

Building ethograms for interpreting research —A technical report

Du Zhiqiang

1. Introduction

Analyzing human behavior is a powerful and arguably essential means of studying minds engaged in multilectal mediated communicative interactions. Human communication engages multiple sensory and cognitive systems, making it highly unlikely for any single indicator to support a comprehensive understanding of behavior in these scenarios. To address this complexity, researchers increasingly rely on multimodal analyses that integrate visual, auditory, and physiological data streams. These approaches allow for a richer understanding of human communication by capturing its inherent multimodality.

To account for this complexity, cognitive translation and interpreting studies (CTIS) have broadened their observational scope to generate diverse multimodal datasets that go beyond oral and written records. This shift is motivated by at least two key factors. First, language itself is inherently multimodal and becomes even more so when used (Kokkola & Ketola 2015; Seeber 2017; Chmiel et al. 2020; Halverson & Muñoz 2020; Doherty 2021). Second, interpreting is characterized by its multitasking demands and constant cognitive shifts. To better understand these demands, researchers have adopted multimodal approaches that capture the complexity of interpreting behaviors.

However, and despite the growing sophistication of multimodal methodologies in CTIS, behavioral studies often lack transparency in how multimodal data is aligned. While many researchers gather data through

audio, video, and screen recording, few explicitly document their alignment processes, which complicates replication and critical analysis. Aligning multimodal data still requires labor-intensive manual techniques that demand significant time and expertise.

This report focuses on the data cleaning process and its further elaboration to illustrate practical methods for aligning multimodal data in naturalistic interpreting environments. The process will be exemplified with my dissertation. While my study was exploratory in nature and limited in generalizability, the alignment approaches presented here that were developed for it have broader implications, particularly for remote data collection settings. By aligning multimodal data effectively, researchers can support more rigorous, ecologically valid research in CTIS.

Data alignment in this report involves results from data collection methods such as speech recording, screen recording, and keylogging. These methods are widely used due to their ease of implementation and unobtrusive nature (Ahrens & Janikowski 2025; Muñoz & Rojo 2025; Enríquez et al. 2025). However, despite their popularity in empirical interpreting studies, clear guidance on how to align these diverse data streams remains largely absent from the literature, with only occasional references and few examples. This report aims to address this gap by exploring how to align data across different devices and even across time zones at both the individual and group levels. Eyetracking is not included in this report. Eyetracking has indeed proven to be an excellent tool in research domains such as reading and usability, even though the link between eye movement and cognition is not always as straightforward as the eye-mind hypothesis might suggest (Walker et al. 2025). Furthermore, in interpreting it requires substantial effort to integrate it into empirical studies (Hooge et al. 2025; Niehorster et al. 2025), especially those concerned with naturalistic conditions. It is thus excluded from this technical report, but aligning eyetracking data will follow the same patterns and readers may extend these guidelines to it too.

2. General considerations before aligning data

This section includes a few general considerations that would help you prepare before embarking in full-scale data alignment. These include under-

standing when to rely on temporal information and chronological order, and when it may be appropriate to depart from them. You should read and familiarize yourself with your data, choose the right tools for alignment, and test your workflow on a small dataset first.

2.1 Considering aligning data through timestamp

Some studies successfully employ measures based on a single data source—e.g., keylogging in Van Waes et al. (2021) and screen-recording in Angelone (2021)—as they are sufficient to shed light on specific research questions. Yet, often in interpreting research cross-referencing or triangulating two or more multimodal data sources provide contextually rich information to compare and validate hypotheses. Speech recordings capture both source-speech and interpreter output, often in timestamped verbatim transcripts—that is, they include language and temporal information, contributing to interpreting research corpora (Aston 2018; Wehrmeyer 2019). When combined with screen recordings, keylogging data offers comprehensive data that helps profile interpreters' behaviors and responses in context. Such multimodal combination is common in written communication research (for instance, Leijten & Van Waes 2013; Wang 2022), but rare in interpreting research.

Despite their different natures, these data streams share a crucial common feature, namely temporal information. Audio transcripts can be timestamped, screen recordings align with video time, and keylogging naturally tracks keystroke timing. This temporal element allows all data types to be synchronized along a unified timeline, making time a central reference point for data analysis. Your alignment methods should be shaped by your research question. In some cases, alignment may focus on comments or qualitative data. In others, it may be based on quantitative elements like time codes, source-speech excerpts, or pauses in the interaction. Here we use chronological order, which means organizing multimodal data by time. Time-based alignment, common in CTIS, may reveal potential patterns in informant behavior and communication. However, the approach is not one-size-fits-all. Research methods should be considered and adapted to the specific context of the study. The goal of this paper is to support researchers who are interested in aligning quantitative multimodal data by offering a flexible framework—not to promote a single solution for everyone.

Multimodal data can be converted into written formats (such as transcripts from speech recordings) and organized chronologically, enabling researchers to observe consistent reactions among informants over tasks. This approach is particularly useful for investigating how people respond to identical actions or stimuli across multiple data modalities (Borg et al. 2025). We will call this combined and time-aligned record of at least three sources of behavioral data from different modes an *ethogram*. The term is borrowed from zoology, where Huxley (1914) coined it to refer to a comprehensive catalog or inventory of all the **behaviors or actions exhibited by an animal species**, typically recorded in a systematic way during observational studies. Here it is meant to underscore the combination of heterogeneous sources and the inherent interdisciplinarity of the approach. It also seeks to remind readers that an ethogram merely aims to capture (complex, communicative) behavior and that cognition still needs to be inferred.

2.2 Anonymizing data sets

Data anonymization is often carried out before data cleaning, in line with the requirements from institutional ethics committees or relevant regulations such as the General Data Protection Regulation (GDPR). Anonymizing data protects and ensures informant privacy and is a crucial step in preparing data for analysis. In addition to anonymization, data backup is another important aspect of data management. A careful data management plan should include backing up the original and version control of the data. This allows you to recover your work if something goes wrong or if alignment mistakes occur. Backups can be stored locally or in secure private storage that is easily and permanently accessible.

2.3 Getting to know your data and tools

Deciding on the order of alignment and the choice of applications to build it are crucial steps before you begin your work. Take time to familiarize yourself with the data—just look through your sources, whether they consist of documents, speech recordings, video, screen recordings, or a combination thereof. Listen or watch audio and video clips. Let the data inspire you and remember that our task is to create information out of data but also that what counts as data depends on how we perceive and frame the

world—that is, the initial and apparently messy data is already filtered by our theoretical lens.

Understanding the data formats also helps you select appropriate tools and develop effective alignment methods. For instance, screen recordings frequently capture both the source audio and the informant's voice (i.e., the interpreters' renderings), making them a primary resource for alignment, as both streams are synchronized on the same timeline. To refine the timing—such as adjusting the position of the source-speech soundtrack—researchers should use a video editor that allows precise, frame-level manipulation. While many video editors support such capabilities (see Ahrens & Janikowski 2025, 195 for an overview of speech recording applications), Adobe Premiere Pro often stands out as a good option.¹ Although it is a subscription-based tool, it provides an intuitive interface with features like drag-and-drop soundtrack positioning, zoomable waveforms, and fine-grained control over both audio and video timelines—making it well-suited for research requiring millisecond-level precision. ELAN is a powerful free alternative to Adobe Premiere Pro.² Originally designed for linguistic annotation, ELAN supports high-precision timing, multi-tier annotations, and synchronized audio-video playback, making it a valuable tool for detailed alignment tasks in interpreting research.

Another key application we need is a speech-to-text tool for transcribing recordings. Instead of relying on manual verbatim transcription, you can choose from a wide range of automatic speech recognition (ASR) tools, such as Microsoft Stream (MS Stream), OpenAI Whisper, and IBM Watson Speech to Text. Among these, open-source OpenAI Whisper stands out for its high transcription accuracy and multilingual support—though it requires some technical setup unless used with a GUI wrapper like Whisper Web.³ IBM Watson offers cloud-based services with customizable models, but its free tier is limited. MS Stream is particularly useful for its interactive features: users can not only view and edit transcripts directly but also

1 Adobe Premiere Pro: <https://www.adobe.com/products/premiere.html>

2 ELAN: <https://archive.mpi.nl/tla/elan>

3 Whisper Web: <https://huggingface.co/spaces/Xenova/whisper-web>

click on individual transcript segments to replay the corresponding audio, facilitating the detection and correction of speech-to-text errors. Moreover, the platform offers timestamp precision down to the millisecond, making it suitable for research tasks that require detailed alignment. More on transcription tools in Ahrens & Janikowski (2025, 198).

2.4 Trial alignment on a sample dataset

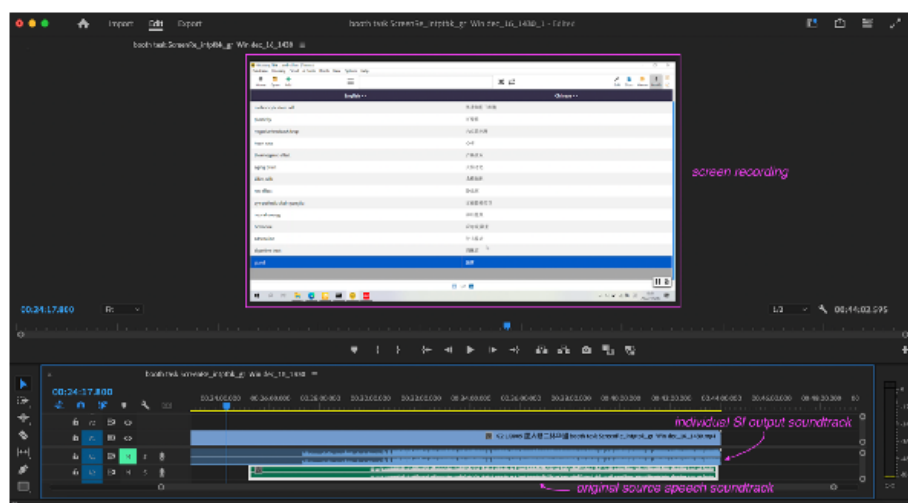
Before starting the full data alignment process, see if it is possible to carry out the full data cleaning and alignment process on-site on a sample dataset to test whether your alignment method works, whether the tools perform as expected, and whether the results are up to your expectations. It is not a pilot test, because you can go back to square one as many times as you need (as long as you did backup your data!). Think of it as a practice run, like preparing a recipe by mixing all the ingredients in a small batch to see how it turns out. Try to imagine each step of the process, consider what could go wrong, and ask yourself if there is a backup plan for each part.

3 Mapping audio to video data

In interpreting research, audio data often registers either the source-speech soundtrack or informants' interpreting renderings. Here, *video data* refers to screen records capturing informants' on-screen activities, both image and sound. Mapping the source-speech soundtrack directly onto the screen recording is a more integrated and efficient approach than using separate applications to handle audio and video files. Screen recording applications generally capture users' voice output (i.e., the interpreters' renderings), so aligning the source-speech soundtrack to the screen recording simultaneously aligns it with the interpreters' renderings, ensuring all audio-visual elements share the same temporal framework and scale. In other words, this method leverages the screen recording's capture of the original source-speech playback interface, allowing for mapping when importing the source-speech soundtrack. This alignment is essential and should be the first step in data cleaning, as it establishes the foundation for subsequent alignment processes described below.

This method offers some advantages—particularly, when compared to other multimodal alignment methods that we will later address—because it is relatively easy to implement and complete. Using a video editing tool like Adobe Premiere Pro, you can import both the screen recording video and the source-speech soundtrack, then zoom in on the video timeline to the millisecond level, *frame-to-frame*. This level of precision mitigates the human error for accurate mapping between the source-speech soundtrack and visual cues in the recording. For example, you can identify the single video frame when the informant clicks to play the source speech and use it to synchronize the corresponding source-speech track (see Figure 1 for example). This process helps minimize human error in approximately locating the precise start point of the imported soundtrack.

Figure 1. Importing video and audio into a video editing tool.



The above method may not be feasible in all scenarios. One common limitation arises when the click-and-play interaction with the web-based source speech player is not captured in the screen recording. This might occur if the informant began the screen recording after source file playback already started in applications like MS Stream, or if the click was simply not visible due to task-focused behavior. In such cases, alternative validation methods can be employed to achieve the same goal. You can listen to the

screen recording (i.e., the voice-over SI output audio) to pinpoint a sound of a distinct word from the source speech at its onset timestamp. You can then use this word as a new anchor point, aligning it with the corresponding moment in the source-speech soundtrack in Adobe Premiere Pro. Another option is to identify signal sounds (e.g., the ending signal sound of the source speech) within the individual screen recording. Once these points are located, they are mapped to their respective positions in the source-speech soundtrack using Adobe Premiere Pro.

Table 1. Calibration for precise alignment of source speech and screen recordings.

method	reference points	steps	when to use
visual observation	timestamp when informant clicks to play source speech	Identify click-and-play action visually. Frame-by-frame inspection in Adobe Premiere Pro. Precisely mark timestamp/frame as anchor point.	click-and-play action is visible in screen recording.
auditory references	distinct word from source speech heard in screen recording	Listen to screen recording to identify onset of a distinctive word. Mark timestamp of this word as reference point. Align with source-speech audio in Adobe Premiere Pro.	visual click-and-play action isn't visible; distinct words from source speech are audible in recording.
signal references	signal sounds (e.g., ending signal of source speech)	Locate specific auditory signals (e.g., ending sounds) in recording. Mark these signals as anchor points. Align with source-speech audio in Adobe Premiere Pro.	visual click-and-play action isn't visible; clear signal sounds from source speech are audible.

The choice among the three alignment methods (Table 1) depends on each informant’s specific scenarios. Once aligned, these elements share the same timeline, resulting in transcripts with streamlined, consistent timestamps. The screen recording video can now be exported separately, with individual interpreter’s renderings and source-speech soundtrack maintaining timeline references shared across the sample. Video files (or those converted to audio) can be easily transcribed using automatic speech recognition software and platform. After processing, you obtain four perfectly aligned data streams (i.e., source-speech sound, screen recording, their transcripts).

4. Align keylogging with screen recording data

Keylogging is an unobtrusive data collection method that allows observing user behavior without interferences (Muñoz et al. 2025). Keyloggers register actions at the mouse and keyboard, including reactions to stimuli, with millisecond accuracy, helping to measure response times on the computer. In the last decades, it has been increasingly adopted in writing studies and in CTIS as well (e.g., Lindgren & Sullivan 2019 for writing research and Muñoz & Apfelthaler 2022 in translation research). Software repositories like GitHub offer a wide range of open-source keyloggers. Selecting the appropriate keylogger depends largely on your research goals—whether it involves capturing just key-down events or both key-down and key-up events (which combined are often referred to as keystrokes), the later widely accepted in human-computer interaction (HCI) research (e.g., Messerman et al. 2011 and Katerina & Nicolaos 2018). Two important factors when choosing a keylogger are cross-platform compatibility and timestamp precision. Using different keyloggers on Windows and macOS, for instance, may result in inconsistencies in data (i.e., timestamps). For HCI studies, a timestamp resolution at the millisecond level is often considered best, so it is essential to verify that the selected tool supports this level of precision.

When research involves informants across different geographical locations and time zones, it becomes equally important to adopt a standardized time registry format. In such cases, Unix time or *Epoch time* is highly recommended. This system represents time as the number of seconds (and optionally milliseconds) elapsed since January 1, 1970, at 00:00:00 UTC, excluding leap seconds. Unix time is widely supported across operating systems, file systems, and programming languages, making it especially suitable for remote data collection. Whenever possible, Unix time should be used throughout the data collection process, with conversions to standard human-readable time formats (hours, minutes, seconds, milliseconds) reserved for the final stages of analysis. For my dissertation's research project, I used the `pynput` library, which is compatible across both Windows and

macOS and records timestamps using the Unix time format, which are not featured by Inputlog and Translog.⁴

The chronological sequence of keypresses registered by keyloggers is a great advantage but is also limited, because we do not really know how information is processed after receiving stimuli. To do so, we need contextual information, and a new way to add it is aligning keystroke data with screen recording. Screen data can make up for this blind spot by providing context, so we need to align these two data sources. The screen recording data has already been processed in the previous alignment but, before we proceed to the new alignment, we still need to clean the keylogged data. Identify and remove irrelevant keystroke records, particularly the spans recorded before and after the relevant study period. Finally, verify the remaining data before importing it for alignment.

Keyloggers like pynput generate keystroke data that displays individual keystrokes in chronological order, each with a Unix timestamp. The principle here is to convert these timestamps into a format that aligns with the screen recording's timeline. First, the keystroke data is imported into a spreadsheet. This is what happens in Column F in Figure 2 with the Unix timestamps placed in Column B. These timestamps are then converted into human-readable datetime format (hours, minutes, seconds, milliseconds) using Excel formulas—see Column C for this converted format. You can find the conversion formula in the Zenodo repository.⁵

The screen recording is then imported into a software application supporting video import and frame-level zooming, such as Adobe Premiere Pro or Adobe Audition. This feature allows researchers to pinpoint the exact video frame at which a character appears or a deletion event (e.g., a backspace) occurs as a result of a keystroke. This visible change on screen acts as a reference point. For instance, if a character is seen to be deleted at video time 00:00:27.447 (cell D2 in Figure 2), it can be aligned with the cor-

4 pynput: <https://pypi.org/project/pynput/>

Inputlog: <https://www.inputlog.net/>

Translog: <https://sites.google.com/site/centretranslationinnovation/translog-ii>

5 Conversion steps and formula: <https://zenodo.org/records/10520445>

responding *key.backspace* in cell F2, with its Unix timestamp in cell B2 and converted date time in cell C2).

Figure 2. Screenshot of conversion table from Unix time to chrono time by IKIs.

	B	C	D	E	F
1	Unix time	date time	video time	chrono time	keystroke
2	1671431700.113	06:35:00.113	00:00:27.447	27.447	Key.backspace
3	1671431700.313	06:35:00.313	00:00:27.647	27.647	Key.backspace
4	1671431701.896	06:35:01.896	00:00:29.230	29.23	Key.caps_lock
5	1671431702.034	06:35:02.034	00:00:29.368	29.368	Key.caps_lock
6	1671431703.013	06:35:03.013	00:00:30.347	30.347	Key.shift
7	1671431703.608	06:35:03.608	00:00:30.942	30.942	'E'
8	1671431703.776	06:35:03.776	00:00:31.110	31.11	'n'
9	1671431703.997	06:35:03.997	00:00:31.331	31.331	'g'
10	1671431704.260	06:35:04.260	00:00:31.594	31.594	'l'
11	1671431704.428	06:35:04.428	00:00:31.762	31.762	'i'
12	1671431704.611	06:35:04.611	00:00:31.945	31.945	's'
13	1671431704.691	06:35:04.691	00:00:32.025	32.025	'h'

Inter-keystroke intervals (IKIs) represent the time elapsed between successive keystrokes (e.g., the timespan differences between cells C3 ↔ C2 and D2 ↔ D3). The same time interval can be applied to the video timeline: for instance, $D3 = D2 + (C3 - C2)$, where Column D stores the aligned video timestamps for each keystroke. Once this initial reference point is manually established, we can use IKIs to synchronize the remaining keystrokes. This formula can then be autofitted down Column D to compute the precise video time for the entire sequence of keystrokes in the format hh:mm:ss.000. To facilitate further numerical calculations, the video time in Column D can be converted into chrono time in seconds (e.g., 27.447 in cell E2) by using the formula $=D2*86400$ in Column E. Please note that, for the formulas to function correctly, all values in Columns C and D need to be properly set as time values, not plain text, as cell properties in the spreadsheet.

This approach focuses on individual keystrokes rather than word-level analysis, generating fully synchronized timestamps that accurately align keystroke data with the screen recording and creating a unified temporal framework for both data streams.

5. Align problem triggers in the audio with keylogged data

The previous section §4 focused on character-level alignment, but we can apply similar approaches for word-level analysis. In CTIS, researchers often assess interpreters' accuracy in rendering specific terms, particularly examining accurate rates and error frequencies when potential problem triggers (e.g., technical terms, entities, and numbers) are interspersed in the source speech (Mankauskienė 2016; Frittella 2022; Prandi 2023). These problem triggers in the source speech can be aligned to interpreters' keystroke logs, enabling the analysis of interpreter reactions during term consultation or information seeking activities on the screen. The primary challenge lies in accounting for temporal variations across informants, as each informant may, as in my sample, remotely begin playback of the source speech at different times. Efficient word level alignment is feasible with the following methods. First, potential problem triggers are identified in the source-speech transcript. Second, interpreters' records are checked to spot the corresponding renditions.

Since in many cases the original source-speech soundtrack is the same for all informants, we can use the onset and offset times from the original source speech as a master reference. That is, there is no need to repeatedly identify the onset and offset times of problem triggers for each interpreter's individual aligned source speech. Similar to the IKI method for keystroke alignment, these time spans can be used to calculate problem trigger positions in each interpreter's timeline. After completing the initial alignment of source speech with screen recordings, researchers use the source-speech transcript to identify onset and offset timestamps for each problem trigger. This can be done with Adobe Audition 2024, due to its capabilities in multi-track recording and millisecond-level zooming. Here Audition is preferred to Premiere because this mapping task does not require adjusting soundtrack locations; Audition's zoom capability is ideal for millisecond precision, though similar tools enabling detailed audio visualization would also be suitable.

Figure 3. Partial alignment list showing timestamps using in a spreadsheet.

timestamp type	timestamp in original source speech	potential problem triggers	timestamp in the video file with individual source speech soundtrack		
onset	30:29.582	hormones	30:39.913		
offset	30:30.224	hormones	30:40.555		identified by Adobe Audition
onset	30:35.205	immune system	30:45.536		
offset	30:36.119	immune system	30:46.450		
onset	31:13.553	cortisol	31:23.884		
offset	31:14.154	cortisol	31:24.485		
onset	31:14.669	epinephrine	31:25.000		
offset	31:15.322	epinephrine	31:25.653		
onset	31:18.885	estrogen	31:29.216		calculated by MS Excel
offset	31:19.624	estrogen	31:29.955		
onset	31:21.233	cholesterol	31:31.564		
offset	31:21.879	cholesterol	31:32.210		
onset	31:26.742	dietary cholesterol	31:37.073		
offset	31:28.227	dietary cholesterol	31:38.558		
onset	31:50.153	stress hormone	32:00.484		

The process consists of the following steps:

1. Import the screen recording video with the aligned source-speech soundtrack into Adobe Audition.
2. Use a spreadsheet containing all problem triggers, zoom in on the source track to mark the first term with precise onset and offset timestamps within the video timeline.
3. Check the master reference of onset and offset times for all problem triggers from the original source-speech soundtrack.
4. Compare these to the marked onset and offset times of the first trigger in the aligned track.
5. Use the identified time difference, calculate timestamps for the remaining problem triggers, which is similar to the inter-key-interval method used for keystroke timing.

This process enables precise timestamp calculations in a spreadsheet, accurately aligning all triggers with the video timeline (see Figure 3). Then, manually check the corresponding positions in informants' interpreting rendering transcripts to identify how each problem trigger was rendered.

6. Chunk-level alignment

Word-level alignment (e.g., identifying potential problems) provided a benchmark for examining informants' interpreting accuracy. Yet, in practice, interpreters deal with information units larger than single words. To better capture this, we borrow the concept of chunks from psychology (e.g., Cowan 2005). Here *chunks* are both temporary memory containers that allow combining smaller pieces of information into larger structures and the result of segmenting incoming information into manageable units (cf. Muñoz & Cardona 2018, 531–32). The perspective depends on the research needs and, when applied to interpreting studies, chunking mostly refers to the way interpreters segment the input into smaller fragments that can be encoded without waiting for an entire sentence to unfold (Seeber 2011, 194).

Chunk-based analysis in interpreting studies enables to explore cognitive dynamics and measure metrics like ear-voice span (Mizuno 2017; Chmiel et al. 2023; Shao & Defrancq 2024; Chmiel et al. 2024), speech production and perception (Christodoulides 2016), delivery rate of source speech (Korpál & Stachowiak-Szymczak 2020), task complexity and directionality (Huang et al. 2023), and working memory (Shao & Defrancq 2024). These studies have generally compared two critical elements in terms of chunks: the segmentation of the source speech and the corresponding informant transcripts.

From the alignment process described in §3, screen recordings with both source-speech soundtracks and interpreting rendering outputs have been processed, each with an associated transcript. If you used automatic transcription tools, you need to manually correct errors in both the source speech and interpreting renderings transcripts. Once the corrections are made, the transcripts should be segmented into corresponding units, resulting in chunked versions of both the source speech and the interpreting renderings with their timestamps (hours: minutes: seconds.000). Depending

on your goals, you may need to list onset and offset times for each chunk. This prepares the data for precise chunk-level alignment.

Both transcripts originate from the same screen recording video, so there is no challenge in aligning them based on time. Nevertheless, the source speech is embedded within individual screen recordings. Hence, if informants play the source speech themselves while recording their screens, there will be variations in the timestamps across different informants’ recordings. We can adjust this with an approach similar to the one described in §4. It involves using the first timestamp in the speech soundtrack as the reference point and locating the same point in the individual source speech transcript timeline (for instance, in Figure 4, Cell B2 in numerical value after converting to numerical format in video timeline). By using time intervals similar to inter-keystroke intervals (IKIs), the spreadsheet can autofit the remaining cells in Column B for chunk onset times and Column D for chunk offset times in the source speech. However, the timestamps (onset and offset) for corresponding chunks in the interpreting renderings require manual checking using audio editors such as Audition 2024.

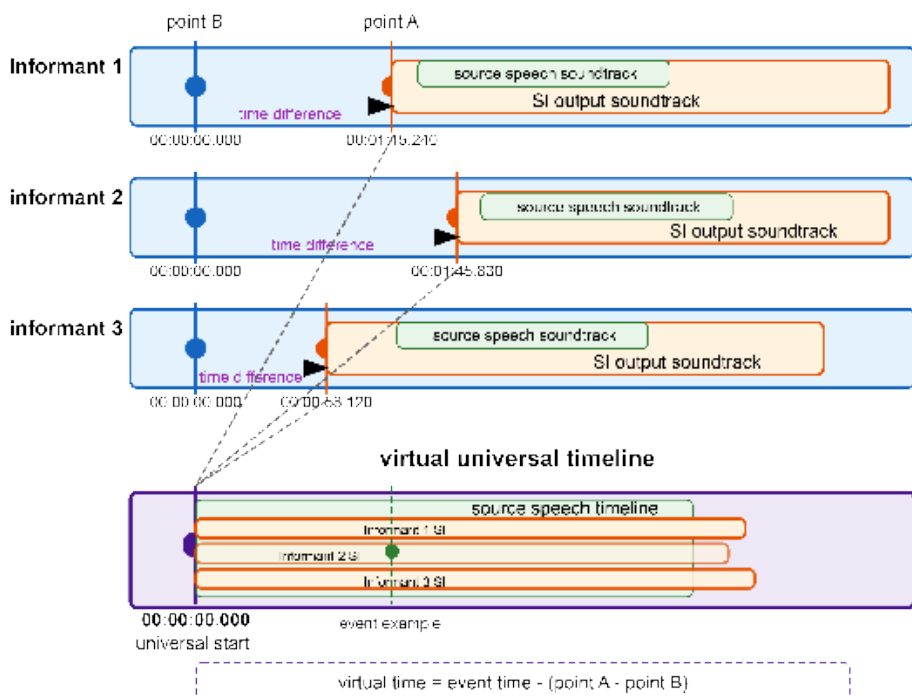
Figure 4. Screenshot of chunks alignment in a spreadsheet.

A	B	C	D	E	F	G	H
No.	timestamps (onset)	source speech	timestamps (offset)	No.	timestamps (onset)	interpreting renditions	timestamps (offset)
s1	1827.531239	Today, we're going to discuss hormones and focus on how particular hormones can influence our energy levels and immune system.	1836.07343	t1	1828.45287	今天我们要讨论的是激素，我们会关注特定激素如何影响我们的能量和免疫系统。	1838.584381
s2	1837.516743	Let's get started talking about how to increase your energy and immunity.	1842.277336	t2	1839.316779	我们首先谈谈如何增加能量和免疫力。	1843.873925
s3	1843.438712	I think it's fair to say that most people would like to feel energetic during the day, especially if they are daytime workers.	1851.695362	t3	1844.667356	我想我们可以公平地说，大多数人会在白天觉得有能量，尤其是那些在白天工作的人。	1853.51717
s4	1852.451377	In other words, you want to have energy, and you want your immune system to function well, to ward off infections of various kinds, bacterial infections, viral infections, etc.	1863.994535	t4	1854.208879	换句话说，如果你希望你的能量和免疫系统较好可以抵御细菌等等其他的感染。	1867.71756
s5	1864.931207	And it turns out that the two key hormones in the processes of having enough energy and having a healthy immune system are cortisol and epinephrine.	1875.340305	t5	1869.487523	有两个关键的激素会影响我们的能量和免疫系统一个是皮质醇。一个是，肾上腺素。	1882.040017
s6	1875.989777	First, cortisol is a hormone much like estrogen and it is derived from cholesterol.	1881.905234	t6	#N/A		#N/A
s7	1882.554705	This cholesterol could be produced by the liver.	1884.468013	t7	1884.847544	胆固醇是由肝产生的。	1887.966385

Since chunks from the source speech soundtrack serve as identical stimuli for each informant, the approach described above facilitates the analysis of differences in how informants process these stimuli, including analyses of HCI searches at different task stages (frequency, distribution, and rate) and of the number of dropped chunks across groups.

To synchronize unmatching timelines among informants, you can use universal time (see Figure 5). Two reference points are identified: *Point A*, marking the first timestamp of the source-speech soundtrack, and *Point B*, representing the zero time at the start of the individual screen recording video. By calculating the time difference between *Points A and B*, you can create a standardized reference frame. This allows you to recalculate the timestamp of any event on an individual video timeline by subtracting this time difference, resulting in a new virtual timestamp that effectively removes the varied starting portions. This approach ensures that all informants' screen recordings with SI output soundtracks are synchronized from the same onset point, regardless of the moment when each one individually began their interpreting process.

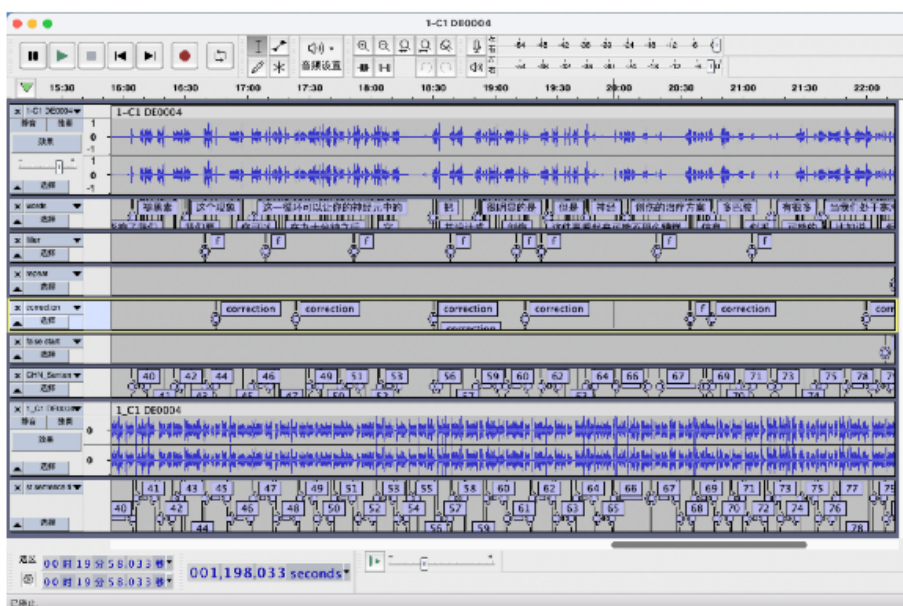
Figure 5. **Universal time alignment for multiple informants.**



7. Further actions: beyond alignment

Although multimodal data processing involves time-intensive work, alignment is merely the first foundational step toward further processing, such as annotation. After successful alignment, you obtain a master spreadsheet, the ethogram, showing the chronological order of events including source-speech transcripts, interpreting output transcripts, and keystroke logs. Rather than combining these into a single scheme, you should keep your ethogram as your primary reference source. From this aligned dataset, you can extract sub-classified data according to your research interests, either as independent spreadsheets or grouped together. These targeted extracts may include, for instance, chunk-based ear-to-key span measurements, numbers of dropped chunks, or accuracy rates of problem triggers—all derived directly from the aligned data.

Figure 6. **Layers for coding indicators in Audacity.**



Processing aligned audio data requires further steps. For example, you can import the source-speech soundtrack into audio editors like Audacity

and add the individual interpreting output as an extra soundtrack layer.⁶ This enriched ethogram allows you to annotate fluency indicators, for instance, false starts, corrections, fillers, and repetitions, as illustrated in Figure 6. Detailed steps for annotation can be found in Du (2024).

I acknowledge that substantial human correction and review are required during the alignment process, even before undertaking alignment, which demands additional steps for data cleaning. Currently, there appears to be no verified shortcut for this challenge. While AI might help with certain aspects of processing, it is not particularly effective for alignment tasks involving audio data displayed in waveform format. Besides, AI may accelerate workflow in the annotation phase through machine learning-supported platforms such as Label Studio, Dataturks, and BRAT Rapid.⁷ These tools require pre-trained datasets—in other words, you need good, labeled data to train the platform on pattern recognition before applying these rules to larger multimodal datasets.

8. Conclusion

This technical report presented a cognitive-situated approach (Robbins & Aydede 2009) to aligning multimodal data for researching aspects of multilectal mediated communication.

This human-centric orientation recognizes the informant as an embedded cognitive agent interacting dynamically with both physical and digital environments. Stimuli such as visual cues from on-screen activities and source-speech audios are not only informative by themselves: they also point to cognitive demands that may shape the interpreter's processing efforts strategy. These indicators are extracted from keylogging, screen recording, and speech recording, enabling cross-referencing and triangulating representations of informants' cognition activities and supporting the

⁶ Audacity: <https://www.audacityteam.org/>

⁷ Label Studio: <https://github.com/HumanSignal/label-studio#try-out-label-studio>

Dataturks: <https://docs.dataturks.com/>

Brat rapid: <https://brat.nlplab.org/>

discussion of informants' cognition in digitally embedded environments. In my case, a quantitative profiling of informants' HCI behavior, using performance metrics such as fluency and accuracy clusters (e.g., Plevoets & Defrancq, 2016), was integrated with a qualitative behavioral analysis.

Interpreters do not process abstract codes. They process information from sound, visuals, and digital sources all at once, and cannot adjust what has been rendered (Zhang 2016). How interpreters handle different information streams may contribute to revealing their underlying cognitive processes, which are influenced by what they already know, how difficult the task is, and fluctuations in the cognitive demands informants may receive. This technical report aimed at better capturing the intertwined reality of interpreters' work and to better support researchers who study challenging multimodal, multilectal communication tasks like translation and interpreting.

The time and efforts devoted to implementing multimodal alignment are still burdensome and may contain human errors in most cases. A further aim is to explore streamlined and semi-automated alignment solutions, integrating keylogging, screen recording, audio capture, and eyetracking into single-platform environments. However, we need to depart from good, correct training data sets, so these painful first steps are inevitable. Tailoring the alignment methods to specific *research questions* and *individuals* is essential, as is thorough documentation to ensure methodological transparency and reproducibility. These were the goals I tried to foster here.

References

- Ahrens, Barbara, and Przemysław Janikowski. 2025. "Speech Recording". In *Research Methods in Cognitive Translation and Interpreting Studies*, edited by Ana María Rojo López and Ricardo Muñoz Martín, 183–212. Amsterdam: John Benjamins. <https://doi.org/10.1075/rmal.10.08ahr>.
- Angelone, Erik. 2021. "Broadening the Scope of Error Categories in Translation Assessment through Screen Recording". *Across Languages and Cultures* 22 (2): 143–57. <https://doi.org/10.1556/084.2021.00021>.
- Aston, Guy. 2018. "Acquiring the Language of Interpreters: A Corpus-Based Approach". In *Making Way in Corpus-Based Interpreting Studies*, edited by Mariachiara Russo,

- Claudio Bendazzoli, and Bart Defrancq, 83–96. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-10-6199-8_5.
- Borg, Claudine, Carmen Heine, and Hanna Risku. 2025. “Observations and Diaries”. In *Research Methods in Cognitive Translation and Interpreting Studies*, edited by Ana María Rojo López and Ricardo Muñoz Martín, 108–32. Amsterdam: John Benjamins. <https://doi.org/10.1075/rmal.10.05bor>.
- Chmiel, Agnieszka, Przemysław Janikowski, Danijel Koržinek, Agnieszka Lijewska, Marta Kajzer-Wietrzny, Dariusz Jakubowski, and Koen Plevoets. 2023. “Lexical Frequency Modulates Current Cognitive Load, but Triggers No Spillover Effect in Interpreting”. *Perspectives*, (June): 1–19. <https://doi.org/10.1080/0907676X.2023.2218553>.
- Chmiel, Agnieszka, Przemysław Janikowski, and Agnieszka Lijewska. 2020. “Multimodal Processing in Simultaneous Interpreting with Text: Interpreters Focus More on the Visual than the Auditory Modality”. *Target* 32 (1): 37–58. <https://doi.org/10.1075/target.18157.chm>.
- Chmiel, Agnieszka, Marta Kajzer-Wietrzny, Danijel Koržinek, Dariusz Jakubowski, and Przemysław Janikowski. 2024. “Syntax, Stress and Cognitive Load, or on Syntactic Processing in Simultaneous Interpreting”. *Translation, Cognition & Behavior* 7 (1): 22–47. <https://doi.org/10.1075/tcb.00091.chm>.
- Christodoulides, George. 2016. “Effects of Cognitive Load on Speech Production and Perception”. PhD dissertation, UCL - Université Catholique de Louvain. <https://dial.uclouvain.be/pr/boreal/object/boreal:177060>.
- Cowan, Nelson. 2005. *Working Memory Capacity*. Essays in Cognitive Psychology. New York: Psychology Press.
- Doherty, Stephen. 2021. “Multisensory Integration in Audiovisual Translation”. In *Multilingual Mediated Communication and Cognition*, edited by Ricardo Muñoz Martín and Sandra Halverson, 155–70. London: Routledge. <https://doi.org/10.4324/9780429323867-7>.
- Du, Zhiqiang. 2024. “Bridging the Gap: Exploring the Cognitive Impact of Interpret-Bank on Chinese Interpreting Trainees”. PhD dissertation, Alma Mater Studiorum - Università di Bologna. <https://doi.org/10.48676/unibo/amsdottorato/11584>.
- Enríquez Raído, Vanessa, Erik Angelone, and Christian Olalla Soler. 2025. “Screen Recording”. In *Research Methods in Cognitive Translation and Interpreting Studies*, edited by Ana María Rojo López and Ricardo Muñoz Martín, 133–56. Amsterdam: John Benjamins. <https://doi.org/10.1075/rmal.10.06rai>.

- Frittella, Francesca Maria. 2022. "CAI Tool-Supported SI of Numbers: A Theoretical and Methodological Contribution". *International Journal of Interpreter Education* 14 (1): 32–56. <https://doi.org/10.34068/ijie.14.01.05>.
- Halverson, Sandra L., and Ricardo Muñoz Martín. 2020. "The Times, They Are a-Changin": Multilingual Mediated Communication and Cognition". In *Multilingual Mediated Communication and Cognition*, edited by Ricardo Muñoz Martín and Sandra L. Halverson, 1–17. London: Routledge. <https://doi.org/10.4324/9780429323867-1>
- Hooge, Ignace T. C., Antje Nuthmann, Marcus Nyström, Diederick C. Niehorster, Gijs A. Holleman, Richard Andersson, and Roy S. Hessels. 2025. "The Fundamentals of Eye Tracking Part 2: From Research Question to Operationalization". *Behavior Research Methods* 57 (2): 73. <https://doi.org/10.3758/s13428-024-02590-2>.
- Huang, Dan Feng, Fang Li, and Hang Guo. 2023. "Chunking in Simultaneous Interpreting: The Impact of Task Complexity and Translation Directionality on Lexical Bundles". *Frontiers in Psychology* 14 (August): 1252238. <https://doi.org/10.3389/fpsyg.2023.1252238>.
- Huxley, Julian S. 1914. "The Courtship - Habits of the Great Crested Grrebe (*Podiceps Cristatus*); with an Addition to the Theory of Sexual Selection." *Proceedings of the Zoological Society of London* 84 (3): 491–562. <https://doi.org/10.1111/j.1469-7998.1914.tb07052.x>.
- Katerina, Tzafilkou, and Protogeros Nicolaos. 2018. "Mouse Behavioral Patterns and Keystroke Dynamics in End-User Development: What Can They Tell Us about Users' Behavioral Attributes?" *Computers in Human Behavior* 83 (June): 288–305. <https://doi.org/10.1016/j.chb.2018.02.012>.
- Kokkola, Sari, and Anne Ketola. 2015. "Thinking Outside the 'Methods Box': New Avenues for Research in Multimodal Translation". In *Rajojen Dynamiikkaa, Gränsernas Dynamik, Borders under Negotiation, Grenzen Und Ihre Dynamik*, edited by Danel Rellstab and Nestori Siponkoski, 4: 219–228. Vaasa, Finland: Vakki Publications. <https://doi.org/10.70484/vakki.149182>.
- Korpala, Paweł, and Katarzyna Stachowiak-Szymczak. 2020. "Combined Problem Triggers in Simultaneous Interpreting: Exploring the Effect of Delivery Rate on Processing and Rendering Numbers". *Perspectives* 28 (1): 126–43. <https://doi.org/10.1080/0907676X.2019.1628285>.
- Leijten, Mariëlle, and Luuk Van Waes. 2013. "Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes". *Written Communication* 30 (3): 358–92. <https://doi.org/10.1177/0741088313491692>.

- Lindgren, Eva, and Kirk P. H. Sullivan, eds. 2019. *Observing Writing: Insights from Keystroke Logging and Handwriting*. Studies in Writing, volume 38. Leiden; Boston: Brill. <https://doi.org/10.1163/9789004392526>
- Mankauskienė, Dalia. 2016. "Problem Trigger Classification and Its Applications for Empirical Research". *Procedia-Social and Behavioral Sciences* 231 (October): 143–48. <https://doi.org/10.1016/j.sbspro.2016.09.083>.
- Messerman, Arik, Tarik Mustafić, Seyit Ahmet Camtepe, and Sahin Albayrak. 2011. "Continuous and Non-Intrusive Identity Verification in Real-Time Environments Based on Free-Text Keystroke Dynamics". In *2011 International Joint Conference on Biometrics (IJCB)*, 1–8. <https://doi.org/10.1109/IJCB.2011.6117552>.
- Mizuno, Akira. 2017. "Simultaneous Interpreting and Cognitive Constraints". In *Journal of College of Literature*, 58:1–28. Aoyama Gakuin University.
- Muñoz Martín, Ricardo, and Matthias Apfelthaler. 2022. "A Task Segment Framework to Study Keylogged Translation Processes". *The International Journal of Translation and Interpreting Research* 14 (2): 8–31. <https://doi.org/10.12807/ti.114202.2022.a02>.
- Muñoz Martín, Ricardo, and José Ma. Cardona Guerra. 2018. "Translating in Fits and Starts: Pause Thresholds and Roles in the Research of Translation Processes". *Perspectives* 27 (4): 525–51. <https://doi.org/10.1080/0907676X.2018.1531897>.
- Muñoz Martín, Ricardo, and Ana María Rojo López. 2025. "Introduction". In *Research Methods in Cognitive Translation and Interpreting Studies*, edited by Ana María Rojo López and Ricardo Muñoz Martín, 1–20. Amsterdam: John Benjamins. <https://doi.org/10.1075/rmal.10.intro>.
- Muñoz Martín, Ricardo, Sanjun Sun, Zhiqiang Du, and Sara Puerini. 2025. "Keylogging". In *Research Methods in Cognitive Translation and Interpreting Studies*, edited by Ana María Rojo López and Ricardo Muñoz Martín: 157–82. Amsterdam: John Benjamins. <https://doi.org/10.1075/rmal.10.07mar>.
- Niehorster, Diederick C., Marcus Nyström, Roy S. Hessels, Richard Andersson, Jeroen S. Benjamins, Dan Witzner Hansen, and Ignace T. C. Hooge. 2025. "The Fundamentals of Eye Tracking Part 4: Tools for Conducting an Eye Tracking Study". *Behavior Research Methods* 57 (1): 46. <https://doi.org/10.3758/s13428-024-02529-7>.
- Plevoets, Koen, and Bart Defrancq. 2016. "The Effect of Informational Load on Disfluencies in Interpreting: A Corpus-Based Regression Analysis". *Translation and Interpreting Studies. The Journal of the American Translation and Interpreting Studies Association* 11 (2): 202–24. <https://doi.org/10.1075/tis.11.2.04ple>.

- Prandi, Bianca. 2023. *Computer-Assisted Simultaneous Interpreting: A Cognitive-Experimental Study on Terminology*. Berlin: Language Science Press. <https://doi.org/10.5281/zenodo.7143055>.
- Robbins, Philip, and Murat Aydede, eds. 2009. *The Cambridge Handbook of Situated Cognition*. Cambridge: Cambridge University Press.
- Seeber, Kilian G. 2017. "Multimodal Processing in Simultaneous Interpreting". In *The Handbook of Translation and Cognition*, edited by John W. Schwieter and Aline Ferreira, 1st ed., 461–75. Hoboken, NJ: John Wiley & Sons. <https://doi.org/10.1002/9781119241485.ch25>.
- . 2011. "Cognitive Load in Simultaneous Interpreting: Existing Theories — New Models". *Interpreting. International Journal of Research and Practice in Interpreting* 13 (2): 176–204. <https://doi.org/10.1075/intp.13.2.02see>.
- Shao, Zhangminzi, and Bart Defrancq. 2024. "Fundamental Frequency as an Acoustic Mirror of Interpreters' Cognitive States". *Interpreting* 27 (1): 1–27. <https://doi.org/10.1075/intp.00107.sha>.
- Van Waes, Luuk, Mariëlle Leijten, Jens Roeser, Thierry Olive, and Joachim Grabowski. 2021. "Measuring and Assessing Typing Skills in Writing Research". *Journal of Writing Research* 13 (1): 107–53. <https://doi.org/10.17239/jowr-2021.13.01.04>.
- Walker, Callum, Kristian Tangsgaard Hvelplund, and Victoria Lei. 2025. "Screen Eyetracking". In *Research Methods in Cognitive Translation and Interpreting Studies*, edited by Ana María Rojo López and Ricardo Muñoz Martín, 213–34. Amsterdam: John Benjamins. <https://doi.org/10.1075/rmal.10.09wal>.
- Wang, Fuxiang. 2022. "Impact of Translation Difficulty and Working Memory Capacity on Processing of Translation Units: Evidence from Chinese-to-English Translation". *Perspectives* 30 (2): 306–22. <https://doi.org/10.1080/0907676X.2021.1920989>.
- Wehrmeyer, Ella. 2019. "A Corpus for Signed Language Interpreting Research". *Interpreting* 21 (1): 62–90. <https://doi.org/10.1075/intp.00020.weh>.
- Zhang, Xiaojun. 2016. "Semi-Automatic Simultaneous Interpreting Quality Evaluation". *International Journal on Natural Language Computing* 5 (5): 1–12. <https://doi.org/10.5121/ijnlc.2016.5501>.

Building ethograms for interpreting research —A technical report

Abstract

By establishing precise temporal relationships across modalities, researchers can connect behavioral indicators with cognitive processes, revealing otherwise hidden patterns in interpreter performance. This paper presents a hands-on approach for aligning multimodal data in interpreting studies. It introduces four strategies for synchronizing audio, video, and keystroke data along a unified temporal framework: visual observation, auditory identification, signal sound recognition, and words and chunk levels. This human-centered approach envisions interpreters as cognitive agents interacting with physical and digital environments. Though time-intensive, these alignment protocols provide essential foundations for capturing interpreting's multimodal nature and support more rigorous, ecologically valid research in cognitive translation and interpreting studies.

Keywords

Data cleaning, multimodal data alignment, interpreting research

Author

Du Zhiqiang

Alma Mater Studiorum Università di Bologna

zhiqiang.du2@outlook.com

<https://orcid.org/0000-0002-6659-250X>

How to cite

Zhiqiang, Du. 2026. "Building ethograms for interpreting research — A technical report." *MeTI: Methods in Translation and Interpreting* 0: xxx-xxx.