"Try, Try, Try Again:" Sequence Analysis of User Interaction Data with a Voice User Interface

Chelsea M. Myers
Drexel University
Philadelphia, United States
chel.myers@gmail.com

Luis Fernando Laris Pardo
IT University of Copenhagen
Copenhagen, Denmark
luisferlaris@gmail.com

Ana Acosta-Ruiz
Drexel University
Philadelphia, United States
ava48@drexel.edu

Alessandro Canossa
IT University of Copenhagen
Copenhagen, Denmark
acan@kadk.dk

Jichen Zhu
IT University of Copenhagen
Copenhagen, Denmark
jichen.zhu@gmail.com

ABSTRACT

Voice User Interfaces (VUIs) pose challenges for users to learn the system's supported features and commands. Users often rely on trial-and-error to navigate VUI dialogues and complete desired tasks. The order in which users try different commands contains vital information about how they learn. In this paper, we explore using sequence analysis techniques to reveal the patterns of tactics our participants (n = 50) used when interacting with an unfamiliar multi-modal VUI. We present the results of our sequence analysis, the top sequences used, and a cluster analysis of our participants on their usage of the top sequences. Our results indicate participants initially struggled with understanding the acceptable utterance structure and entities more so than utterance keywords. Additionally, we discuss our participants' behavior differences and discuss the usefulness of this methodology of sequence analysis for HCI VUI research.

CCS CONCEPTS

• **Human-centered computing → Natural language interfaces**.

KEYWORDS

voice user interfaces, sequence analysis, user data analytics, trial-and-error

ACM Reference Format:
Chelsea M. Myers, Luis Fernando Laris Pardo, Ana Acosta-Ruiz, Alessandro Canossa, and Jichen Zhu. 2021. "Try, Try, Try Again:" Sequence Analysis of User Interaction Data with a Voice User Interface. In *3rd Conference on Conversational User Interfaces (CUI '21), July 27–29, 2021, Bilbao (online), Spain.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3469595.3469613

1 INTRODUCTION

Voice User Interfaces (VUIs) are becoming more common in our homes, but users are not exploring the full potential of available features [4, 5]. Researchers have identified that VUIs' low learnability imposes the second most frequent obstacles for users [35], immediately following Natural Language Processing (NLP) errors [18, 22, 27, 36]. New VUI users are often not familiar enough with the system to proficiently execute available features (i.e., intents) and verbal commands (i.e., utterances) [35]. Even frequent users cannot always keep up with the regularly added new features of VUIs without much notice. In order to improve VUIs' learnability, it is important to look closely at how users interact with unfamiliar VUIs, analyze where they struggle, and then devise design solutions.

A growing body of recent research has analyzed usage data of VUI users to identify users' expectations, command request strategies, and conversational roles of users engaging in a dialogue with a VUI [8, 10, 35–39]. With a few exceptions [35], the vast majority of these works manually review (e.g., conversation analysis) VUI usage data in order to find common patterns. While this approach is effective to identify trends, it is difficult for researchers to quantify the sequential patterns of users when using the VUI. For example, when a user tries to learn a new VUI command, she may fail the first few attempts before succeeding. The sequence in which she carries out these attempts — what she tries first, whether and how she uses the VUI's feedback, and what she tries next — contains invaluable information about a user's expectations and mental models and can thus inform design solutions to improve VUI learnability. So far, the VUI research community has only limited understanding of the sequence of steps that users take to learn VUIs.

In this paper, we propose a new approach for conducting sequence analysis on VUI usage data. Our method is informed by sequence analysis for gameplay data in user research for computer games [41] and we adapted it to VUI usage data. To illustrate the usefulness of our approach, we applied it to a dataset of how users (n = 50) interact with a multi-modal VUI calendar app *DiscoverCal*. Specifically, we analyze the sequence of trial-and-error tactics users employ when trying to learn VUI commands. As a result of our sequence analysis, we identify the 16 most common sub-sequences and the different usage exhibited by participants through a cluster analysis. Overall, we found that participants typically first relied on editing the entities in their utterances when encountering obstacles instead of keywords. We believe this indicates participants initially

struggled with understanding the acceptable utterance structure and entities more so than keywords. We also found participants differed in what tactics they used initially more frequently when encountering obstacles. With our results, we make recommendations for future multi-modal VUI designs, especially adaptive VUIs, and review the usefulness of sequence analysis in this domain.

Our paper extends our previous work [35], in which we analyzed the usage data of 12 users interacting with a VUI over three sessions. We identified the common tactics new users employed and the transition probability that a user applies one given tactic immediately after another one. Our previous paper's contribution focuses on the differences of VUI tactic usage *between types of obstacles* while this paper's contribution centers on the differences of VUI tactic sequence usage *between participants*. Previously, we did not analyze the differences in participant behavior and instead looked at the overall behavior of the VUI users. This current paper advances the previous work in several crucial ways. First, in our previous study [35], we analyzed the sequence of only 2 tactics whereas this analysis expands the notion of sequence to any number of them. Second, this analysis identifies the sequences of tactics most frequently used by the users whereas our previous work [35] focuses on how likely any tactic will be proceeded by another. Third, in this paper we identify different user clusters based on how they use the top sequences. This information leads to new design implications. Last but not least, the results in this current paper are based on a significantly larger dataset of 50 users. In summary, this paper makes the following contributions:

- We propose a novel approach to treat VUI usage data as sequential events and adopt a new method from a related domain of user research. To our knowledge, this is one of the first papers applying sequence analysis on VUI trial-and-error behavior from an HCI perspective. HCI VUI research more commonly reviews and discusses the aggregated behavioral data of how users interact with a VUI [13, 31, 36] or observations of VUI interaction [10, 17, 33].
- We identified the top 16 most frequently used sequences of tactics, revealing how users learn to use an unfamiliar VUI step-by-step.
- Our sequence analysis allowed us to identify four distinct user clusters and derive design implications for these groups.

2 RELATED WORK

To bridge the gap between users' expectations and VUI design, research often empirically evaluates VUI design techniques to correct users' expectations or scrutinizes those expectations through means such as conversation analysis of VUI dialogues and thematic analysis of user interviews. Studies proposing or comparing VUI design techniques typically evaluate the techniques' impact on user satisfaction and performance. These studies test various ways to improve the usability of VUIs through prompt and menu designs for discoverability of commands [13, 17, 31, 42], adaptive user models to adjust the system to users' predicted needs [23, 43], recovery and repair strategies [30]; to name a few. Observation-based research interviews VUI users and reviews dialogues to identify patterns in users' expectations and behavior. For example, studies interview users to reflect on their VUI experience and behavior [4, 14, 33], conduct in-situ interviews to elicit user utterances [9, 10, 32], or label a collected corpus of VUI dialogues to analyze [4, 5, 19, 35, 38]. Our work falls into the latter category– relying on labeling a corpus of VUI interactions. This labeling process allows researchers to explore the sequential nature of VUI dialogues more so than interviews since user utterances are coded and can be processed manually or algorithmically. While these related works use conversation analysis and Markov Chain Modeling to analysis VUI dialogues, we employ sequence analysis to generate the top sequence in our corpus. Of course, this technique requires us to select *what* to label our corpus with to generate sequences from. We labeled our corpus with the trial-and-error strategies VUI users exhibit when learning a new VUI.

Several studies analyze different strategies users employ when trying to learn a VUI. Studies categorizing user strategies from observations have broken down strategies into individual tactics (e.g., hyperarticulation, relying on a menu) [35] or broader user mental models (e.g., "push" and "pull" models) [9]. Research has not only reviewed how user strategies impact dialogues, but the impact of VUI feedback strategies on the users being able to complete their tasks with a system. For example, Cho et al. found that VUI feedback that provides even *incorrect* feedback as a result of an incorrect utterance (compared to a generic response), gives users more information on why their command failed and provides users with a better sense of "progress" [10]. Other research has found that contextual prompts [42] and discovery strategies in general [31] can help users learn VUIs. Our work builds upon these previous studies by analyzing not only how users try to learn correct utterances but the sequence of these efforts.

Most similar to our work is our previous study [35] where we categorized tactics users employed when encountering VUI obstacles and calculated the probability of these tactics being used initially when encountering specific obstacles. However, we focused what type of tactics participants used for the types of obstacles they encountered. In contrast, our current paper focuses on the sequences of a subset of these tactics and the types of VUI users who employ them. Additionally, in [35], we employ Markov Chain modeling to see the probability of one tactic following another. In this current paper, we employ sequential analysis to see the frequency of trial-and-error tactics being used across participants. We build upon our previous work by using its defined tactics to label a different multi-modal VUI dataset. Few studies employ sequence analysis for VUI researchers to analyze user behavior and inform VUI designs. VUI sequence analysis studies more commonly process usage data to train models to automatically detect a user's intent (desired command) [6, 21], emotion [3], and satisfaction [43], to name a few examples. Sequence analysis where VUI researchers review sequences to inform designs have done so to identify verbal (e.g., humming) and non-verbal cues (e.g., nodding) that signal affective states [20] or turn-taking behaviors [1]. Another study reviews the sequences of users' intents when accomplishing tasks with a car's VUI [29]. Unlike these studies, our work algorithmically reviews the sequence of trial-and-error tactics users employ when trying to learn an unfamiliar VUI.

2.1 Sequential User Behavior Modeling

Sequence analysis of users' interaction data is more commonly seen in other domains such as computer games research. Analyzing player behavior, also referred to as game analytics and player modeling, has become an essential pillar both for game development and games research [47]. Close analysis of player behavior can inform design decisions, streamline quality assurance, optimize monetization strategies, and support personalization of game experience [49]. More importantly, it allows detailed, large-scale investigations of human behavior and psychology in the context of play. For example, researchers have shown that modeling player behavior in educational games can shed light on their cognitive, affective, and behavioral patterns, which can inform designing better scaffolding to support learning [25, 26, 44, 47, 48].

A promising sequence analysis method to build player profiles is based on technique originally developed to investigate DNA sequences in biology. Developers at *Massive Entertainment* noticed the similarity and adapted it to their player base [34]. Sequential profiling provides a more accurate approach to segmentation, categorization, or profiling of users as it integrates a temporal viewpoint of the actions thus providing more context to understand the intent of users. Similar to existing VUI user research, traditional snapshot profiling is utilized to analyze players' gameplay data. Snapshot profiling is based on aggregate data but only provides information about the state of the users across the period of time covered by the snapshot [16, 40]. Sequence analysis has proven effective at finding patterns in users action space [24, 28, 45, 46] but has, to our knowledge, not be applied to understanding the sequences for VUI user interactions.

3 METHODS

For our VUI sequence analysis, we used the dataset (n = 50) collected in our previous study of a multi-modal VUI calendar named *DiscoverCal* [36]. Below we summarize the information necessary for our sequence analysis. *DiscoverCal* was created with Google's Dialogflow[1] NLP and the Web Speech API for ASR[2]. The system's intents and utterances where based on Google Assistant and Alexa's calendar interaction design. Further details of the original user study can be found in [36]. *DiscoverCal* is a web-based calendar application controlled solely through voice interaction with visual output to display the calendar. *DiscoverCal* visual's includes a static list of intents and utterances and a visual display of a calendar. Our dataset is from a previous study [36] using *DiscoverCal* in an online format where participants completed 10 prescribed tasks, along with a pre- and post-test questionnaire. Participants were given 10 tasks to complete. Tasks were ordered by complexity with the "easier" tasks being given first. We selected this design for two reasons: 1) to let VUI users familiarize themselves with the system and 2) to have participants first create appointments with basic commands to manipulate later with more complex tasks. This way participants manipulated appointments that they created themselves. A total of 50 participants spent an average of 14.16 ± 5.28 minutes on these tasks. The dataset consists of transcripts of dialogues with *DiscoverCal*, containing 2800 utterances provided by participants and their identified intents through NLP. For our sequence analysis, we reviewed the transcripts collected from participants completing the tasks with *DiscoverCal* and the post-test questionnaire responses for participants' perceived usability of the systems via the SUS instrument [7]. The participants' transcripts with *DiscoverCal* were augmented by applying labels to utterances demonstrating specific user behavior. In our previous study [35], we categorized tactics for users overcoming obstacles (i.e., VUI not reacting as desired) when learning an unfamiliar VUI. We took a subset of these tactics that were observable without audio/video data (since we only have the participants' written transcripts) and applied them to this new dataset. These tactics are VUI users' attempts at unblocking VUI dialogues to accomplish their tasks through trial-and-error.

Our sub-set of tactics applied along with definitions can be seen in Table 1: *Repetition* (repeating the same utterance), *Simplification* (removing words or entities from utterance), *New Keyword* (changing the keyword), *Use More Info* (adding more entities or entity details), and *Restarting* (aborting the dialogue and restarting the task). To apply these tactics and check for agreement, two of our researchers took 10% of participants' *DiscoverCal* transcripts of the original dataset and applied tactic codes to each utterance for an inter-coder reliability check. All transcripts but one resulted in an inter-coder reliability check above 90% agreement. The final transcript was reviewed, discussed, and re-coded by the researchers until over 90% agreement was reached. Afterward, one of the researchers coded the remaining transcripts.

3.1 Sequence Analysis & Segmentation

We adopt a sequence analysis method developed to analyze player behavior [34]. We chose this method because it allows VUI researchers to identify the top sequences in a corpus. This method can also be generalized to other sequential data of VUIs. As outlined in Makarovych et al. [34], the effectiveness of sequence analyses rests on four factors: the selection of the events used to define the sequences, the method used to process those sequences, the method utilized to analyze and cluster those sequences, and the relevance of the resulting clusters for the stakeholders.

3.1.1 Event List Definition. In defining events, we reviewed our available dataset, composed of both user utterances and system responses, and selected *events* to build our participants' sequences seen in Table 1. The majority of our events categorize the users' utterances (e.g., *Cancels* is exiting the dialogue). Two events were selected as *start signals* to provide potential further context for the sequences. For example, our *Attempt Start* event indicates what attempt the user is on for accomplishing a task. A team of computer scientist and VUI researchers reviewed to see if they were meaningful in terms of observing the desired user behavior (e.g., trial-and-error), that they had sufficient variance over the population, and that they were minimally correlated with each other (so that each sub-sequence did not show a behavior already captured in another sub-sequence) [2, 41, 46]).

In addition to the VUI trial-and-error tactics from our previous work [35], we included events for *Novelty* utterances (participants saying something outside of requirements for the task such as "How

[1]More information on Dialogflow can be found at https://dialogflow.cloud.google.com/
[2]More information on Web Speech API can be found at https://developer.mozilla.org/en-US/docs/Web/API/SpeechRecognition

Table 1: Selected events defined with their codes

Event	Code	Definition
Command	C	An utterance given to *DiscoverCal*
Novelty	Nv	Utterance not related to the tasks given (e.g. "How are you?")
Cancels	Ca	Executes the cancel command
Tactics (Can be combined in an utterance)		
Repetition	Rp	Repeats the same utterance
Simplification	S	Removes words or entities from utterance
New Keyword	N	Uses a new keyword (a likely synonym) in their utterances
Use More Info	Um	Adds more details to their utterance to be more specific
Restarting	Rs	Restarts their attempt from scratch
Start Signals		
Attempt Start	A#	# is the attempt number for a task (e.g., A1 is attempt #1)
Task Start	T#	# is the task number (e.g., T1 is the start of task #1)

are you, *DiscoverCal*?"), *Cancels* (participants evoking the cancel command and aborting a dialogue), and start signals for when a user begins a task or begins an attempt at that task. The *Command* event describes utterances devoid of any of the other events. The full list of events can be seen in Table 1. With these events and their codes, our participants' *DiscoverCal* usage data can be represented as sequences of events. For example, in Table 2, Example 1 could be represented as such: C → Um. In this example, the phrase *"Plan an event"* is not recognized by *DiscoverCal* for creating an event and the user tries an new approach to accomplish her task. Here, *C* would represent a participant executing an utterance with no trial-and-error tactic (*Command*) while → *UM* represents it was sequentially followed by an utterance using the *Use More Info* tactic. Since multiple tactics can occur in a single utterance, Table 2 dialogue Example 2 would be represented as C → N+Um since it is a *Command* utterance followed by an utterance using both *New Keyword* and *Use More Info*.

3.1.2 Sequences Pre-Processing. Once the events were defined, participants' sub-sequences were generated in a format for SPMF[3], an open-source software and data mining library specialized in the discovery of patterns in data. Then, we used an apriori algorithm (a bottom-up approach) commonly used to find related actions in sequence datasets. Specifically, we used the CM-SPAM algorithm [45] since this algorithm allows us to preserve the order of the actions in the dialogues since the goal was to discover sequences appearing often in our dataset.

This process generated our participants' top sub-sequences. The next step was to identify the most information-rich sub-sequences; for this, we applied Shannon Entropy [11] and an entropy score was calculated for each sequence. Following recommendations set by Makarovych et al. [34], the 30 highest-scoring sub-sequences were selected. The number 30 was selected to reduce the dimensionality of the clustering process. Lastly, we needed to remove redundant sub-sequences that were heavily correlated and essentially contained the same trial-and-error patterns. Each of the 30 sub-sequences received a normalized score between 0 and 1 that

[3]SPMF software can be found at https://www.philippe-fournier-viger.com/spmf/

Table 2: Example *DiscoverCal* dialogues with events labeled

Dialogue Example 1
User: Plan an event [C]
DC: Sorry, I don't understand
User: Plan an event tomorrow at 3 PM [Um]

Dialogue Example 2
User: Plan an event [C]
DC: Sorry, I don't understand
User: Create an event tomorrow at 3 PM [N+Um]

represented the frequency of occurrences of that sequence in all sessions. To decide what sub-sequences to remove, we used the following heuristics: the sub-sequences' correlation p-values, how meaningful were the sub-sequences were for VUI researchers, and the length of sub-sequences. We determined to keep the shortest sub-sequence of two heavily correlated pairs. For example, if C → Um → C and C → Um → C → C were correlated, we would keep the shorter sequence since the longer provides no new information. If a top sub-sequence was a shorter version of another sub-sequence, but *not* correlated, it was not removed.

The top 16 sub-sequences can be seen in Table 3. In this table, the *Ref.* column is the sub-sequence's code for us to refer to. The sub-sequences are in descending order of total frequency; meaning S1 was the most frequently used sub-sequence. *Start* is the sub-sequence's start signal, if any. *Sub-Sequence* shows a representation of the events in the sub-sequence in chronological order. The remaining columns show the average of the normalized usage in total for our participants and per cluster.

3.1.3 Top Sequence-based Clustering. The last step was clustering the users according to the frequency with which they employ any of the 16 sub-sequences selected. For this, three algorithms were tried to have a point of comparison between methods. These algorithms were the agglomerative clustering, DBSCAN, and k-means. To analyze the results, a close inspection of the results was done through

Table 3: Top sub-sequences and their average normalized usage overall and per cluster. Tactics are in *blue and marked with an asterisk.

Ref.	Start	Sub-Sequence	Total $n = 50$	*Explicits* $n = 12$	*Repeaters* $n = 16$	*Divergers* $n = 13$	*Guessers* $n = 9$
S1	A1	*Um → C	0.70	0.74	0.72	0.57	0.78
S2	T2	C → C	0.63	0.68	0.68	0.45	0.72
S3	A1	C → C → C	0.55	0.61	0.61	0.35	0.65
S4		C → *Um → C	0.55	0.60	0.59	0.38	0.66
S5		C → C → C → C	0.45	0.48	0.51	0.26	0.57
S6	A1	C → *Um	0.41	0.45	0.39	0.18	0.73
S7	A1	C → C → C → C	0.39	0.46	0.40	0.22	0.52
S8	A1	C → *S	0.39	0.29	0.38	0.17	0.86
S9		C → C → *Um	0.38	0.43	0.35	0.15	0.67
S10	A1	C → C → *Um	0.35	0.40	0.33	0.15	0.62
S11		C → C → C → *Um	0.35	0.40	0.31	0.13	0.64
S12		C → C → *S	0.34	0.19	0.34	0.13	0.81
S13	A1	C → *RP	0.32	0.12	0.56	0.12	0.47
S14		C → C → C → C → C	0.28	0.40	0.29	0.11	0.36
S15		C → C → *RP	0.28	0.08	0.51	0.08	0.40
S16		C → C → *N+Um	0.27	0.10	0.33	0.25	0.42

an analysis of how the clusters were confirmed. For each cluster, we checked which sub-sequences were frequently used and how this compared to the other clusters. In the end, after running the 3 algorithms several times with different parameters, k-means was chosen as the best model with $k = 4$ based on the elbow method. Starting from the 4 clusters selected, we generated Table 3, showing our top sub-sequences, their start signals, and the average of the normalized values of how many times the sequence was executed by the participants in total and per cluster. Finally, we performed the nonparametric Kruskal-Wallis H test [15] (since our dataset did not have a normal distribution) and Conover post-hoc squared ranks test [12] between clusters for our participants SUS score, total time, and total utterances to check for statistically significant differences.

4 RESULTS

Our analysis focused on isolating behaviors that could not be identified by looking at the aggregated data of our participants. In this section, we review our top sub-sequences and the clusters formed from them to find patterns in our participants' behavior when initially learning *DiscoverCal*.

4.1 Top Sub-Sequences

When comparing these top sub-sequences to our participants' aggregated data, the most interesting sub-sequences are those with a start signal. The sub-sequences with no start signal have no context for what task or attempt the sub-sequence was a reaction to and tell us only that an event or tactic occurred (which can be observed in aggregated data as well). For example, S9 only shows that the *Use More Info* tactics at some point in the task's sequence. This information can be gleaned through aggregated data since we can see how often each participant used each tactic. Meanwhile, S10 contains the same events and order but has a start signal of A1. This tells us that this sub-sequence was used often and early during the first attempts at tasks. This temporal data (i.e., *when* tactics are used) cannot be seen in our participants' aggregated data. Looking at our top sub-sequences, we see our participants' often commonly relied on *Use More Info* in initial attempts at tasks.

The only start signals in our top sub-sequences are Attempt 1 (A1 in S1, 3,6-8, 10, 13) and Task 2 (T2 in S2). We speculate this is caused by participants' behavior diverging from each other after the initial attempt, reducing the likelihood of shared sub-sequences. In S1, the most frequently used sub-sequence, we see the popularity of *initially* applying *Use More Info* when making the first attempt at a task. Participants were quick to *Use More Info* when starting a new task (probably informed by their previous task experience). In S2, we also see participants were likely to make a successful initial utterance for Task 2, but we do not see this for any other task. Task 2 is a simple task, similar to Task 1. This could imply participants were quick to apply their experience in Task 1 to successfully invoke the commands for Task 2. We also see that several tactics are coupled with an Attempt 1 start signal and occur towards the beginning of the sequence (e.g., S6, 8, 13). This aligns with expectations since a participant's first attempt at a task would more likely result in more obstacles since it is the participant's first attempt at the VUI's feature. Finally, we also see the *lack* of two tactics in our top sub-sequences for a first attempt: *New Keyword* and *Restarting*. From this, we see in a first attempt, participants were more likely to first rely on tactics that slightly modify their failed utterance by maintaining the keywords (e.g., *Use More Info* and *Simplification*) or restating the utterance (e.g., *Repetition*). Exploring new keywords or abandoning the attempt occurred less frequently early in the participants' attempts.

4.2 Participant Clusters

The top 16 sub-sequences were used to cluster our participants. These clusters show differences between what tactics each clusters initially and overall preferred to use. Cluster 1, the *Explicits*, used sub-sequences with more steps than the other clusters, breaking down tasks into more utterances and used sub-sequences providing more explicit information. Cluster 2, the *Repeaters*, used sub-sequences that contained *Repetition* more than the others. Cluster 3, the *Divergers*, diverged from using the top sub-sequences the most. Finally, *Guessers* used the sub-sequences containing tactics the most of the clusters (besides *Repetition*).

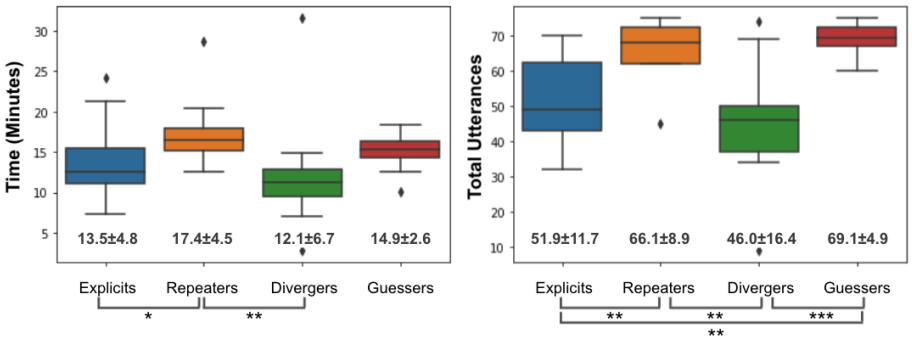

Figure 1: Boxplot Distributions for Total Time (minutes) and Total Utterances. Statistically significant differences between clusters shown as * p < 0.05, ** p < 0.01, and *** p < 0.001.

When comparing these clusters, our sequence analysis allows us to see what tactics each group used more when *first* attempting to complete a task. S1, 6, 8, 10, and 13 all have both a start signal of Attempt 1 and a trial-and-error tactic. From S1, we see all groups commonly relied on the *Use More Info* tactic when starting a new task. However, S6, 8, 10, and 13 show us what clusters applied the *Use More Info*, *Simplification*, or *Repetition* tactic more so when initially experiencing an obstacle in their first attempt at a task. We see the *Explicits* relied on *Use More Info* initially the most and *Simplification* the least. *Repeaters* used *Repetition* more than *Use More Info* and *Simplification*. *Divergers* show no common pattern, while *Guessers* relied slightly more *Simplification* but seemed to rely on almost all the tactics initially.

With these clusters, we next checked for statistically significant differences for their SUS scores, total time spent on all tasks, and total utterances given. For their SUS scores, we found no statistically significant difference ($H = 3.99, p = 0.264$). *Divergers*, who showed no common pattern in what tactics to apply initially when encountering an obstacles, were more distributed in their satisfaction with *DiscoverCal*. *Explicits* had a SUS score average of 64.21 ± 20.46, *Repeaters* with 48 ± 14.99, *Divergers* with 53.46 ± 29.02, and *Guessers* with 58.75 ± 19.91. *Repeaters* used *Repetition* early in their first attempts at tasks and had a lower overall SUS score (although not statistically significant). This could indicate that participants who relied on *Repetition* earlier were less satisfied with the VUI.

Additionally, we found statistical significant differences when comparing the clustered participants' total time to complete the tasks ($H = 12.33, p = 0.0063$) and total utterances given during their session ($H = 19.03, p = 0.0002$). Their mean values and standard deviation are shown in Figure 1. Here we see the *Explicits* were consistently the more successful group along with the more distributed *Divergers*. The *Repeaters* spent more time on the tasks than the *Explicits* ($p = 0.0456$) and *Divergers* ($p = 0.0036$). *Repeaters* executed more utterances than *Explicits* ($p = 0.0041$) and *Divergers* ($p = 0.0016$). *Guessers* also executed more utterances than *Explicits* ($p = 0.0019$) and *Divergers* ($p = 0.0007$). Similar to their SUS scores, the *Divergers* had a wider distribution for total utterance executed and outliers in both directions seen in Figure 1. We believe the *Explicits*' success aligns with the trend in VUI performance since groups who are more explicit in their commands with the system can achieve better results. While the *Divergers*, who did not rely on our top sub-sequences as much as the other clusters, consisted of participant who found success while others found difficulties with their alternative behavior.

5 DISCUSSION

When reviewing our results, we see what tactics participants relied on first overall, what tactics our clusters of participants relied on first more frequently, and the benefits and weaknesses of sequence analysis for VUI usage data.

5.1 Relying Initially on Entity Revisions

We saw that participants first relied more on tactics that changed the entities of their original utterance when encountering errors over changing the keywords. For example, we see *Repetition*, *Use More Info*, and *Simplification* were commonly used in first attempts at tasks overall as an immediate response to the participant's previous utterance not working. While *Repetition* is the participant repeating the same utterance, *Use More Info* and *Simplification* change the number of entities provided or details of the entities in the utterances. We do not see our participants relying on *New Keyword* as frequently overall or in their first attempt.

Design Implications. We speculate this is because users were more confident that their utterance's keyword was correct and struggled more with finding correct wording for entities and the structure of the utterance overall. Multi-modal VUIs such as the Echo Show, Google Hub, and *DiscoverCal* provide example utterances in a menu for users to refer to. These example utterances start with their keyword (e.g., *Create* an event tomorrow) but the available entities users can edit (e.g., date and time) may be harder to discern. Multi-modal VUI research can explore more ways to clarify how to structure utterances and the options for entities. For example, VUI research can prototype a variety of entity design such as highlighting visually, highlighting placeholder entities (e.g., Create an event on [date]), and providing example entities (e.g., Create an event on Monday).

5.2 Comparing Cluster Behaviors

Additionally, we saw our clusters differed by what tactics they used more frequently when first attempting a task. Seeing differences in the initial interaction with a VUI could inform adaptive multi-modal

VUI guidance [17] or even generate personalized guidance based on user modeling. Looking at the *Divergers*, it is hard to see any patterns since they diverge from the top sub-sequences the most. We speculate this cluster is the "catch-all" and groups participants that did *not* use tactics because 1) they did not need them since they encountered fewer obstacles and 2) encountered obstacles but did not rely on tactics the same way the other clusters did. This could account for the *Divergers* large distribution of SUS score and total utterances executed since it would be combining proficient participants and participants struggling but not using common sub-sequences.

Design Implications. This highlights that while detecting tactics for adaptive multi-modal guidance could benefit participants such as those in the *Guessers*, *Explicits*, and *Repeaters* clusters, it would leave a sub-group of the *Divergers* without further aid. Multi-modal VUI adaptive guidance could attempt to detect the *absence* of tactics early in users' attempts at tasks. We speculate that this could isolate *Divergers* while other metrics, such as time spent completing a task or the success of a task (e.g., was the task accepted by the user with explicit confirmation), could further separate the more struggling users in the *Diverger* group. Once isolated, more guidance could be provided.

5.3 Recommendations for Multi-modal VUI Sequence Analysis

Reviewing our VUI sequence analysis, we highly recommend this method for VUI with contextual signals available. Our two contextual signals were attempt count and task number. We believe that VUI with even more contexts, for example, a multi-modal system with different screens, could benefit from sequence analysis even more. The labor required to label transcripts was intensive, but systems that can automatically inject contextual signals such as users switching display views and opening up menus can greatly benefit from this type of analysis and reduce the labor required. Other automatically generated actions such as ASR and NLP confidence scores could be explored as well. Datasets already labeled from conversation analysis methods, such as [10] could also apply this method. However, this method struggles with handling simultaneous events and we do not recommend for multi-user dialogues. We do not believe that VUI data that lacks contextual data would greatly benefit from this method. We speculate that without contextual data the sub-sequences would show a pattern of actions without the crucial information of when it started and what the pattern may be a reaction to.

6 LIMITATIONS AND FUTURE WORK

In this paper, we analyze performance metrics from a single-context, multi-modal VUI. Future studies could apply sequence analysis to VUIs with additional modalities (e.g., touch) or solely audio VUI for other domains. Additionally, we only label the participants' utterances in our corpus and not the responses of our VUI. We did this to focus on our participants' trial-and-error behavior but believe sequence analysis would also be a valid method to analyze users' responses to VUI feedback. As discussed, our work highlights participants' uncertainty in selecting correct entities. Future research can explore what approaches, or combination of approaches, can assist users in learning correct entities. Another approach could be for multi-modal VUI guidance to explore adapting to this pattern to detect obstacles and provide users with more guidance. Users' utterances could be compared to detect entity revisions and not keywords. By detecting this change, multi-modal VUIs could increase verbal or visual guidance instructing users on correct entities.

7 CONCLUSION

This paper is among the first to apply sequence analysis on the tactics users employ when testing utterances with an unfamiliar VUI. From usage data collected from a user study ($n = 50$), we present participants' top 16 sub-sequences and 4 clusters grouping participants by their top sub-sequence usage. Our results indicate participants relied more on tactics than editing the utterance structure and entities. We believe this indicates they were less confident in understanding correct entities over keywords initially. Additionally, we found participant clusters differed by what tactics they were more likely to use first when attempting a task or diverged from using the top sub-sequences. Based on these results, we propose VUI design implications for supporting users in better understanding the supported of entities for a VUI and adapting multi-modal VUIs to users who do not show trial-and-error behavior. Finally, we recommend sequence analysis as a method to analyze VUI user interaction and suggest that it will be most beneficial when augmented with rich contextual information.

REFERENCES

[1] Ágnes Abuczki. 2011. A multimodal analysis of the sequential organization of verbal and nonverbal interaction. *Argumentum* 7 (2011), 261–279.

[2] Rakesh Agrawal and Ramakrishnan Srikant. 1995. Mining sequential patterns. In *Proceedings of the eleventh international conference on data engineering*. IEEE, 3–14.

[3] Hua Ai, Diane J Litman, Kate Forbes-Riley, Mihai Rotaru, Joel Tetreault, and Amruta Purandare. 2006. Using system and user performance features to improve emotion detection in spoken tutoring dialogs. In *Ninth International Conference on Spoken Language Processing*.

[4] Erin Beneteau, Yini Guan, Olivia K Richards, Mingrui Ray Zhang, Julie A Kientz, Jason Yip, and Alexis Hiniker. 2020. Assumptions Checked: How Families Learn About and Use the Echo Dot. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–23.

[5] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottrjdge. 2018. Understanding the Long-Term Use of Smart Speaker Assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–24. https://doi.org/10.1145/3264901

[6] A. Bhargava, A. Celikyilmaz, D. Hakkani-Tür, and R. Sarikaya. 2013. Easy contextual intent prediction and slot detection. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 8337–8341. https://doi.org/10.1109/ICASSP.2013.6639291

[7] John Brooke. 1996. SUS A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[8] Minsuk Chang, Anh Truong, Oliver Wang, Maneesh Agrawala, and Juho Kim. 2019. How to design voice based navigation for how-to videos. *Conference on Human Factors in Computing Systems - Proceedings* (2019). https://doi.org/10.1145/3290605.3300931

[9] Janghee Cho and Janghee. 2018. Mental Models and Home Virtual Assistants (HVAs). *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18* (2018), 1–6. https://doi.org/10.1145/3170427.3180286

[10] Janghee Cho and Emilee Rader. 2020. The Role of Conversational Grounding in Supporting Symbiosis Between People and Digital Assistants. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–28.

[11] Ronald R Coifman and M Victor Wickerhauser. 1992. Entropy-based algorithms for best basis selection. *IEEE Transactions on information theory* 38, 2 (1992), 713–718.

[12] William Jay Conover and Ronald L Iman. 1979. On multiple-comparisons procedures. *Los Alamos Sci. Lab. Tech. Rep. LA-7677-MS* 1 (1979), 14.

[13] Eric Corbett and Astrid Weber. 2016. What Can I Say?: Addressing User Experience Challenges of a Mobile Voice User Interface for Accessibility. *Proceedings*

of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services (2016), 72–82. https://doi.org/10.1145/2935334.2935386

[14] Benjamin R Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "What can i help you with?": Infrequent Users' Experiences of Intelligent Personal Assistants. *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services - MobileHCI '17* (2017), 1–12. https://doi.org/10.1145/3098279.3098539

[15] Wayne W Daniel et al. 1990. Applied nonparametric statistics. (1990).

[16] Anders Drachen, Rafet Sifa, Christian Bauckhage, and Christian Thurau. 2012. Guns, swords and data: Clustering of player behavior in computer games in the wild. In *2012 IEEE conference on Computational Intelligence and Games (CIG)*. IEEE, 163–170.

[17] Anushay Furqan, Chelsea Myers, and Jichen Zhu. 2017. Learnability through Adaptive Discovery Tools in Voice User Interfaces. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '17* (2017), 1617–1623.

[18] Shiyoh Goetsu and Tetsuya Sakai. 2019. Voice input interface failures and frustration: Developer and user perspectives. *UIST 2019 Adjunct - Adjunct Publication of the 32nd Annual ACM Symposium on User Interface Software and Technology* (2019), 24–26. https://doi.org/10.1145/3332167.3357103

[19] Shiyoh Goetsu and Tetsuya Sakai. 2019. Voice Input Interface Failures and Frustration: Developer and User Perspectives. In *The Adjunct Publication of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 24–26. https://doi.org/10.1145/3332167.3357103

[20] Joseph F. Grafsgaard. 2014. Multimodal analysis and modeling of nonverbal behaviors during tutoring. *ICMI 2014 - Proceedings of the 2014 International Conference on Multimodal Interaction* (2014), 404–408. https://doi.org/10.1145/2663204.2667611

[21] Yuki Irie, Shigeki Matsubara, Nobuo Kawaguchi, Yukiko Yamaguchi, and Yasuyoshi Inagaki. 2004. Speech intention understanding based on decision tree learning. *8th International Conference on Spoken Language Processing, ICSLP 2004* August 2014 (2004), 2185–2188.

[22] Jiepu Jiang, Wei Jeng, and Daqing He. 2013. How do users respond to voice input errors? *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '13* (2013), 143. https://doi.org/10.1145/2484028.2484092

[23] Kristiina Jokinen, Kari Kanto, Antti Kerminen, and Jyrki Rissanen. 2004. Evaluation of Adaptivity and User Expertise in a Speech-Based E-Mail System. *Proc. of the 20th ACL International Conference on Computational Linguistic* March 2016 (2004), 44–52.

[24] Shin Jin Kang, Young Bin Kim, and Soo Kyun Kim. 2014. Analyzing repetitive action in game based on sequence pattern matching. *Journal of real-time image processing* 9, 3 (2014), 523–530.

[25] Pavan Kantharaju, Katelyn Alderfer, Jichen Zhu, Bruce Char, Brian Smith, and Santiago Ontanón. 2018. Tracing player knowledge in a parallel programming educational game. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 14.

[26] Pavan Kantharaju, Katelyn Alderfer, Jichen Zhu, Bruce Char, Brian Smith, and Santiago Ontanon. 2020. Modeling Player Knowledge in a Parallel Programming Educational Game. *IEEE Transactions on Games* (2020).

[27] Clare-Marie M Karat, Christine Halverson, Daniel Horn, and John Karat. 1999. Patterns of entry and correction in large vocabulary continuous speech recognition systems. *Proceedings of the SIGCHI conference on Human factors in computing systems the CHI is the limit - CHI '99* (1999), 568–575. https://doi.org/10.1145/302979.303160

[28] E Kastbjerg. 2011. *Combining sequence mining and heatmaps to visualize game event fl ows (working title)*. Ph.D. Dissertation. Master's thesis, IT University of Copenhagen, Copenhagen.

[29] Nobuo Kawaguchi, Shigeki Matsubara, Itsuki Kishida, Yuki Irie, Hiroya Murao, Yukiko Yamaguchi, Kazuya Takeda, and Fumitada Itakura. 2005. Construction and analysis of a multi-layered in-car spoken dialogue corpus. In *DSP for In-Vehicle and Mobile Systems*. Springer, 1–17.

[30] Jihyun Kim, Meuel Jeong, and Seul Chan Lee. 2019. " Why did this voice agent not understand me?" error recovery strategy for in-vehicle voice user interface. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings*. 146–150.

[31] Philipp Kirschthaler, Martin Porcheron, and Joel E Fischer. 2020. What Can I Say? Effects of Discoverability in VUIs on Task Performance and User Experience. In *Proceedings of the 2nd Conference on Conversational User Interfaces*. 1–9.

[32] Dounia Lahoual and Myriam Frejus. 2019. When Users Assist the Voice Assistants: From Supervision to Failure Resolution. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3290607.3299053

[33] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16* (2016), 5286–5297. https://doi.org/10.1145/2858036.2858288

[34] Sasha Makarovych, Alessandro Canossa, Julian Togelius, and Anders Drachen. 2018. Like a DNA string: Sequence-based player profiling in Tom Clancy's The Division. In *Artificial Intelligence and Interactive Digital Entertainment Conference*. York.

[35] Chelsea M Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for How Users Overcome Obstacles in Voice User Interfaces. 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, 1–7. https://doi.org/10.1145/3173574.3173580

[36] Chelsea M Myers, Anushay Furqan, and Jichen Zhu. 2019. The impact of user characteristics and preferences on performance with an unfamiliar voice user interface. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–9.

[37] Chelsea M Myers, Furqan Anushay, David Grethlein, and Jichen Zhu. 2019. Modeling Behavior Patterns with an Unfamiliar Voice User Interface. In *Proceedings of the 27th Conference on User Modeling, Adaptation and Personalization (UMAP '19)*. ACM, New York, NY, USA.

[38] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. *CHI '18* (2018), 1–12. https://doi.org/doi.org/10.1145/3173574.3174214

[39] Martin Porcheron, Joel E. Fischer, and Sarah Sharples. 2017. "Do Animals Have Accents?": Talking with Agents in Multi-Party Conversation. (2017), 207–219. https://doi.org/10.1145/2998181.2998298

[40] Rafet Sifa, Anders Drachen, and Christian Bauckhage. 2018. Profiling in games: Understanding behavior from telemetry. *Social interactions in virtual worlds: An interdisciplinary perspective* (2018).

[41] Rafet Sifa, Anders Drachen, Christian Bauckhage, Christian Thurau, and Alessandro Canossa. 2013. Behavior evolution in tomb raider underworld. In *2013 IEEE Conference on Computational Inteligence in Games (CIG)*. IEEE, 1–8.

[42] Arjun Srinivasan, Mira Dontcheva, Eytan Adar, Seth Walker, Ann Arbor, and Seth Walker. 2019. Discovering Natural Language Commands in Multimodal Interfaces. *Proceedings of the 24th International Conference on Intelligent User Interfaces - IUI '19* (2019), 661–672. https://doi.org/10.1145/3301275.3302292

[43] Stefan Ultes, Robert ElChab, and Wolfgang Minker. 2014. Application and evaluation of a conditioned hidden markov model for estimating interaction quality of spoken dialogue systems. In *Natural Interaction with Robots, Knowbots and Smartphones*. Springer, 303–312.

[44] Josep Valls-Vargas, Santiago Ontanón, and Jichen Zhu. 2015. Exploring player trace segmentation for dynamic play style prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 11.

[45] G Wallner. 2002. Sequential Pattern Mining Using Bitmaps. In *Proceedings of the Eigth ACM SIGKDD International Conference of Knowledge Discovery and Data Mining*. 429–435.

[46] Guenter Wallner. 2015. Sequential analysis of player behavior. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*. 349–358.

[47] Georgios N. Yannakakis, Pieter Spronck, Daniele Loiacono, and Elisabeth André. 2013. *Player Modeling*. Technical Report. 59 pages.

[48] Jichen Zhu, Katelyn Alderfer, Brian Smith, Bruce Char, and Santiago Ontañón. 2020. Understanding Learners' Problem-Solving Strategies in Concurrent and Parallel Programming: A Game-Based Approach. *arXiv preprint arXiv:2005.04789* (2020).

[49] Jichen Zhu and Santiago Ontañón. 2020. Player-Centered AI for Automatic Game Personalization: Open Problems. In *International Conference on the Foundations of Digital Games*. 1–8.