions for onscreen facial visibility, or to use an external camera to keep track of facial expressions when they are not visible (enough) on the screen. We hope that this method will inspire future work, especially CMC settings for language learning, including telecollaboration.

6.2 Multimodal alignment

Our second research question was: how do instances of alignment in different modalities relate to each other? As far as we know (see also Rasenberg et al., 2020), our study is the first to research multimodal alignment as it occurs at three different levels, and the first to take an interest in facial expressions. We therefore have no other study to which we can compare our findings. Three observations emerged from this study. The first observation is that our method allows for the study of how alignment emerges in different modalities at the same time. We identified 2,262 instances of alignment, broken down into 264 lexical, 654 structural, and 1,344 facial expression instances of alignment. Of these, 206[8] occurrences happened at the same time within a one-second frame, of which only 18 occurred across all three modalities. In other words, only 15% of the occurrences of alignment we detected are at least bi-modal, and 1% are trimodal. This result will need to be compared with results from larger datasets. Yet, our data already suggest that alignment does not often percolate simultaneously (i.e., within a one-second window) across different modalities, that is at different levels of communication (Pickering & Garrod, 2021). How alignment percolates will need to be studied empirically in wider timeframes to understand if alignment in one modality leads to subsequent alignment in other modalities. We think this is an important finding in shaping our understanding of alignment.

The second observation is the importance of facial visibility in videoconferencing settings. When facial expression data are available to us, mainly in the TT session, analysis seems to indicate that this semiotic resource is pivotal in multimodal alignment; in fact, facial expressions are involved in 102 out of 124 multimodal instances of alignment (82%), followed by structural alignment with 83 cases (67%) and finally lexical alignment with 69 out of 124 (56%). The study of alignment therefore seems to confirm previous observations about the importance of facial visibility in language learning in videoconferencing settings, which has already been noted for instance by Guichon (2017) for social presence in telecollaboration, or by Satar and Wigham (2017) for instruction-giving sequences.

6.3 Toward a comparison of telecollaboration settings

The analysis of our data also showed some differences and similarities of how alignment emerges in the two telecollaboration settings. Given that we only consider two sessions for

[8] This number represents the total multiplied by 2 for bimodal cases of alignment and multiplied by 3 of trimodal alignment presented in table 12.

each telecollaboration setting, we will need to explore larger datasets. In other words, the observations in the present section are aimed at orienting future research on larger datasets rather than providing substantial evidence of a given dynamic.

That being stated, our results show a much greater potential of the TT setting to be more conducive to alignment. Different reasons can be advanced to explain this. The first is that we studied the first session for the F1L group, whereas it was the fifth for the TT pair. The fact that the F1L group was meeting for the first time means that they could have still been getting to know each other and that alignment might have still been in its early stages. A future study comparing the first session of the TT pair with the last session of the F1L group may provide further insight to this hypothesis. As for lexical alignment, results show that in French, the NNSs tend to align more to the NS than the other way around. In other words, the learners incorporate more words used by their interlocutor in their own speech, which, following Coste and Cavalli (2015), could be interpreted as an indicator of socialization in the foreign language. However, the figures are different for the English part of the TT session, in which the NS aligns significantly more to the NNS than the other way around (40 vs. 23 times). This challenges the hypothesis by Lewis (2020), who suggests that in tandem, learners mostly adopt the language of the NSs. In our opinion, this difference between the two parts of the TT session that we analyzed indicates that for this pair, there might be a consistent dynamic in which the relationship between interlocutors drives alignment, with Tabitha aligning more to Océlia, no matter which language is used. This is also visible in structural alignment in the English part, in which Tabitha aligns 47 times to Océlia, while the latter aligns to Tabitha only 31 times. This, however, is partly contradicted by structural alignment in the French part, in which Océlia aligns more to Tabitha (49 vs. 40 times). As for facial expressions, it is not possible to compare the two settings given that we could not analyze the large amounts of data in the F1L session. This absence of facial expression analysis for part of the F1L session, coupled with the observation that facial expressions play a prominent role in multimodal alignment, could be one of the reasons for the striking difference in multimodal alignment we find in our data, which is almost absent in the F1L session. All in all, a pedagogical implication to our work might be that the TT session shows a higher potential for alignment for all the modalities considered, and therefore for multimodal alignment as well.

7. Limitations

Several limitations of the current study have been highlighted throughout the text, the main one being that we worked with a small dataset that is based on two telecollaboration sessions. Moreover, the limitations include the difference in familiarity between the interlocutors, due to the fact that we focused on the first session for the F1L setting and the fifth one for TT. Methodologically, it could be questioned how exactly alignment should be coded after primes, especially for lexical and multimodal alignment. In particular, the size of the search window is open to debate, as this calls into the question the extent to which mental representations remain active in an interlocutor's mind after priming. We hope that insights to come from neuropsychology and neurolinguistics will offer support to how applied linguists methodologically implement the study of alignment.

8. Conclusions

The main aim of our article was to illustrate and test a methodological framework to study alignment in desktop videoconferencing-based telecollaboration from a multimodal perspective. In this sense, it provided an answer to Rasenberg et al.'s (2020) call for more research into alignment taking into account the multimodality of communication. The article drew on existing methodological frameworks and presented methodological innovations to study lexical and facial expression alignment, and to articulate the study of alignment across different modalities, which we hope will be useful for researchers and practitioners interested in studying alignment. The analysis of our small dataset showed that alignment rarely happens simultaneously across two or three modalities. We concluded that if alignment percolates, it does so in a sequential way, which will also need to be taken into account in future studies. Moreover, our study confirmed the importance of facial visibility in desktop videoconferencing-based interactions. Finally, the TT session we observed showed a greater amount of alignment than the F1L session. Future studies will continue to explore these dimensions, in order to paint a more complete picture of multimodal alignment in telecollaboration.

References

Akiyama, Y. (2019). Using Skype to Focus on Form in Japanese Telecollaboration: Lexical Categories as a New Task Variable. In I. Management Association (Eds.), *Computer-Assisted Language Learning: Concepts, Methodologies, Tools, and Applications* (pp. 617-647). IGI Global.

Aljaafreh, A. & Lantolf, J.P. (1994). Negative feedback as regulation and second language learning in the zone of proximal development. *The Modern Language Journal, 78(4),* 465-483. https://doi.org/10.2307/328585

Avgousti, M. I. (2018). Intercultural communicative competence and online exchanges: A systematic review. *Computer Assisted Language Learning, 31*(8), 819-853. https://doi.org/10.1080/09588221.2018.1455713

Belz, J. A. (2003). From the special issue editor. *Language Learning & Technology*, *7*(2), 2–5. http://dx.doi.org/10125/25193

Bezemer, J., & Kress, G. (2016). *Multimodality, learning and communication: A social semiotic frame*. Routledge. https://doi.org/10.4324/9781315687537

Bigi, B. (2015). SPPAS-multi-lingual approaches to the automatic annotation of speech. *The Phonetician, 111*(112), 54-69.

Bigi, B., Bertrand, R., & Guardiola, M. (2014). Automatic detection of other-repetition occurrences: Application to French conversational speech. *Proceedings of the Ninth Conference on Language Resources and Evaluation (LREC)* (pp. 836-842). Reykjavik.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory and cognition, 22*(6), 1482-1493. https://doi.org/10.1037/0278-7393.22.6.1482

Cappellini, M., & Azaoui, B. (2017). Sequences of normative evaluation in two telecollaboration projects: A comparative study of multimodal feedback through desktop videoconference. *Language Learning in Higher Education, 7*(1), 55-80. https://doi.org/10.1515/cercles-2017-0002

Cappellini, M., & Hsu, Y. Y. (2020). When future teachers meet real learners through telecollaboration: An experiential approach to learn how to teach languages online. *Journal of Virtual Exchange, 3,* 1-11. https://doi.org/10.21827/jve.3.35751.

Cappellini, M., & Hsu, Y. Y. (2022). Multimodality in webconference-based language tutoring: An ecological approach integrating eye tracking. *ReCALL Journal,* 34(3), 255-273. https://doi.org/10.1017/S0958344022000076

Costa, A., Pickering, M. J., & Sorace, A. (2008). Alignment in second language dialogue. *Language and cognitive processes, 23*(4), 528-556. https://doi.org/10.1080/01690960801920545

Coste, D., & Cavalli, M. (2015). *Education, mobility, otherness. The mediation functions of schools*. Council of Europe.

Cunningham, J. (2017). Second language pragmatic appropriateness in telecollaboration: The influence of discourse management and grammaticality. *System, 64*, 46-57. https://doi.org/10.1016/j.system.2016.12.006

Develotte, C., Guichon, N., & Vincent, C. (2010). The use of the webcam for teaching a foreign language in a desktop videoconferencing environment. *ReCALL*, *22*(3), 293-312. https://doi.org/10.1017/S0958344010000170

Develotte, C., Mangenot, F. & Zourou, K. (2007). Learning to teach online: 'Le français en (première) ligne' project. In R. O'Dowd (Ed.). *Online Intercultural Exchange. An Introduction for Foreign Language teachers* (pp. 276-280). Multilingual Matters.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3/4), 169–200. https://doi.org/10.1080/02699939208411068

European Centre for Modern Languages (2021, April 27th). *The future of language education – learning lessons from the pandemic* [Webinar]. ECML/CELV. https://www.ecml.at/ECML-Programme/Programme2020-2023/Thefutureoflanguageeducation/tabid/5491/Default.aspx

Gass, S. M. (1997). *Input, Interaction, and the Second Language Learner*. Routledge. https://doi.org/10.4324/9780203053560

Guichon, N. (2009). Training future language teachers to develop online tutors' competence through reflective analysis. *ReCALL*, *21*(2), 166-185. https://doi.org/10.1017/S0958344009000214

Guichon, N. (2017). Se construire une présence pédagogique en ligne. In N. Guichon, & M. Tellier (Eds.), *Enseigner l'oral en ligne. Une approche multimodale* (pp. 31-61). Didier.

Guichon, N. & Tellier, M. (Eds.) (2017). *Enseigner l'oral en ligne. Une approche multimodale*. Didier.

Guichon, N., & Wigham, C. R. (2016). A semiotic perspective on webconferencing-supported language teaching. *ReCALL*, *28*(1), 62-82. https://doi.org/10.1017/S0958344015000178

Jackson, C. N. (2018). Second language structural priming: A critical review and directions for future research. *Second Language Research, 34*(4), 539-552. https://doi.org/10.1177/026765831774620

Kern, R. (2006). La communication médiatisée par ordinateur en langues: recherches et applications récentes aux USA. *Français dans le monde. Recherches et applications, 40*, 17-31.

Kim, Y., Skalicky, S. & Jung, Y. (2020). The role of linguistic alignment on question development in face-to-face and synchronous computer-mediated communication contexts: A conceptual replication study. *Language Learning*, *70*(3), 643-684. https://doi.org/10.1111/lang.12393

Kötter, M. (2003), Negotiation of meaning and codeswitching in online tandems. *Language Learning and Technology*, *7(2)*, 145-172.

Lantolf, J. P. & Thorne, S. L. (2006). *Sociocultural Theory and the Genesis of Second Language Development*. Oxford University Press.

Lewis, T. (2020). From tandem learning to e-tandem learning: How languages are learnt in tandem exchanges. In S. Gola, M. Pierrard, E. Tops, & D. Van Raemdonck (Eds.), *Enseigner et apprendre les langues au XXIe siècle* (pp. 107-128). Peter Lang. https://doi.org/10.3726/b16391

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. Ritchie & T.K. Bhatia (éds). *Handbook of language acquisition: Vol. 2. Second language acquisition* (pp. 413–468). Academic Press of San Diego.

McDonough, K., Kim, Y. L., Uludag, P., Liu, C., & Trofimovich, T. (2022). Exploring the relationship between behavior matching and interlocutor perceptions in L2 interaction. *System*, *109*. https://doi.org/10.1016/j.system.2022.102865.

Michel, M., & Cappellini, M. (2019). Alignment during synchronous video versus written chat L2 interactions: A methodological exploration. *Annual Review of Applied Linguistics*, *39,* 189-216. https://doi.org/10.1017/S0267190519000072

Michel, M., & O'Rourke, B. (2019). What drives alignment during text chat with a peer vs. a tutor? Insights from cued interviews and eye-tracking. *System*, *83*, 50-63. https://doi.org/10.1016/j.system.2019.02.009

Michel, M., & Smith, B. (2018). Measuring lexical alignment during L2 chat interaction: An eye-tracking study. In S. M. Gass & P. Spinner & J. Behney (Eds.), *Salience in second language acquisition*. (pp.244-268). Routledge.

Michel, M., & Stiefenhöfer, L. (2019). Priming Spanish subjunctives during synchronous computer-mediated communication: German peers' classroom-based and homework interactions. In M. Sato &S. Loewen (Eds.), *Evidence-based second language pedagogy: A collection of instructed second language acquisition studies*. (pp. 191-218). Routledge.

Nassaji, H. (2016). Anniversary article Interactional feedback in second language teaching and learning: A synthesis and analysis of current research. *Language Teaching Research*, *20*(4), 535-562. https://doi.org/10.1177/1362168816644940

O'Dowd, R. (2018). From telecollaboration to virtual exchange: state-of-the-art and the role of UNICollaboration in moving forward. *Journal of Virtual Exchange*, *1,* 1-23. https://doi.org/10.14705/rpnet.2018.jve.1

O'Dowd, R., & O'Rourke, B. (2019). New developments in virtual exchange in foreign language education. *Language Learning & Technology, 23*(3), 1-7. https://doi.org/10125/44690

Pallaud, B., Rauzy, S., & Blache, P. (2013). Auto-interruptions et disfluences en français parlé dans quatre corpus du CID 1. *TIPA. Travaux interdisciplinaires sur la parole et le langage*, 29, 1-19.

Pickering, M. J., & Ferreira, V. S. (2008). Structural priming: a critical review. *Psychological bulletin*, *134*(3), 427-459. https://psycnet.apa.org/doi/10.1037/0033-2909.134.3.427

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, *27*(2), 169-190. https://doi.org/10.1017/S0140525X04000056

Pickering, M. J., & Garrod, S. (2009). Prediction and embodiment in dialogue. *European Journal of Social Psychology*, *39*(1), 1162–1168. https://doi.org/10.1002/ejsp.663

Pickering, M. J., & Garrod, S. (2021). *Understanding dialogue: Language use and social interaction*. Cambridge University Press.

Rasenberg, M., Özyürek, A., & Dingemanse, M. (2020). Alignment in multimodal interaction: An integrative framework. *Cognitive Science*, *44*(11): e12911. https://doi.org/10.1111/cogs.12911

Satar, M. H., & Wigham, C. R. (2017). Multimodal instruction-giving practices in webconferencing-supported language teaching. *System*, *70*, 63-80. https://doi.org/10.1016/j.system.2017.09.002

Sloetjes, H., & Wittenburg, P. (2008). Annotation by category-ELAN and ISO DCR. In *6th international Conference on Language Resources and Evaluation (LREC 2008).*

Suffil, E., Kutasi, T., & Pickering, M. J. (2021). Lexical alignment is affected by addressee but not speaker nativeness. *Bilingualism: Language and cognition*, First View, 1-12. DOI: https://doi.org/10.1017/S1366728921000092

Swain, M. (2000). The output hypothesis and beyond: Mediating acquisition through collaborative dialogue. In J.P. Lantolf (Ed.). *Sociocultural theory and second language learning* (pp. 97-114). Oxford University Press.

Tekin, O.,Trofimovich, T., Chen, T.-H., & McDonough, K. (2022). Alignment in second language speakers' perceptions of interaction and its relationship to perceived communicative success, *System, 108,* https://doi.org/10.1016/j.system.2022.102848.

Telles, J. A. (Ed.) (2009). Teletandem. Um contexto virtual, autônomo, colaborativo para aprendizagem das linguas estrangeiras no século XXI. Pontes Editores.

Van der Zwaard, R., & Bannink, A. (2019). Towards a Comprehensive Model of Negotiated Interaction in Computer-mediated Communication. *Language Learning & Technology*, *23*(3), 116-135. https://doi.org/10125/44699

Van Ek, J. A., & Trim, J. L. M. (1991). *Threshold level 1990*. Council of Europe.

Varonis, E. M., & Gass, S. (1985). Non-native/ non-native conversation. A model for negotiation of meaning. *Applied Linguistics, 6*(1), 71-90. https://doi.org/10.1093/applin/6.1.71

Ware, P. & O'Dowd, R. (2008). Peer feedback on language form in telecollaboration. *Language Learning & Technology, 12*(1), 43-63. http://dx.doi.org/10125/44130

Wigham, C. R. (2017). A multimodal analysis of lexical explanation sequences in webconferencing-supported language teaching. *Language Learning in Higher Education*, *7*(1), 81-108. https://doi.org/10.1515/cercles-2017-0001

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162–185. https://psycnet.apa.org/doi/10.1037/0033-2909.123.2.162

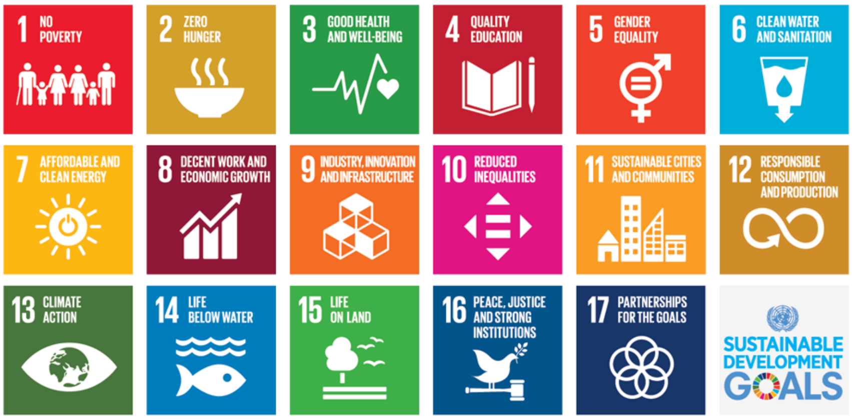
Appendix. Pedagogical material

Activité : Le développement durable

Vous commencerez cette session en langue anglaise[9].

Pour démarrer : qu'est-ce que vous connaissez sur la thématique du développement durable ? Quelles sont vos opinions sur le sujet ?

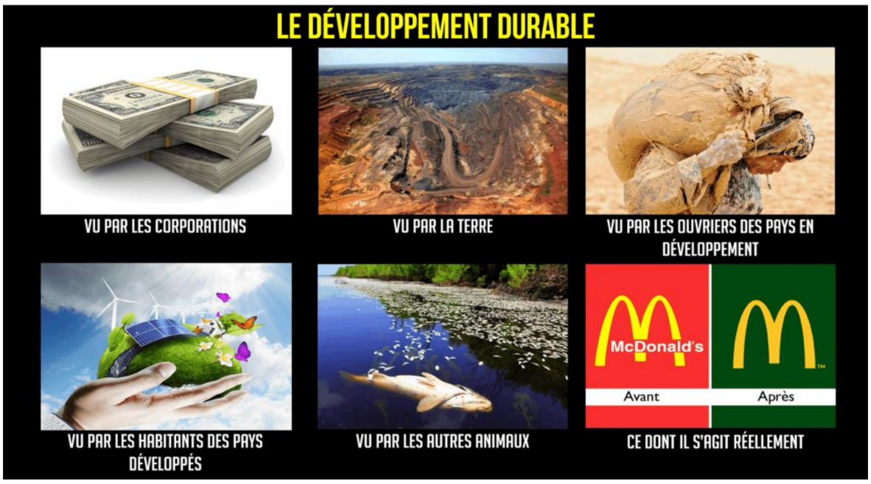
Observez l'infographie suivante (tirée de https://en.unesco.org/sdgs).


1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS
SUSTAINABLE DEVELOPMENT GOALS

1. Discutez pour répondre ensemble à ces questions : qui est l'auteur ? Qui sont les destinataires ? Quelle est le but de ce document ?
2. Chacun regarde les différentes cases et explique à l'autre une initiative écoresponsable qu'il/elle connaît en relation avec une des bases.
3. Chacun choisit un objectif parmi ceux des cases et imagine un changement qu'il/elle peut faire dans son quotidien pour contribuer à cet objectif.
4. Chacun choisit un objectif parmi ceux des cases et imagine un changement que son université peut faire pour contribuer à cet objectif.
5. Allez sur les sites web de vos universités concernant le développement durable
   https://sustainability.asu.edu/
   https://developpement-durable.univ-amu.fr/
   Est-ce que les initiatives que vous avez imaginées existent déjà ?

[9] We reproduce the task instructions for the teletandem setting. In the F1L, the tutor had the same instructions and the whole interaction was conducted in French.

Dans la partie en langue française, commencez par regarder ce document :


LE DÉVELOPPEMENT DURABLE
VU PAR LES CORPORATIONS
VU PAR LA TERRE
VU PAR LES OUVRIERS DES PAYS EN DÉVELOPPEMENT
VU PAR LES HABITANTS DES PAYS DÉVELOPPÉS
VU PAR LES AUTRES ANIMAUX
McDonald's
Avant
Après
CE DONT IL S'AGIT RÉELLEMENT

1. Discutez pour répondre ensemble à ces questions
2. Qui est l’auteur ? Qui sont les destinataires ? Quelle est le but de ce document ?
3. Quelle est la structure du document ? Quels sont les moyens utilisés pour exprimer le message ?
4. Quels contre-arguments peut-on imaginer face à ce document ?
5. Quelle est votre opinion concernant l’idée exprimée dans ce document ?

Figures and tables

Tables and figures

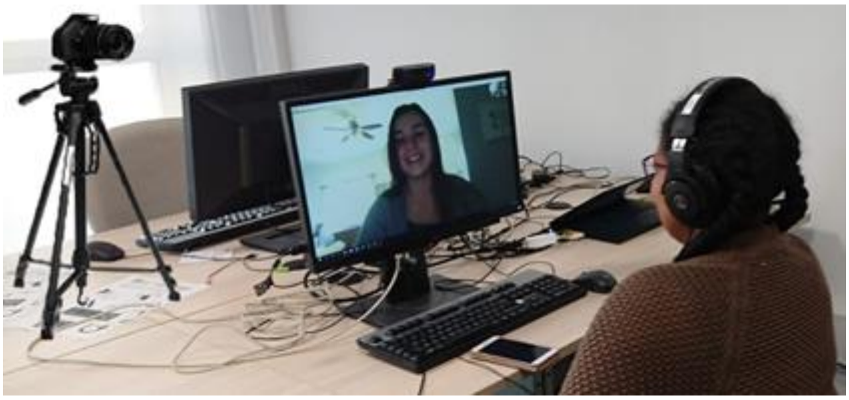


Figure 1. Data collection setting.

Interaction	Time	Participant	Number of IPUs	Number of words	Total IPUs	Total words
F1L	38:16	Alain[1] (tutor)	933	2844	1153	4185
		Stephen	158	1056		
		Isa	62	285		
TT	1:04:35	Océlia	2750	5218	4074	8100
		Tabitha	1324	2882		

Table 1. Corpus of study.

[1] All names are pseudonyms.

Other repetition source	This is the name of the IPU that is automatically created containing a word or series of words that are repeated by the other speaker. They are named S1, S2, S*n* for *n* sources that are repeated.
Source strain	This is the first word that is repeated.
Source Len	This is the number of words in the series that are repeated.
Source type	This indicates whether the repetition is strict or has variations such as being split or reduced.
Echo	This is the name of the IPU that is created containing the word or series of words that are repeated. They are named R1, R2, R*n* for *n* repetitions.
Token Strain	These are the tokenized words that are repeated.
Stop word	This tier indicates whether the tokenized word is part of list of stop words. If a word is considered a stop word, then it is not counted as a repetition.
Token strain	echo These are the tokenized words that are repeated.

Table 2. Set of annotations for lexical alignment.

Telecollaboration setting	*Interaction*	*Accuracy rate of 10% check*

F1L	Isa repeats Alain in French	100% for 3 manual verifications
	Alain repeats Isa in French	100% for 2 manual verifications, 1 automatic annotation is mislabeled
	Stephen repeats Alain in French	100% for 4 verifications, 1 is mislabeled
	Alain repeats Stephen in French	100% for 2 verifications
TT	Tabitha repeats Océlia in French	100% for 7 verifications, 1 is mislabeled
	Océlia repeats Tabitha in French	100% for 4 verifications
	Océlia repeats Tabitha in English	100% for 3 verifications
	Tabitha repeats Océlia in English	100% for 4 verifications, 2 are mislabeled.

Table 3. Manual verification of automatic annotations.

Aller + complément	To go + complement

Avoir + complément	To have + complement
Il y a existentiel	There is / there are
Construction infinitive	/
Participe présent	/
Passif	Passive voice
Futur proche	Going to
Subordonnée relative	Relative clause
Pouvoir	Modal verbs
Devoir	
Vouloir	
J’aime / j’aime pas	Like / dislike
Être + complément de lieu	To be + place
Être copule	Copular be
État mental	Mental state

Table 4. Categories for structural alignment.

Interaction	Number of lexical alignment
Isa repeats Alain in French	29
Alain repeats Isa in French	8
Stephen repeats Alain in French	38
Alain repeats Stephen in French	20
Tabitha repeats Océlia in French	68
Océlia repeats Tabitha in French	38
Océlia repeats Tabitha in English	23
Tabitha repeats Océlia in English	40

Table 5. Occurrences of lexical alignment per SPPAS' interaction

Telecollaboration setting	NNS-NS alignment	NS-NNS alignment
F1L	67	28
TT French	68	38
TT English	23	40

Table 6. Lexical alignment NNS-NS vs. NS-NNS

	prime NS	primeNNS	NS aligns NNS	NNS aligns NS	Nonalignment	Total occurrences	Alignment score	Nonalignment score
aller + compl	0	0	0	0	0	0	0	0
avoir + compl	0	2	2	0	11	15	0,13	0,73
il y a existentiel	3	0	1	4	19	27	0,18	0,70
construction infinitive	0	1	1	1	4	7	0,28	0,57
participe présent	0	0	0	0	0	0	0	0
passif	0	0	0	0	0	0	0	0
futur proche	1	0	0	1	16	18	0,05	0,89
subordonnée relative	0	0	0	0	0	0	0	0
pouvoir	1	2	4	1	12	20	0,25	0,6

devoir	0	0	0	0	0	0	0	0
vouloir	0	0	0	0	1	1	0	1
j'aime / j'aime pas	0	0	0	0	0	0	0	0
ê + compl lieu	0	0	0	0	1	1	0	1
ê copule	1	2	5	1	18	27	0,22	0,67
état mental	1	0	0	3	17	21	0,14	0,81
	7	7	13	11	99	137	0,17	0,72

Table 7. Structural alignment in the F1L session (French only)

	fr_primeNS	fr_primeNNS	fr_NS aligns NNS	fr_NNS aligns NS	Nonalignment	Total occurrences	Alignment score	Nonalignment score
aller + compl	0	0	0	0	7	7	0	1
avoir + compl	2	1	1	2	19	25	0,12	0,76

il y a existentiel	2	0	0	5	25	32	0,16	0,78
construction infinitive	0	0	0	0	9	9	0	1
participe présent	0	0	0	0	0	0	0	0
passif	0	0	0	0	0	0	0	0
futur proche	1	0	0	2	12	15	0,13	0,8
subordonnée relative	0	0	0	0	0	0	0	0
pouvoir	0	0	0	0	12	12	0	1
devoir	0	0	0	0	2	2	0	1
vouloir	0	0	0	0	1	1	0	1
j'aime / j'aime pas	0	0	0	0	4	4	0	1

ê + compl lieu	0	0	0	0	4	4	0	1
ê copule	7	5	32	17	58	119	0,41	0,49
état mental	5	1	16	14	31	67	0,45	0,46
	17	7	49	40	184	297	0,30	0,62

Table 8. Structural alignment in the teletandem session – French

	eng_primeNS	eng_primeNNS	eng_NS aligns NNS	eng_NNS aligns NS	Nonalignment	Total occurrences	Alignment score	Nonalignment score
go + compl	0	0	0	0	3	3	0	1
have + compl	0	1	4	1	15	21	0,29	0,71
there is/there are	1	1	8	2	12	24	0,42	0,5
passive voice	0	0	0	0	0	0	0	0

going to	0	0	0	0	0	0	0	0
relative clause	0	0	0	0	8	8	0	1
modal verb	3	1	7	3	24	38	0,26	0,63
like/dislike	0	0	0	0	0	0	0	0
be + place	0	0	0	0	4	4	0	1
copular be	5	2	17	15	38	77	0,41	0,49
mental state	3	1	11	10	20	45	0,47	0,44
	12	6	47	31	124	220	0,35	0,56

Table 9. Structural alignment in the teletandem session – English

	Tutor	Learner 1	Learner 2
Happy	533 (25%)	22 (7%)	95 (5%)

Neutral	1357 (64%)	292 (90%)	1688 (94%)
Sad	106 (5%)	6 (2%)	10 (1%)
Fear	8 (0%)	1 (0%)	6 (0%)
Surprise	104 (5%)	1 (0%)	1 (0%)
Angry	0 (0%)	4 (1%)	0 (0%)
Total (4234)	2108	326	1800

Table 10. Counts (and percentages) of emotion types for each speaker in the F1L session

	Tutor	Learner 1
Happy	1655 (44%)	864 (24%)
Neutral	1075 (29%)	462 (13%)
Sad	664 (18%)	2156 (59%)
Fear	2 (0%)	33 (0%)
Surprise	2 (0%)	34 (0%)
Angry	12 (0%)	75 (2%)
Disgust	316 (8%)	50 (1%)

Total (7400)	3726	3674

Table 11. Counts (and percentages) of emotion types for each speaker in the TT session

	Tutor-Learner1	Tutor-Learner2
Happy	11 (5.5%)	64 (5.5%)
Neutral	189 (94.5%)	1080 (94.5%)
Sad	0	0
Fear	0	0
Surprise	0	0
Total	200	1144

Table 12. counts (%) of tutor-learner emotion alignment in the F1L session

	Tutor-Learner
Happy	347 (51%)
Neutral	55 (8%)
Sad	275 (41%)

Fear	0
Surprise	0
Angry	0
Disgust	0
Total	677

Table 13. Counts (%) of tutor-learner emotion alignment in the TT session

	F1L		TT	Total
	Tutor-learner1	Tutor-learner2		
FL	0	0	41	41
FS	3	9	55	67
LS	0	1	22	23
FLS	0	0	6	6
Total	3	10	124	137

Table 14. Multimodal alignment