Building a multimodal corpus to study the development of techno-semio-pedagogical competence across different videoconferencing settings and languages

Benjamin Holt, Marco Cappellini, Brigitte Bigi, Christelle Zielinski

▶ **To cite this version:**

Benjamin Holt, Marco Cappellini, Brigitte Bigi, Christelle Zielinski. Building a multimodal corpus to study the development of techno-semio-pedagogical competence across different videoconferencing settings and languages. 2021. halshs-03476577

HAL Id: halshs-03476577
https://shs.hal.science/halshs-03476577

Preprint submitted on 13 Dec 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Building a multimodal corpus to study the development of techno-semio-pedagogical competence across different videoconferencing settings and languages

Abstract

Our article aims to describe the constitution and annotation of a multimodal and multilingual (French, English, and Mandarin Chinese) corpus for the study of skills development in online language teaching as part of a research project. Starting from the scientific objective of observing the development of techno-semio-pedagogical competence, we explain the theoretical framework adopted, characterised by an ecological approach to the interactive environment. We subsequently illustrate the procedures for the collection of audio, video and eye-tracking data. We also provide details concerning the annotation of raw data to produce a corpus of analysis drawing on different automatic and semi-automatic annotation tools.

Keywords: multimodality, videoconferencing, eye-tracking, second language acquisition

Introduction

The development of desktop videoconferencing (DVC) tools and broadband Internet has allowed for the emergence of innovative pedagogical practices in foreign language education. In fact, DVC has been widely integrated into pedagogical environments since the-mid 2000s (O'Dowd, 2016) to allow students in different countries to communicate and collaborate online and has since become a common tool for Computer-Assisted Language Learning (CALL) (O'Dowd & O'Rourke, 2019). These pedagogical environments, often named "telecollaboration" (O'Dowd, 2016) or "virtual exchange", make way for a variety of pedagogical models (Mangenot, 2013). Among these models, the most widespread is e- or tele-tandem (Little & Brammerts, 1996; Telles, 2009), in which pairs of students of different mother tongues interact in both languages to improve their learning. This model does not aim to develop learners' teaching skills and there is therefore no training or counselling provided to students for this purpose. On the contrary, another model, based on the *Français en (Première) Ligne* (F1L) project (Develotte et al., 2008), consists of trainee teachers developing learning activities and implementing them with language learners during DVC sessions as part of their professional development. The objective of F1L-based projects is the development of techno-semio-pedagogical (TSP) competence (Guichon, 2012, defined below). We present a corpus that allows direct comparison of these two models.

We first explain the objectives and the theoretical framework of the research project within which the corpus was created. Second, we present a detailed account of the audio, video and eye-tracking data collection procedures, including ethical considerations. Third, we explain the data annotation process that was adapted to build the corpus.

1. Objectives and theoretical framework of the VAPVISIO project

The project *Vers une approche comparative de l'apprentissage/enseignement des langues étrangères par visioconférence pour développer les compétences techno-semio-pédagogiques d'enseignants en formation*[1] (henceforth VAPVISIO) aims to promote advancements in two related fields:

1. The modelization of collaborative language learning and teaching in online innovative pedagogical practices called *telecollaborations* or *virtual exchanges*;

1 Towards a comparative approach to second language learning/teaching through desktop videoconferencing to develop trainee teachers' techno-semio-pedagogical skills. https://anr.fr/Projet-ANR-18-CE28-0011

2. The construction of a methodological framework that combines the benefits of in-depth case studies and broad-based comparisons of general learning dynamics across learning environments.

The main objective of our project is to study the development of TSP competence across two pedagogical models. This will allow us to understand which skills develop naturally through DVC interaction (telecollaboration) and which skills need formal training with an instructor to emerge (F1L).

1.1 Definition of techno-semio-pedagogical competence

The literature presents several models of pedagogical competence to teach languages with Information and Communication Technologies (ICT) (Dooly 2010, Kessler 2016 among others). One of the most widespread models for online language teaching is Hampel & Stickler's pyramid model (2005, 2015), which identifies basic skills for teaching language with ICTs and then moves up to identify more individual and creative skills. However, despite a specific study on DVC (Hampel & Stickler, 2012), this model is neither specifically conceived for, nor largely adapted to this type of Computer-Mediated Communication (CMC). On the contrary, Guichon's model of techno-semio-pedagogical competence (2012) was initially designed to describe integration of ICT in language teaching in general, and subsequently adapted to teaching through DVC (Guichon & Tellier, 2017). This model defines techno-semio-pedagogical competence as "knowledge and skills about:

- communication tools available (forum, wiki, videoconference, etc.) that are most suitable for the objectives of a pedagogical sequence;
- taking into account the appropriate modes (written, oral, video, or a combination thereof) for a given activity and for the development of linguistic competences;
- the pedagogical management of learning activities with and peripheral to CMC tools (planning, regulations during task completion, learning assessments)."[2] (Guichon, 2012: 187).

TSP *competence* as a whole can therefore be subdivided into discreet TSP *skills* for the purpose of analysis. This general definition was subsequently adapted to the specific context of teaching French as a Foreign Language through DVC in the F1L-based ISMAEL telecollaboration project (Guichon & Tellier, 2017), making it possible to operationalize the model and identify specific recommendations for online teachers and teacher trainers.

1.2 Theoretical framework and limitations

A review of the existing literature shows that there is only a small amount of studies on teacher education in CALL and telecollaboration, relying not on trainee teachers' perceptions but on analysis of actual pedagogical practices (Arnold & Ducate, 2015). Within this small body of research, TSP competence is observed at specific moments and not longitudinally (see Guichon & Tellier 2017, Kurek & Muller-Hartmann 2019 for recent examples), making it difficult to study its evolution over time. Moreover, most of these studies focus on only one telecollaborative model, which hinders a comparative approach that would enable one to determine which skills need formal training to develop.

Another gap in the literature that the VAPVISIO project aims to fill concerns the definitions of TSP competence and the methodological tools to observe it. In fact, most of the authors agree that this competence includes not only the ability to effectively use relevant modes and strategies for communication (and possibly teaching), but also knowledge (see Guichon's definition above) and awareness (Hauck, 2010) of the semiotic modes available. This knowledge/awareness has been

2 Our translation.

studied through retrospective introspection, especially through learning logs (i.e. Fuchs et al. 2012) and more rarely through stimulated recall (Cohen, 2017), but never within (inter)action itself. Our project aims to surpass this approach and its limits by trying to gain insights into the process of selection of modes in DVC interactions during the communication itself. This is made possible with the recent development of eye-tracking data collection and methodologies (i.e. Stickler & Shi 2017). More precisely, our aim is to inscribe the observation of the development of TSP competence within a wider ecological framework (Van Lier, 2004; Cappellini & Combe, 2017) allowing us to comprehend the DVC environment in terms of affordances (Blin, 2016), meaning a set of possibilities and constraints for action (Cappellini & Hsu, 2018).

1.3 Scientific objective and hypotheses

The study of this corpus will allow us to understand, through comparison of two pedagogical models over the course of a semester, which TSP skills require training to be developed. This implies that we address the scientific and technical barriers defined as 'limits' above. In our project we build a corpus and a methodological framework allowing comparison not only of the two pedagogical models (teletandem and F1L-based telecollaboration), but also of the learning of different languages (French, English and Mandarin Chinese). Our main assumption is that since teletandem lacks any guidance on developing TSP competence, such skills emerge naturally through online communication (Cappellini, 2014). On the contrary, in F1L-based telecollaboration, where professional guidance is present by definition, we expect tutors to develop more advanced levels of TSP competence. The comparison between the TSP competence developed in teletandem and in F1L-based telecollaboration would therefore allow us to identify the TSP skills that need particular guidance and instruction to emerge. This corpus will allow two hypotheses to be tested concerning the development of TSP competence across pedagogical configurations and over time. We predict that future studies based off of our corpus will find that:

1. Students' TSP skills are more advanced at the end of a telecollaborative project than at the beginning.
2. Students' TSP skills at the end of a F1L project are superior to those of students at the end of a teletandem project.

2. Data collection

Data collection took place during the spring semester in 2019. The first teletandem project paired students from Aix-Marseille University (AMU) with students from Arizona State University (ASU) for learning French and English. The second teletandem project paired students from AMU with students from Shenzhen Foreign Language University (深圳外国语学院) for learning French and Mandarin Chinese. All of the students for these teletandem projects were at the undergraduate level and most were pursuing language-related degrees. As for the F1L-based telecollaboration, the first project paired graduate students enrolled in a master's program for teaching French as a foreign language at AMU with undergraduate students of French at the University of California Berkeley. The second project paired graduate students enrolled in a master's program for teaching Mandarin Chinese as a foreign language at the Hong Kong Polytechnic University (香港理工大学) with undergraduate learners of Chinese at AMU. All the learners in the four telecollaboration projects had a proficiency level ranging between B1 and B2 on the *Common European Framework of Reference for Languages* scale (Council of Europe, 2001). In each telecollaboration group, five pairs participated in the data collection, except for the AMU-ASU teletandem in which only four pairs were present.

	French-(English)	*Chinese-(French)*
Teletandem setting	4 pairs Aix-Marseille University – Arizona State University	5 pairs Shen Zhen University – Aix-Marseille University

F1L setting	5 groups Aix-Marseille University – University California Berkeley	5 groups Hong Kong Polytechnic University – Aix-Marseille University

Table 1. Composition of groups for data collection

2.1 Ethical issues and consent

All of the students provided informed consent before data collection, and participation in the data collection was voluntary. Depending on the country or university involved, ethical issues were dealt with in a slightly different way in order to obtain the necessary authorizations complying with local regulations. For instance, at the French university, the coordinator of the project filled out the *Formulaire d'inscription au registre* of the Delegate to data protection at the CNRS. This document was presented and given to each of the French students during a 30-minute individual meeting before they signed the informed consent form. The document specifies, among other things: the types of data and how they are collected, stored, and shared, participants' rights to access or delete data, and the objectives of the project.

2.2 Audio-visual recording

As Popescu-Belis (2010: 211) points out, "the capture of multimodal corpora requires complex settings such as instrumented lecture and meetings rooms, containing capture devices for each of the modalities that are intended to be recorded, but also, most challengingly, requiring hardware and software for digitizing and synchronizing the acquired signals." In the following sections we provide details on these challenges.

The resolution of the audio recording devices (microphones, file format, software) and the environment have a determining influence on the quality of the corpus as well as on the annotations. These factors are important to ensure that researchers will be able to exploit the recordings. The following recommendations were considered based on previous data collection within the research team:

1. One channel per speaker, i.e. one microphone per speaker;
2. The use of a professional grade head-worn cardioid microphone with a maximum audio frequency bandwidth;
3. The use of an anechoic room, or an environment with no/low noise;
4. The use of uncompressed file formats for the recordings, commonly WAV;
5. The use of sampling rates of at least 16000 Hz – ideally 48000 Hz, with 16 bits;
6. Synchronization of audio signals when multiple speakers or recording devices are present. A clap in the beginning of the recording can facilitate this task.

As for point 3, data were collected at the *Centre de Formation et Autoformation en Langues* of AMU in a dedicated room. We complied with point 4 by using the audio recording software Audacity[3], and with point 5 by using a Roland Rubix 22 audio interface linked to an AKG C520 headset microphone. Points 1, 2 and 6 can be problematic for videoconferencing environments because it is not easy to obtain one channel per speaker, but these points were addressed by collaborating with the *Centre d'Expérimentation sur la Parole*[4] (for 6, see below). The first step consisted in dividing the audio streams of the two interlocutors. The person present in the recording room was equipped with the above-mentioned headset microphone that was subsequently split and directed in parallel to:

1. The computer used for DVC so that the interlocutor on the other end could hear;

3 https://audacity.fr/

4 http://cep.lpl-aix.fr/

2. The Roland Rubix 22 external sound card connected to a second computer for recording using Audacity.

The distant interlocutor's output audio stream was also split and directed to both the DVC computer so that the interlocutor in France could hear, and to the separate computer for recording. This procedure allowed us to obtain a separate sound channel for each. However, the distant partners on the other end did not duplicate this process, and had only one sound channel.

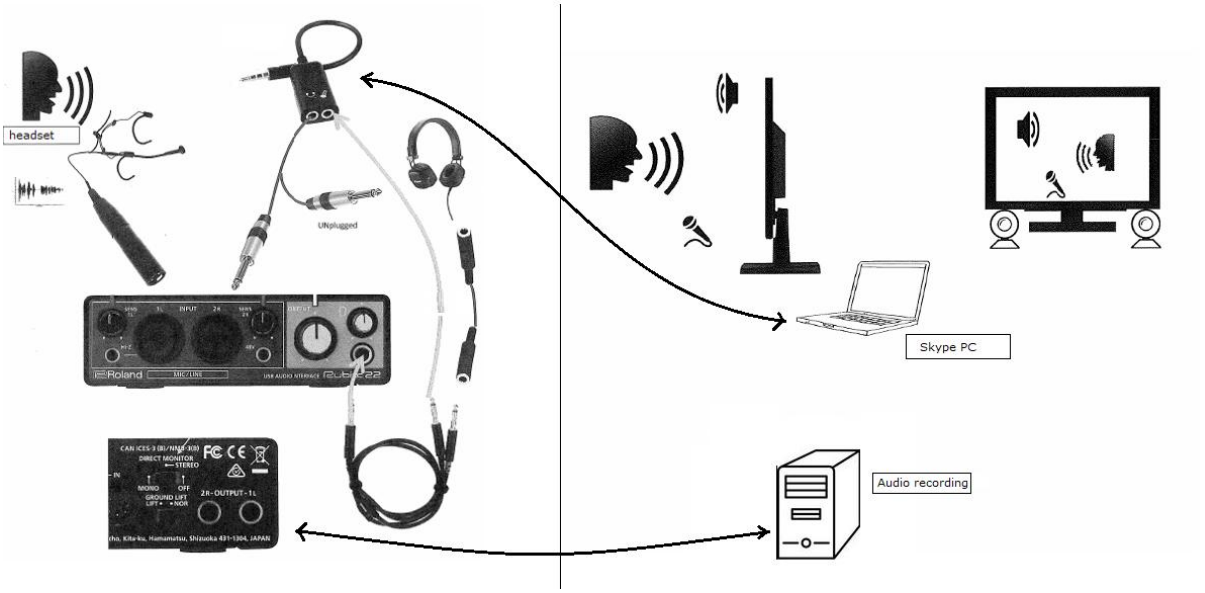

Figure 1. Audio recording setting

Video was recorded from multiple sources. As mentioned, the interactions took place in a separate classroom with two computers: a laptop for the videoconferencing and a desktop for sound recording. During the interactions, the students sat in front of an external monitor connected to the laptop. The external monitor had a Tobii eye-tracking bar attached to the bottom, as well as an external keyboard and mouse. The laptop's screen and keyboard were used solely by the researchers to calibrate the eye-tracker and launch the recording at the beginning of each interaction. Students manipulated the external keyboard and mouse to use the videoconferencing platform, an Internet browser, and various pedagogical documents that were displayed on the external monitor.

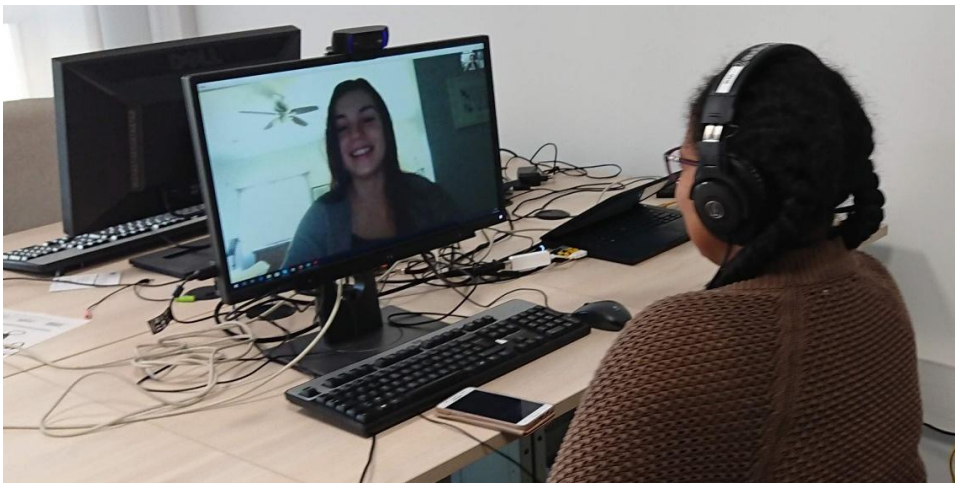

Figure 2. Videoconference environment

Video data come from two main sources: dynamic screen recordings that were recorded directly from the Tobii Studio software[5] (including the user's microphone recording), and video recordings from an external camera. In addition to recording eye movements, the Tobii software recorded everything that happened on the screen during the entirety of each interaction. The external camera was placed on a tripod on a table behind the external monitor slightly offset to the left. The camera's field of view therefore included the interlocutor's torso, hands and face, as well as possible use of objects on the table such as notepads and smartphones, allowing us to record all hand gestures produced, including those that were not visible to the webcam.

Figure 3. View from the external camera

Sound recordings come from three sources. Aside from the audio recording procedure described above with the desktop computer, audio recordings were also retrieved from the dynamic screen captures with Tobii Studio on the laptop and from the external camera. The audio tracks from the Tobii video exports and from the external camera were used only to synchronize the different video recordings, and were ultimately discarded. Only the sound files from the external sound card were used for the audio transcription and for the final video exports. The table below summarizes the different types of audio-visual data that were collected.

Recording device	**Sound**	**Video**	**Eye-tracking**
Tobii Studio software on laptop computer	Local interlocutor's voice, one channel	Screen recording	Eye-tracking data were collected (see below)
Audacity software on desktop computer with external sound card	Local and distant interlocutors' voices, split into separate channels		
External camera	Ambient sound in the room (mostly local interlocutor's voice)	View of the local interlocutor's head and torso, the desk, and immediate surroundings	

Figure 4. Audio-visual channels recorded

Data export and synchronization

After each recording, video and audio files were exported using the Tobii Studio software on the laptop computer, and the Audacity software on the desktop. Using the Tobii Studio program, two screen recording video files per interaction were exported, one with eye movements and one without,

[5] https://www.tobiipro.com/fr/produits/tobii-pro-studio/

each at 30 frames per second. The video with eye movements represents each fixation as a red dot superimposed onto the video, with red lines connecting successive fixations. Each red dot represents the average position of the estimated gaze position of each eye. The red dots increase in size according to the fixation duration. These two video files, as well as the video data from the external camera, were subsequently compressed into MP4 video files using the H264 video codec in the VSDC free video editor in order to reduce their size and to make them usable in Adobe Premiere Pro and in ELAN (see below). The stereo sound file (one channel per interlocutor) recorded using Audacity on the desktop computer was split and exported as two separate mono files in the waveform audio file format (WAV), one for each interlocutor, at 16 bits/44.1 kHz.

The MP4 video files and the WAV sound files were then synchronized using Adobe Premiere Pro video editing software. The WAV sound files and the external camera video files were adjusted (stretched or compressed due to different playback speeds) to match the exact length of the video files that were exported from Tobii, because the raw eye-tracking data are aligned with these. Five new files were exported per interaction: two sound files in WAV format (one per interlocutor, 44.1 kHz, mono, 16-bit) and three video files in MP4 format (one video file with the eye-tracking dots, one without, and one video file for the external camera, 1920x1080, 30fps). The three video files were mixed with the sound from the external sound card, meaning that these final video exports contain the sound of all interlocutors but not the sound from the external camera. The synchronization of these video and sound allows them to be played back together using the ELAN annotation software (see below).

2.3 Eye-tracking

The eye-tracking data are used to estimate the coordinates of the fixations that occur on the screen during the interaction. The fixation corresponds to a period of time of relatively stable gaze position, and according to the eye-mind hypothesis (Conklin et al., 2018), can be interpreted as an indication of where the person's attention might be directed during the online interactions (O'Rourke, 2012).

Eye-tracking data were collected at a sampling rate of 120 Hz using the Tobii Pro X3-120 device, a compact, horizontal bar-shaped remote video-based eye-tracker. The accuracy and precision of the gaze position estimation as indicated in the specification sheet of the eye-tracker are respectively 0.4° and 0.24°. The operating distance ranges from 50 to 90 cm with a tracking box dimension (linked to the robustness to head movement) of 50 x 40 cm (width x height). The eye-tracker was fixed magnetically to the bottom center of the screen, a 21.5" LED Dell monitor (1920 x 1080 pixel resolution, 60 Hz refresh rate), and was connected to an External Processing Unit (EPU) dedicated to the eye-tracking related calculations.

The Tobii Studio 3.4.8 software installed on a Dell Latitude 7490 laptop connected to the EPU was used to record, manage and export the eye-tracking data. The experimental design was defined using the Screen Recording item with a frame rate of 5 fps, with the option of also recording the sound from the participant's microphone. The default 5-point calibration method was used; the participant was asked to follow a red dot moving around the screen. Once the calibration was validated by the experimenter, the screen and gaze movements were recorded until the end of the interaction.

The eye-tracking data were exported as a TSV file using the data export functionality of Tobii Studio. Each row of the file represents the gaze data recorded every 8 milliseconds. These data include the timestamp from the beginning of the recording in ms, the estimated gaze coordinates relative to the left-top corner of the screen (x, y) in pixels, and the fixation index that specifies if the data point is a part of a fixation, as determined by the Tobii I-VT fixation filter with standard settings (in particular, a velocity threshold for the classifier of 30 degrees/second and a minimum fixation duration of 60 ms). Additional columns indicate the specific region of the screen in which the fixation point is located.

This information relies on the definition of static or dynamic areas if interest (AOI), detailed below (section 3.2).

The table below shows the duration of each recorded interaction. An asterisk indicates that eye-tracking data are not available for the interaction because it was not recorded using our equipment at the language center.

Telecollaboration	Name of participant in France	Duration S1	Duration S2	Duration S3	Duration S4	Duration S5
Polytechnic University Hong Kong – Aix Marseille University (F1L)	Lina	00:40:20	01:03:12	01:01:33	01:07:03	01:27:23
	Lola	00:41:20	00:36:38	00:35:37	00:39:55	00:43:36
	Bérénice	00:38:03	00:58:34	00:53:29	00:53:53	00:52:05
	Kévin	00:37:18	00:45:47	00:51:17	00:43:02	00:47:50
	Eddy	00:43:40	00:42:10	00:51:40		
Aix Marseille University – University of California Berkeley (F1L)	Alain	00:38:16	00:51:59	01:14:09*		
	Carole	00:40:28*	00:34:04*	00:43:23*		
	Léa	00:44:10*	00:33:12*			
	Marie	00:48:37*				
	Natacha	00:22:05*	00:22:38*	01:00:43*		
	Tabatha	00:31:37				
Aix Marseille University – Shenzhen University (Teletandem)	Jean-Philippe	00:52:38	00:27:28	01:02:36	00:51:20	00:46:27
	Kévin	00:54:58	00:59:57	01:01:39	00:58:03	
	Mario	00:30:06	00:05:13	01:00:17	01:05:50	
	Nathan	01:01:50[6]	00:45:06	01:06:54	00:26:58	
	Fahima	00:28:55	00:54:56	01:00:25	00:56:38	00:55:05
Aix Marseille University – Arizona State University (Teletandem)	Darie	01:16:53	01:15:42	01:27:02	01:19:07	01:06:57
	Claudine	00:49:19	00:52:15	00:38:25	00:51:19	
	Bérénice	01:02:44	00:44:36	01:16:48	01:05:07	01:30:07
	Océlia	00:51:48	01:07:47	01:12:39	00:58:55	01:05:31

Figure 5. Durations of recordings

3. Annotations

Annotation allows the researcher to add interpretative multimodal information to a corpus according to research questions (Leech, 1997; Guichon, 2017). Software is often used to facilitate annotation and retrieval of data, and this should be preferred when it can save time, especially for large corpora, and when its error rate has been considered. Furthermore, data must be temporally synchronized and standardized to facilitate sharing (Bert et al., 2010). Here we describe the automatic and semi-automatic annotation that is currently being performed on the VAPVISIO corpus. Our methodology was developed through work on previous corpora at the LPL, including the *CID - Corpus of Conversational Data* (Bertrand et al. 2008) with its associated annotation scheme (Blache et al. 2010), and several subsequent corpora.

The following free software tools were selected to annotate the VAPVISIO corpus:

- Praat (Boersma and Weenink, 2001) is a tool for manually annotating sound files. It provides different visualizations of audio data – waveform or spectrogram display – and, among others, enables pitch contour and formant calculation and visualization. Annotations can be created on multiple layers, called tiers.

[6] This recording is unexploitable because of a problem in sound recording.

- ELAN (Sloetjes & Wittenburg, 2008) is a tool for the creation of complex annotations of video (and audio) files. Annotations can be created on multiple layers, which can be hierarchically interconnected and can correspond to different levels of linguistic analysis. The annotation files are in a specific XML format. Annotation files can be imported from and exported to several other formats, including Praat-TextGrid.
- SPPAS (Bigi, 2015) is a tool for automatic speech annotation and analysis. Among other features, it allows for the automatic segmentation of speech, whose annotations can then be visualized and searched by the analyst. The annotation files are in a specific XML format, but can be converted to and from a variety of other formats, including Praat, ELAN, and CSV.

We describe in the next section the ways in which these tools were used to annotate the corpus.

3.1 Semi-automatic annotation of speech

Step 1: Determination of IPUs (SPPAS)

First, SPPAS was used to automatically segment each sound file into Inter-Pausal Units (IPUs) by setting the threshold volume for speech and silence (Bigi and Priego-Valverde 2018). This is much faster than using Praat or ELAN to manually define IPUs. As in previous studies, the minimum duration of an annotation was set at 100 ms and the minimum duration for a silence was set to 200 ms.

Step 2: Checking IPU boundaries and transcribing (Praat or ELAN)

Checking IPUs mainly consists of the following actions (Bigi & Priego-Valverde, 2018): adding missing IPUs; deleting irrelevant IPUs; merging two continuous IPUs; splitting an automatically-generated IPU that should contain different turns; and adjusting boundaries. Adding IPUs and adjusting boundaries can be performed with any annotation software displaying a timeline, but if speech segmentation at various phonetic levels is expected (see next section), the IPU boundaries must be carefully examined and corrected. Praat is therefore most appropriate for this step, even though other options exist such as ELAN, Phonedit (Teston and al. 1999) or AnnotationPro (Klessa et al. 2013).

Orthographic transcription is also performed during this step using Praat or ELAN. Orthographic transcription is often the minimum requirement for a speech corpus, and as such it is at the top of our annotation procedure and the entry point for most of the automatic annotations. Orthographic transcription of French and English was guided by the SPPAS convention[7] (Bigi et al. 2012). For Mandarin Chinese, we followed the transcription convention elaborated in Cappellini (2014). At a minimum, the orthographic transcription must include: filled pauses, short pauses, truncated words, repetitions, noises and laugh items; i.e. it must include all events present in the speech signal in order to properly perform the forced-alignment task of the speech segmentation process. For the same reason, indication of irregular pronunciations, liaisons and elisions is recommended.

Step 3: Automatic annotations (SPPAS)

One of the main features of SPPAS is its automatic speech segmentation. Speech segmentation is the process of taking the orthographic transcription text of an audio speech segment and determining where particular phonemes/words occur in the speech segment. Text normalization is the first step of this process. It "normalizes" the text by removing punctuation (if any), converting numbers to letters, tokenizing the text, etc. The normalized text is then phonetized – converted into its phonetic constituents. The forced-alignment task first locates phonemes in time, then tokens. Finally, an automatic syllabification and several other automatic annotations can be performed.

[7] Available at https://www.ortolang.fr /market/corpora/sldr000873

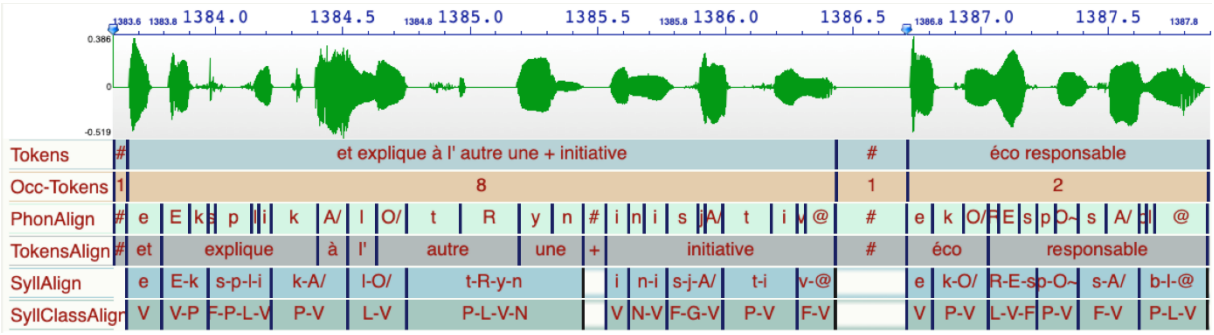

Figure 6: Speech segmentation

The table below summarizes the steps taken to annotate our corpus. Blue boxes represent automatic annotations, yellow boxes represent manual ones, and white boxes represent documents.

3.2 Semi-automatic annotation of gaze

The eye-tracking data captured by the Tobii system contain information about the location and duration of fixations on the screen. The Tobii Studio software allows the researcher to define areas/regions of interest (AOIs or ROIs) to automatically create annotations that indicate when and for how long a participant's gaze is directed at that part of the screen. An AOI is defined by placing a rectangle around a specific part of the screen such as a chat box, a browser window, an open document, or a participant's webcam image.

Participants' webcam images are of special interest due to the rich information transmitted through facial expressions, gazes, postures, and gestures. These can be useful for measuring personal involvement in the exchange (Guichon, 2013), giving instructions (Satar & Wigam, 2017), explaining lexical items (Holt, in press), and more. Although many studies have looked at visual communication during videoconferencing, few have been able to say for sure whether the participant was actually looking at the webcam image. Eye-tracking technology helps to resolve this shortcoming.

Semi-automatic definition of AOIs for interlocutors' faces

Defining AOIs is straightforward for objects that are generally fixed on the screen such as chat windows, but can be problematic for images that move about the screen such as a participant's webcam image. This is compensated for by a tool called OpenFace 2.2.0 (Baltrušaitis et al., 2018), which is an open source tool for facial behaviour analysis, including facial landmark detection, head pose estimation, facial action unit recognition and eye-gaze estimation. OpenFace first detects the face using a tool provided by the Multi-Task Convolutional Neural Network (MTCNN) face detector (Zhang et al., 2016). Facial landmark detection is subsequently carried out on the detected face using the Convolutional Experts Constrained Local Model (CE-CLM) algorithm (Zadeh et al., 2017), resulting in the automatic positioning of markers on the contours of the face, eyebrows, eyes, nose, and mouth on the detected face.

The FaceLandmarkVidMulti executable was used to analyze the dynamic screen recording. It can detect multiple faces, allowing us to detect the distant partner's face and possibly the user's own face and any other face displayed on open document or web page. The output file consists of a CSV file containing the pixel coordinates (x, y) of 68 facial landmark points detected for each video frame and for each detected face, and the associated confidence percentage.

Some post-processing was carried out in Matlab to define the AOI of the distant partner's face from this output file. First, a rectangular box was defined to delimit the face based on the positions of the facial markers. A horizontal margin of 25 pixels was added from the left and rightmost extreme coordinates, as well as a vertical margin of 25 pixels under the chin and 50 pixels above the eyebrows

in order to account for possible small inaccuracies in the spatial determination of fixations by the eye-tracker. Second, the distant interlocutor's face was selected by considering its size. The minimum expected dimensions were set at 150 x 150 pixels. In the case of multiple faces detected in one frame meeting the required size, only the face with the larger size was kept. Finally, a video with the superimposed AOI of the distant interlocutor's face and the fixation was generated for the annotator to control the AOI definition.

Fixations annotations

Each AOI corresponds to a tier with a specific label name, for instance "fixaoi_openface_main" for the distant interlocutor's face detected by OpenFace, or "fixaoi_instant_msg" for a manually defined chat window or "fixaoi_web_page" for a browser window. The fixation occurrence inside one of these AOIs is marked as a time interval with the duration of the fixation. The resulting annotations file is formatted as a tab-separated TXT file displaying one row per annotation with the associated data: tier name (AOI label), begin time, end time, duration, mean confidence level of the eye detection by the eye-tracker, and mean confidence level of the face detection by OpenFace during the fixation. These annotations are subsequently imported into ELAN. The following screen capture shows the AOIs for the OpenFace detection, the chat window and the web browser page. On the bottom, the ELAN annotation shows tiers for the fixation occurrences in the different AOIs. The annotation values correspond to the fixation durations in milliseconds.

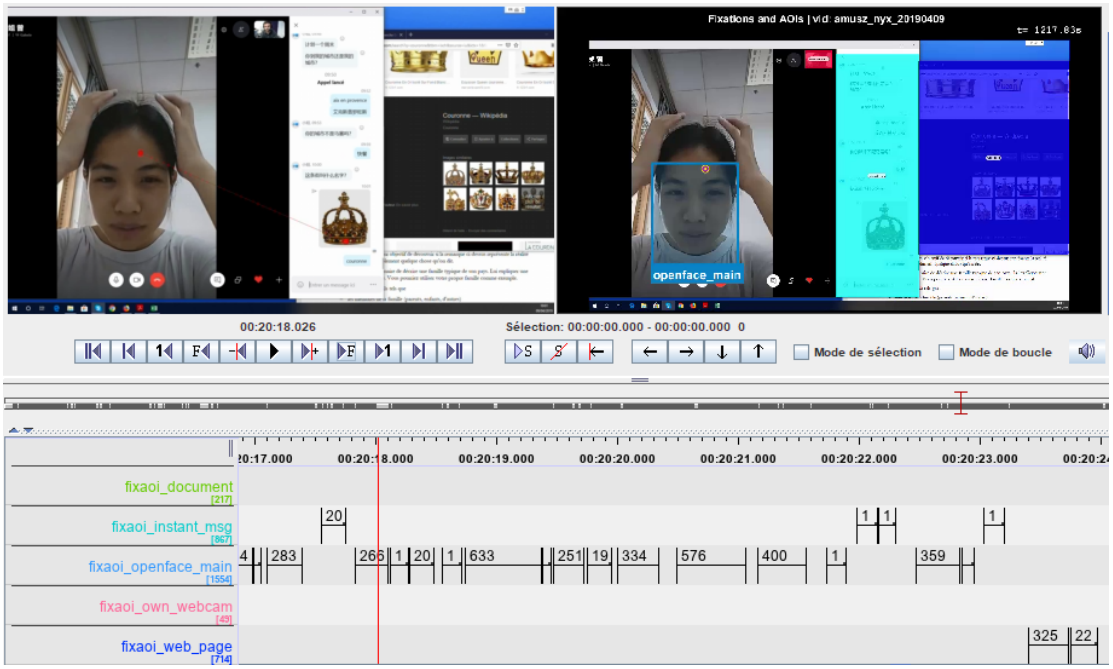

Figure 8: Screen capture of ELAN

3.3 Multi-layered annotation

ELAN was chosen for its versatility and for its ability to combine multiple annotations from the previous steps on separate tiers. Analysis will be based on speech transcriptions (following the SPPAS convention) and written chat messages (following on Cappellini, 2014) for each interlocutor, eye-tracking data, and any other multimodal phenomena deemed necessary such as gestures. Eye-tracking data will include gaze data from the local interlocutor, with a focus on manually-defined AOIs such as chat windows, and on automatically-annotated AOIs such as the distant participant's face.

Conclusion

This article has described the construction of a large multimodal corpus that will serve as the foundation of ongoing and future studies on telecollaboration. We hope that this description will enable other researchers and engineers to replicate or to ameliorate this methodology. The particularity of the VAPVISIO corpus is that it is multimodal, multilingual, includes two types of videoconferencing-based telecollaborative settings, and makes use of eye-tracking technology. The construction of this corpus has allowed us to mobilize an ecological approach that was developed on small-scale case studies that used eye-tracking data to study interlocutors' techno-semio-pedagogical competence (Cappellini & Hsu, 2018). This new corpus will allow us to perform comparisons across telecollaborative projects (Cappellini & Azaoui, 2017) and platforms (Cappellini & Combe, 2017).

In order to answer our research questions, the VAPVISIO corpus will be used to analyze three types of variation: (1) individual evolution between the first and last videoconferencing session of each of the four groups; (2) variation between the two telecollaborative settings (teletandem and F1L); and (3) variation between the three languages (French, English and Mandarin Chinese). Studying the possible development of techno-semio-pedagogical skills within each telecollaborative setting over time will allow us to discern which of these skills develop naturally and which ones require formal training. Comparison of the languages will enable us to understand which of the characteristics are related to them and how they impact learning through DVC. These analyses will enable us to verify our hypotheses concerning the development of techmo-semio-pedagogical skills across different telecollaboration settings.

Annotation of the interactions is ongoing, and the VAPVISIO corpus will be made available on the Ortolang platform by 2022.

References

Arnold N. & Ducate L. (2015). Contextualized views of practices and competencies in CALL teacher education research. *Language Learning & Technology, 19*(1), 1-9.

Baltrušaitis T., Zadeh A., Lim Y. C. & Morency L.-P. (2018). OpenFace 2.0: Facial behavior analysis toolkit. *IEEE International Conference on Automatic Face and Gesture Recognition*.

Bertrand R., Blache P., Espesser R., Ferré G., Meunier C., Priego-Valverde B. & Rauzy S. (2008). Le CID-Corpus of Interactional Data-Annotation et exploitation multimodale de parole conversationnelle. *Traitement automatique des langues, 49*(3): 1-30.

Bigi B. (2015). SPPAS - Multi-lingual Approaches to the Automatic Annotation of Speech. *The Phonetician, 111-112*, 54-69.

Bigi, B., Péri P. & Bertrand R. (2012). Orthographic Transcription: which enrichment is required for phonetization?. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, 1756-1763.

Bigi B. & Priego-Valverde B. (2018). Search for Inter-Pausal Units: application to *Cheese!* corpus. *9th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, 289-293.

Bigi B. & Saubesty J. (2015). Searching and retrieving multi-levels annotated data. *Proceedings of Gesture and Speech in Interaction - 4th edition*, 31-36.

Blache, P., Bertrand R., Bigi B., Bruno E., Cela E., Espesser R., Ferré G., Guardiola M., Hirst D., Magro E.-P., Martin J.-C., Meunier C., Morel M.-A., Murisasco E., Nesterenko I., Nocera P., Pallaud B., Prévot L., Priego-Valverde B., Seinturier J., Tan N., Tellier M. & Rauzy S. (2010). Multimodal Annotation of Conversational Data. *The Fourth Linguistic Annotation Workshop, ACL 2010*, 186-191.

Bert, M., Bruxelles, S., Etienne, C., Jouin-Chardon, E., Lascar, J., Mondada, L., Teston, S., & Traverso, V. (2010). Grands corpus et linguistique outillée pour l'étude du français en interaction (plateforme CLAPI et corpus CIEL). *Pratiques*, *147‑148*, 17‑34.

Blin, F. (2016). The theory of affordances. In C. Caws & M.-J. Hamel (Eds.), *Learner Computer Interactions: New insights on CALL theories and applications* (pp. 41-64. Amsterdam: John Benjamins.

Boersma P. & Weenink D. (2001). Praat: doing phonetics by computer [Computer program], Version 6.0.37, retrieved 14 March 2018 from http://www.praat.org/

Cappellini, M. (2014). *Modélisation systémique des étayages dans un environnement télétandem pour le français et le chinois langues étrangères. Une étude interactionniste et écologique du soutien au développement de la compétence de communication*. (Original PhD dissertation). Université Lille 3 SHS, Lille.

Cappellini, M. & Azaoui, B. (2017). Sequences of normative evaluation in different pedagogical settings through desktop videoconference. *Language Learning in Higher Education*, *7*(1), 55-80.

Cappellini, M. & Combe, C. (2017). Analyser des compétences techno-sémio-pédagogiques d'apprentis tuteurs dans différents environnements numériques : résultats d'une étude exploratoire. *Alsic 20*.

Cappellini, M. & Hsu, Y.Y. (2018). Ce que l'oculométrie peut apporter à une approche écologique aux échanges en ligne. Une discussion épistémologique et une étude de cas. *EPAL - Echanger Pour Apprendre en Ligne.*

Cohen, C. (2017). Former à l'enseignement en ligne. In N. Guichon & M. Tellier (Eds.), *Enseigner l'oral en ligne: Une perspective multimodale* (pp. 218-242). Paris: Didier.

Conklin, K., Pellicier-Sanchez, A., & Carrol, G. (2018). *Eye-tracking. A guide for applied linguistics research*. Cambridge: Cambridge University Press.

Develotte, C., Guichon, N. & Kern, R. (2008). « Allo Berkeley ? Ici Lyon… Vous nous voyez bien ? » Etude d'un dispositif de formation en ligne synchrone franco-américain à travers le discours de ses usagers. *Alsic*, *11*(2), 129-156.

Dooly, M. (2010). Teacher 2.0. In S. Guth & F. Helm (Eds.), *Telecollaboration 2.0* (pp. 277-303). Bern: Peter Lang.

Fuchs, C., Hauck, M. & Müller-Hartmann, A. (2012). Promoting learner autonomy through multiliteracy skills development in cross-institutional exchanges. *Language Learning & Technology*, *16*(3): 82–102.

Guichon, N. (2012). *Vers l'intégration des TIC dans l'enseignement des langues*. Paris: Didier.

Guichon, N. (2013). Une approche sémio-didactique de l'activité de l'enseignement de langue en ligne : Réflexions méthodologiques. *Éducation & Didactique*, *7*(1), 101- 116.

Guichon, N. (2017). Sharing a multimodal corpus to study webcam-mediated language teaching. *Language Learning & Technology*, *21*(1), 55- 74.

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

Hampel, R. & Stickler, U. (2005). New skills for new classrooms: Training tutors to teach languages online. *Computer-Assisted Language Learning*, *18*(4), 311-326.

Hampel, R. & Stickler, U. (2012). The use of videoconference to support multimodal interaction in an online language classroom. *ReCALL*, *24*(2), 116-137.

Hampel, R. & Stickler, U. (Eds.) (2015). *Developing Online Teaching Skills*. New York: Palgrave Macmillan.

Hauck, M. (2010). Telecollaboration: At the interface between multimodal and intercultural communicative competence. In S. Guth & F. Helm (Eds.). *Telecollaboration 2*.0 (pp. 219-244). Bern: Peter Lang.

Holt (in press). Le rôle des gestes dans les explications lexicales par visioconférence. *Travaux Interdisciplinaires sur la Parole et le Lgnage, 36*.

Kessler, G. (2016). Technology standards for language teacher preparation. In F. Farr & L. Murray (Eds.), *The Routledge handbook of language learning and technology* (pp. 57-70). London: Routledge.

Klessa, K., M. Karpiński & Wagner A. (2013). Annotation Pro-a new software tool for annotation of linguistic and paralinguistic features. *Proceedings of the Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop,* 51-54.

Kurek, M., & Müller-Hartmann, A. (2019). The formative role of teaching presence in blended Virtual Exchange. *Language Learning & Technology*, *23*(3): 52-73.

Leech, G. (1997). Introducing corpus annotation. In R. Garside, G. Leech & A.M. McEnery (Eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora* (pp. 1-18). London: Routledge.

Little, D. & Brammerts, H. (Eds.) (1996). A guide to language learning in tandem via the Internet. *CLCS Occasional paper, n° 46.* Dublin: Center for Language and Communication Studies.

Mangenot, F. (2013). Les échanges en ligne comme secteur de pratiques et de recherches en ALAO : quelles problématiques, quelles évolutions ?. *Les Cahiers de l'ILOB, 5*, 3-21.

O'Dowd, R. (2016). Emerging trends and new directions in telecollaborative learning. *CALICO Journal*, *33*(3), 291-310.

O'Dowd, R. & O'Rourke, B. (2019). New developments in virtual exchange in foreign language education. *Language Learning & Technology*, *23*(3), 1–7.

O'Rourke, B. (2012). Using eye-tracking to investigate gaze behaviour in synchronous computer-mediated communication for language learning. In M. Dooly & R. O'Dowd (Eds.), *Researching online foreign language interaction and exchange* (pp. *305-341*). Bern: Peter Lang.

Popescu-Belis, A. (2010). Chapter 11 - Managing multimodal data, metadata and annotations: Challenges and solutions. In J-P Thiran, F. Marqués & H. Bourland (Eds.), *Multimodal Signal Processing* (pp. 207-227). Oxford: Academic Press.

Satar, H. M., & Wigham, C. R. (2017). Multimodal instruction-giving practices in webconferencing-supported language teaching. *System, 70*, 63‐ 80.

Sloetjes, H. & Wittenburg, P. (2008). Annotation by category – ELAN and ISO DCR. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Stickler, U. & Shi, L. (2017). Eyetracking methodology in SCMC: A new tool for empowering learning and teaching. *ReCALL*, *29*(2), 160-177.

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

Teston B., Ghio A. & Galindo B. (1999). A multisensor data acquisition and processing system for speech production investigation. *International Congress of Phonetic Sciences (ICPhS),* 2251-2254.

Van Lier, L. (2004). *The Ecology and Semiotics of Language Learning: a Sociocultural Perspective*. Dordrecht: Kluwer Academic Publishers.

Zadeh, A., Baltrušaitis T., & Morency L.-P. (2017). Convolutional experts constrained local model for facial landmark detection. *Computer Vision and Pattern Recognition Workshops*.

Zhang, K., Zhang Z., Li Z. & Qiao Y. (2016). Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, *23*(10), 1499-1503.