Typist Experiment: an Investigation of Human-to-Human Dictation via Role-play to Inform Voice-based Text Authoring

CAN LIU*, School of Creative Media, City University of Hong Kong, Hong Kong SAR, China
SIYING HU†, School of Creative Media, City University of Hong Kong, Hong Kong SAR, China
LI FENG, School of Creative Media, City University of Hong Kong, Hong Kong SAR, China
MINGMING FAN, Computational Media and Arts Thrust, The Hong Kong University of Science and Technology (Guangzhou), China and Division of Integrative Systems, Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China

Voice dictation is increasingly used for text entry, especially in mobile scenarios. However, the speech-based experience gets disrupted when users must go back to a screen and keyboard to review and edit the text. While existing dictation systems focus on improving transcription and error correction, little is known about how to support speech input for the entire text creation process, including composition, reviewing and editing. We conducted an experiment in which ten pairs of participants took on the roles of authors and typists to work on a text authoring task. By analysing the natural language patterns of both authors and typists, we identified new challenges and opportunities for the design of future dictation interfaces, including the ambiguity of human dictation, the differences between audio-only and with screen, and various passive and active assistance that can potentially be provided by future systems.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: dictation, speech, text input, authoring, role-play, intelligent interface

ACM Reference Format:
Can Liu, Siying Hu, Li Feng, and Mingming Fan. 2022. Typist Experiment: an Investigation of Human-to-Human Dictation via Role-play to Inform Voice-based Text Authoring. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 338 (November 2022), 33 pages. https://doi.org/10.1145/3555758

1 INTRODUCTION

The great potential of voice interfaces has been widely recognized over decades, with research showing that text input via speech is much faster than typing [13, 23, 49]. Recent breakthroughs in Speech Recognition [20, 42] and Natural Language Processing (NLP) [9] have dramatically improved the ability of machine intelligence to understand speech [9, 20]. Nowadays, Speech recognition is available on most mobile devices, yet speech is far from being as widely accepted as typing

*Corresponding Author
†Student first author

Authors' addresses: Can Liu, canliu@cityu.edu.hk, School of Creative Media, City University of Hong Kong, Hong Kong SAR, China; Siying Hu, siyinghu-c@my.cityu.edu.hk, School of Creative Media, City University of Hong Kong, Hong Kong SAR, China; Li Feng, feliciafeng35@gmail.com, School of Creative Media, City University of Hong Kong, Hong Kong SAR, China; Mingming Fan, Computational Media and Arts Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China and Division of Integrative Systems, Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China, mingmingfan@ust.hk.

2573-0142/2022/11-ART338 $15.00
https://doi.org/10.1145/3555758

for text input and even the state of the art of voice-based text authoring still faces limited usage. Existing dictation systems focus on transcription and error correction [24, 37, 51]. There is a lack of support for the entire loop of interaction of text input, including text composition, review and editing [13, 15, 16]. With existing dictation systems, users start by generating text by speaking, but then typically have to fall back to the keyboard for reviewing and editing it [15]. This switch of modality interrupts the voice-based experience, leading to breakdowns especially in mobile scenarios when people are walking, cooking or driving [16].

To alleviate this problem, future dictation interfaces need to not only improve the accuracy of transcription, but also support a more well-rounded experience of text authoring via voice. While it is almost impossible to write an essay purely using voice today with existing dictation systems, people were writing books in the old time with professional typists. How can we design dictation systems that can support all the fundamental tasks in authoring a piece of text, including composing, reviewing and editing? To inform the design of such interfaces, it is important to have a holistic and in-depth understanding of users' natural speech patterns for doing that.

One commonly used method to study natural user behaviors for designing future systems is Wizard-of-Oz [43]. Wizard-of-Oz studies use human to simulate a system in order to evaluate the user experience before developing a functional system that is intelligent enough to understand free speech and distinguish different modes of operations. In this case the simulated machines are often constrained by only responding to a predefined set of commands and rules given to the participants, thus the results of the studies are limited by technical constraints and design choices made by experimenters. On the other hand, existing studies for understanding natural speech focus on improving speech recognition by addressing problems caused by disfluencies [26, 31] and how users address speech recognition errors by changing the way they speak [1, 49]. Therefore it remains unclear how users would naturally use dictation to author an entire piece of text.

To fill this gap, we adopt role-play as a new method for studying natural dictation. Previously role-play has been used as a design method for experiencing low-fidelity prototypes [47, 52] and a teaching method [34, 54] for engaging students. This work invents a new use for it: by learning from how humans naturally perform a task with the assistance of another human, we generate insights and inspirations for the design of natural user interfaces potentially powered by machine intelligence. We conducted an experiment in which participants dictated text to another human, akin to speaking to a typist in the old time. By analyzing the natural speech patterns of both the authors and the typists as well as how they interacted with each other, we aimed to derive insights and opportunities for designing future dictation interfaces for text authoring. In the experiment, 10 pairs of participants took the roles of Authors and Typists to perform tasks of composing and editing text until it is polished and ready for publication. Afterwards, Authors and Typists switched their roles for the second half of the experiment. In addition, as dictation interfaces are used in both eyes-free or with-screen situations, we tested both conditions in our experiment and investigated the differences.

The main contributions of this work are the empirical findings and design implications, generated from the quantitative and qualitative analysis of natural speech patterns in human-to-human dictation. Specifically, we extracted a comprehensive list of behavior patterns from the Authors and Typists. We found the Authors composed and edited text by instructing the Typists with the following behavior patterns: *creating new content, re-speaking, explicitly locating and editing text, reviewing text, delegating task* and *thinking aloud*. Meanwhile, the Typists provided the Authors with the following verbal assistance alongside typing: *passively responding to requests*, *actively correcting and preventing errors, proposing reviews or ideas*, and *taking over by making unsolicited edits*. Among these, we discovered how *re-speaking* was used for five different purposes implicitly by the Authors and two other purposes by the Typists. Moreover, we identified their deictic references

and communicating strategies used for locating text and resolving misunderstandings, as well as how the Authors and Typists synchronized and adapted to each other quickly throughout the experiment. In addition, we found that while seeing the text made it easier for Authors to review and edit efficiently, not seeing the text was sometimes preferred as it let their thoughts flow freely without distraction.

Beyond the particulars, we are the first to use a role-play method to understand how users could naturally interact with another human, who acts as a service provider comparable to an intelligent system, to inform the design of future interfaces. Despite the likely differences between human-human and human-machine interaction, our findings provides inspirations for system designers and engineers by unpacking new understandings of natural user behaviors, which can be potentially supported by future systems.

2 RELATED WORK

Our work was inspired and informed by related works in three areas: Voice-based Dictation for User Interfaces, Understanding Natural Speech Patterns, and the use of role-play method.

2.1 Voice-based Dictation User Interfaces

Automatic speech recognition (ASR) based dictation has been studied for different tasks (e.g., composing and transcribing) and for different target user groups (e.g., people with learning disabilities [11], blind users [3], or doctors and nurses [8, 38, 55]). Previous work found that blind users used speech input more often than sighted users on mobile devices [3]. Their study showed that blind users who perform text input with speech could do it 5 times faster than with a keyboard, but editing recognition errors was frustrating and it took 80% of the task completion time. Another research on the use of speech recognition software, which studied user groups with and without physical disabilities, found that on average, users could spend around 66% of their time on correcting dictation errors [45]. These findings show that dictation and correction are two main activities when using ASR-based dictation software, and that correcting dictation errors remains a key challenge. This inspired researchers to investigate users' expectations and strategies for correcting dictation errors. Basapur et al. studied users' expectations regarding dictation on mobile devices, finding that users would prefer to correct the errors by using voice commands instead of typing [50]. Sear et al. studied voice commands for navigating to the errors in the text and found that the direction-oriented navigation (e.g., move up two lines) was less effective than the target-oriented navigation (e.g., select target) [44].

Researchers also investigated various interaction techniques to help users correct their errors. In a desktop environment, Suhm et al. showed that multi-modal error correction that combines techniques of respeaking, handwriting, pen-based gestures, and keyboard input, is more efficient than uni-modal error correction [50]. For mobile devices, Kumar et al. designed Voice typing that uses a marking menu with touch gestures for faster error corrections [27]. Ghosh et al. designed EDITalk to support quick identification of sentence boundaries and speech commands [16]. Recently, Ghosh et al. also designed a technique called VoiceRev to support 2 common types of eye-free editing: commanding and re-dictation [18].

Although these interactive techniques enable users to better locate and correct diction problems, they were designed around the current ASR technique, which has little to no ability in understanding users' intents (e.g., composition, correction, or task-unrelated utterances). As ASR continues to improve, we envision that in the near future, it would be able to understand the nuances in users' dictation (e.g., intention, satisfaction) and even have human-like conversations to collaboratively resolve errors in dictation. Towards this future, we seek a more holistic understanding of natural

human dictation to inform future dictation interfaces. We achieve this by observing *how the Author who dictates, and the Typist who writes, can work collaboratively to compose and edit a given text.*

2.2 Understanding Natural Speech Patterns

Many previous studies have observed and analyzed how people speak, either to a speech recognition system or to other humans. Disfluencies, such as self-repair - people correcting their speech immediately after an erroneous phrase is spoken - have been studied extensively in these literature. By analyzing the characteristics of speech disfluencies, researchers have been trying to identify and correct them automatically. For example, Nakatani and Hirschberg [35] created a predictive model - the Repair Interval Model (RIM) that uses lexical, prosodic and acoustic cues - to detect and correct self-repair in spontaneous speech. This model has been commonly used in transcription systems [25]. Human-to-human speech has also been studied to facilitate speech recognition. For instance, Yang Liu et al. [31] studied the repetitions in human conversations, such as "On Monday I- On Monday I am going to...", and the use of fillers like "so", "anyway", "I mean". They proposed a computational method to extract such disfluencies in order to correct the speech recognition output to improve its readability. Research also found hyper articulation, where user tries to recover from speech recognition error by elongating utterances, pausing, or by altering the pitch. Furthermore, Large et al. [29] suggested that people speak to intelligent agents as if they were human. When tasked to speak to a simulated driving assistant, users were observed to be using polite words, deictic references, as well as giving vague instructions. However, these previous works focused on analyzing speech patterns to either improve speech recognition accuracy or to inform the design of a conversational agent. Their goals were not for generating texts.

In a previously mentioned work, VoiceRev, the authors conducted a Wizard-of-Oz study to observe users' eyes-free speech patterns while composing and editing text via a human-simulated voice interface [18]. Although some of the behaviors observed in their study also appeared in our work, their study focused on identifying editing strategies that can be immediately implemented, including the use of commands and respeaking. While Wizard-of-Oz studies can simulate some intelligent system behaviors, they are limited by predefined interface and functions, and their findings are subject to participants' assumptions and biases about what "an intelligent system" is capable of. Our role-play study complements that by unpacking a much richer and interactive process between humans while covering all phases of text authoring including composition, review and editing.

2.3 The use of role-play method

Role-play is a simulation technique to deliberately construct an experience under controlled conditions as designed by experimenters or therapists. It was used in social psychology experiments for studying group dynamics, attitude changes, and in clinical uses for therapies [34]. In HCI, role-play has been mainly used as a design method, particularly for early stage prototyping. Blinder [5] used role-play with low-fidelity prototypes to get users involved in the design of a PDA-based system. Vogiazou et al. [56] used this method to understand which features a new IT system needs by acting out their standard work processes. Simsarian described how role-play was used at the design company IDEO, where they let clients and end-users assume various roles in bodystorming activities [47]. Howard et al. [21] introduced professional actors in scenario-based participatory design to enhance immersion. Brandt and Grunnet [6] explored using elements in drama, such as settings, scenarios and props, in a collaborative user-centered design process. Svanaes et al. [52] developed a theoretical framework and a format of process for running workshops with role playing and low-fidelity prototyping, to allow simultaneous exploration of future technology use and design. Seland had system designers assessing the role-play method, who found it beneficial

for active participation of end users, faster ideation in the early stages of product design, and for enhancing developers' understanding of the context-of-use [46]. More recently, Buruk and Özcan added wearable devices in role-play games to facilitate movement-based play in game research [7]. Furturemore, role-play is also being used as a teaching method in system development courses. For instance, Moroz-Lapin and Maxim [33, 34] asked students to act as potential users of a system to enhance their understanding of the requirements as well as the use of the system. In addition, Stokoe used role-play for communication skills training across a number of workplace settings [48].

All the above mentioned uses of role-play methods focused on engaging users or stakeholders in early stages of participatory design processes in order to create immersion or empathy. In this work, we create a novel use of a different role-play method for designing future systems, with a particular focus on observing natural human behaviors in interacting with an "intelligent party", which in our case is another human who is not hiding behind the "wizard's" curtain. While we acknowledge the differences between human-to-human interactions and human-computer interactions, this approach aims to unpack unknown natural behaviors of users and to provide new inspirations for future intelligent systems.

3 THE TYPIST EXPERIMENT

The major advantages of dictation include the high speed of text input and the possibility to perform dictation eyes-free and hands-free. This makes its use scenarios on mobile devices very compelling, where staring at a screen is inconvenient and typing on a keyboard is relatively slow. In such scenarios, text of different lengths need to be created, from short to long messages, emails, memos, diaries, blogs, etc. In this work, we seek to understand how users might use natural dictation to compose and edit whole paragraphs of text. We designed a role-play experiment – the Typist Experiment, to achieve this by observing natural human-to-human interaction, in which one participant (i.e., the Typist) provides a "smart dictation service" by typing what the other participant (i.e., the Author) dictates to him/her.

While speech is the main modality of input for dictation, users rely on visual or auditory feedback to understand how the task is being executed. Therefore we believe one important factor that would affect users' behavior and experience is the modality of feedback. In fact, using dictation software eyes-free is compelling for many use cases especially in the mobile scenarios, such as while walking, driving or doing other tasks. In order to understand the differences between eyes-free and seeing the text, we make modality as a main independent factor in this experiment. With it, we sought to answer the following research questions (RQs):

- **RQ1**: How do Authors dictate to Typists?
- **RQ2**: How do Typists assist Authors?
- **RQ3**: How do Authors and Typists coordinate and collaborate?
- **RQ4**: How does the communication modality (*Audio only* and *Audio+Screen*) affect Authors, Typists and their cooperation?

3.1 Experiment Design

The experiment followed a within-subject design featuring one main factor with two independent variables: Communication Modality [*Audio only*, *Audio+Screen*]. Pairs of participants were recruited and asked to co-create written text on given topics, each taking the role of an Author and a Typist respectively. In *Audio+Screen* condition, the Typist shared his/her screen showing the editor interface where text was being typed in and the real-time word count was shown. Whereas in *Audio only* condition the screen was not shared, and the Author could not see the text. Cameras were all off, to keep speech as the only communication modality besides the editor interface in

Fig. 1. Example images provided for participants for text composition, from the Dixit game.

Audio+Screen condition. Each participant took the role of Author or Typist to begin with, and then switched roles with their partner. Thus half of the participants took the role of Author first and the other half were Typist first. We expected that the experience of being Author and Typist may influence their behaviors in their second role. This was an intentional choice to collect more data and potentially observe richer behavior patterns in their collaboration strategies.

As an Author, the participant was asked to compose a piece of text by describing a given image. The image (Fig. 1) is randomly chosen from an image set of the Dixit game[1], which is an ambiguity board game where every image card is an abstract illustration designed for having multiple interpretations. In our experiment, Dixit cards were used as the source of inspiration or description basis in authoring tasks. This choice was to ensure certain control on the task difficulty without limiting the creative space or introducing biases by particular topics. As the experiment focused on observing the interaction patterns between Authors and Typists, the actual content they composed was left to their choice to ensure they felt comfortable with the task. The Author was asked to create a text of 70-80 words by speaking to the Typist and instructing him/her to modify the content as the Author wanted. We chose a paragraph-length text generation task to cover the writing situations in most writing tasks on mobile, such as for messages, emails and memos, without making the experiment too long. A task was finished when the text reached the requested range of length, free of mistakes and the Author was satisfied with it so as to be willing to publish it on social media or send it to friends.

As a Typist, the participant needed to type in what the Author composed and edit it according to the requests and instructions from the Author. The Typist cannot see the image given to the Author. Both participants were told to communicate freely while trying to complete the task as fast as possible.

3.2 Apparatus

The experiment was conducted online, with the Author, the Typist and the experimenter in the same virtual meeting session. Zoom[2] or Skype[3] or TencentMeeting[4] was used as the meeting platform, considering the participants' preferences and their regional network situations. Standard Google Doc[5] or Tencent Doc[6] was used as the editor, and the real-time word count view was

[1]https://www.libellud.com/
[2]https://zoom.us/
[3]https://www.skype.com/en/
[4]https://www.tencent.com/en-us/responsibility/combat-covid-19-tencent-meeting.html
[5]https://support.google.com/a/users/answer/9282664?hl=en
[6]https://docs.qq.com/

configured to be visible in all conditions. Participants were asked to disable their camera for reasons introduced above. Video and audio were recorded using the experimental platform built-in software function (e.g., Zoom and Skype recording) combined with screen recording software.

3.3 Procedure

Before starting the experiment, the experimenter introduced the study, collected the informed consent and sent designated images to the Author. The roles of the Author and Typist were introduced at the beginning of the experiment, as the Author being the writer and the Typist being the service provider. 24 images were provided to each Author (three per trial for them to choose one) during the entire experiment. Each set of images for a group were randomly selected from a fixed set of 84 Dixit cards. We chose to let participants type in words instead of using speech to text on Google Doc. This was to avoid introducing the voice typing interface as a third player in the interaction, which would distract the interaction between the two parties with a cooperative error correction task caused by the limitation of current STT products. The experimenter checked the function settings of the online typewriting platform used by the Typist before the test started. The document platform turned off all notification-related functions, such as grammatical error display and spelling suggestion function, to provide participants with pure documentation tools. The purpose was to reduce the influence of machine assistance and visual reminders' interference, which helped to observe and record the participants' natural voice communication behaviour in the experiment. The participants began with a training session before starting each condition to familiarize their role and the task condition. For each trial, the experimenter provided three different images for the Author to choose one for composing text. They were told to freely describe the image or talk about experiences or opinions inspired by the image. Images were randomly chosen from the set, and we ensured that different images were provided for each trial of one experiment.

Each participant was assigned a role as the Author or the Typist to begin with. Each measured trial required the completion of composing and editing one piece of text. Two repetition trials per Communication Modality condition were performed in the first round. Then the participants switched their roles and performed another two trials per Communication Modality condition. The order of Communication Modality was counterbalanced across participants. The participants began with a training session before starting each condition to become familiar with their role and the task condition. For each trial, the experimenter provided three different images for the Author to choose one for composing text. They were told to freely describe the image or talk about experiences or opinions inspired by the image. Each participant completed four trials as Author and four other trials as Typist in total. The participants were strangers to each other, and were randomly paired. There was no restriction on communication, but participants were asked to complete the task as fast as possible while ensuring a sufficient quality.

The experiment in total collected 2 Communication Modality $\times$ 2 roles $\times$ 2 repetition $\times$ 10 groups = 80 measured trials. After finishing all the trials, we conducted a semi-structured interview with each group. The experiment took around two hours for each group, with a break in the middle.

3.4 Participants

We recruited 20 participants (8 females and 12 males) from local universities to form 10 groups (Table 1). None of them had professional typewriting or transcription training. To reduce biases of personal relationships, we paired the participants so that each pair did not know each other. The tasks were performed in English. The participants included three groups of native English speakers and seven groups non-native English speakers. All the non-native English speakers studied in English-taught university programs and a score of at least 6.5 in IELTS tests. They

Participant	Group	Gender	Education/ Profession	Native Speaker
P1	G1	F	Postgraduate	No
P2		M	Postgraduate	No
P3	G2	M	Postgraduate	No
P4		M	Postgraduate	No
P5	G3	M	College Lecturer	Yes
P6		F	College Lecturer	Yes
P7	G4	F	Undergraduate	No
P8		F	Undergraduate	No
P9	G5	F	Postgraduate	No
P10		M	Postgraduate	No
P11	G6	F	Postgraduate	No
P12		M	Postgraduate	No
P13	G7	M	Undergraduate	No
P14		F	Undergraduate	No
P15	G8	M	Undergraduate	No
P16		M	Undergraduate	No
P17	G9	F	Postgraduate	Yes
P18		M	Postgraduate	Yes
P19	G10	M	Postgraduate	Yes
P20		M	Postgraduate	Yes

Table 1. Summary of participants' demographic information.

had diverse backgrounds and education levels, holding bachelor, master and doctoral degrees in communication, arts, finance, law, chemistry and biomedical engineering. For typewriting the content of the composition task, eight groups used Google Doc, and two groups used Tencent Doc. Although some of our participants had typewriting capacity issues, we focused on the nature of the speech patterns between the Typist and the Author in dictation-based text authoring practice. We discussed this point about participant's ability in the limitation part.

3.5 Data Collection and Analysis

The following data was collected: 1) screen and audio recordings of the online experiment sessions; 2) audio recordings of the participant interviews; 3) observational notes from the experimenter. For the experiment sessions, we transcribed all the conversations between Authors and Typists into text and performed a Thematic Analysis on the utterances. The analysis was done with the aid of the audio and screen recordings as references. Two researchers independently coded four randomly selected trials of data from two groups, which covered 5% of the entire data set. They discussed their codes to gain a consensus. After rounds of discussion, their codes reached a substantial level of inter-rater reliability (Cohen's kappa: k = 0.78). Fig 2 shows a few examples of how the codes were applied. After that, one of them continued to code the rest of the data. All interview sessions were transcribed into text manually and coded by two researchers after reaching consensus about the themes. During the coding process, the observational notes were also analyzed.

4 FINDINGS

This large section elaborates in detail the behavior patterns we identified from the utterances of both the Authors and Typists, as well as their communication and cooperation strategies. While some examples are given in the main text in prose, we supplement in the Appendix A our coding scheme of all transcribed utterances, together with their descriptions and examples. Each utterance can be coded with one or more codes in the coding scheme. The rest of this section label our participant IDs in this format: [Group Number]-[Participants ID]-A/T(Author/Typist), i.e., "G10P20-A". We add AO (Audio Only) and AS (Audio+Screen) when comparing modalities.

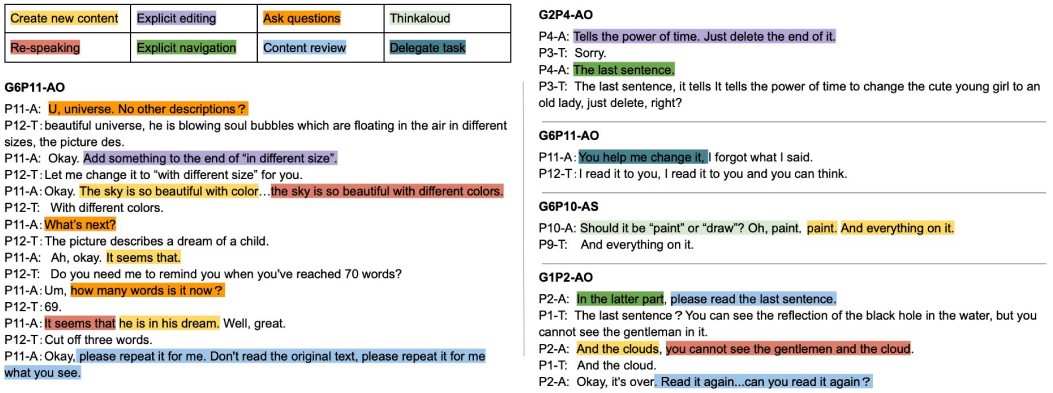

Fig. 2. Examples of how Authors' utterances are coded with the 8 categories.

4.1 How The Authors dictated to The Typists? (RQ1)

The analysis of Authors' utterances focus on extracting behavior patterns of the Authors, including how they composed and edited the text by giving implicit and explicit instructions, how they switched between the different types of instructions as well as what other requests they made. The findings presented in this section are based on our categorization of all the utterances from the authors. Fig. 2 shows examples of how these categories are coded in Authors' utterances. We calculated the occurrences of each category and normalized the numbers into percentages by dividing them with the sum of coded Author utterances for each group, which gives us an estimation of frequency of each behavior pattern. Seven categories emerged in total: Create new content (44.22%), Re-speaking (37.45%), Explicit editing (6.06%), Explicit navigation (2.15%), Content review (1.76%), Ask questions (4.52%), Delegate task (0.64%) and Think aloud (3.20%). The numbers here are average percentages among 20 Authors, and indicate that the primary requests made by Authors were Creating new content and Re-speaking.

4.1.1 *Ambiguity between composition and edits.*

Creating new content. We use New Content as the code for composition of new text. It is coded in chunks of words within utterances. One utterance often includes both newly composed words and previously composed words, which are repeated with or without modifications.

Re-speaking. Re-speaking is coded as Authors repeating chunks of content after it being composed. Existing literature studied speech repair [10], which we see as one type of re-speaking behavior, typically observed as a sort of stuttering. In our study, we identified five types of re-speaking behaviors from the Authors, depending on their different intent.

- Overwrite to modify: re-speaking the text to overwrite the different part of it, as an implicit request to modify the text. Example: *"G5P9-A: The **haircut**...oh...the **hair style**."* and Example 7 in Appendix A shows the longer overwrite.
- Confirm or repeat for Typist: re-speaking to make sure the Typist clearly understood and took down what they composed. Example: *"G5P9-A: En...sorry. When we were little child. When we were little child."*
- Continue composition: re-speaking a few words as a way to continue composing after it, so that the Typist knows where to continue typing. Example: *"G8P16-A: **Everyone**, no, no, a new sentence, **everyone** was telling her."*

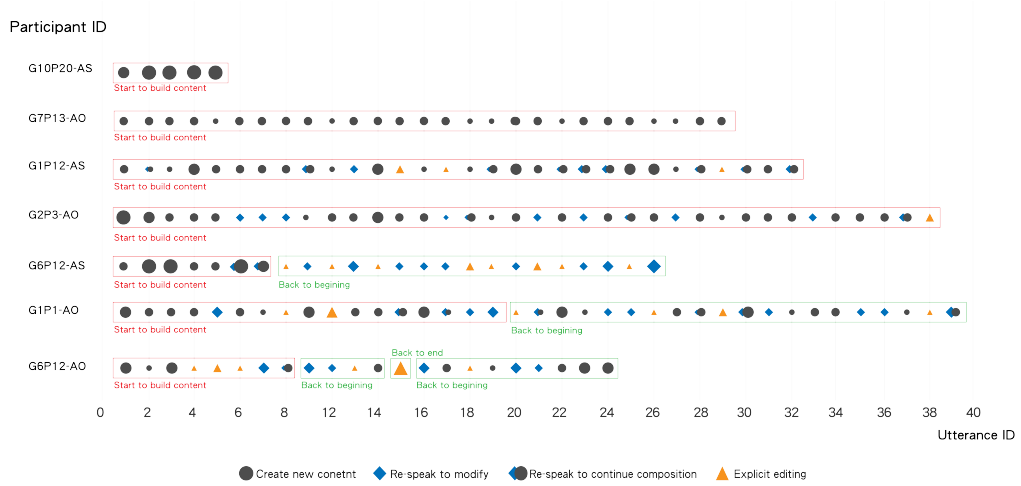

Fig. 3. Example visualizations of how text got composed and edited over the timeline of a trial. Each symbol represents the content development in *one* utterance: *Create new content, Re-speak to modify, Re-speak to continue composition* and *Explicit editing*. AO (*Audio only*) and AS (*Audio+Screen*) are communication modalities between Authors and Typists. Four sizes of the symbols represent the unit size of text operated in the utterance: *word, phrase, clause* and *multiple clauses*. A red contour contains the first composition pass in the trial and a green contour contains one revision pass.

- Locate / refer to text position: calling keywords as a deictic reference to communicate a text position, for making a request then. Example: *"G6P11-A: Front, back to front, back to top.* ***'boy'*** *is it? a* ***'boy'*** *walking on the grass."*
- Natural speech repair: repeating words in subconscious stuttering for repairing one's own speech. Example: *"G1P2-A: Is* ***the, the*** *paper man becomes loosening."*

The average percentages of each type of Re-speaking in all the Re-speaking utterances are: Confirm/repeat for Typist (42.80%), Overwrite to modify (22.25%), Continue composition/ Anchor new content (20.05%), Natural speech repair (6.96%), Locate/ Refer to previous content (5.44%).

Explicit editing. The previous section introduced Re-speaking as a major way of making implicit editing requests. The other way Authors made editing requests to the Typist was to be explicit, closer to what existing dictation software could support, such as the Google Docs Voice Typing[7]. We observed six types of requests: Add, Delete, Replace, Format, Organize and Punctuate. Different from the voice commands in fixed syntax supported by existing systems, the Authors' requests were made in free speech with various syntax. For instance *"G5P10-A: success, no no no no no no"* was used to express a "Delete" request. Organizing content can be said in many ways, such as, *"G8P16-A: Put all the previous words in a quotation mark. There is one more sentence after the quotation mark. This is Richard."*

4.1.2 Mixed composition and edits. Based on the transcriptions, we analyzed how the text evolved over the timeline of each trial. We observed three types of composition strategies. A few examples are visualized in Fig. 3 to show this process. There are only four operations that affect the text: Create new content, Re-speak to modify, Re-speak to continue composition and Explicit edit. The first strategy was to focus on generating new content first and then go back to the beginning to

[7]https://support.google.com/docs/answer/4492226?hl=en

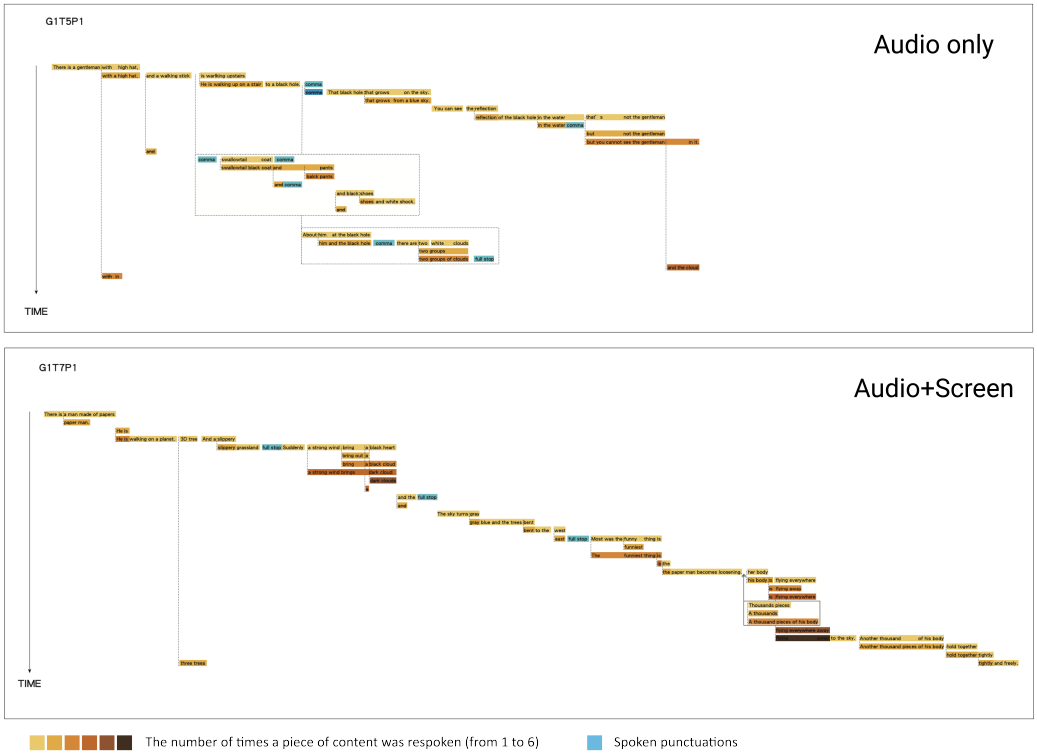

Fig. 4. Example visualizations of the content development process. Overwritten text by re-speaking are aligned vertically across lines. The timeline of text generation runs from left to right and then from the top down. Each coloured text block is generated from one utterance. Darker orange color indicates the same content block being re-spoken more times. Blue blocks are spoken punctuations. Edits by explicit requests are annotated with crossovers or insertion marks.

edit, as seen in a typical example G1P2-AO. In this case, the tasks were finished after two or more passes (see example G1P1-AO, G6P12-AS). The second strategy was to edit as they compose, the editing instructions were a mixture of explicit and implicit (via Re-speaking) requests. The task was finished with one pass. The third strategy happened rarely with Authors who were able to organize their thoughts and articulate in a smooth sequence in one go without needing to edit (see example G7P13-AO, G10P20-AS). Every above-mentioned strategy appeared in both *Audio only* and *Audio+Screen* conditions. These strategies also get mixed up often, leading to a process with less order (example G6P12-AO). Random jumps of editing locations could cause the Typist to get lost, or needed more synchronization effort.

With a more elaborated visualization, Fig. 4 illustrates in detail how text evolved in time. In this figure we ignored other types of utterances from the transcription and focused on *creating new content, re-speaking and explicit editing* as they were the only ones modifying the text. We can see each chunk of text the Author generated or modified by each utterance and over time. Overall we can see many edits done by respeaking, there is no clear mode distinction between composition and editing in nature dictation, nor an easy way to make it. This example is from G1P1 in *Audio only* and *Audio+Screen* condition in comparison. This Author did extensive editing in

both trials. However the edits were more in order in *Audio+Screen* condition, compared to in *Audio only* condition there were larger chunks of text being inserted and modified at multiple locations. We provide this visualization for all the trials in the supplementary material of this paper as a data set. From this data set we can see most edits were done in one or multiple passes, with some jumps between editing locations, which happened in both *Audio only* and *Audio+Screen* conditions. Perhaps this is due to our limited working memory [4], in *Audio only* condition, the edits tended to be more extensive after jumping locations. Further research is needed to investigate this.

4.1.3 Other requests and behaviors of Authors.

Explicit navigation. We introduced in the previous section that Authors navigated the text by simply re-speaking a few keywords. In contrast with that implicit navigation, there were also explicit instructions for navigation. For instance, Authors would say: *"first line, at the end (G6P11-A)"*, or *"The last sentence, after the last sentence (G2P4-A)"* or *"after 'black hole' (G1P2-A)"*. More about navigation can be found in Section 5.1.1

Content review. Content review requests were issued when Authors needed to know what had been written. We also noted a small amount of content review requests were asking the Typists to summarize the written content instead of reading it. For instance, *"G6P11-A: Tell me what you understood from the writing, don't just read the sentences."*

Ask questions. The Authors occasionally asked questions to their Typist. The types of questions include: asking how to spell or translate a word; asking for suggestion of words; checking whether the Typist finished typing; checking word counts. For example, an Author tracking the typing progress would ask: *"G4P8-A: are you finished?"* or *"G5P9-A: So can we move on?"*. We observed that some Authors ask for suggestions. For instance, G4P8-A: *" Do we need to add something like 'Anyway, I'm very so mad' in the end?."*

Delegate task. Very occasionally, some Authors asked the Typist to help them with the task when they faced difficulties. They asked the Typist to make the sentence better or compose something new. For instance, *"G6P11-A: Please help me to change it."* or *"G4P7-A: Can you help me think about what else to say?"*.

Think aloud. Think aloud has been observed as a natural behavior. For instance Authors spoke their mind when they were unsure about the use of a word, or how to express an idea, or indicating they were taking time to think. This behaviour was also observed in previous Wizard-of-Oz studies [19].

4.1.4 Authors' dictation strategies.

The participants were asked to describe their dictation strategies in the interview. The following findings describe the conscious efforts they made when being an Author.

Consciously reducing uncertainty and repetition. One participant reported she was automatically spelling the words that may introduce confusion or need to double check, *"G1P1-A: When I was the Typist, there were words I wasn't sure about. Then when I became the Author, I would actively spell those words, to avoid repeating it"*. She also tried to tell the Typist exactly where she was uncertain, to avoid having to reading back the whole text: *"When I wasn't sure about something, I would pick a sentence for him to read, wouldn't review the whole thing, which is too time-consuming."*

Like telling stories to children. One participant said he was doing the Author's job pretending he was telling stories to children: making up things quickly and keeping things simple. G3P5-A echo that *"I spoke more slowly than normal. Normally, I speak much faster. I think my strategy was I was*

imagining that I was telling a story to my niece or nephew. So two little children. So that was my main strategy. Yeah, story, telling two little children. Well, because when you're tying stories to children, you don't have time to edit. And you have to come up with everything very quickly. So you have to keep everything kind of simple. And they always end up asking questions, and you always have to think of something right away. So that's why I think my strategy was pretending like, Yeah, one of like, a little child was asking me to tell a story."

Formulating thoughts. One participant reported he would formulate his thoughts to make it clear, consistent and avoid major changes. *"G10P20-A: When I was authoring, I was a lot more conscious about what I was going to say. So formulating my thoughts a lot more as opposed to when I'm writing for myself, (...). Because I'm conscious that I need to transfer his information to him. So I want to make sure that it's clear on the one hand. On the other hand, I thought that the exercise was such that you have to put things on paper that are very consistent. And I want to prevent anything like major changes that we need to make to it in order to every time efficient so to speak."*

Making orderly speech. Some Authors explained how they managed to speak in an orderly manner. One Author who composed in one go without much editing explained that was because he was highly trained in writing and had a very good verbatim memory. G3P6-A: *"I think I come from a place of being highly trained in writing (...) I focus more on keeping everything in memory, sort of remembering everything I had said and having a clear picture of everything that I did said so, so that then I didn't need a lot of addition (...). So I basically remembered everything I said."* Another participant intentionally slowed down the pace of composition to leave pauses for mentally structuring the whole text. As mentioned by G9P18-A *"And my main strategy for composing was to take pauses between sentences to think. So, it was quite slow. Building up the story in piece by piece slowly was my strategy."*

4.1.5 Authors' challenges and their wishlists for Typists.

Feeling rushed and disorganized. Participants mentioned the difficulty of composing text with voice was the stress and lack of structure. G10P19-A:*" I think being an Author just speaking the story I felt a bit rushed, I guess. Like how it's kind of not panicking, but like trying to think that way the story changed because I was kind of maybe like blurting things out without thinking so thoroughly."* Another participant articulated the challenge in tracking down the non-linear thinking that comes with free speech. G1P2-A: *"You have lots of free space. Maybe you finish one sentence and suddenly thought about another point, but you are still speaking about this point, I wish the Typist could understand. Maybe you have to let him choose one point or note down both and then ask you to confirm your choice. Because you are quite free when you speak. The most difficult thing, is you think about many things and just pour them out randomly."*

Giving up control. It was also reported that giving up control to another person was a challenge. G3P5-A:*" So the main challenge, I think, is having to give up some control. Which I think is fine. It was fun. But you know, it's like the, you know, you get we get so used to typing our own words. "*

Communication, rhythm and punctuation. Further challenges reported of being an Author doing text authoring via a Typist echoed our observations and other parts of the finding, included communication problems caused by homonyms, the need to keep stopping and repeating their speech, how to control the speed of speech, how to express the segmentation of sentences, and how to make Typists understand better the meaning of the sentence. As G1P2-A echoed *"The main problem is communication obstacles. It is possible that I am giving some piecemeal information. Then there may be problems with his record. He doesn't know what to record, and may need further communication."* The same challenge also reported from another participant *"G1P1-A: It will cause*

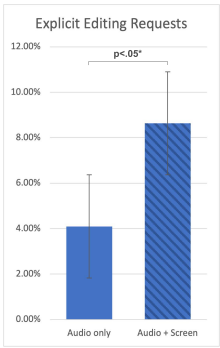

Fig. 5. Percentages of occurrences of Explicit Editing Requests in *Audio only* and *Audio+Screen*.

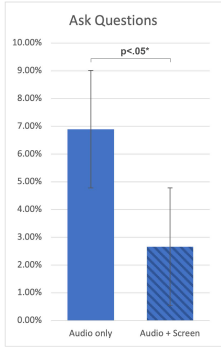

Fig. 6. Percentages of occurrences of Ask Questions in *Audio only* and *Audio+Screen*.

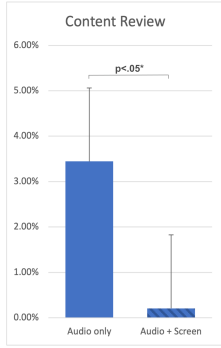

Fig. 7. Percentages of occurrences of Content Review in *Audio only* and *Audio+Screen*.

the lack of context." and *"G6P12-A: He didn't let me check in the end. I think he wrote a lot of error."* And G5P9-A said *"Because I prefer to have frequent sentence breaks in the process of speaking a sentence. I will say a long sentence, and there may be many sentence breaks in it. Maybe this is a problem of my personal language habits. But the other party may just break it, and it may be a little different from what I said."*

Authors' Wish-list. We asked the Authors what they wished the Typists could have done. Most participants mentioned they hoped their Typists help them check and fix the grammar and typos. Three groups mentioned rather than recording silently, they hoped the Typists could continuously give feedback while they compose. As G5P9-A mentioned *"Besides, I don't know if he can keep up with me, I don't know where it stops, and then I just stop for a while. I hope he will give me feedback after he finishes typing."* More requests on the wish-list include: Typists marking their mistakes but not fixing them; checking the format and structure of their text; learning their preferences and habits in order to help them fill missing words. In addition, an Author preferred the Typists to not modify any of their dictation but mark them out. G5P10-A explained that *"For example, punctuation or grammar is to make this paragraph of my content more complete, but changing words is to modify the description and change my original intention. This may not be acceptable to me. I can accept that you help me fixing grammar and punctuation, but only if my original intention is accurately expressed and paraphrased."*

4.1.6 Effects of modality for Authors (RQ4). To investigate the impact of Communication Modality on Authors' operations, we ran a T-test on each category of Author utterances between the *Audio only* (AO) and *Audio+Screen* (AS) condition. To remind, the categories are: Create new content, Re-speaking, Explicit editing, Explicit navigation, Content review, Ask questions and Delegate Task. The results showed significant differences on three categories: Explicit editing, Content review and Ask questions. No significant difference was identified in other categories. We will elaborate in the respective sections. We also performed a T-test between *Audio only* and *Audio+Screen* condition for each type of Re-speaking, as identified in Section 4.1.1, and identified no significant difference. In terms of subjective preference, the participants were asked for their preferred condition between *Audio only* and *Audio+Screen* as Authors. Nine Authors preferred the *Audio+Screen* condition while five preferred *Audio only*. Six Authors had no clear preference.

Statistically significant differences. Fig. 5 shows Explicit Editing were more frequently requested in *Audio+Screen* than *Audio only*. A two-sample t-test was performed to compare the occurrences of Explicit editing requests showed a significant difference between *Audio only* (M = 4.1%, SD = 4.7%) and *Audio+Screen* (M = 8.6%, SD = 8.0%); t(19) = 2.1, p = 0.012. Fig. 7 shows significantly more content review requests occurred in *Audio only* condition compared to *Audio+Screen*. A two-sample t-test performed on the occurances of Content review in *Audio only* and *Audio+Screen* showed a significant difference between *Audio only* (M = 3.3%, SD = 5.6%) and *Audio+Screen* (M = 2.7%, SD = 1.0%), t(19) = 2.1, p = 0.035. Fig. 6 showed questions were more frequently asked in *Audio only* than *Audio+Screen* condition. A two-sample t-test was performed on the occurances of Ask questions in *Audio only* and *Audio+Screen* showed a significant difference between *Audio only* (M = 6.9%, SD = 6.0%) and *Audio+Screen* (M = 0.3%, SD = 3.2%); t(19) = 2.1, p = 0.002.

Not seeing text lets the thoughts flow. Participants explained their reasons of preferring *Audio only*. Two Authors said it allowed them to focus and keep their train of thought while composing, whereas looking at the text could be distracting. G5P9-A-AS articulated well: *"When I can see him typing, sometimes my train of thoughts get interrupted, because I had to watch where he was segmenting sentences and where he put punctuation, I needed to make space in my head to think about those, it was distracting. When I cannot see, I just let him handle it and trust him."* G9P18-A-AS said something similar,*"because it allowed me to focus more on the image and, composing the text in my head. I knew that [Typist name] could see what I was writing. So I didn't have to think about the cues and things like that. So it was in a way more relaxed."* In addition, Authors mentioned it was less stressful and they had less self-doubt in the *Audio only* condition. One Author said it felt more free, *"G5P9-A-AO: I like not to see it. It feels more free, I don't get constrained."* Interestingly, the modality also affected the trust between Authors and Typists. *"G3P5-A-AS: With the screen, I rely less on [Typist name]. But with the eyes free, then it's more collaboration. More trust in him."*

Seeing text supports and encourages editing. Authors explained that having a shared screen helped them to keep track and make sure there was no misunderstanding. *"G2P3-A-AS: I wish to see, because this is an interactive process, when I input something, it can be checked in time, it's an interface."* One participant preferred *Audio+Screen* because the screen "helped him to think" (G9P17-A-AS). Beyond these, seeing the spoken text also makes Authors more compelled to edit. This partially explains why Authors made more explicit edits in *Audio+Screen* condition (Fig. 5). G3P6-A-AO explained he was more easily satisfied when he could not see the text, but he cared more about visual details when seeing it. *"well the eyes free condition. It just hindered my desire to edit, made it easier for me to be satisfied because the editing process was a bit cumbersome. (...) I'm happy with it. I noticed that they said visual components to text when you're writing, you're not, you don't only care about how things sound, but also how things look. So it's not the same to leave a line in between paragraphs or to go to the next paragraph, or to make a line longer or shorter. And that dimension disappears, when we have somebody reading to you. So you don't care about that. But if you see the text, then you have these, these other axes that you care about, and it's something that you're going to incorporate in the text edition."*

4.2 How the Typists assisted the Authors? (RQ2)

4.2.1 Categories of passive and proactive assistance. Fig. 8 shows all the categories of Typist behaviors emerged from our analysis. Based on our observation, we summarized a higher level of categories to represent the types of "services" provided by the Typists, including: Respond to request, Error correction, Error prevention, Propose review, Propose ideas and Take over. While *Respond to request* was observed to be passive, all the other categories describe active feedback or assistance initiated by Typists. As we can see from Fig.10, active assistance accounted for the

Respond to request		Error correction and prevention			Propose Review		Propose ideas	Take over
Passive read back	Answer questions	Error report	Double check content or request	Active feedback on progress	Active read back	Ask author to check	Suggest	Unsolicited edits
1.87%	3.67%	2.62%	21.80%	64.79%	0.82%	1.87%	1.72%	0.82%

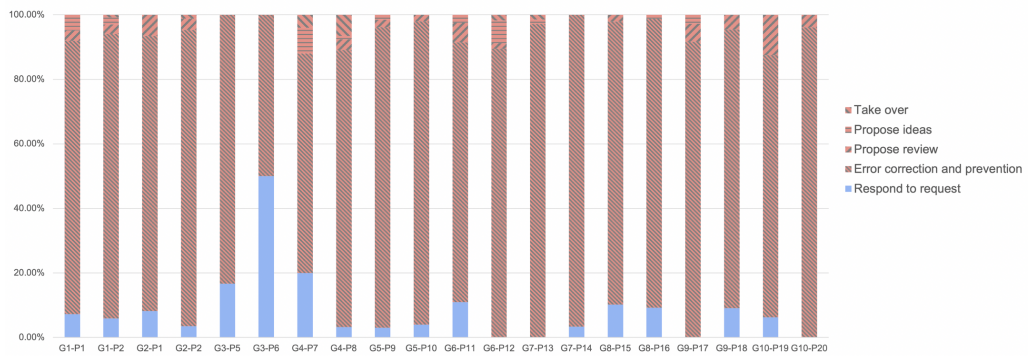

Fig. 8. Top: The categories of verbal assistance provided by the Typists and their occurrences in percentages. Blue categories were passive responses. Pink categories were actively initiated by the Typists. Bottom: Visualization of the style of each Typist by stacking the percentage of each category of their utterances.

majority of Typist utterances, leaving passive verbal responses only 5.5% in a sum. Out of the active assistance, *Error correction and prevention* was the most frequently observed behavior category, within which providing "Active feedback on progress" accounted for 64.8% of the total utterances. We elaborate on each category in the next subsection. Furthermore, we observed various group dynamics, which affects the passive/active level of the assistance. Fig. 8 illustrates the stacked percentages of utterances in categories from each Typist. While most Typists were active in their verbal assistance, we can see G3P6 was the most passive, with half of the time only responding to the Author's requests and never made suggestions.

4.2.2 "Services" provided by the Typists. The Typists demonstrated the following behaviors when assisting the Authors with the task.

Respond to request. Two types of passive behaviors of the Typists emerged. One is Answer questions: Typists directly answer Authors' questions. The types of questions being asked are elaborated in Section 4.1.3 and Appendix B. The other one is Passive read back, which happened when Authors asked Typists to read back part of the transcribed text for reviewing it.

Error correction and prevention. The Typists demonstrated three types of behaviors to help with error correction or prevention. It included Error report behaviors where Typists brought it up when something seemed wrong, Double check content or request when Typists were unsure, and Active feedback on progress to help Authors keep track and synchronize.

The Typists reported to the Authors when it was clear something went wrong in the process, coded as Error report. The types of reported errors include: the Typist did not understand which words to type, e.g., *"Which word? I didn't hear it clearly"*, *"What do you mean?"*; the Typist could not catch the speed of the Author, e.g., *"I couldn't catch up!" "I forgot what's after ..."*; the Typist did not know how to spell the word, e.g., *"... I forgot how to spell it"*.

Interestingly, the Typists also exhibited various types of Re-speaking behaviors for error correction or prevention. They were not re-speaking what they said by themselves, but what the Authors said. The Typists repeated some part of the content in a question when the Typists missed or misheard the words, e.g., *"G10P19-T: 'Very' what, sir?"* More implicitly, when the Typists misheard some words in the end of a sentence, they simply repeated the last correctly-taken word using a question tone, instead of explicitly saying they misheard something. Take a look at this example. *"G2P3-A: I would name the picture as a cat and a hard crystal ball. G2P4-T: A cat and a hard?"* In this case it was clear the Typist missed the last few words. Not only for correcting mistakes, Re-speaking were also used to prevent errors when something did not sound right. The Typists double-checked content by re-speaking the problematic words with a question tone, e.g., *"G1P2-A: she is... she's trashing. G1P1-T: trashing?"*

In order to prevent errors from occurring, the Typists did Double check content or request, following an instruction given by the Authors that appeared unclear or incorrect to the Typists. When it was about content, besides Re-speaking, they asked questions like,*"G10P20-T: Maybe somewhere 'to' the future or 'in' the future?" "G10P20-T: Did you say 'it seemed to lead down' or ...? G10P19-A: Oh, 'lead him'."* They also asked questions based on the grammatically structure, e.g., *"G1P1-T: What was the subject of this sentence? 'at the bottom of the books'."* Sometimes they simply asked the Authors to repeat the sentence they just said, e.g., *"G1P1-T: Please repeat this sentence."* When it was about the request, what was unclear could be about the location, e.g., *"G1P1-A: em, at the second sentence, next to 'the soldiers'. G1P2-T: The second?" "G6P11-A: not 'wine', I meant 'wire'. G6P12-T: In which line?"* The confusion could also be about the editing operation, e.g., *"G6P11-A: Continue deleting! G6P12-T: You don't want the following anymore?"* Sometimes it became unclear whether it was about the content or the editing operation: *"G6P11-A: Change the earlier thing to be 'a boy is walking in the beautiful universe'. G6P12-T: beautiful universe, do I remove the 'grass'?"*

The Typists verbally provided Active feedback to the Authors to help them keep track of their typing progress. There are three types of them. The first one observed was that the Typists constantly read while typing, in order to provide real-time feedback of where he/she was. The second verbal cue was a brief signal indicating typing just finished, such as "em". Similarly the third verbal cue was phrased as a suggestion to continue, "ok, and then?". Last but not least, since we set a word limit for each task, the Typists sometimes reported the word count when they saw appropriate. This was rather necessary in the *Audio only* condition.

Propose Review. There were times when the Typists initiated a text review without being asked by the Authors. One way of doing this was Active read back, which was the Typists proactively reading back the entire or a large part of the content without being asked by the Authors. G5P9-T explained for some unsure things, like the tense, he would read back instead of asking questions, *"When it comes to tense, I wouldn't ask him which tense, I would read it back to him. If he thought it wasn't right he would correct it."* G1P1 also said he/she would read it back instead to double-check.

The other way of proposing a text review was to ask the Authors to check the content, coded as Ask Author to check. In the *Audio+Screen* condition, it was straightforward that the Authors were asked to look at the text, e.g., *"G6P11-T: Please check if there is anything wrong."* In the *Audio only* condition, the Typists had to read back the text to be checked, e.g., *"G6P12-T: Let me read it back to you and see how you feel..."*

Make suggestions. The Typists provided active Suggestions to the Authors to help with the task. Content suggestions included grammatical corrections, e.g., *"G5P9-T: I think the 'and' should be changed to 'which', right?"*, better wording, e.g., *"G6P12-T: How about I change it to 'with different size'?"*, or about the style, e.g., *"G6P11-T: Another 'and'? Let's not use 'and' again, too many 'and'!"*

Besides suggestions about the content, the Typists also gave procedural comments, for instance, *"G4P8-T: Or let's edit this first, fix the capitalization and then come back to think about the last sentence."*

Unsolicited edits. Unsolicited editing is the behavior of the Typists editing the text without consulting the Author. This happened mostly for correcting obvious mistakes or the mistakes of the Typist him/herself, occasionally for the meaning of the sentences. When being asked whether they did things the Authors didn't ask for, 8 out of 20 participants reported that they took the initiative and corrected minor grammatical errors, such as the use of singular and plural, prepositional and conjunctions, etc. four of them took the initiative to revise the rationality or meaning of the sentence.

In the interview, participants elaborated why and how they made unsolicited edits. For instance, G9P17-T corrected segmentation of sentences and captions, *"... I was probably having a little bit more ... control over say things like where the sentences ended and full stops because like [AuthorName] wasn't able to see that."* G5P9-T explained in detail her mental process while making segmentation: *"I couldn't be sure whether he finished the sentence, so I would segment it subconsciously (with period). But then I felt the sentence after seemed to connect to the previous one, I would change the period to comma. I would follow my own judgement and I didn't ask him."* G5P10-T corrected his own mistake once realized, *"I heard 'honor', then I felt something was wrong, ..., then I realized I spelt it wrong, so I corrected it later. At that time, [AuthorName] could not see."* The same participant also mentioned him adding plural, *"... I added 's' when I felt there should be one... But sometimes I would ask her when I'm not sure."* In addition, the Typists added punctuation. *"G7P13-T: Because I am not only able to listen, but also to think, to understand its content, and then make a judgement whether it should be a question mark or a comma."*

Some Typists refrained from making unsolicited changes, or even content suggestions. G9P18-T said, *"I tried to be faithful to the Author. So I tried not to add it. I was only edited in my own errors."* G3P6-T said, *"the only thing I did was focusing on, on writing and being accurate. And then adding the commas and full stops wherever I felt they should be... because [AuthorName] didn't provide any information for that. ... I refrain from making any comments or suggestions about the text because it would have been inappropriate."* Interestingly, G3P5-T described her mental model and experience of being a Typist as "killing part of the brain": *"I used to have to take minutes as a teaching assistant or for work. So it's like, you kill part of your brain, like you stop thinking and you just type. But ... the problem is that sometimes [AuthorName] would say something and my head would say, but, like grammar wise that doesn't make sense. Or like, sometimes the grammar tried to override the thinking, the 'not thinking'. It's like, if you get so used to hearing the way certain things are said for many years. It's hard to not change things. It's like an instinct."*

4.2.3 Typists' choice of modality (RQ4). As Typists, eleven participants preferred the *Audio+Screen* condition while four preferred *Audio only*. The rest six of them had no preference. The main reason for preferring the screen was to get real-time feedback from the Authors to be more efficient. *"G2P4-T-AS: I felt a sense of safety, feel like what I wrote can receive realtime feedback. Sometimes there was mistyping, maybe that word is pronounced the same but spelled differently. If he can see it he can tell me." "G5P10-T-AO: I hope he could see me typing, because it could avoid many problems, like homonyms or punctuation." "G9P18-T-AS: So the screen was helpful that I could be sure to know that I was doing what she wanted me to do."* G1P1 thought it highly depended on the tasks: if she was recording exactly the Author's words, then *Audio+Screen* was preferred. If she needed to be creative in editing the content, *Audio only* was preferred. Furthermore, two groups of participants thought being able to see made it easier to locate text. As the reason to prefer *Audio only*, G2P4-T-AS reported having mixed feelings about *Audio+Screen* because she felt uncomfortable and nervous

that the Author could see her typing mistakes immediately; yet she found it effective for completing the task.

4.2.4 Typists' perceived challenges. The Typists also faced challenges in the process of assisting Authors. Four main challenges were identified in the interview. (1) Mishearing words or wrong spellings. As G2P4-T said, *"Listening is also difficult, there are a lot of words that are pronounced similarly and then the network is not particularly clear, so it's easy to get mixed up and not know what words are there anyway."* (2) Misjudging punctuation due to a lack of understanding of the content being expressed. As mentioned by G5P10-T, *"After reading the sentence, as a Typist, I thought he had finished the sentence, but in the context of [AuthorName], she hadn't finished the sentence, so she decided to go on. But at this place I typed a full stop, but he was probably prepared to put a comma."* (3) Unsure about how to provide feedback. For instance G10-P2 was worried about providing constant feedback being annoying. (4) Losing track of the Author. This was sometimes due to the typing speed slower than the Authors' speech. Some other times it was due to Typists' memory, as G4P8 said, "I might forget what the second half of what he said was when I was typing. Just asked him to say the second half of the sentence."

5 BETWEEN AUTHORS AND TYPISTS (RQ3)

The interaction between Authors and Typists demonstrated rich behavior patterns in terms of how they cooperatively navigated and located text, how misunderstandings occurred and were resolved, how they coordinated their closely-coupled collaboration and how they adapted to each other's work style.

5.1 Coordination

This dictation task is a closely-coupled collaborative task [14, 39, 57], which requires the Authors to keep track of the Typists and the Typists needed to be able to follow the Authors all the time. While this is relatively easy when visual feedback is provided, it can be a challenge in the *Audio only* condition.

5.1.1 Navigating and locating text. When Authors asked Typists to edit, read back or address a question about a part of the text, they needed to specify a deictic reference to help locate the target. We identified the following methods they used for locating text.

Deictic References. Calling keywords or reading a phrase was the most frequently observed way to locate content. If the keywords are not exactly the target, the reference is commonly accompanied by a temporal deixis, namely "before" or "after". More observed temporal deixes include "the sentence/words I just said", "at the beginning" or "at the end". Actually, most of these temporal deixes used by our participants can be seen as spatial deixes as well, given the sequential nature of textual content. Numeric references were used in combination with a text unit, such as, "the second sentence", "the last word", "the first line". Locating with the order of lines was only observed in *Audio+Screen* conditions. Example 1 below shows a few occurrences of temporal and spatial deixes. More examples can be found in Appendix A.

Communicating with mouse pointer. In *Audio+Screen* condition, apart from the text and the moving cursor at the editing position provided visual feedback, we also noticed one more visual cue - the mouse pointer of the Typist, played an important role in the communication. Example 4 below illustrates how the moving pointer of the Typist reflected how she was lost in searching for a keyword, where she was searching, and how the Author guided her smoothly from locating the sentence to the part of the sentence in subtle steps. Example 5 shows a situation when such visual feedback is missing and how it can become difficult to resolve misunderstanding. The Typist got

lost at some point and asked a few questions to seek for help locating the editing position with keywords and numeric indexes of the sentence: *"Do you mean after the sentence with 'white socks'? Or 'walking up a stair?' After the second sentence? Or after the first sentence?"* But the Author could not directly answer these questions, probably because she could not see which sentence that was, thus continued to repeat the keyword 'black hole' to locate. That did not solve the problem - the Typist responded with something even more confusing. The Author then gave up the precision and said 'whatever', until later simply started a new sentence.

Selected example dialogues between Authors and Typists:

Example 1. G6, *Audio only.*
P11-A: *"**Grass**"*
P12-T: *"Grass, okay."*
P11-A: *"'And blow soul bubbles, it seems that', **to the end,** don't use 'it seems that', it seems that, **Put this sentence what I just said, add a little thing before this sentence, go back to its beginning first.** 'The picture describes a dream of a child, it is so simple for a child to get happiness.'"*
P12-T: *"So simple for a child."*

Example 2. G6, *Audio+Screen.*
P12-A: *"Comma, appearance, they give out, full stop, yes. they give out that they knew how to weave scarves of the most beautiful colors, scarves of, of the most beautiful colors, and elaborate patterns, **ELABROATE, ELA**, yes,**BR** , Oh, **BO-RATE** , no more."*
P11-T: *"**You can check to see if there is anything wrong.***
P12-A: "Change the full stop of the 'leaf' to a comma."
P11-T: "What line?"
*P12-A: "**Second line**. Capitalize the 's'.""*

Example 3. G5, *Audio only.*
P10-T: *"OK. A toy sitting on a chair manipulated by someone we don't know. Does it sound familiar? It's just like my life. When I was a little boy If was manipulated by my parents..."*
P9-A: [interrupting by barging in] *"**No, sorry. I have to correct. It seems that you didn't change the words I said.** 'Just does it sound familiar and actually it's my real life."*
P10-T: *"OK OK. You want to change the third sentence into 'actually'?"*

Example 4. G1, *Audio+Screen.*
P1-A: *"Em, at the second sentence, near 'the some soldiers'. [P2's **pointer** starts to **travel rapidly** through the whole text.] Em.."*
P2-T: *"Second? [P2's pointer is **scanning the second sentence**.]"*
P1-A: *"... From the second sentence, next sentence, add a sentence. [P2's pointer **stops at the beginning** of the second sentence.] ... **After** the second sentence. [P2's pointer finally **moved to the end** of the second sentence.] Em. "*
P2-T: *"Em."*
P1-A: *"'they hold the guns'."*

Example 5. G1, *Audio only.*
P2-A: *"'Above his'... wait, the 'walking'."*
P1-T: *"[Do you mean here] 'He is walking up a stair to a black hole'?"*
P2-A: *"Then you say 'above him at the black hole', 'above' as on the top."*
P1-T: *"Do you mean **after the sentence with 'white socks'? Or 'walking up a stair?' After the second sentence? Or after the first sentence?**"*
P2-A: *"**'black hole', after 'black hole'.**"*
P1-T: *"'black socks'? After 'white socks'? Ah black hole, okay."*
P2-A: *"Either is fine, **whatever**."*
P1-T: *"Ok. Then, 'above the' ...?"*
P2-A: *"'above'... **How about this, after the 'black hole', after that, start a new sentence:** 'Above him, him and the black hole,'"*
P1-T: *"'and the black hole'"*
P2-A: *"'comma, there are two'"*
... [P1 continues repeating after typing and P2 continues composing]

Example 6. G2, *Audio only.*
P3-T: *"Glass. OK, I need to read again: 'It tells the power of time to change the cute young girl to...'"*
P4-A: [interrupting by barging in] *"'Tells the power of time', **remove after that**."*
P3-T: *"Sorry?"*
P4-A: *"**The last sentence.**"*
P3-T: *"**The last sentence:** 'It tells the power of time to change the cute young girl to an old lady.' **Just delete, right?**"*
P4-A: *"Uh, **'it tells the power of time.'**"*
P3-T: *"OK. 'It tells the power of time', **nothing**."*

Example 7. G1, *Audio only.*
P1-A: *"'Hold by his hand. And **since he is walking to the mountain, since he is climbing the mountain.'**"*
P2-T: *"**'walking to the mountain', then, 'climbing ... ' right?** "*
P1-A: *"Em... **'Since he is climbing the mountain.'**"*
P2-T: *"oh, oh, 'since he is climbing the mountain', and then?"*
P1-A: *"And 'some leaves is falling down from his body.'"*

Example 8. G8, *Audio only.*
P16-T: *"OK, there is one sentence I didn't understand, **what do you mean by 'they look like very carefully'?**"*
P15-A: *"**They open their eyes extremely, very extremely, ... so big. Their eyes are so big.** [This was intended to overwrite the sentence asked by the Author.]"*
P16-T: *"**Which sentence exactly?** Which sentence? 'Their eyes are so big.'"*
P15-A: *"Actually **just there, at 'they, they'** [he meant where the Author was at]."*
P16-T: *"...[A few seconds of typing sound]... Alright?"*
p16-T: *"Alright."*

Example 9. G8, *Audio+Screen.*
P16-A: *"'The direction of sunset. He is wearing a cloth made from leaves. Leaves.'"*
P15-T: *"Yeah."*
P16-A: *"'In front of him are three', **no, this is the second sentence, a new sentence.** [Then the Typist added period and changed the 'i' of 'in front of' to 'I'.]"*
P15-A: *"'three big mountains.'"*

Barging in readback. In *Audio only* condition, the above-mentioned deictic references are not sufficient. Barging in readback emerged as the Authors interrupted the Typists during his/her readback of the written text and suggested something near the interruption point. This was occasionally observed for locating problematic area, as seen in Example 3 and Example 6.

5.1.2 Synchronizing speed and matching rhythm. In the Section 4.2.2 we described how the Typists provided active feedback to inform Authors where they were. This behavior was indeed helpful

according to our interviews. For instance, G9P17-T said, *"I think just reading through at the end was useful for making sure that you got everything."* G4P7-A said, *"She would read after every sentence. And she would read it again after finishing."* If the Typists did not automatically give active feedback, some Authors would ask for it, even defined their own verbal cue for it. *"G4P8-A: I told him to make a 'zhi' sound every time he finished typing."* Furthermore, Authors paid attention to the non-verbal cues, namely the sound of the keyboard to make a judgement and adapt their speaking speed. A few more proactive Authors was actively asking the Typists whether each keyword had been taken down, in order to track progress. In addition, there were Authors who did not pay active attention to track the Typists' progress, relying on error calls by the Typists if they could not follow. G7P13-A simply slowed down her speech based on her own estimation: *"I would compare to my own typing speed."* G10P19-A had full confidence for his Typist: *"I mean [AuthorName]'s English level was phenomenal, and I feel that we were concentrated during the exercise."*

Typists' self-correction in idle time. We observed incidences where Typists reviewed the text spontaneously and corrected errors while the Authors were pausing to think or had completed. G9P17-T corrected segmentation of sentences and captions: *"So thinking of something I might go back and correct where I put a full stop of sentence or correct the type while [AuthorName] was thinking. "* G7P13-T explained the use of idle time to correct her own mistakes, *"Oh, most of the mistakes I made when typing... [I remember them], and I would go back and correct it when I have a chance during task or when I have time."*

5.2 Confusion and Misunderstanding

The participants generally felt their communication was smooth and misunderstandings were rare and only about small things. Yet, four types of misunderstandings between Typists and Authors were observed during the experiment and reflected in the interviews.

5.2.1 Mishearing words. The primary cause of misunderstandings was Typists misheard words. They were particularly prone to this error when homonyms were involved in the text or words were pronounced incorrectly. G10P19-T said, *"I think there was also one moment where I didn't pronounce something correctly or came up with a name like 'blob', so I felt like these were either slight misunderstandings or they were potential misunderstanding. So that's why I tried to correct them immediately and to be more clear for [AuthorName] to help him understand and bring some order to the chaos."* This quote also explained why Authors repeated themselves as a measure for error prevention. One participant explained this challenge was sometimes due to the lack of context: *"G1P1-T: When you (the Author) are talking, the Typist doesn't know the context. If you are talking about a garden, then talking about flowers and trees is normal, he may know this word is flower or tree. Then if you suddenly say in the garden there is a Humanoid robot, maybe he couldn't associate immediately and just guessed it's another word more related to nature. So it's a problem of lost of context."* G3P5-T thought their team had minimal misunderstanding and attributed that to their similar background and writing skill. She said, *"I think the misunderstandings ... when we pronounce certain words differently from each other. And to our idea of where to put comma, or period is different. I don't think there is any, like, major misunderstanding. And I know, this is not done on purpose. But the thing is, in terms of our background, you know, [AuthorName] and I are kind of similar. You know, we both have PhDs, we both were in [university name], we both have to do a lot of writing. And we both have to do a lot of creative thinking. [...] I'm guessing he and I ended up reading a lot of the same kinds of things. So, you know, when he says something, it makes sense to me. And when I said something to him, yes, it made him laugh, but he seemed to understand why I said it, and how I said it."*

5.2.2 Misunderstanding of editing location or scope. The location of requested edits could be misunderstood, especially when Authors and Typists got out of sync. When Authors moved their attention to a different area, they sometimes expected the Typists to follow, which could be unrealistic. Misunderstandings in location also occurred when the Author gave an abstract location that covered multiple sentences. In Section 5.1.1 - Communicating with mouse pointer, we elaborated examples of such misunderstandings and how they got resolved in *Audio+Screen* and *Audio only* conditions. Even when the Typist correctly located the editing position, misunderstandings of which exact words to edit still happened. Example 6 shows one case of confusion and clarification about the scope of the deletion. The Author wanted to delete half of a sentence, but the Typist thought the intention was about the whole sentence. The Author corrected it by re-speaking the final text:"it tells the power of time.", which then needed to be double confirmed with a repetition plus "nothing" by the Typist.

5.2.3 Composing or editing? Misunderstandings also occurred when the Author switched between composition and editing. This happened more often to some Authors who preferred to re-speak for overwriting content instead of giving an explicit editing request. Re-speaking utterances could be easily misunderstood as a new composition. For instance, Example 7 shows how an implicit editing of replacing "walking to" with "climbing" was misunderstood as new composition. Example 8 shows an example of even more confusing behaviors from Authors. When the Typist posted a question about a problematic sentence, the Author did not explicate anything, but directly spoke a new sentence to replace it. To make it worse, in the same utterance of speaking the "replacement sentence", he overwrote part of it three times and repeated the sentence once. Yet, the confusion was resolved surprisingly easily, he simply answered with keywords to help the Typist locate the editing position, and the Typist then understood everything. Here it must be due to the semantic similarity of the erroneous sentence and the replacement sentence.

5.2.4 Unclear sentence segmentation. The last source of misunderstanding was reflected on Authors and Typists having different ideas about sentence segmentation. Example 9 shows one case of Authors fixing the sentence segmentation when they could see the text. The ambiguity of sentence segmentation was also reported in the interview, G10P19-T said, *"I guess my main misunderstanding was just I think the sentence was ending or it was just like pausing."* G10P20-A explained this could also be caused by Authors being indecisive about where to end the sentence: *"I guess when I was speaking, I kind of sometimes started a sentence and I realized I didn't wanna continue. So maybe I was a little indecisive when I was trying to tell some of the stories."* The misunderstanding of sentence segmentation appeared rather acceptable considering most participants thought their communication was smooth. G8P16-T said, *"This does not affect understanding of the story, unless there are a couple of obscure words we couldn't understand. (...) There wasn't any big problem, just that we may wish the format looked the same as we imagined, (...), but it's just a matter of one sentence versus two."*

5.3 Co-adaptation

The cooperation strategies between Authors and Typists developed over time as they got familiar with each other throughout the experiment. The participants also applied strategies after they switched roles to better facilitate the partner, as they experienced some challenges on the other role.

5.3.1 Adjusting their speed and speech. The co-adaptation between Authors and Typists was first reflected on how they adjusted their speed to fit each other. Three groups mentioned in the interview that they slowed down the dictation speed as an Author and became more patient. One participant

reported he started to pay more attention to speak out punctuation after being a Typist. Some Authors started out dictating faster than what the Typists could follow, and slowed down over time to adapt to the typing speed. G7P13-A mentioned this was easier with the shared screen: *"If [I] can see it, I wouldn't need to speak very slowly, I could automatically adjust and match my speed."*

5.3.2 Adjusting feedback. As we explained earlier, both Authors and Typists needed to consciously provide feedback to each other. Authors needed to be informed of the typing progress in the *Audio only* condition. Typists wanted effective feedback from the Authors when they made typing mistakes or misunderstood the dictation. A few Typists started out by passively providing feedback while being frequently asked by the Authors "Have you finished (typing)?", then gradually became active in reporting it without being asked. One participant explained how he adapted to the Author over time by reducing disturbing behaviors and refraining from making suggestions as well as getting more comfortable with typing. G10P19-T: *"At the beginning, when I was at this automatically, I started kind of reciting after [AuthorName] so I was telling him what I was doing, which later on, I thought that's pretty annoying, but he should be in his own kind of creative space. Towards the edge, I felt that I needed to shut up, not to give suggestions. Plus I got used to the voice (...), his way of kind of coming up with the words. I feel more comfortable with all this thing to what do you want to say in terms of me being able to get it on paper as well."*

5.3.3 Learning habits and preferences. The participants learned their partner's preferences and wording over time, such as their vocabulary, whether they wanted more or less feedback, etc. As described in Section 4.2.2, Typists sometimes performed unsolicited actions and changed the text without consulting the Authors. This happened more in the later stages of the experiments, when participants were confident that such unsolicited actions would be accepted. G10P20-T: " *at the 4th story, I felt like I could maybe create a little bit more."* Verbal coordination also reduced over time, as participants began to understand what their partner meant or wanted without a thorough explanation. In the interview, one participant mentioned that getting used to her partner's vocabulary and wording habit would help reduce misunderstanding. *"G9P17-T: (...) it's probably just a matter of getting used to someone's vocabulary. And then, if I listen to [AuthorName] speaking more (...) if we did this long term, I think there would probably be less misunderstandings. So just getting used to what sort of words you tend to use for things. I might be probably better at guessing. "*

5.3.4 Influences of the modality. Regarding how the Authors and Typists collaborated, modality first influenced how they helped each other navigate and locate text. In the *Audio only* condition, the Authors either barged in the readback from Typists or relied on vague temporal / spatial memory about their composed text. In *Audio+Screen* they could locate precisely with keywords and other positional references. The Typists' moving mouse pointers were used as an effective communication tool for resolving misunderstandings in location and editing scope. Second, modality affected how Authors could keep track of the progress of the Typists. When Authors could not see the text, constant verbal feedback was needed from the Typists side. Having the screen also made it easier for the Authors to notice mistakes made by the Typists and provide feedback.

6 DISCUSSION

In this section we summarize our findings in terms of how they answer our research questions, and discuss them in light of developing future systems.

6.1 RQ1: How do Authors dictate to Typists?

Our findings answer RQ1 with a detailed understanding of the Authors' intentions, expressions, strategies and challenges when completing a dictation task assisted by another human. Without

implying necessity, they suggested a set of experiences that could be potentially supported in future dictation interfaces.

6.1.1 Diving into the ambiguity. Our analysis dived into the ambiguity of human speech for dictation and visualized how implicit and explicit instructions were made throughout the development process of the written text. This approach differs from the data collection and analysis of natural speech with the goal of training a machine for speech recognition and repair [35]. We identified both explicit and implicit ways that the Authors instructed the Typists to create or edit text. Explicit editing included requesting *Add, Delete, Replace, Format, Organize, Punctuate* operations via informal and free speech. Implicit editing was instructed via *Re-speaking*, which was a highly ambiguous behavior with five possible intentions including *editing*, *confirming*, *locating*, *continuing composition*, and *repairing speech*. Furthermore, re-speaking was also observed on the Typists' side, but in the context of repeating the Authors' words, as an implicit expression for *error correction or prevention*. Recent state-of-the-art techniques support re-speaking phrases for eyes-free text editing [12, 18] or in multimodal interfaces [17, 58]. While these make important steps towards supporting natural dictation, our research reveals there is much more to understand about re-speaking and contributes a detailed understanding of the demands and noises around it. More research is warranted to explore how to support implicit user instructions.

Another source of ambiguity was found in distinguishing composition and editing. We observed how the Authors went back and forth between composition and editing, by *globally making several passes of the text* and *locally mixing text composition and editing*. Our visualizations (Fig. 3 and Fig. 4) illustrate how entangled these two modes are in natural dictation. It is even harder to clearly distinguish between composition and editing. For instance, Re-speaking as an implicit editing behavior to continue the half-sentence spoken before, could also be considered as an edit of that sentence with an addition at the end. Existing dictation systems handle text input and editing with two different modes that need explicit switching by users, which apparently does not adapt well with natural dictation. Our appendix and supplement material of this work provides details of how the Authors and Typists communicated and how each piece of text got composed and edited over a timeline. They could serve as assets for system designers, engineers and researchers for further investigation or design consideration.

6.1.2 Dictation is not only about transcription. Besides composing and editing, we also identified other behavior patterns of the Authors supporting the task, including *explicit navigation, content review, ask questions, delegate task and think aloud*. While we expected the frequencies of behaviors being affected by individual choices in a semi-controlled experiment, we believe the list provides new perspectives of Authors' needs for natural dictation. We provide categories of expressions with examples used by the Authors, which can serve as inspirations for the design of new dictation interfaces. For instance, a conversational interface could potentially be provided to help users review content and ask questions. Moreover, our interview revealed a number of dictation strategies the Authors developed, including consciously reducing uncertainty and repetition, formulating and organizing their thoughts before speaking to avoid major changes, speaking slowly in simple words, etc. These can be seen as areas where users could learn or be willing to be trained on.

6.1.3 Help Authors organize and remember thoughts. We identified a number of challenges the Authors faced, even with the intelligent assistance provided by Typists. The primary one was how they felt rushed and disorganized when composing text with speech. Written text is a linear representation of the often arbitrary human thoughts. Human speech is produced as we speak, not before [30]. Some Authors reported the challenge of organizing their free speech into a linear story, and wished the Typists could help. One Author suggested that the Typist could note down

the Author's multiple composition ideas and later check with the Author. Furthermore, as the composition process developed, it appeared to be hard for most Authors to remember their own wordings without seeing the composed text. They used keywords as reminders or developed a spatial memory of the sentence or part of the text their target might be in. These findings could serve as inspirations for future intelligent dictation system features.

6.2 RQ2: How do Typists assist Authors?

Our findings answer RQ2 by listing the types of assistance provided by the Typists and elaborating their intentions, expressions and encountered challenges.

6.2.1 Error correction and prevention are major. The majority of Typist utterances (over 85%) were for correcting or preventing errors. Among them, about 65% provided active feedback about their typing progress. It appeared that the active verbal and non-verbal feedback from Typists was crucial for Authors to keep track of the composed texts and coordinate with Typists. Automatic reading back, which is a feature being supported by previous work in eyes-free dictation [15], was shown to be an effective measure for both error correction and prevention. Future interfaces should improve on the timing and choices for voice readback. More than reading back, the Typists frequently double-checked with the Authors about their intention, the editing location and their actual editing request. Re-speaking potentially erroneous words with a questioning tone seemed to be an effective expression for double-checking. This could be implemented as a system feature.

6.2.2 Proactive AI features are promising. Except one participant being relatively passive, we found all the Typists were proactive in more than 80% of their utterances (Fig. 8). Specifically, their proactive behaviors included *correcting and preventing errors*, *proposing review*, *proposing ideas* and *taking over*. Except for taking over with unsolicited edits that might cause a sense of losing control, this proactive assistance was largely appreciated and desired by the Authors, especially when Authors could not see the text. Many of the corresponding system features are already supported by existing systems by highlighting wrong words or automatic correction, such as Grammarly[8]. Moreover, Typists' self-correction appeared to be a needed feature. This is comparable to the automatic correction features based on semantic context in existing speech recognition engines such as the Google Cloud API[9]. In addition, we found it was not easy for the Typists to segment sentences as how the Authors wanted without explication, even with the ability of understanding context as humans. Future systems could focus on engaging users in explicating segmentations.

6.2.3 Effective Typist assistance. Participants mentioned that it was helpful when Typists provided active feedback to Authors to inform their progress. One group with low level of misunderstanding thought it was because of their similar background and writing skill. Some apprecated the help from the Typist in coming up with better wording and segmenting sentences. Furthermore, participants appreciated the experience getting smoother as they adapted to each other over time and needed less explicit coordination. Compared to machines, humans have superior empathetic skills to detect what the other person needs during their interaction. The information comes from not only the words they hear but also the tone, the context of the task and their understanding of the other person. An even more powerful skill is that humans adjust their behaviors quickly based on the other person's reaction. If a Typist tried to provide an assistance and noticed it was not helpful, he/she would stop it quickly. If some critical support was missing, the Author would have asked for it. We believe the intrinsic assistance provided by most Typists can serve as a good reference for effective support to Authors.

[8]https://www.grammarly.com/
[9]https://cloud.google.com/speech-to-text

6.3 RQ3: How do Authors and Typists coordinate and collaborate?

Our findings answer RQ3 by analysing how the Authors and Typists kept track of each other, how they communicated their intentions and resolved misunderstandings, as well as how they adapted to each other.

6.3.1 Coordinating location and communicating intention. By observing how Authors and Typists coordinated their actions in this highly synchronous collaborative text composition task, we identified 3 important factors at play: *how to locate and select text with deictic references*, *in what scope the Authors were operating (word, phrase, clause or multi-clause)* and *the attention locale of both parties*.

To locate text, both Authors and Typists referred to one-dimensional temporal or spatial references for locating text, together with calling keywords and naming numeric indexes of sentences. Verbal barging-in was observed for locating text when Authors could not see the text. We believe these text selection concepts are fundamentally different from how we search and locate text in Graphical User Interfaces (GUI). However, our dictation interfaces today look very much similar to a traditional text editor, which is a GUI legacy centered around the cursor position. However, speech is more about words, semantics and subtleties in voice. We need to rethink the fundamental concepts of an editing interface adapted to speech input.

Problems occured when it was unclear where the typing or editing location should be, how much of the original text should be removed or replaced, or when the Author and the Typist were looking at a different place in the text. To coordinate, the Typists and Authors gave active feedback to each other to synchronize their speed of speaking and typing, sometimes even establishing their own verbal cues to constantly signal the completion of typing. One effective communication strategy we observed when the Author could see the text, was that the Typist used the mouse pointer to circle around where she was editing location. Future systems could adopt similar strategies to assist error recovery interactively. For example, gaze input can be considered in such scenarios to help detect users' editing intention.

6.3.2 Unavoidable misunderstandings. Resolving misunderstand is important for both human-human and human-computer communication. First, we learned the types of misunderstandings, which were *mishearing words*, *misunderstanding editing location and scope*, *confusion between composing and editing* and *sentence segmentation*. Considering humans' superior ability of understanding context compared to machines, we believe misunderstandings are unavoidable, and more research is needed to understand how humans resolve these misunderstandings, such as what and how they use subtle behavior signals, to inform the design of intelligent machine support.

6.3.3 Co-adaptive features. The cooperation between Authors and Typists were like dancing, where they adjust to each others' speed and rhythm in real-time. They also adjusted the frequencies and types of feedback to each other by learning their habits and preferences from the reactions. Because of this, the participants felt there were fewer misunderstandings and smoother experiences developed over time. Future dictation systems could consider learning such rhythmic dynamics from the users' choices of words, composition speed, and feedback.

6.4 RQ4: How does the communication modality affect Authors, Typists and their cooperation?

Answering RQ4, we found both quantitative and qualitative differences between *Audio only* and *Audio+Screen* conditions. Authors made more explicit edits with a screen compared to audio only, but similar amounts of implicit edits via re-speaking. One explanation indicated in the interview was that Authors were more tolerant to errors when they did not see the composed text. We also

found the Authors asked fewer questions and requested less content readback from the Typists when seeing text. Participants also felt *Audio+Screen* helped them to be more efficient in finding and correcting errors, while in *Audio only* Typists had to provide more feedback about the progress. Coordinating locations were also easier with a shared screen. It allowed creative communication strategies to happen, such as used the moving pointer to resolve navigation issues. Interestingly, despite it being frustrating to not be able to see the composed text, some Authors preferred the *Audio only* condition. This was because not seeing the text can be liberating and allowed their thoughts to develop more freely. Seeing the text all the time was considered to be distracting. In addition, we found modality was associated with emotion. Not seeing the composed text could be less stressful and even increase the Author's trust in the Typist.

Based on our findings, we believe dictating eyes-free and with-screen should be both supported in future interfaces. These two situations have different benefits and drawbacks that complement each other. Composing text in an eyes-free modality can be liberating, yet being able to see the system status is efficient and much needed for error correction. Too much visual demand would interrupt the eyes-free experience and break down some mobile user scenarios. Therefore we believe, how to strike a balance between seeing and not seeing the text, and how to provide visual and acoustic feedback, are major design and research challenges for future dictation systems.

6.5 In the context of writing support research

Previous research has supported writing tasks in various ways. Numerous works have developed AI systems for automatically generating scripts or stories [41]. Researchers have developed intelligent tools to support users in creative writing. For instance, ReQUEST [40] is an intelligent tool for authoring plots, not by providing content, but by acting like an audience to provide users with constructive feedback. The system asks "why" and "Consequence" questions to direct the users through the process. Furthermore, computational analysis has taken place in extracting writing strategies from alargeamount of articles [2]. New writing strategies such as microwriting has been studied in the context of gamified writing [22], or to support collaborative writing [53] and crowdsourced writing [36].

Although our research supports text composition tasks, we need to clarify that our focus is on dictation for *text input*, which is different from other existing research on writing support that focuses on *creative writing*. However, our findings contribute to the understandings of using speech to write, including the entangled composition and revision process, the ambiguity and disorganized nature of speech, and the liberating experience of speaking without looking at the text. Future AI systems could build on these understandings to develop creativity support for dictation interfaces.

Our findings on coordination and communication between Authors and Typists align with those in the literature of collaborative writing. For instance, previous research showed that collaborative writing can benefit from tools providing conversational grounding (e.g., enhanced coordination and group awareness) [32]. To enhance conversation grounding in collaborative writing, Kütt et al. developed a tool to visualize their partner's gaze information, as a form of shared visual information, and showed that it helped to improve mutual understanding and flow of communication [28].

6.6 Limitation

6.6.1 English fluency. Our study might be affected by the fact most participants are non-native English speakers and they performed English writing tasks. Some exchanges between the Authors and Typists were about correcting grammatical mistakes and spelling words, which appeared less in native speakers. However, our native speaker groups of participants exhibited similar behavior categories. We did not observe visible biases to our main findings. Future studies will investigate potential effects of native versus non-native language use. While English fluencies

and writing skills may affect quantitative ratios of different types of Author requests or Typist behaviors, we are confident the categories of phenomena are not affected. We draw our focus on the types of communication behaviors, which are in a way orthogonal to the actual content of the communication.

6.6.2 Ecological validity. This study explores human-to-human dictation, where Typists typed much slower than a speech recognition engine, and had limitations in memory capacity and English vocabulary. These potentially affected the behaviors of Authors in a number of ways. For instance we observed the Authors slowing down their composition to wait for the Typists, and the Authors occasionally spelled the words to the Typists. The slower typing speed and limited memory of Typists probably affected the efficiency of completing tasks, the frequencies of the Authors repeating content for the Typists, and the report from Typists for not following the Authors. In addition, the lack of English vocabulary probably led to more questions from the Typists and spelling help from the Authors, and maybe choices of simpler wordings.

We acknowledge that humans interact with computers differently from interacting with each other. There are emotional and social interactions between humans, which may not transfer to human-computer interaction in the same way. In this study we coded the data focusing on the actions and information flow related to task productivity only, not other aspects of the interaction. By finding out what support features Authors found effective, we hope to inform and inspire the design of future dictation interfaces. However, the actual effectiveness of the potential features still needs to be tested in a human-to-computer setting.

7 CONCLUSION

This paper presents an experiment that investigated natural human dictation for text composition by observing pairs of participants dictating text to each other in the roles of Typists and Authors. Our results unpacked the ambiguity of natural dictation behavior and showed that supporting dictation is not just about improving speech recognition accuracy. Based on a comprehensive understanding of how Authors dictated, how Typists assisted and how they collaborated with each other, this work informs the design of future dictation interfaces by uncovering design opportunities and providing inspiration from human-to-human interaction. We are the first to use a role-play method for this purpose by observing how people provide an "intelligent service" to each other. It provides new insights that complement existing findings from Wizard-of-Oz studies and opens up discussions for future research on natural human behavior.

ACKNOWLEDGMENTS

This research was supported by the Hong Kong Research Grants Council - ECS scheme under the project number CityU 21209419. We thank our reviewers for their constructive feedback.

REFERENCES

[1] Xiang Ao, Xugang Wang, Feng Tian, Guozhong Dai, and Hongan Wang. 2007. Crossmodal Error Correction of Continuous Handwriting Recognition by Speech. In Proceedings of the 12th International Conference on Intelligent User Interfaces (Honolulu, Hawaii, USA) (IUI '07). Association for Computing Machinery, New York, NY, USA, 243–250. https://doi.org/10.1145/1216295.1216339

[2] Tal August, Lauren Kim, Katharina Reinecke, and Noah A Smith. 2020. Writing strategies for science communication: Data and computational analysis. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 5327–5344.

[3] Shiri Azenkot and Nicole B Lee. 2013. Exploring the use of speech input by blind people on mobile devices. In Proceedings of the 15th international ACM SIGACCESS conference on computers and accessibility. 1–8.

[4] Alan D Baddeley and Graham J Hitch. 1994. Developments in the concept of working memory. Neuropsychology 8, 4 (1994), 485.

[5] Thomas Binder. 1999. Setting the stage for improvised video scenarios. In CHI'99 extended abstracts on Human factors in computing systems. 230–231.
[6] Eva Brandt and Camilla Grunnet. 2000. Evoking the future: Drama and props in user centered design. In Proceedings of Participatory Design Conference (PDC 2000). 11–20.
[7] Oğuz Turan Buruk and Oğuzhan Özcan. 2016. WEARPG: Game Design Implications for Movement-Based Play in Table-Top Role-Playing Games with Arm-Worn Devices. In Proceedings of the 20th International Academic Mindtrek Conference (Tampere, Finland) (AcademicMindtrek '16). Association for Computing Machinery, New York, NY, USA, 403–412. https://doi.org/10.1145/2994310.2994315
[8] Janice Carter-Wesley. 2009. Voice recognition dictation for nurses. JONA: The Journal of Nursing Administration 39, 7/8 (2009), 310–312.
[9] C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani. 2018. State-of-the-Art Speech Recognition with Sequence-to-Sequence Models. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 4774–4778. https://doi.org/10.1109/ICASSP.2018.8462105
[10] Mark G. Core and Lenhart K. Schubert. 1999. A Syntactic Framework for Speech Repairs and Other Disruptions. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics (College Park, Maryland) (ACL '99). Association for Computational Linguistics, USA, 413–420. https://doi.org/10.3115/1034678.1034742
[11] Susan De La Paz. 1999. Composing via dictation and speech recognition systems: Compensatory technology for students with learning disabilities. Learning Disability Quarterly 22, 3 (1999), 173–182.
[12] Jiayue Fan, Chenning Xu, Chun Yu, and Yuanchun Shi. 2021. Just Speak It: Minimize Cognitive Load for Eyes-Free Text Editing with a Smart Voice Assistant. In The 34th Annual ACM Symposium on User Interface Software and Technology. 910–921.
[13] Margaret Foley, Géry Casiez, and Daniel Vogel. 2020. Comparing Smartphone Speech Recognition and Touchscreen Typing for Composition and Transcription. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3313831.3376861
[14] Susan R Fussell, Robert E Kraut, and Jane Siegel. 2000. Coordination of communication: Effects of shared visual context on collaborative work. In Proceedings of the 2000 ACM conference on Computer supported cooperative work. 21–30.
[15] Debjyoti Ghosh, Pin Sym Foong, Shengdong Zhao, Di Chen, and Morten Fjeld. 2018. EDITalk: Towards Designing Eyes-Free Interactions for Mobile Word Processing. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (Montreal QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3173574.3173977
[16] Debjyoti Ghosh, Pin Sym Foong, Shengdong Zhao, Can Liu, Nuwan Janaka, and Vinitha Erusu. 2020. EYEditor: Towards On-the-Go Heads-Up Text Editing Using Voice and Manual Input. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376173
[17] Debjyoti Ghosh, Pin Sym Foong, Shengdong Zhao, Can Liu, Nuwan Janaka, and Vinitha Erusu. 2020. Eyeditor: Towards on-the-go heads-up text editing using voice and manual input. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 1–13.
[18] Debjyoti Ghosh, Can Liu, Shengdong Zhao, and Kotaro Hara. 2020. Commanding and Re-Dictation: Developing Eyes-Free Voice-Based Interaction for Editing Dictated Text. ACM Transactions on Computer-Human Interaction (TOCHI) 27, 4 (2020), 1–31.
[19] Debjyoti Ghosh, Can Liu, Shengdong Zhao, and Kotaro Hara. 2020. Commanding and Re-Dictation: Developing Eyes-Free Voice-Based Interaction for Editing Dictated Text. ACM Trans. Comput.-Hum. Interact. 27, 4, Article 28 (Aug. 2020), 31 pages. https://doi.org/10.1145/3390889
[20] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Processing Magazine 29, 6 (Nov 2012), 82–97. https://doi.org/10.1109/MSP.2012.2205597
[21] Steve Howard, Jennie Carroll, John Murphy, and Jane Peck. 2002. Using'endowed props' in scenario-based design. In Proceedings of the second Nordic conference on Human-computer interaction. 1–10.
[22] Shamsi T Iqbal, Jaime Teevan, Dan Liebling, and Anne Loomis Thompson. 2018. Multitasking with play write, a mobile microproductivity writing tool. In Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology. 411–422.
[23] Clare-Marie Karat, Christine Halverson, Daniel Horn, and John Karat. 1999. Patterns of Entry and Correction in Large Vocabulary Continuous Speech Recognition Systems. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Pittsburgh, Pennsylvania, USA) (CHI '99). ACM, New York, NY, USA, 568–575. https:

//doi.org/10.1145/302979.303160
[24] Clare-Marie Karat, Christine Halverson, Daniel Horn, and John Karat. 1999. Patterns of entry and correction in large vocabulary continuous speech recognition systems. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems. 568–575.
[25] Tatsuya Kawahara. 2007. Intelligent transcription system based on spontaneous speech processing. In Second International Conference on Informatics Research for Development of Knowledge Society Infrastructure (ICKS'07). IEEE, 19–26.
[26] Tedd Kourkounakis, Amirhossein Hajavi, and Ali Etemad. 2020. Detecting Multiple Speech Disfluencies Using a Deep Residual Network with Bidirectional Long Short-Term Memory. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 6089–6093.
[27] Anuj Kumar, Tim Paek, and Bongshin Lee. 2012. Voice typing: a new speech interaction model for dictation on touchscreen devices. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2277–2286.
[28] Grete Helena Kütt, Kevin Lee, Ethan Hardacre, and Alexandra Papoutsaki. 2019. Eye-write: Gaze sharing for collaborative writing. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 1–12.
[29] David R Large, Leigh Clark, Annie Quandt, Gary Burnett, and Lee Skrypchuk. 2017. Steering the conversation: a linguistic exploration of natural language interactions with a digital assistant during simulated driving. Applied ergonomics 63 (2017), 53–61.
[30] W Levelt. 1999. Producing spoken language. The neurocognition of language (1999), 83–122.
[31] Yang Liu, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf, and Mary Harper. 2006. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. IEEE Transactions on audio, speech, and language processing 14, 5 (2006), 1526–1540.
[32] Paul Benjamin Lowry and Jay F Nunamaker. 2003. Using Internet-based, distributed collaborative writing tools to improve coordination and group awareness in writing teams. IEEE Transactions on Professional Communication 46, 4 (2003), 277–297.
[33] Bruce R Maxim, Stein Brunvand, and Adrienne Decker. 2017. Use of role-play and gamification in a software project course. In 2017 IEEE frontiers in education conference (FIE). IEEE, 1–5.
[34] Kristina Moroz-Lapin. 2009. Role play in HCI studies. HCI Educators 2009-playing with our education (2009), 64–67.
[35] Christine Nakatani and Julia Hirschberg. 1993. A Speech-First Model for Repair Detection and Correction. In Proceedings of the 31st Annual Meeting on Association for Computational Linguistics (Columbus, Ohio) (ACL '93). Association for Computational Linguistics, USA, 46–53. https://doi.org/10.3115/981574.981581
[36] Michael Nebeling, Alexandra To, Anhong Guo, Adrian A de Freitas, Jaime Teevan, Steven P Dow, and Jeffrey P Bigham. 2016. WearWrite: Crowd-assisted writing from smartwatches. In Proceedings of the 2016 CHI conference on human factors in computing systems. 3834–3846.
[37] Jun Ogata and Masataka Goto. 2005. Speech repair: quick error correction just by using selection operation for speech input interfaces. In Ninth European Conference on Speech Communication and Technology.
[38] John A Pezzullo, Glenn A Tung, Jeffrey M Rogg, Lawrence M Davis, Jeffrey M Brody, and William W Mayo-Smith. 2008. Voice recognition dictation: radiologist as transcriptionist. Journal of digital imaging 21, 4 (2008), 384–389.
[39] Ilona R Posner, Ronald M Baecker, and M Mantei. 1993. How people write together. In Proceedings of the Hawaii International Conference on System Sciences, Vol. 25. IEEE INSTITUTE OF ELECTRICAL AND ELECTRONICS, 127–127.
[40] Mark Riedl, Jonathan Rowe, and David K Elson. 2008. Toward intelligent support of authoring machinima media content: story and visualization. (2008).
[41] Mark O Riedl. 2010. Story planning: Creativity through exploration, retrieval, and analogical transformation. Minds and Machines 20, 4 (2010), 589–614.
[42] Haşim Sak, Andrew Senior, and Françoise Beaufays. 2014. Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition. arXiv:1402.1128 [cs.NE]
[43] S. Schlögl, G. Doherty, and S. Luz. 2015. Wizard of Oz Experimentation for Language Technology Applications: Challenges and Tools. Interacting with Computers 27, 6 (2015), 592–615.
[44] Andrew Sears, Jinhuan Feng, Kwesi Oseitutu, and Claire-Marie Karat. 2003. Hands-free, speech-based navigation during dictation: Difficulties, consequences, and solutions. Human-computer interaction 18, 3 (2003), 229–257.
[45] Andrew Sears, Clare-Marie Karat, Kwesi Oseitutu, Azfar Karimullah, and Jinjuan Feng. 2001. Productivity, satisfaction, and interaction strategies of individuals with spinal cord injuries and traditional users interacting with speech recognition software. Universal Access in the information Society 1, 1 (2001), 4–15.
[46] Gry Seland. 2006. System Designer Assessments of Role Play as a Design Method: A Qualitative Study. In Proceedings of the 4th Nordic Conference on Human-Computer Interaction: Changing Roles (Oslo, Norway) (NordiCHI '06). Association for Computing Machinery, New York, NY, USA, 222–231. https://doi.org/10.1145/1182475.1182499

[47] Kristian T Simsarian. 2003. Take it to the next stage: the roles of role playing in the design process. In CHI'03 extended abstracts on Human factors in computing systems. 1012–1013.
[48] Elizabeth Stokoe. 2011. Simulated interaction and communication skills training: The 'conversation-analytic role-play method'. In Applied conversation analysis. Springer, 119–139.
[49] Bernhard Suhm, Brad Myers, and Alex Waibel. 2001. Multimodal Error Correction for Speech User Interfaces. ACM Trans. Comput.-Hum. Interact. 8, 1 (March 2001), 60–98. https://doi.org/10.1145/371127.371166
[50] Bernhard Suhm, Brad Myers, and Alex Waibel. 2001. Multimodal error correction for speech user interfaces. ACM transactions on computer-human interaction (TOCHI) 8, 1 (2001), 60–98.
[51] Bernhard Suhm and Alex Waibel. 1997. Exploiting repair context in interactive error recovery. In Fifth European Conference on Speech Communication and Technology.
[52] Dag Svanaes and Gry Seland. 2004. Putting the users center stage: role playing and low-fi prototyping enable end users to design mobile systems. In Proceedings of the SIGCHI conference on Human factors in computing systems. 479–486.
[53] Jaime Teevan, Shamsi T Iqbal, and Curtis Von Veh. 2016. Supporting collaborative writing with microtasks. In Proceedings of the 2016 CHI conference on human factors in computing systems. 2657–2668.
[54] Rhinedd Toms. 1985. The Effective Use of Role-Play: A Handbook for Teachers and Trainers. By Morry van Ments. London: Kogan Page. 1983. Pp. 186.£ 12.50. The British Journal of Psychiatry 146, 3 (1985), 340–340.
[55] Johanna Viitanen. 2009. Redesigning digital dictation for physicians: A user-centred approach. Health Informatics Journal 15, 3 (2009), 179–190.
[56] Yanna Vogiazou, Jonathan Freeman, and Jane Lessiter. 2007. The use of improvisational role-play in user centered design processes. In International Conference on Human-Computer Interaction. Springer, 262–272.
[57] Dakuo Wang, Haodan Tan, and Tun Lu. 2017. Why users do not want to write together when they are writing together: Users' rationales for today's collaborative writing practices. Proceedings of the ACM on Human-Computer Interaction 1, CSCW (2017), 1–18.
[58] Maozheng Zhao, Wenzhe Cui, IV Ramakrishnan, Shumin Zhai, and Xiaojun Bi. 2021. Voice and Touch Based Error-tolerant Multimodal Text Editing and Correction for Smartphones. In The 34th Annual ACM Symposium on User Interface Software and Technology. 162–178.

A CODES FOR AUTHOR UTTERANCES

B CODES FOR TYPIST UTTERANCES

Received April 2021; revised November 2021; accepted March 2022

Codes for Author utterances

Large categories	Sub-types	Description	Examples
Navigating/ Locating text	Visual references	using indexes of lines	*First line, in the end. (G6P11)*
			Second line. (G6P12)
	Sequential deixis	using 1D chronological / spatial references	*Is it after the phrase "while socks"? Or does it say "working up a stair"? After the second sentence, right? After the first sentence? (G1P1)*
			Put this sentence what I just said, add a little thing before this sentence. (G6P1)
			Next sentence "It is open." (G1P2)
			Put this sentence what I just said, add a little thing before this sentence. (G6P1)
			go back to its beginning first. (G6P1)
			and then, "below the gap". (G1P2)
	Numeric references	using indexes of semantic units, e.g., sentence	*The last sentence, it tells "It tells the power of time to change the cute young girl to an old lady". just delete, right? (G2P3)*
			Em, the second sentence, at "the soldiers". (G1P1)
			The same sentence. (G5P10)
	Keyoword references	refering to exact words or expression, ideally with few occurances	*Em, the second sentence, at "the soldiers". (G1P1)*
			after 'black hole'. (G1P2)
Re-speaking	Overwrite to modify	to overwrite the re-spoken text , as an implicit editing request	*The haircut...oh...the hair style...OK...the hairstyle. (G5P9)*
			Yes...yes. was recycling the stars...was taking, was taking, was collecting, was collecting those stars. (G1P2)
	Confirm or repeat for Typist	to make sure the Typist clearly understood what to be typed	*En...sorry. When we were little child. When we were little child. (G5P9)*
			Twinkling star. Twinkling star. T-w-i-i-n-k-l-i-n-g. (G1P2)
	Continue composition	to resume composition of unfinished chunk of text, also to help Typist locate	*Everyone, no, no, a new sentence, everyone was telling her. (G8P16-AS)*
			instead, she died at twelve. A new sentence, ah no, let me continue my sentence. She died at twelve and this pair of ballet shoes. (G8P16)
	Locate/ refer to text position	to locate a position with keywords and request operations from there	*Front, back to front, back to top. "a boy" is it? a "boy" walking on the grass. (G6P11)*
			After "black hole", a new sentence, that is "about him, him and the black hole". (G1P2)
	Natural speech repair	subconscious repetition of words to repair one's own speech	*Is the, the paper man becomes loosening. (G1P2)*
			This, this, this image contains many kinds of face mask.(G8P15)
Explicit editing	Add	traditional editing operations requested with informal and contextual instructions	*write "which is waving it is this" after "pigeon". (G8P15)*
			add a new sentence. (G8P16)
	Delete		*enter. (G4P8)*
			success, no,no,no (G5P10)
			to, not not "to" (G5P9)
	Replace		*stars, don't need "stars", delete "stars" (G6P11)*
			flying everywhere, replace it with away. (G1P2)
			change to "that". (G6P11)
	Format	helping Typist split or organize sentences by either telling them what actions to perform or how the result should look like	*mountains, start a new sentence. (G8P16)*
			and let a new sentence. (G7P3)
	Organize		*put all the previous words in one quotation mark...There is one more sentence after the quotation mark. "This is Richard." (G8P16)*
	Punctuate	dictating punctuation marks	*comma / dot / question mark*
Reviewing content	Visually review	telling the Typist he/she is reading the screen	*A bit inside, let me see. (G1P1)*
	Verbally review	asking the Typist to read part of the text, often giving a reference	*OK, so please read it again(G5P9)*
			Can you read it out? (G9P17)
	Summarize content	asking the Typist to summarize the composed text in their own words	*Tell me what you understood from the writing, don't just read the sentences. (G6P11)*
Asking questions	Check word count	asking for the word count when there is a word limit	*How many words are there? Are there enough words? (G4P7)*
	Track Typist process	Asking for feedback of typing being finished	*Did you write it all down? (G1P1)*
			Uh, have you finished typing? (G4P8)
			So can we move on? (G5P9)
	Ask for suggestion	Asking Typist for suggestions and evaluation of their wording	*Is it "to earn my living"? (G4P7)*
			How about "the table full stop"? and let a new sentence. (G7P3)
			What else needs to be changed? (G5P9)
			Do we need to add something like "Anyway, I'm very, so mad'"in the end? (G4P8)
	Ask for spelling / grammar correction	Asking Typist to double check potential errors	*How should I say "Tuanju" in English? (G6P7)*
			Add "ly" to strange, is there such a usage? "very strangely". (G6P11)
	Check typist understanding	Asking whether Typist understood the content	*... how should I pronounce that one? ...and you do know I'm talking about the compass? (G5P10)*
Delegating task	Delegate editing	Asking Typist to help improving a wording	*Please help me to change it. (G6P11)*
	Delegate composition	Asking Typist to help composing when the Author is out of idea	*Can you help me think about what else to say? (G4P7)*
Thinking aloud	Speak aloud one's thought process	Exposing one's thinking process to enhance mutual understanding	*"Let me think, em, she is, she is ah, she is... (G1P2)"*

Codes for Typist utterances

Large categories	Sub-types	Description	Examples
Respond to request	Passive read back	read back text following Author's request	*G7P14-A: OK. Can you repeat?* *G7P13-T: So like..So some words are announced not important like and.. on the other hand. I repeat the whole story for you.*
	Answer questions	answer Authors' question	*G4P8-A: What did I say in my last sentence?* *G4P7-T: You said "considering her super beauty and kind personality.She will obtain a great success in the next year."*
Error report	Problem in comprehension	did not understand Author's wording or intention	*Which word? I didn't hear it clearly. (G1P2)* *What do you mean?*
	Problem following Author's speed	not fast enough to follow Author	*I couldn't catch up! (G6P12)* *I forgot what's after. (G6P12)*
	Problem in language capability	did not know how to spell a word	*I forgot how to spell it. (G5P10)*
Re-speaking	for error correction	After misheard some words, Typist re-speak the last correctly-taken word in a question, to indicate a need for hearing the rest again.	*'Very' what, sir? (G10P19)* *G2P3-A: I would name the picture as a cat and a hard crystal ball.* *G2P4-T: A cat and a hard? 'Very' what, sir? (G10P19)*
	for error prevention	double-check postentially wrong words with a question tone	*G1P2-A: she is... she's trashing.* *G1P1-T: trashing?*
Double check content or request	Check content	asking about a unclear part or asking Author to repeat	*G10P20-T: Did you say 'it seemed to lead down'or ...?* *G10P19-A: Oh, 'lead him'.*
			What was the subject of this sentence? 'at the bottom of the books'. (G1P1)
			Please repeat this sentence. (G1P1)
	Check location	asking about an unclear location with navigation references	*G1P1-A: em, at the second sentence, next to 'the soldiers'.* *G1P2-T: The second?*
			G6P11-A: not 'wine', I meant 'wire'. *G6P12-T: In which line?*
	Check request	not sure about the edting request	*G6P11-A: Continue deleting!* *G6P12-T: You don't want the following anymore?*
Active feedback on progress	Read loud while typing	constantly reading the text being typed to indicate progress	*G1P1-A: since he is climbing the mountain.* *G1P2-T: since he is climbing the mountain,*
	Give constant verbal cues	established verbal cue to signal typing is finished or ready to continue	*em.*
			ok, and then?
	Report word count	reporting the word count whenever they deemed appropriate	*It's 79 words. (G9P18)*
Active readback	Readback a large piece automatically	proactively reading back the entire or a large part of the content when it felt appropriate, without being asked by Author	*I'll read it back...(G1P1)*
Ask author to check	Suggest visual review	asking Author to read on screen	*Please check if there is anything wrong. (G6P11)*
	Suggest auditory review	suggesting Author to hear it	*Let me read it back to you and see how you feel. (G6P12)*
Make Suggests	Minor language fixes	suggesting grammatical corrections	*I think the 'and' should be changed to 'which', right? (G5P9)*
		suggesting better wording	*How about I change it to 'with different size'? (G6P12)*
	Writing style	criticizing bad language style	*Another 'and'? Let's not use 'and' again, too many 'and'! (G6P11)*
	Managerial suggestion	suggesting writing procedure	*Or let's edit this first, fix the capitalization and then come back to think about the last sentence. (G4P8)*
Unslicited edits	Correcting errors	making edits without consulting Author	*Observation note:"Typist returns to correct leftover spelling errors in recorded text when Author pauses"*