SUMM-RE: A corpus of French meeting-style conversations

Julie Hunter, Hiroyoshi Yamasaki, Océane Granier, Jérôme Louradour, Roxane Bertrand, Kate Thompson, Laurent Prévot

SUMM-RE: A corpus of French meeting-style conversations

Julie Hunter[1] Hiroyoshi Yamasaki[2] Océane Granier[2]
Jérôme Louradour[1] Roxane Bertrand[2] Kate Thompson[1] Laurent Prévot[2]
(1) LINAGORA Labs, Toulouse, France
{jhunter, jlouradour, cthompson}@linagora.com
(2) Aix Marseille Université & CNRS, LPL, Aix-en-Provence, France
{hiroyoshi.yamasaki, oceane.granier, roxane.bertrand,
laurent.prévot}@univ-amu.fr

RÉSUMÉ

Nous présentons le corpus SUMM-RE un ensemble de données d'environ 95 heures de conversations spontanées de type réunion en français. Le corpus est conçu pour servir de base à des tâches en aval telles que le résumé de réunions. Dans son état actuel, il offre 25 heures de transcriptions corrigées manuellement et alignées sur le signal audio, ce qui en fait une ressource précieuse pour l'évaluation des systèmes d'ASR et de reconnaissance du locuteur. Il comprend également des transcriptions automatiques et des alignements de l'ensemble du corpus qui peuvent être utilisés pour des tâches de NLP en aval. L'objectif de cet article est de décrire la conception, la production et l'annotation du corpus jusqu'à l'étape de transcription, ainsi que de fournir une description quantitative du corpus permettant de comprendre ses principales caractéristiques linguistiques.

ABSTRACT

We present the SUMM-RE corpus, a dataset of roughly 95 hours of spontaneous meeting-style conversations in French. The corpus is designed to serve as a foundation for downstream tasks such as meeting summarization. In its current state, it offers 25 hours of manually corrected transcripts that are aligned with the audio signal, making it a valuable resource for evaluating ASR and speaker recognition systems. It also includes automatic transcripts and alignments of the whole corpus which can be used for downstream NLP tasks. The aim of this paper is to describe the conception, production and annotation of the corpus up to the transcription level as well as to provide statistics that shed light on the main linguistic features of the corpus.

MOTS-CLÉS : Corpus, Réunions, Conversation spontanée, Français, Dialogue, Transcription.

KEYWORDS: Corpus, Meetings, Spontaneous conversation, French, Dialogue, Transcription.

1 Introduction

Spontaneous, multiparty conversation poses particular problems for both speech and natural language processing tasks. Disfluent speech, non-standard or incorrect grammar, idiosyncratic speaker styles, and overlapping speech and interruptions — not to mention transcription errors introduced when transcripts are automatically generated — complicate the task of understanding conversation transcripts, especially for NLP models that have been trained largely on text (Rennard *et al.*, 2023).

These linguistic features together with complicated acoustic conditions, shared microphones, different accents, and rapid speech aggravate the problem for speech processing tasks such as transcription and speaker identification (Yamasaki *et al.*, 2023).

Progress on these topics is hindered by a lack of data especially for specific tasks such as meeting summarization. Few people are comfortable sharing recordings of their discussions and even if they are, preparing the data for training poses a significant hurdle. Data scarcity is even more serious when we look to languages other than English.

The principal goal of the SUMM-RE project is to develop models for conversation and meeting summarization with a particular focus on French. A critical step in accomplishing this goal has been to collect and prepare a corpus of roughly 100 hours of meeting-style conversations in French. The aim of this paper is to present this corpus and to detail its conception and production.

2 Related corpora

AMI (Augmented Multi-party Interaction ; Carletta *et al.*, 2005) and ICSI (International Computer Science Institute ; Janin *et al.*, 2003; McCowan *et al.*, 2005) are the most well-known meeting corpora (and until recently, were the only meeting corpora at all). AMI contains 137 scenario-driven meetings that last from 15 to 45 minutes each, for a total of around 65 hours of conversations. In each meeting, four participants play roles in a fictitious electronics company and participate in a sequence of four meetings. The roles and scenarios are well developed and always the same, which facilitates the task of getting four strangers to carry out structured discussions on topics for which they have little background knowledge and also avoids privacy concerns triggered by real meanings. On the other hand, while the language remains spontaneous, the heavy corpus design engenders conversational styles and interactions that are arguably much cleaner than real-life meetings and also leads to a homogeneity of content and vocabulary. SUMM-RE by contrast, is designed to encourage more fluid discussion on a range of topics ; the focus is on trying to elicit particular types of discursive interactions that are characteristic of meetings without insisting that the participants play employee-like roles.

ICSI consists of natural, weekly meetings that last about one hour each for a total of roughly 72 hours of recordings. On average, the meetings involve six participants that can include undergraduates, graduate students, and professors who meet to discuss technical topics related to natural language processing, computational linguistics and even the ICSI corpus itself. As ICSI contains real meetings between people who were actively collaborating on projects at the time of recording, the interactions are more natural than those in AMI. At the same time, they draw on technical vocabulary and specialized subjects as well as a considerable amount of shared knowledge between participants, which can complicate the interpretation of the content for models that do not have this knowledge. While such a scenario is undeniably realistic, the SUMM-RE corpus is designed to strike a balance between naturalness and feasibility for automatic summarization. As such, it includes light guidance of the meeting structure and limits the impact of shared background knowledge by bringing in participants who in many cases did not know each other before participating in our corpus.

While both AMI and ICSI are entirely in English, there have been recent efforts to expand to other languages. VCSum (Versatile Chinese Meeting Summarization Dataset Wu *et al.*, 2023) is a collection of transcripts and videos of roundtable meetings in Chinese found on the internet. A total of 239 meetings were selected for a total of over 230 hours of recording time. The corpus is called “versatile”

because it contains a variety of annotations that can be relevant for different summarization tasks. The ELITR corpus (Nedoluzhko *et al.*, 2022) contains transcripts for 113 technical project meetings in English but also for 53 meetings in Czech, for a total of over 160 hours of content. Like ICSI, the meetings in ELITR are natural, leading to many of the same advantages and drawbacks. Unlike ICSI, AMI, VCSum and SUMM-RE however, the ELITR audio files have not been released and parts of the meetings are censored for privacy.

There are also a variety of smaller, conversational corpora in French (for a recent list, see Hunter *et al.*, 2023). CID (The Corpus of Interactional Data Blache *et al.*, 2017), which contains eight one-hour dialogues between friends, has notable similarities with SUMM-RE in that a major effort was involved in adding different levels of annotation, including dialogue-central information that can be exploited by downstream NLP models. With only eight hours of recording, however, it remains very small and is not focused specifically on meetings.

3 Corpus Design

To bypass the concern of sharing personal information, the SUMM-RE corpus does not contain real meetings. The conversations were nevertheless designed to imitate certain important features of meeting-style conversation, developing situations in which participants have to make decisions or plan out a list of action items or report on things they have done. They were also designed to have some continuity with past meetings, as real meetings often do : almost every one of the 96 individual experiments in the corpus is made up of a series of three 20 minute meetings that develop a given topic. [1] In general, [2] the participants were asked to plan an event during the third meeting, while the second meeting focused on deciding what kind of event to plan, and the first meeting involved participants going around the table to discuss their experience or opinions about certain topics. Each one-hour experiment contains the same set of participants—usually four but sometimes fewer—throughout.

In some cases, meeting participants knew each other before the experiment. This condition adds an arguably realistic element to the interactions, as many meetings are held by people who have worked together before. However, because our meeting scenarios were artificial, we also feared that it would encourage playful interactions full of jokes and laughter. While such interactions are certainly possible in professional meetings, many meetings involve a higher level of seriousness and personal distance. In an effort to vary the interactions and imitate different levels of professionalism, we aimed to balance the number of groups in which all participants knew each other, some participants knew each other or no participants knew each other.

To encourage participants to speak naturally and spontaneously while also pushing them to stick to a meeting-like agenda, we had to strike a balance between role playing and freedom for the participants to talk about their own experiences. To encourage freedom, we defined a set of topics that we assumed most participants would be comfortable discussing, including a) internet and technology-related topics like social networks and the societal influence of companies like Google and Amazon, b) films and TV series, c) fundraising, [3] d) music, e) environment-related topics such as global warming and renewable energy. Within these larger topics, participants were allowed to choose a subtopic that

1. Due to technical problems, we had to remove five meetings from the final corpus.
2. In the early pilot studies, we tried different types of organization but ultimately decided upon the one described here.
3. We abandoned this topic, which is discussed in 4.4% of the overall data, because participants struggled to develop it.

interested them.

To add structure to the conversations, we proposed a series of subtopics or responsibilities that each participant could choose to lead as well as a series of points, questions, or tasks (depending on the format of the meeting) that the participants might want to pursue. For each subtopic, the participants received an individual card or a collective document listing the points so that they had a "cheat sheet" if they struggled to develop their contribution to the conversation.

Here, for example, is the set of questions (translated from French to English here) given to a participant who chose to discuss Amazon during a reporting meeting :

> *You have chosen to be responsible for leading the discussion about Amazon. You will need to be able to talk about this subject for 3-4 minutes. To help guide you, below is a list of questions to which you might respond (but you are free to choose other questions if you find them more suitable) :*
> — What is Amazon ? How do we interact with Amazon in our daily lives
> — What are the positive and negative sides of Amazon ?
> — What do you think about their approach to package delivery, the Prime video platform, employee conditions, etc ?
> — How do you think that Amazon will evolve in the future ? Will they become more influential ? Will other actors replace them ?
> — ...
> *You can draw from your personal experience to respond to these questions or provide concrete examples.*

Finally, for each 20 minute conversation, the group was asked to choose a moderator among them. In addition to managing the discussion for a 20 minute conversation (also following suggested guidelines), each moderator was assigned the task of taking notes and of making an oral summary at the end of the meetings based on these notes. To preserve the continuity of the conversation, the summary was recorded immediately after the conversation on the same record. The summary file was later extracted in post-processing.

4 Data acquisition

The original plan was to record the entire SUMM-RE corpus in the H2C2 [4]. Unfortunately, corpus collection began in 2020 and ran through 2021 when the Covid pandemic was still a threat, so in-person meetings were not always an option. We thus adopted two different strategies for corpus collection : in person recordings and Zoom.

In all cases, the recordings shared certain characteristics. For instance, each one hour experiment involved the same set of participants. In general, there were four participants but sometimes, we were only able to find three and on very rare occasions, two. Each participant had their own microphone for recording, as explained below. Finally, each experiment was led by one of the co-authors of this paper, who would begin by giving instructions to the participants and making sure they understood the task, but would leave the room during recording to avoid interaction with the participants.

Most in-person recordings took place in the H2C2, though there are a few exceptions in which it was

4. https://plateformeh2c2.fr/

easier to go into peoples' homes. Participants for these recordings were generally recruited through our platform, calls for participants in social media or announcements made in university courses. In total, 248 out of 283 meetings were recorded in person (see Table 3 for more detail).

The H2C2 studio is composed of two rooms :

— an experimental room where participants interact during the recordings
— an observation and recording room where the person in charge of managing the experiment goes during the experiments and can observe participants through a one-way mirror.

Each participant was equipped with an individual headset with microphone (AKG 520). An additional microphone was used to capture the streams of all speakers at once. All microphones were recorded with a Zoom H8 handy recorder.

While recording conditions in the H2C2 were near ideal, we encountered two problems that impacted the quality of the final recordings. First, although the room was spacious enough to leave a fair amount of distance between speakers and each headset microphone was adjusted to the voice of the person wearing it, sometimes a participant's voice would vary between the test conditions and the final recording conditions. A speaker might end up using a higher pitch during recording due to elevated emotion, for example. In such cases, individual microphones would often end up capturing the voices of multiple speakers. Our efforts to isolate the contributions of the main speaker in such cases are described in Section 5. The second complication resulted from efforts to navigate health and safety regulations enforced during the Covid pandemic. To reduce contact between speakers, we first tried installing plexiglass barriers so that participants could see each other's mouths, but this led to an echo in the recordings. Ultimately, we asked the participants to wear masks. This worked in most cases, but idiosyncratic approaches to mask wearing still led to suboptimal recording in some cases.

35 out of 283 of the meetings were recorded through Zoom. For these, participants were recruited through the crowd sourcing platform Prolific. [5] Each person was required to use their own microphone to take part in the experiment in an effort to limit background noise. While Zoom facilitated the task of speaker identification and minimized the impact of phenomena like overlapping speech, dependence on personal equipment and internet connections led to poor recording quality in some cases. See 7.2 for more information on the effect of communication modality on participants' behavior.

5 Annotation and post-processing

The SUMM-RE corpus is partitioned into `train`, `dev` and `test` data sets with proportions at roughly 75%, 12.5% and 12.5%, respectively. Files were assigned one of three classes randomly with the constraint that the ratio of Zoom experiments to in-person experiments as well as relative proportions of scenarios were kept roughly constant. After this automatic process, minor adjustments were made to ensure that certain files that had been manually corrected ended up in the `dev` set. This procedure resulted in 210 files in the train set, 36 files in the dev set and 37 files in the test set.

The entirety of the SUMM-RE corpus has been automatically transcribed and the entirety of the `dev` set, roughly 12 hours, has been manually corrected and annotated. Correction of the `test` set is underway and will soon be complete. Manual correction is performed in two phases. First, the transcripts are manually corrected for transcription errors. Corrections at this stage are made with the software Praat (Boersma & Van Heuven, 2001). In places where Whisper misses a substantial

5. https ://www.prolific.com/

pause (i.e. on the order of hundreds of milliseconds) a pause marker # is added. If the IPU boundaries differ significantly from the true boundary, this is also manually corrected. All files are then manually verified for any obvious errors made during the first correction as well as minor adjustments such as adding non-linguistic annotations for non-speech sounds like laughs and coughs to improve tier boundary accuracy. On the `dev` set, corrections took one month and were carried out by one of the co-authors who was paid for the task and who is a native speaker of French. Manual verification was performed by another co-author (non-native speaker).

While the details of our automatic pipeline are described in Yamasaki *et al.* (2023), we give a brief overview here. At a high-level, the pipeline can be divided into two parts : the detection of *inter-pausal units* (IPUs)—segments of audio in which the principal speaker is speaking—and the transcription of the words that were uttered in the IPU along with their start and end times. IPU detection was necessary because, as explained in Section 4, individual microphones often captured the voices of multiple speakers. For this task, we employed out-of-the-box IPU annotation provided by the SPPAS package (Bigi & Priego-Valverde, 2019; Bigi, 2015) combined with an approach which relies on speaker diarization using the Pyannote package (Bredin *et al.*, 2020; Bredin & Laurent, 2021). This second step was necessary because the voice intensity processed by SPPAS and the voice quality processed by Pyannote contain complementary information. Once the IPUs were identified, we created new audio files in which background voices were replaced with silences and then passed the resulting audio files to Whisper (Radford *et al.*, 2022), OpenAI's speech to text model. [6]

The quality of the annotation and alignment predicted by our pipeline was extensively evaluated in Yamasaki *et al.* (2023) on a 3.3 hour subset of the `dev` set. Table 1 shows the word error rate (WER) for this subset, broken down by deletions (Del), insertions (Ins) and substitutions (Sub). It also includes T-δ, an average (in milliseconds) of the absolute difference in start and end times for each word and an F1 score inspired by Bain *et al.* (2023). See Yamasaki *et al.* (2023) for details.

Pipeline	**F1**	**T-δ**	**WER**	Del	Ins	Sub
Our Reference	**0.81**	108	**18.8**	8.1	5.8	4.8

TABLE 1 – Brief summary of automatic word level evaluation of the SUMM-RE corpus including F1 score for annotation correctness, time-δ for alignment error (in milliseconds) and word error rate (WER) and corresponding deletion, insertion and substitution scores.

A final point is that in order to prevent identification of individual participants we developed a script to remove participant names from both audio files and transcripts and replace the corresponding audio intervals with a beep. This was done by :

— identifying possible candidates for mentions of individual names by calculating the Levenshtein distance of each token to participants' first and second names
— manually filtering false positives
— replacing the resulting list of names by `anon` and the corresponding WAV interval with a single beep

This anonymization scheme assumes the correctness of the transcription, which is not an issue for the

6. There are several variants of Whisper models such as Klein (2023); Bain *et al.* (2023). We chose Louradour (2023) after a detailed comparison of their performance as detailed in Yamasaki *et al.* (2023). We opted for the large v2 model as a starting point but added two custom features. The first employed prompting to encourage Whisper to transcribe disfluencies, yielding a more faithful transcript. The second involved introducing precise word alignment through techniques developed in Louradour (2023) and the Julius forced aligner (Lee *et al.*, 2001). Further details of both the IPU detection algorithm and the Whisper transcription process can be found in Yamasaki *et al.* (2023).

test and dev splits as they are manually corrected, but is for the train set which is not. In cases where an individual's name was not correctly transcribed it may not be anonymized. However, we noticed that it is highly rare for participants to use full names or family names to refer to each other. As first names are not particularly identifiable we believe this anonymization scheme is sufficient.

6 Basic corpus statistics

The SUMM-RE corpus includes a total of 207 unique participants. Due to the recruitment procedure described in Section 4, our data is biased with respect to age, gender and occupation. In particular, the majority of participants were students with a mean age of 28.7, there are nearly three times more female than male participants, and a large portion of participants were monolingual French speakers who grew up in France. See 2 in the Appendix for further metadata on participants.

Our data set consists of 96 sessions with 3 conversations for almost every session, yielding a total of 283 conversations of roughly 20 minutes each. [7] Of these, 8 sessions (22 files) are pilot experiments and the remaining 88 sessions (261 files) are non-pilot experiments. As stated above (Section 4) some meetings (35/283 files) were recorded on Zoom due to Covid restrictions.

7 Linguistic statistics

Dialogue corpora vary greatly depending on a wide range of contextual factors. To better understand the nature of a dialogue corpus, it is crucial to look at its linguistic, and in particular its interactional, properties. In this section, we focus on some core metrics that are are interesting proxies to characterize the corpus. Namely, we look at (i) distribution of Inter-Pausal Unit (IPU) durations (as a proxy to sentence length in the written realm); (ii) automatically extracted backchannels; (iii) amount of overlapping speech; (iv) filled pause count; and (v) speaker dominance. Together these metrics provide an insight about the degree of spontaneity and interaction in the corpus.

7.1 By metric

IPU distribution Figure 1 shows the distribution of IPU duration and number of IPUs for dev, train and test. As the splits were random, both IPU duration and count are similar across splits, though there are slightly more IPUs shorter than 1 second for train than for dev and test. This may be due to erroneous IPUs (e.g., from someone else laughing) removed during correction.

Independent t-tests indicated that while the number of IPUs in meetings from train and dev varied significantly ($p \leq 0.01$), test was consistent with train, meaning that there should not be an issue when using test for evaluation. There was no significant difference for pilot and experiment recordings either ($p = 0.17$), which justifies combining them into a single data set. We also performed t-tests to compare our different scenarios (a-e) and meeting styles (Section 3). For scenarios, 6 pairs reached significance : a-c $p \leq 0.001$, a-d $p \leq 0.05$, a-e $p \leq 0.01$, b-c $p \leq 10^{-4}$, c-d $p \leq 10^{-4}$, c-e $p \leq 10^{-5}$. For tasks, reporting differed significantly from both decision ($p \leq 10^{-4}$) and planning ($p \leq 10^{-5}$), which is to be expected given that reporting meetings were designed to be less interactive.

7. 005b_PBP, 005b_PBR, 012a_EBR, 017b_EBD, 087c_EEP were rejected due to technical issues.

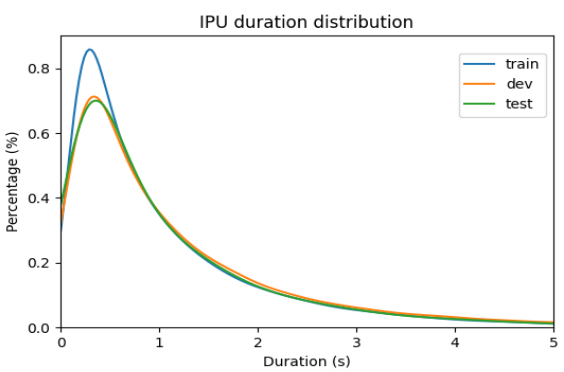


FIGURE 1 – Left : kernel density estimate of IPU duration by splits, Right : number of IPUs by splits

Token distribution. Figure 2 shows the distribution of token duration and speech rate. As expected, there is only a small difference between splits. Average number of tokens per file was 1187 words for `train`, 1107 words for `dev` and 1218 words for `test`. Independent t-test revealed no significant differences between splits (train-dev : $p = 0.11$, dev-test : $p = 0.08$, train-test : $p = 0.55$). Pilot vs experiment comparison was not significant ($p = 0.99$). Comparison by scenarios gave 4 significant pairs : a-c $p \leq 0.01$, b-c $p \leq 0.05$, c-d $p \leq 0.01$ and c-e $p \leq 0.01$. Comparison by task showed that reporting again differed from both decision ($p \leq 0.001$) and planning ($p \leq 0.05$).

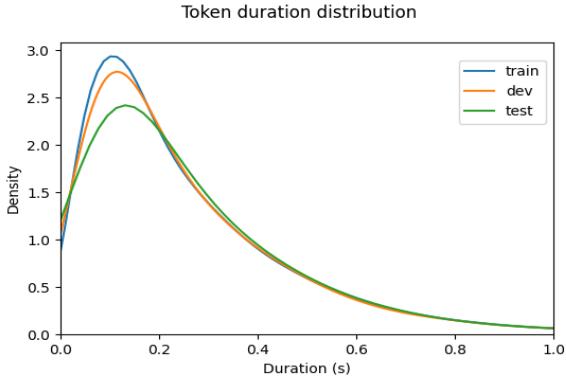


FIGURE 2 – Left : kernel density estimate of token duration by splits, Right : speech rate by splits

Backchannels. Figure 3 (Left) shows the number of backchannels found per minute. For the purpose of this paper we used a rather simplistic definition of backchannel in which we only considered single word utterances that matched "`oui`", "`ouais`", "`hm`", "`mh`", "`non`", "`ok`", "`ah`", "`ben`", "`bien`", "`eh`", "`euh`", "`voilà`". This decision was made for the purpose of simplicity but has a disadvantage of underestimating the actual number.

The number of backchannels did not differ between `train` and `test` ($p = 0.62$) but was slightly lower for `dev` than `train` ($p \leq 0.01$) and `test` ($p \leq 0.01$). Comparison between pilots and experiments did not reach significance ($p = 0.37$). Seven scenario pairs reached significance : a-b $p \leq 0.01$, a-c $p \leq 0.01$, a-d $p \leq 0.01$, b-c $p \leq 0.001$, c-d $p \leq 10^{-4}$, c-e $p \leq 0.001$. The reporting

task again differed from decision ($p \leq 10^{-5}$) and planning ($p \leq 10^{-5}$). Finally, in person meetings had more backchannels than Zoom meetings ($p =\leq 0.001$). See Appendix B.3 for the figures.

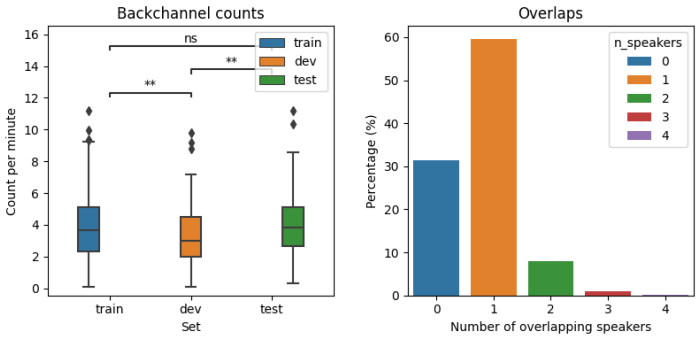


FIGURE 3 – Left : Backchannel counts by sets, Right : Number of speakers at the same time

Overlap. Figure 3 (Right) shows the relative duration of intervals in which $N \in 0, 1, 2, 3, 4$ speakers were speaking at the same time. This was calculated by 1) segmenting the entire conversation into small segments of equal duration (here 0.02 seconds), 2) counting how many speakers are speaking in each segment, 3) aggregating by number of simultaneous speakers, 4) normalizing by the total duration. As expected, in the majority of cases, there is either one speaker or no speaker ; overlaps with more than three participants are extremely rare, in line with previous findings (Çetin & Shriberg, 2006). This suggests that the turn-taking system is robust in multi-party interaction, though overlaps still constitute an important factor (around 10% of speaking time) to consider in downstream processing. We observe that overlaps are even more rare in zoom meetings. See Appendix B.4 for full results.

Filled pause. To assess the degree of conversational fluidity, we considered the number of filled pauses (e.g. "euh", marked as `fp` in the final annotation). The results showed no significant difference for data set, split or task, although there were more filled silences for Zoom meetings than in-person meetings ($p \leq 10^{-5}$). Full results can be found in Appendix B.5.

Dominant speaker. We also looked at whether certain speakers spoke substantially more than others. Results (Appendix B.6) show that on average, there is a dominant speaker and that they tend to speak around 50% of total speaking time followed by the second most active speaker at around 30%. The decrease in percentage is fairly linear. [8] There were no obvious trends across different conditions.

7.2 By condition

In-person vs. Zoom. Figure 4 shows that the number of IPUs ($p \leq 10^{-5}$) and the number of tokens per minute ($p \leq 10^{-5}$) were lower for Zoom than in-person meetings. The number of backchannels per minute was also lower for Zoom ($p \leq 0.001$), while the number of filled pauses was significantly larger ($p \leq 10^{-5}$). Overall, these results indicate that Zoom meetings are less interactive than in person meetings, which is to be expected given the social distance and the fact that participants often cut their microphones when they did not have the floor (as people tend to in real-life online meetings).

8. Some conversations have only three participants meaning the fourth participant's speaking time is set to zero. Thus the data about the fourth participant should be interpreted with care.

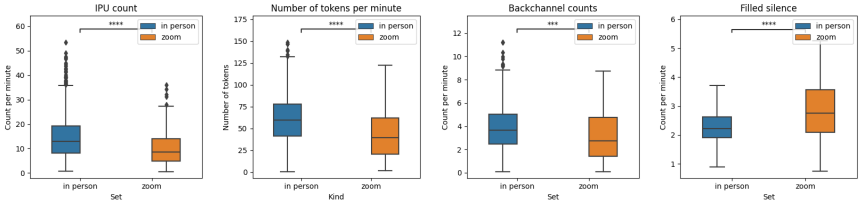


FIGURE 4 – Comparisons of in-person and Zoom meetings from left to right : IPU count, token count, number of backchannels, number of filled pauses

Differences between different scenarios. As noted in Section 3, scenario (c) was too difficult and had to be abandoned. This is consistent with the fact that IPU counts, number of tokens per minute, filled pauses and backchannel counts all differed significantly between scenario (c) and the others (see above and Appendix B for figures and details). Scenario (a) also stood out. IPU counts were smaller for (a) (technology) than (e) (environment, $p \leq 0.01$) and (d) (music, $p \leq 0.05$). Backchannel counts per minute were different for : a-d ($p \leq 0.05$), a-e ($p \leq 0.01$), a-c ($p \leq 0.001$).

Overlaps paint a slightly different picture. While scenario (c) has more pauses and less speech, consistent with the above, scenario (a) has the smallest amount of silence and highest amount of one person talking. This might suggest that scenario (a) is a "more serious" topic, for which people have a tendency to speak longer and in a less chat-like manner (see Appendix B for figures).

Differences between different tasks. Comparison by task revealed that reporting meetings have fewer IPUs than planning ($p \leq 10^{-5}$) and decision ($p \leq 10^{-4}$), fewer tokens per minute than planning ($p \leq 0.05$) and decision ($p \leq 0.001$), and fewer backchannels than planning/decision ($p \leq 0.0001$). Reporting also had highest portion where only one person is speaking (see Appendix B for figures). This is to be expected due to the fact that reporting meetings were designed in a round-table style in order to encourage monologue. There was no difference in the number of filled pauses across tasks.

8 Conclusion

We have presented the SUMM-RE corpus, a new dataset containing 96 hours of spontaneous, multiparty meeting-style conversations in French. The corpus is the only French corpus of its kind, and one of the only large-scale meeting corpora in a language other than English. While the meetings are based on loose role-playing, they remain natural and are designed to elicit basic discursive interactions that we can expect from real meetings. We have offered preliminary analyses of the data to illustrate the level of spontaneity and interactivity in the corpus and have shown how this can vary depending on the type of meeting involved, the subject and—although this was not a part of the initial aim of the corpus—whether the conversation was recorded in-person or on Zoom. The dataset is available on Hugging Face at https://huggingface.co/datasets/linagora/SUMM-RE.

9 Acknowledgements

We gratefully acknowledge support from the ANR grant SUMM-RE (ANR-20-CE23-0017).

Références

BAIN M., HUH J., HAN T. & ZISSERMAN A. (2023). Whisperx : Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv :2303.00747*.

BIGI B. (2015). Sppas-multi-lingual approaches to the automatic annotation of speech. *The Phonetician. Journal of the International Society of Phonetic Sciences*, **111**(ISSN : 0741-6164), 54–69.

BIGI B. & PRIEGO-VALVERDE B. (2019). Search for inter-pausal units : application to cheese ! corpus. In *9th Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics*, p. 289–293.

BLACHE P., BERTRAND R., FERRÉ G., PALLAUD B., PRÉVOT L. & RAUZY S. (2017). The corpus of interactional data : A large multimodal annotated resource. *Handbook of linguistic annotation*, p. 1323–1356.

BOERSMA P. & VAN HEUVEN V. (2001). Speak and unspeak with praat. *Glot International*, **5**(9/10), 341–347.

BREDIN H. & LAURENT A. (2021). End-to-end speaker segmentation for overlap-aware resegmentation. In *Interspeech*.

BREDIN H., YIN R., CORIA J. M., GELLY G., KORSHUNOV P., LAVECHIN M., FUSTES D., TITEUX H., BOUAZIZ W. & GILL M.-P. (2020). pyannote.audio : neural building blocks for speaker diarization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

CARLETTA J., ASHBY S., BOURBAN S., FLYNN M., GUILLEMOT M., HAIN T., KADLEC J., KARAISKOS V., KRAAIJ W., KRONENTHAL M. *et al.* (2005). The AMI meeting corpus : A pre-announcement. In *International workshop on machine learning for multimodal interaction*, p. 28–39 : Springer.

ÇETIN O. & SHRIBERG E. (2006). Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site : Insights for automatic speech recognition. In *Ninth international conference on spoken language processing*.

HUNTER J., LOURADOUR J., RENNARD V., HARRANDO I., SHANG G. & LORRÉ J.-P. (2023). The claire french dialogue dataset. *arXiv preprint arXiv :2311.16840*.

JANIN A., BARON D., EDWARDS J., ELLIS D., GELBART D., MORGAN N., PESKIN B., PFAU T., SHRIBERG E., STOLCKE A. *et al.* (2003). The ICSI meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003 (ICASSP'03)*, volume 1, p. I–I : IEEE.

KLEIN G. (2023). Faster whisper transcription with ctranslate2. *GitHub repository*.

LEE A., KAWAHARA T., SHIKANO K. *et al.* (2001). Julius-an open source real-time large vocabulary recognition engine. In *INTERSPEECH*, p. 1691–1694.

LOURADOUR J. (2023). whisper-timestamped. *GitHub repository*.

MCCOWAN I., CARLETTA J., KRAAIJ W., ASHBY S., BOURBAN S., FLYNN M., GUILLEMOT M., HAIN T., KADLEC J., KARAISKOS V., KRONENTHAL M., LATHOUD G., LINCOLN M., LISOWSKA MASSON A., POST W., REIDSMA D. & WELLNER P. (2005). The AMI meeting corpus. *International Conference on Methods and Techniques in Behavioral Research*.

NEDOLUZHKO A., SINGH M., HLEDÍKOVÁ M., GHOSAL T. & BOJAR O. (2022). ELITR Minuting Corpus : A novel dataset for automatic minuting from multi-party meetings in English and Czech. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC-2022)*, Marseille, France : European Language Resources Association (ELRA). In print.

RADFORD A., KIM J. W., XU T., BROCKMAN G., MCLEAVEY C. & SUTSKEVER I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv :2212.04356*.

RENNARD V., SHANG G., HUNTER J. & VAZIRGIANNIS M. (2023). Abstractive meeting summarization : A survey. *Transactions of the Association for Computational Linguistics*, **11**, 861–884.

WU H., ZHAN M., TAN H., HOU Z., LIANG D. & SONG L. (2023). VCSUM : A versatile Chinese meeting summarization dataset. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Findings of the Association for Computational Linguistics : ACL 2023*, p. 6065–6079, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-acl.377.

YAMASAKI H., LOURADOUR J., HUNTER J. & PRÉVOT L. (2023). Transcribing and aligning conversational speech : A hybrid pipeline applied to french conversations. In *2023 IEEE Automatic Speech Recognition and Understanding*.

A Basic data set information

Age		28.7 ± 13.4 years
Gender	M	N = 56
	F	N = 146
	Other	N = 5
Country	France	N = 158
	Other	N = 49
Occupation	Student	N = 120
	Other	N = 87
Languages spoken		1.6 ± 0.9
Total		207

TABLE 2 – Participants metadata summary

Pilot	pilot	N = 22
	experiment	N = 261
Location	zoom	N = 35
	H2C2	N = 212
	Home	N = 18
	LPL	N = 18
Task	Reporting	N = 95
	Decision	N = 94
	Planning	N = 94
Scenario	A	N = 84
	B	N = 86
	C	N= 12
	D	N = 36
	E	N = 65
Video	yes	N = 241
	no	N = 42
duration		19min 42 ± 3min 8

TABLE 3 – Session metadata summary

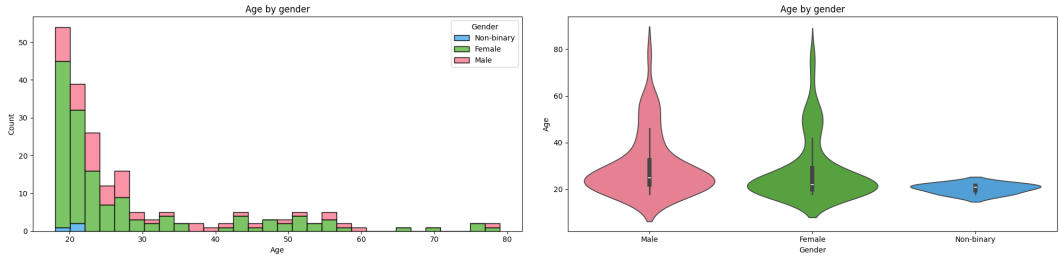
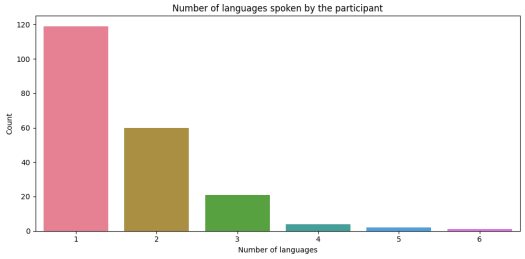


FIGURE 5 – distribution of age by gender



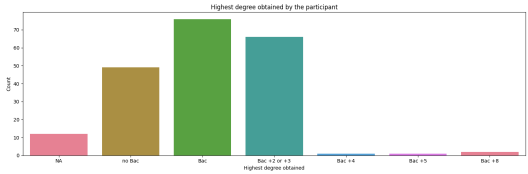


FIGURE 6 – Left : number of languages spoken by the participant, Right : Highest degree obtained by the participant

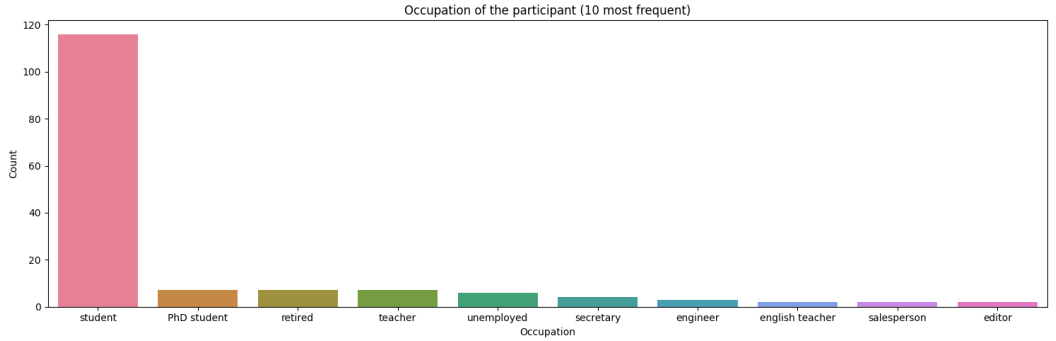


FIGURE 7 – Occupations of the participants

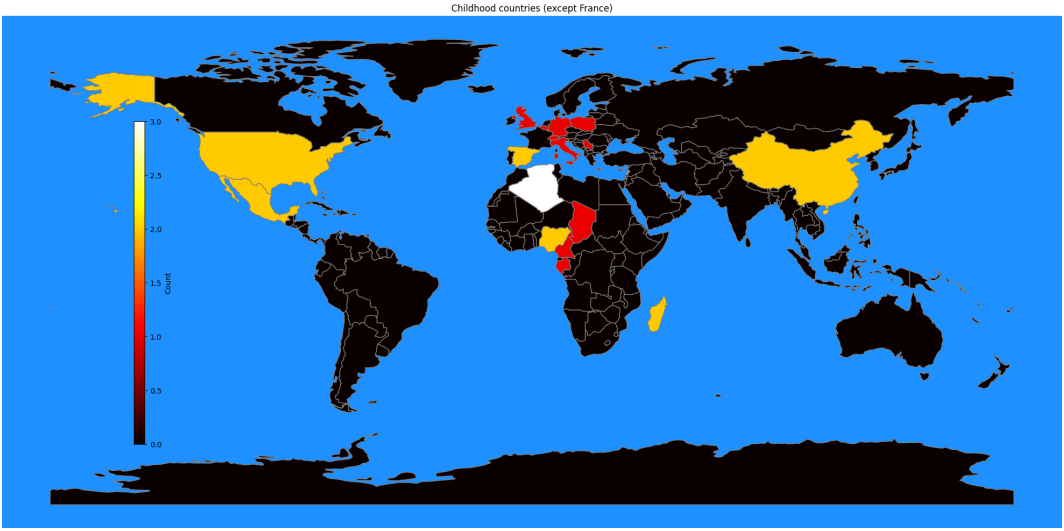


FIGURE 8 – Countries participants are from (apart from France)

B Additional linguistic statistics

B.1 IPU

B.1.1 Distribution

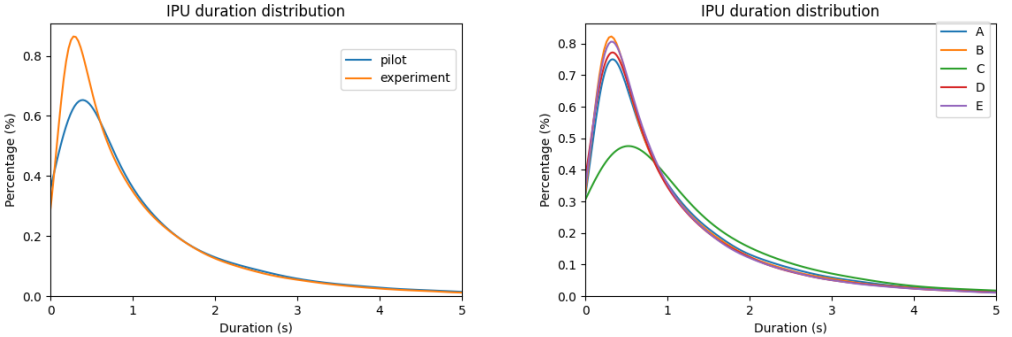


FIGURE 9 – Left : IPU duration distribution by pilot vs experiment, Right : IPU duration distribution by scenario

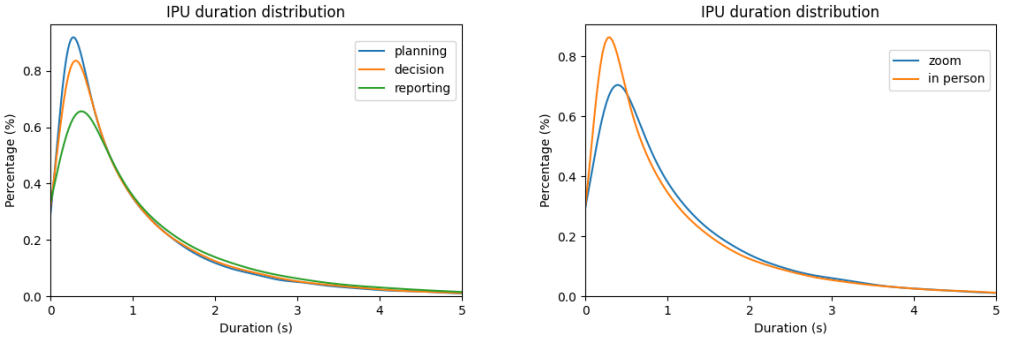


FIGURE 10 – Left : IPU duration distribution by task, Right : IPU duration distribution by place

B.1.2 Count

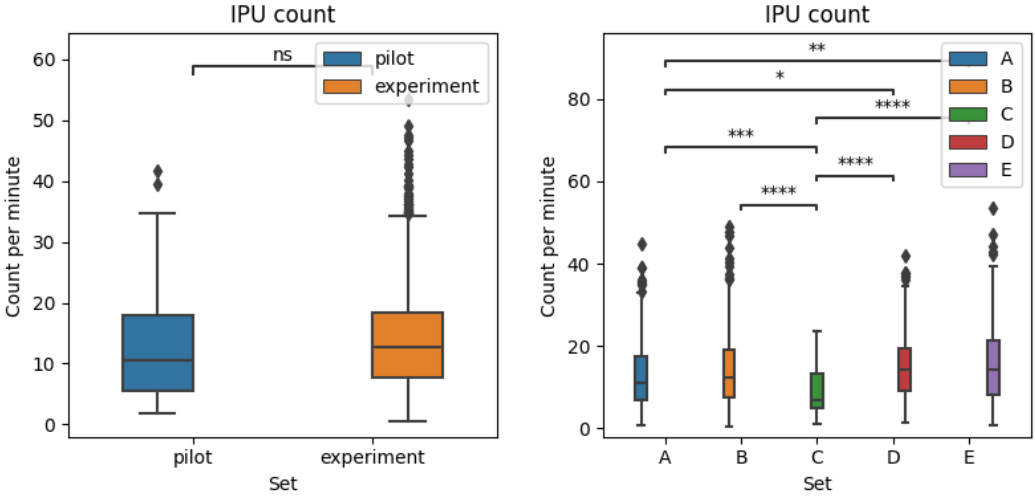


FIGURE 11 – Left : IPU count by pilot vs experiment, Right : IPU count by scenario

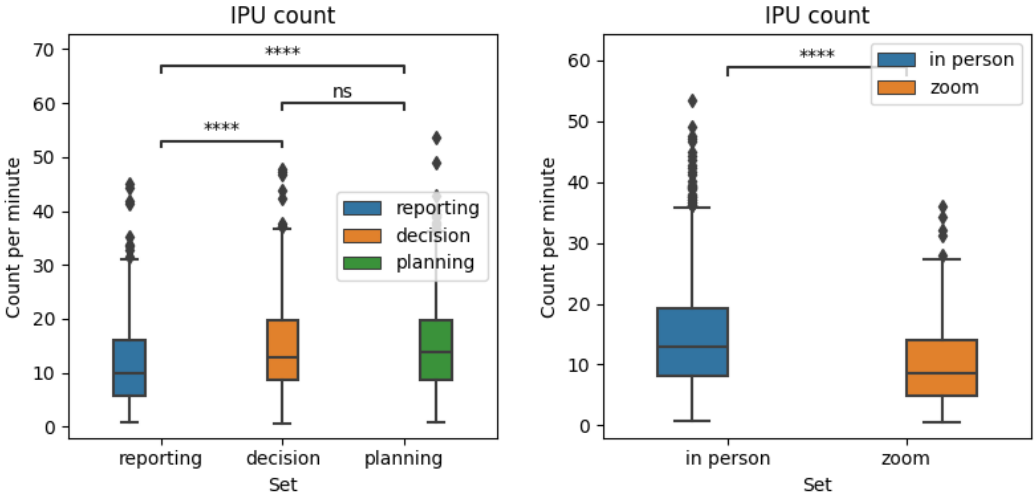


FIGURE 12 – Left : IPU count by task, Right : IPU count by place

B.2 Token

B.2.1 Distribution

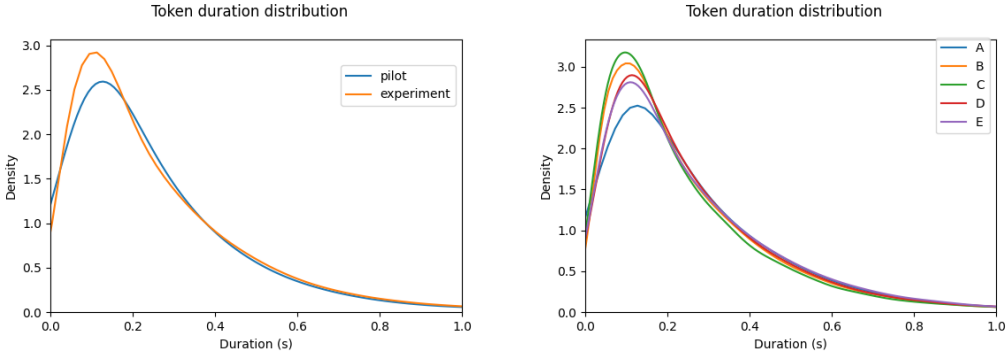


FIGURE 13 – Left : Token duration distribution by pilot vs experiment, Right : Token duration distribution by scenario

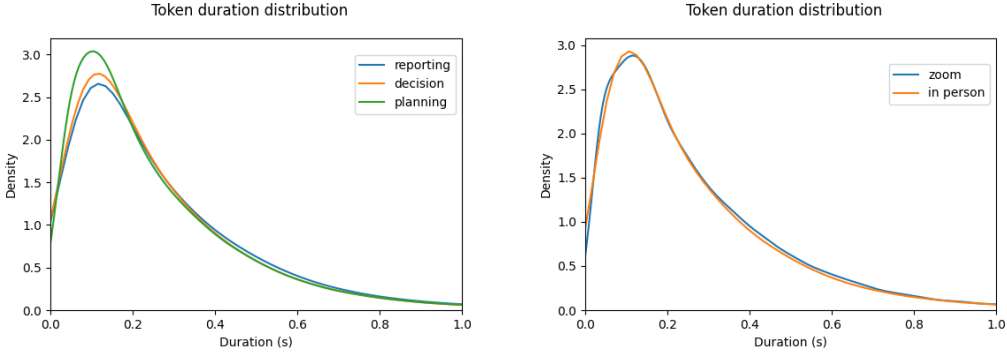


FIGURE 14 – Left : Token duration distribution by task, Right : Token duration distribution by place

B.2.2 Count

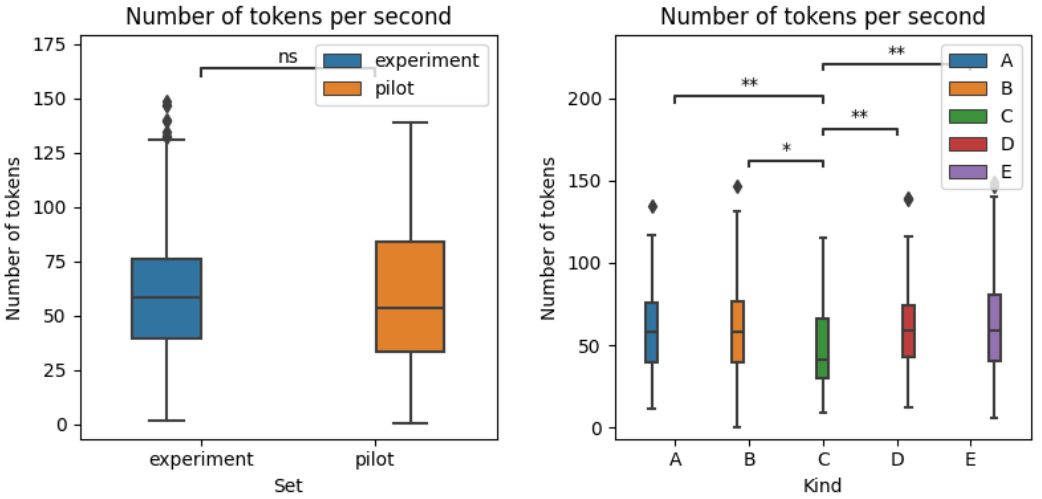


FIGURE 15 – Left : Token count by pilot vs experiment, Right : Token count by scenario

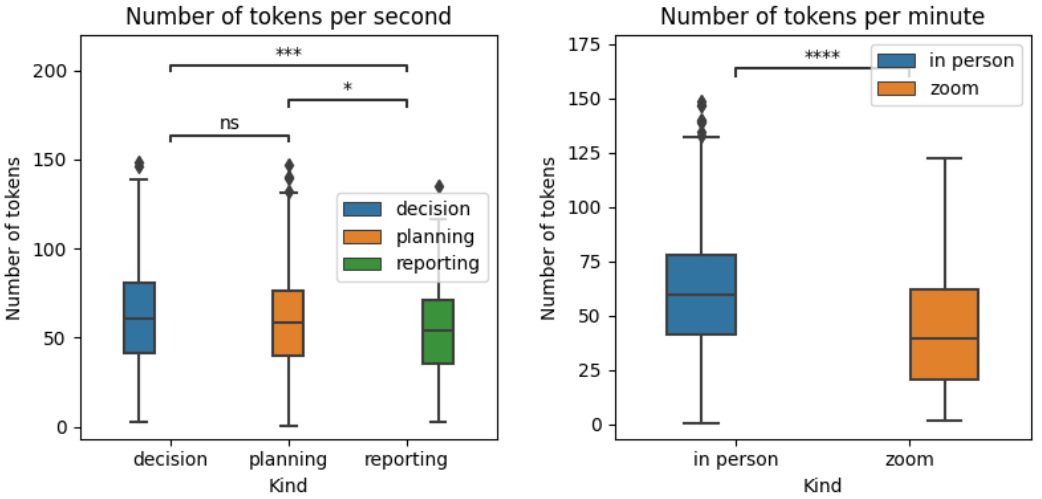


FIGURE 16 – Left : Token count by task, Right : Token count by place

B.3 Backchannel

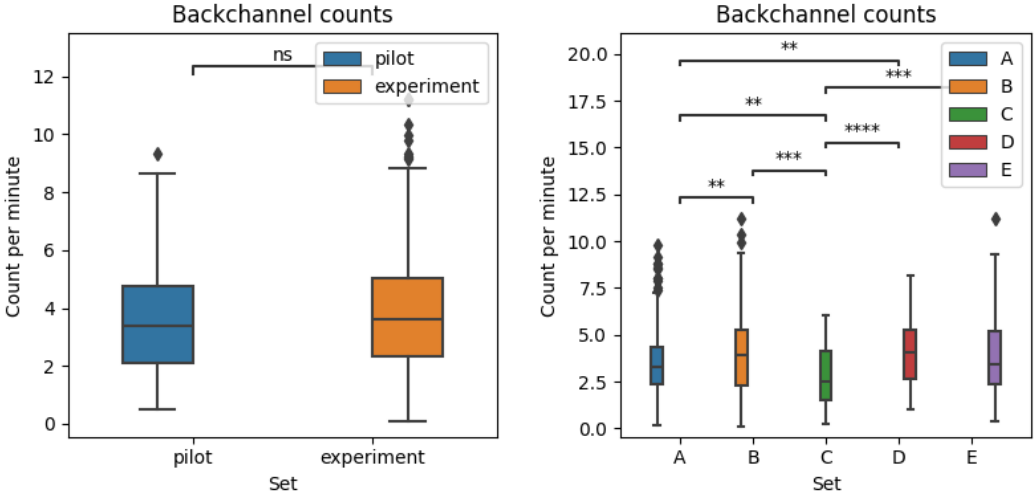


FIGURE 17 – Left : Backchannel count by pilot vs experiment, Right : Backchannel by scenario

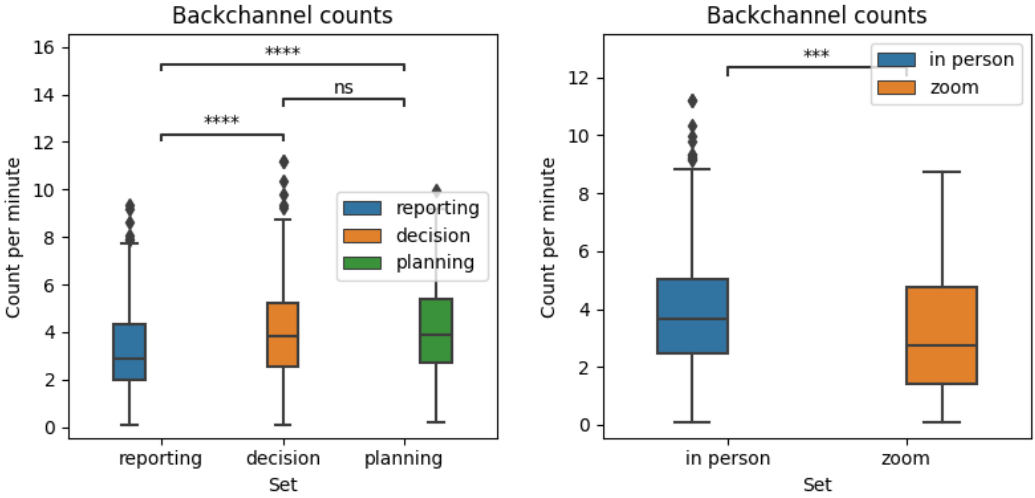


FIGURE 18 – Left : Backchannel count by task, Right : Backchannel by place

B.4 Overlap

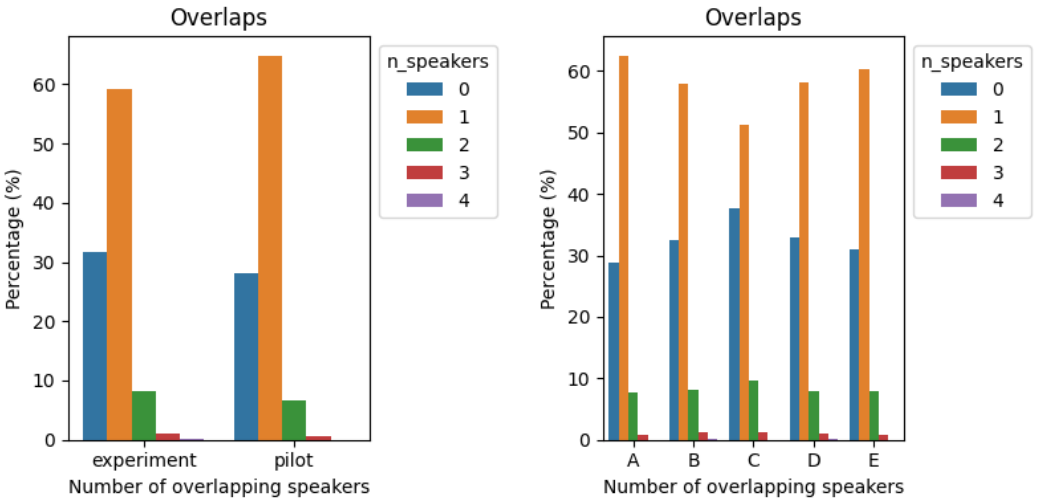


FIGURE 19 – Left : Overlaps by pilot vs experiment, Right : Overlaps by scenario

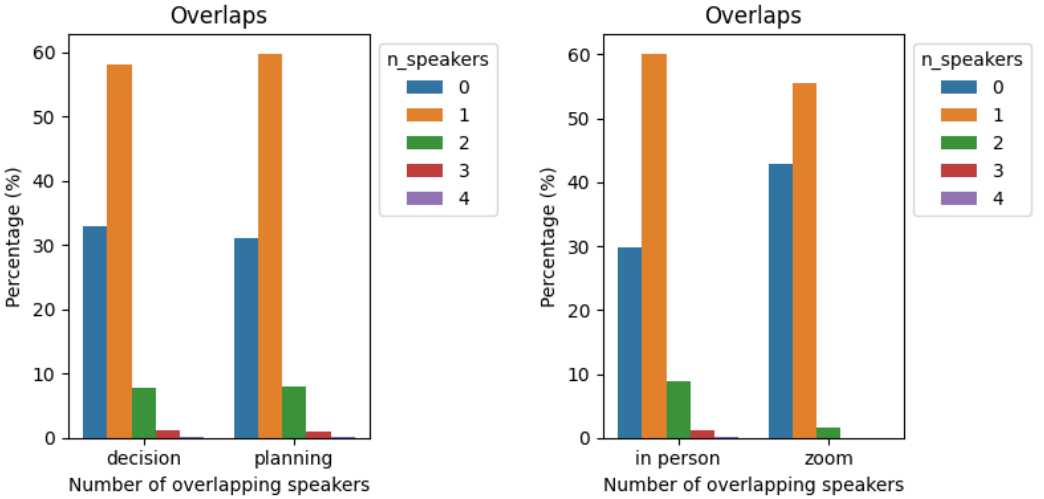


FIGURE 20 – Left : Overlaps by task, Right : Overlaps by place

B.5 Filled pause

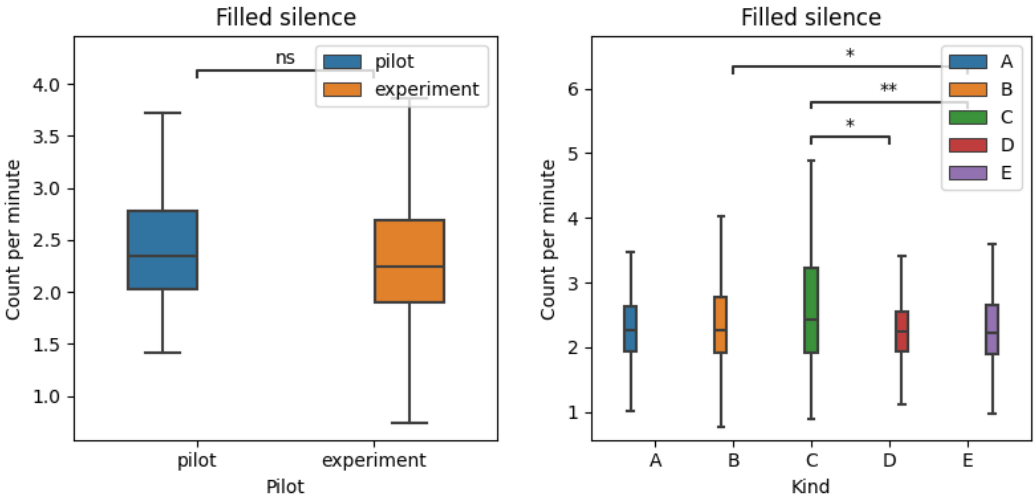


FIGURE 21 – Left : Filled Pause by pilot vs experiment, Right : Filled Pause by scenario

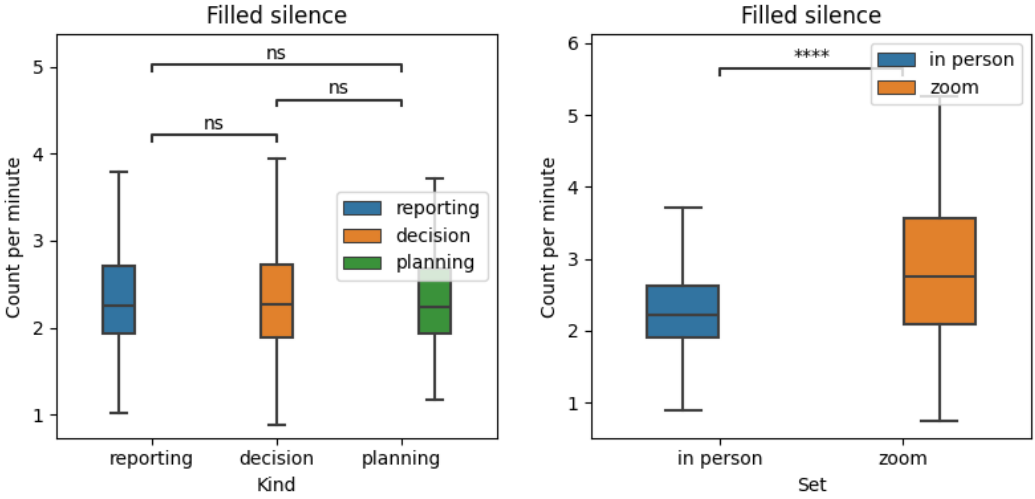


FIGURE 22 – Left : Filled Pause by task, Right : Filled Pause by place

B.6 Dominant speaker

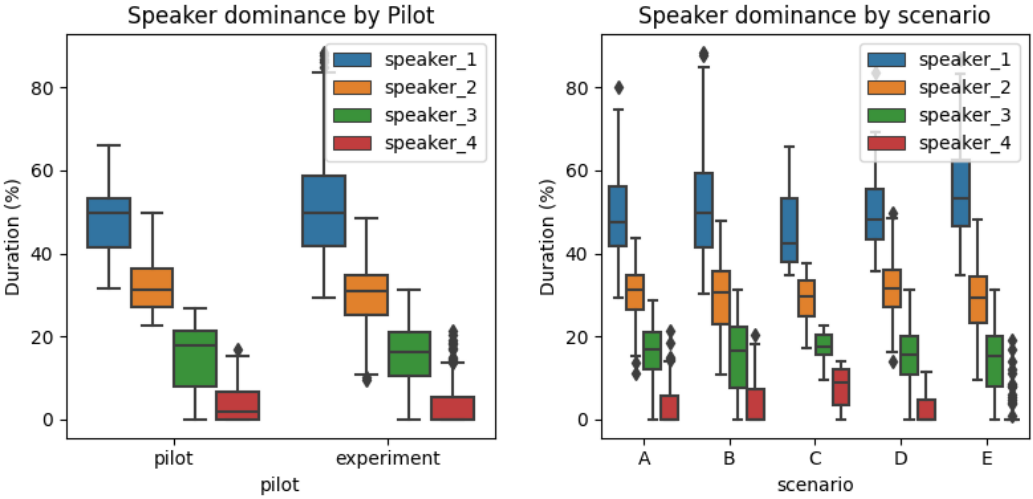


FIGURE 23 – Left : Speaker dominance by pilot vs experiment, Right : Speaker dominance by scenario

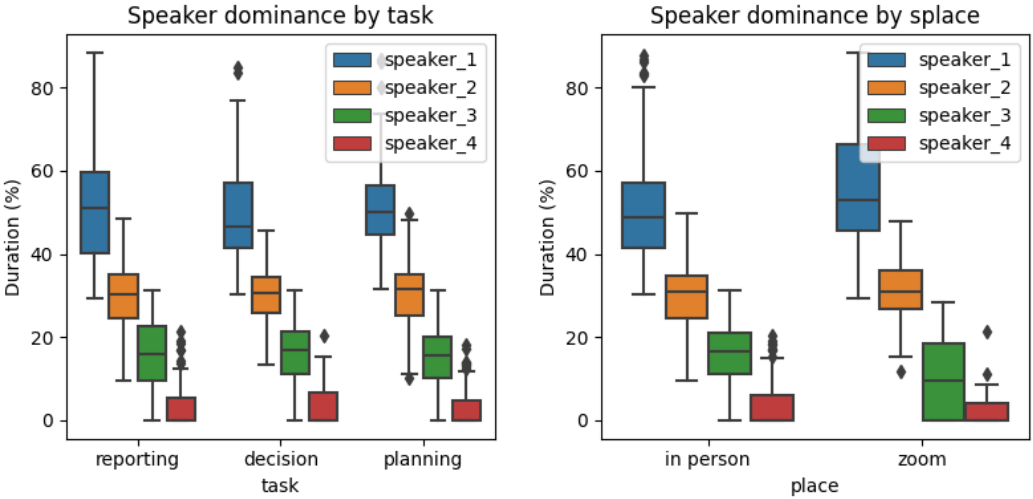


FIGURE 24 – Left : Speaker dominance by task, Right : Speaker dominance by place