Linguistic Cognitive Load Analysis on Dialogues with an Intelligent Virtual Assistant

Mohammad Arvan,[1] **Mina Valizadeh,**[1] **Parian Haghighat,**[2]
Toan Nguyen,[2] **Heejin Jeong,**[3] and **Natalie Parde**[1]
[1] Department of Computer Science, University of Illinois at Chicago, USA
[2] Department of Mechanical and Industrial Engineering, University of Illinois at Chicago, USA
[3] Ira A. Fulton Schools of Engineering, Arizona State University, USA
{marvan3,mvaliz2,phaghi3,tnguy239,parde}@uic.edu
heejin.jeong@asu.edu

Abstract

Virtual assistants have become fixtures in everyday settings, but most research focuses on their development rather than their use following deployment. To facilitate study of their use in office settings, we introduce *OfficeDial*, a multimodal dataset containing audio recordings, transcriptions, eye tracking data, and screen recordings from conversations between humans and virtual assistants in office environments. Conversations are paired with physical and behavioral measures of cognitive load. We study the associations between verbal behavior and noise level and reveal key relationships between verbal redundancy, disfluency, and noise level. We make our new dataset available to interested researchers to inspire further exploration.

Keywords: cognitive load, conversational agents, dataset

Introduction

Virtual assistants are pervasive across day-to-day settings. The ways in which individuals interact with these agents and their motivations for doing so may vary depending on environmental factors such as surrounding company (e.g., exchanges alone in the home may differ from those in a crowded elevator), background task requirements (e.g., exchanges while folding laundry may differ from those while driving on a busy street), and ambient noise level (e.g., exchanges in a library may differ from those in a restaurant). The latter may hold strong implications for office use, since although noise level may ebb and flow throughout the day it is rare to experience long periods of silence. Noise level is known to negatively correlate with task performance (Sundstrom, Town, Rice, Osborn, & Brill, 1994), and it follows that it may negatively influence virtual agent use in office environments.

However, despite the extensive work undertaken to *develop* virtual agents for a wide range of tasks (McTear, Callejas, & Griol, 2016; H. Chen, Liu, Yin, & Tang, 2017; Sarikaya, 2017; Gao, Galley, & Li, 2019), comparatively less work has examined their actual use following deployment, and how that use may be affected by factors associated with the realities of daily living (Luger & Sellen, 2016). It is not known whether and how noise level impacts virtual assistant use and corresponding cognitive load in the office, although common sense suggests an inverse relationship (Sundstrom et al., 1994). This lack of understanding may contribute to negative attitudes towards virtual assistants in the workplace, slowing adoption (Hornung & Smolnik, 2021). A barrier to analyzing use of deployed systems has been the absence of available, relevant data logging interactions between virtual agents and individuals performing workplace tasks, and corresponding measurements of cognitive load.

We set out to fill this gap, by introducing a dataset of recorded, transcribed conversations between a virtual agent and 48 human subjects performing everyday workplace tasks. Each conversation is paired with quantitative multimodal measurements of cognitive load. Our contributions are as follows:

- We systematically collect data across multiple modalities and numerous workplace tasks at varying noise levels, with and without the aid of a virtual assistant, from 48 participants using a randomized experimental design.
- We segment and transcribe each interaction to facilitate downstream language analyses.
- We perform statistical analyses to compare facets of cognitive load between conditions, focusing on linguistic metrics.

It is our hope that this dataset will facilitate much-needed analyses of workplace virtual assistant interactions, positively influencing the development of future systems. We present the findings from our experiments as a benchmark to stimulate additional, more complex analyses of cognitive load. In the following sections we review and compare with other prior work in this domain, describe our data collection, experimental design, and linguistic analysis methods, report our findings, and establish key takeaways to guide future research.

Background

Although prior work has examined dialogue with virtual agents and workplace cognitive load separately, no datasets exist to facilitate joint accomplishment of those goals. We review relevant prior work for both.

Dialogue with Virtual Agents

Intelligent virtual assistants (IVAs) are ubiquitous in many facets of life and are commonly used to improve productivity across various domains (Eberhart, Bansal, & Mcmillan, 2020; Kaelin, Valizadeh, Salgado, Parde, & Khetani, 2021). High-quality datasets documenting their use may aid in our understanding of their roles in our lives. However, publicly available datasets in this area are limited in their size, linguistic diversity, annotation, and domain coverage (Peskov et al., 2019; Farzana, Valizadeh, & Parde, 2020; Valizadeh, Ranjbar-Noiey, Caragea, & Parde, 2021). Previously, Rastogi, Zang, Sunkara, Gupta, and Khaitan (2020) presented the *Schema-Guided Dialogue* (SGD) dataset consisting of over 20k annotated multi-domain, task-oriented conversations between a human and a virtual assistant. These conversations cover twenty domains, ranging from everyday tasks to more domain-specific tasks like banking assistance. *Taskmaster-1* and *Taskmaster-2* are two other datasets that were collected using a Wizard of Oz (WoZ) framework in which crowdsourced workers interacted with human operators playing the "virtual assistant" (Byrne et al., 2019). Taskmaster-1 and Taskmaster-2 contain 13k and 17k dialogues, respectively, in the domains of restaurants, food ordering, movies, hotels, flights, music, and sports.

In addition to these large, general datasets, Eric and Manning (2017) proposed a multi-turn and multi-domain task-oriented dialogue dataset with 3k conversations in three domains appropriate for an in-car assistant, including calendar scheduling, weather information retrieval, and point of interest navigation. There are also other available datasets in the domains of movies (Radlinski, Balog, Byrne, & Krishnamoorthi, 2019; Merdivan et al., 2020) and travel and telecommunications (Beaver, Freeman, & Mueen, 2020). However, there are currently no available datasets collected from workplace environments with varying noise levels. Grover, Rowan, Suh, McDuff, and Czerwinski (2020) presented a productivity virtual agent which helps users schedule and block out time on their calendar to focus on important tasks. They evaluated the performance of this agent through a three-week, within subjects study design with 40 participants, across different work roles in a large organization. Kim, de Melo, Norouzi, Bruder, and Welch (2020) investigated the effects of IVA embodiment on collaborative decision making, in which participants performed a task in three conditions: (1) alone, (2) with a disembodied virtual assistant, and (3) with an embodied assistant. Recently, Li and Yang (2021) also introduced an IVA for the manufacturing industry to handle a variety of complex services, such as order processing and production execution. They presented three scenarios to test the usability and flexibility of the agent regarding the manufacturing environment.

Multimodal Cognitive Load Measurement

The underlying goal when using IVAs in the workplace is to increase workers' performance at office tasks, often by reducing their cognitive load and freeing them to focus more fully on problems requiring their attention. This relationship between cognitive workload, or the quantified demand of a given task on the mental resources to process information (F. Chen et al., 2016), and performance has been found to take on an inverted U-shape (Veltman & Jansen, 2005; Babiloni, 2019). Measuring cognitive load is challenging; although numerous techniques have been explored for doing so, the search for improved measures remains an active area of research. Methods proposed so far that attempt to estimate the cognitive load of a given task include subjective, performance, behavioral, and physiological measures (Sweller, van Merriënboer, & Paas, 2019). Given our focus on dialogue, we center our review on behavioral measures from a linguistic perspective, and subsequently leverage these measures in our work. To the best of our knowledge, ours is the first work that focuses on analyzing the impact of noise on the behaviors associated with workplace interactions with IVAs from this perspective.

Most prior work examining cognitive load using linguistic measures has been done in the context of human-robot interaction. For instance, Schwarz and Fuchs (2017) present the design requirement, conceptual framework, and proof-of-concept implementation of a system, *Real-time Assessment of Multidimensional User State* (RASMUS), for evaluating user performance as it pertains to workload, attentional focus, and fatigue in real time. Sevcenko, Ninaus, Wortha, Moeller, and Gerjets (2021) measure cognitive load using in-game metrics. In their study, users completed a given simulation and rated their corresponding workload. The authors report these subjective workload measures, and demonstrate that gaming performance matches the proposed in-game metrics. Some researchers have turned to wearable technology and physiological sensors to measure workplace cognitive load. For example, Giorgi et al. (2021) utilized wearable technologies in place of laboratory technologies to assess mental workload, stress, and emotional state of users during workplace tasks. In similar work, Planke et al. (2021) developed a cyber-physical human that provides affordable physiological sensors to measure cognitive load.

Our dataset includes physiological measurements paired with transcribed dialogue and computed linguistic measures of cognitive load, enabling use of this information in follow-up studies. It is close in size and in its scope of cognitive load measures to several other datasets focused on multimodal cognitive load assessment. For example, *ZuCo* (Hollenstein et al., 2018), *ZuCo 2.0* (Hollenstein, Troendle, Zhang, & Langer, 2020), and *CopCo* (Hollenstein, Barrett, & Björnsdóttir, 2022) contain eye tracking and electroencephalography (EEG) data recorded from users reading natural sentences, and *CoLoSS* (Herms, Wirzberger, Eibl, & Rey, 2018) contains speech under cognitive load recorded in a learning task scenario. Altogether, *OfficeDial* provides speech, transcriptions, eye tracking data, and cognitive load measures based on linguistic behavior.

Methods

All of our data was collected from in-person experimental participants conducting workplace tasks (Haghighat et al., 2023). We note that while simulated office environments created for lab-based data collection may differ in some ways from actual office spaces, there is a high entry barrier to collecting data from real office environments due to privacy and intellectual property concerns. Because of this, many popular dialogue datasets, including but not limited to *MultiWOZ* (Budzianowski et al., 2018), *MMD* (Saha, Khapra, & Sankaranarayanan, 2018), and *TreeDST* (Cheng et al., 2020), fully or partially simulate aspects of dialogue generation.

Data Collection

Participants. This work was approved by the University of Illinois at Chicago's Institutional Review Board (#2019–1185). Forty-eight participants were recruited using flyers and social media posts. Participants were screened to ensure eligibility based on age (at least 18 years old), auditory and visual acuity (sufficient for typical computer and IVA use), physical ability to use a computer, and English proficiency. Before the experiment, all participants were briefed on the study and were asked to complete the consent form and a demographic survey. Participants ranged from 19 to 63 years old (M=26.1; SD=8.72) and consisted of native and non-native English speakers with different ethnic backgrounds. We surveyed participants on current and prior IVA experience, and found that experience varied widely: 22% of the participants stated that they never used an intelligent virtual assistant in their daily lives, whereas 36% reported having interacted with the specific IVA platform used in our study previously.

Apparatus. The study used the latest version of Microsoft Cortana[1] intelligent virtual assistant, and access to it was provided through the Windows 10 built-in application installed on a Dell desktop computer. Data collection was conducted within a short time period (two months) to prevent the effect of newer versions on users' task performance. An external microphone connected to the computer recognized the voice commands used. Non-verbal (52 dBA) and verbal noises (60 dBA) were collected from our institution's library and campus cafe and were played via a Bluetooth loudspeaker. The speaker was placed at a distance identical for all participants (~150 cm). The audio and screen were recorded during the experiment via the Google Chrome built-in screen recorder.

An eye-tracking system (Dikablis Glasses 3, Ergoneers GmbH[2]) was used to track and measure pupil dilation. The D-lab Ergoneers recorded the index of cognitive activity (ICA) at 60 Hz. The ICA, measured as a function of the number of rapid changes in pupil dilation over a given time frame, is known to be correlated with mental workload (Marshall, 2002). Participants' survey responses were collected using a 10.5-inch tablet (Samsung Galaxy Tab S5e).

Task Scenarios. Three office task scenarios were developed, and each included two sub-scenarios. We intentionally selected simple, common office tasks that we expected participants to easily understand. Removing this aspect of the study design would have reduced experimental control, making it more difficult to conclude whether observed increases in cognitive load originated from task characteristics themselves or from (our variable of interest) introduced acoustic noise. The tasks and sub-scenarios are defined here:

- **Scenario 1:** Asking an IVA to send an email.
  - *S1-1*: Sending an email to [a fictional name] to ask for feedback on a document sent previously.
  - *S1-2*: Sending an email to [a fictional name] to follow up on the previous email.
- **Scenario 2:** Asking an IVA to set up a timer/reminder.
  - *S2-1*: Setting up a timer for 20 minutes.
  - *S2-2*: Setting up a reminder to attend a meeting.
- **Scenario 3:** Asking an IVA to perform an internet search.
  - *S3-1*: Searching the nearest location of a given target store (e.g., UPS).
  - *S3-2*: Searching the phone number of a given target location (e.g., a university student center).

Name	**OfficeDial**	**ZuCo**
Words	24,934	21,629
Sentences	3019	1107

Table 1: Dataset statistics for *OfficeDial* and *ZuCo* (Hollenstein et al., 2018).

For subscenarios incorporating fictional names, participants were assigned a random name from the fictional names: *Kenny Stone*, *Mehdi Lake*, *Dean Rice*, *Samantha Stevens*, *Kevin Silva*, *Sarah Morgan*, *Alma Kerry*, *Luis Brady*, and *Maria Allison*. These names were selected for their diversity and ease of pronunciation across varying linguistic backgrounds.

Experimental Design and Procedure. To investigate performance and cognitive demand posed by interactions with an IVA in a noisy environment, the study employed a 2 (system) × 3 (noise level) design. The two systems were:

- **System A:** Performing tasks with the assistance of an IVA.
- **System B:** Performing tasks using only a keyboard and mouse.

Although comparisons of Systems A and B pose interesting opportunities for non-linguistic exploration (e.g., the collected ICA measures), they naturally offer limited means for between-system linguistic comparison. Thus, we focus our analysis in this paper on System A conditions. We selected three noise levels representing different noise conditions that might exist in an office-like environment:

- **N1:** Silence
- **N2:** Non-verbal noise
- **N3:** Verbal noise

All participants practiced the tasks (the six reported scenarios) before starting the experiment. They were informed that the goal was to achieve a task completion time as short as possible and an error rate as low as possible. The main experiment consisted of two sessions: one session to perform tasks via System A, and another session to perform tasks via System B. The order of the sessions was counterbalanced across participants. The order of tasks in each session was also counterbalanced to avoid learning and order effects.

48 participants were paid for 90 minutes of participation. Since all participants completed all conditions (although in varying orders), this resulted in 48 recordings for each (system) × (noise level) condition. The System A conversations were automatically transcribed, and the transcripts were manually quality-checked and edited for correctness by a member of the study team. Each recorded interaction was matched to its corresponding physical measures of cognitive load. In Table 1, we present our dataset statistics in comparison to ZuCo

[1] https://www.microsoft.com/en-us/cortana

[2] https://www.ergoneers.com/en/mobile-eye-tracker-dikablis-glasses-3/

(Hollenstein et al., 2018), a multimodal cognitive load assessment dataset collected from users reading natural sentences.

Data Analysis

Following data collection, we employed analysis techniques to further understand the recorded and transcribed conversations. We were primarily interested in the extent to which the collected data could provide evidence, either in support of or contrary to, the hypothesis that as noise level increased, so did users' cognitive loads when performing their tasks. Answering this question could guide future roles and technical development of workplace virtual assistants. We focused specifically on analysis using linguistic measures of cognitive load, leaving other analyses possible with this dataset (e.g., detailed investigations of saccades and pupil dilation) outside the present scope. Likewise, more complex comparisons of user behavior during interactions with different systems remain out of scope, with our study focusing on System A interactions.

Linguistic measures of cognitive load are behavioral measures that provide quantifiable assessment of voluntary user activity, as observed via spoken or written language while performing a given task. Some linguistic features associated with cognitive load in prior work include measures of spoken disfluency, articulation rate, and filler and pause rates (Farzana, Deshpande, & Parde, 2022). We selected a variety of measures that are relevant for assessing characteristics of utterances situated in two-party dialogue, summarized in the following subsections. Since our dataset contains conversations between a user and an IVA, the frequencies of some categories of words (e.g., emotion words, swear words, and words pertaining to cognition) were extremely low, leading us to exclude measurements that relied on those words. Broadly construed, our linguistic measures assessed disfluency behaviors, verbosity, and language complexity.

Disfluency Count. Pauses, repetitions, and corrections do not correspond to specific meaning but are often present in spoken dialogue, and their prevalence may correlate with cognitive load (Berthold & Jameson, 1999; Müller, Großmann-Hutter, Jameson, Rummer, et al., 2001; Khawaja, Chen, & Marcus, 2014) or other cognitive disfunction (Farzana et al., 2022). Unfortunately, these markers of verbal disfluency are often absent from transcripts, making empirical studies of these behaviors challenging (Umair, Mertens, Albert, & de Ruiter, 2022). In conversations with IVAs, dialogues exhibit many repetitions and self-corrections of phrases and utterances. We group different forms of verbal disfluency into a single category (*[REP]*) and manually insert these disfluency tags into the transcripts, shown in the following examples:

> *Hi Cortana [REP] Hi Cortana [REP] Hi Cortana [REP] Hi Cortana can you send an email.*

> *Hey Cortana. [REP] Hey Cortana. Set up a reminder. [REP] Hey Cortana. [REP] Hey Cortana. Set up a reminder.*

Since it has previously been observed that increased verbal disfluency is associated with higher cognitive load (Berthold & Jameson, 1999; Müller et al., 2001; Khawaja et al., 2014), we anticipated that disfluency frequency would increase with noise level (N1 → N2 → N3) in our dataset.

Verbosity. Variations in cognitive load may also manifest as changes in verbosity, often as a result of individuals using language as a coping mechanism to manage mental load (Sexton & Helmreich, 2000). Prior work specifically has found evidence that increased verbosity, observable as greater word count and average number of words per sentence, correlates with increased cognitive load. We compute two measures of verbosity for each conversation:

- **Word Count (WC):** The number of words spoken by the human participant while performing their task.
- **Words Per Sentence (WPS):** The average number of words per sentence spoken by the human performing their task.

We anticipated that the trends observed in prior work would hold true in our dataset as well.

Language Complexity. Finally, variations in cognitive load may be associated with changes in language complexity. For example, individuals may alter their word choice or vary their sentence structure due to external distraction or insufficient mental processing resources, both of which may influence verbal fluency or recall (Khawaja et al., 2014). A wide range of measures have been proposed for evaluating language complexity, including but not limited to type-token ratio (Chotlos, 1944; Templin, 1957; Ure, 1971), Gunning Fog Index (Gunning, 1952), Flesch-Kincaid Grade (Flesch, 1948), and SMOG Index (Laughlin, 1969). We describe our selected measures of language complexity below. The measures of language complexity fall under two broad categories: *type-token ratio* provides an estimate of lexical variety or verbal redundancy, whereas the *Gunning Fog Index*, *Flesch-Kincaid Grade*, and *SMOG Index* capture lexical complexity.

Type-Token Ratio (TTR). The simplest and most common measure of lexical variety is type-token ratio (TTR). TTR is the ratio of unique words over total words in the given text:

$$\mathrm{TTR} = \frac{N_{\text{unique words}}}{N_{\text{words}}} \tag{1}$$

We expected TTR to decrease as the noise level increased.

Lexical Complexity. In contrast to TTR which operates without regard to word form or content, several measures consider lexical complexity to be dependent on word length. The Gunning Fog Index (Gunning, 1952) is one such measure that originally sought to estimate the years of formal education required to comprehend a given text, based on the ratio of complex words (assumed to be those with three or more syllables) to words in general:

$$\mathrm{GFI} = 0.4 \times \frac{N_{\text{words}}}{N_{\text{sentences}}} + 100\frac{N_{\text{complex words}}}{N_{\text{words}}} \tag{2}$$

Metric	N1	N2	N3
Disf.[†]	1.01 ± 2.03	1.15 ± 1.98	2.71 ± 2.70
WC[†]	55.63 ± 54.12	57.65 ± 45.64	82.22 ± 69.00
WPS	8.62 ± 3.14	8.20 ± 2.20	8.42 ± 2.03
Terms[†]	30.76 ± 10.47	30.69 ± 10.48	35.25 ± 13.38
TTR[†]	0.67 ± 0.16	0.63 ± 0.16	0.57 ± 0.21
FKG	2.84 ± 2.13	2.50 ± 1.67	2.66 ± 1.54
SMOG	4.81 ± 2.51	4.64 ± 2.31	5.04 ± 2.34
GFI	4.55 ± 1.68	4.12 ± 1.29	4.21 ± 1.25

Table 2: Comparisons between N1, N2, and N3 using all included measures of language complexity and standard deviations. [†] indicates statistically significant differences ($p < 0.05$ with a one-way repeated measures ANOVA) in at least one comparison under a given metric.

The closely related Flesch-Kincaid Grade (Flesch, 1948) also incorporates syllable length as a proxy for complexity:

$$\text{FKG} = 0.39 \times \frac{N_{\text{words}}}{N_{\text{sentences}}} + 11.8 \times \frac{N_{\text{syllables}}}{N_{\text{words}}} - 15.59 \quad (3)$$

One final syllable-based measure of lexical complexity that has gained popularity in numerous settings is the Simple Measure of Gobbledygook Readability Formula (Laughlin, 1969), or SMOG Index. Similarly to the Gunning Fog Index, the SMOG Index assumes that complex words are those containing three or more syllables and computes a score as follows:

$$\text{SMOG} = \frac{N_{\text{complex words}}}{N_{\text{sentences}}} \times 30 + 3 \quad (4)$$

We expected that all measures of lexical complexity would correlate inversely with noise level.

Results

We computed the specified measures for each System A condition (N1, N2, and N3) in our dataset, and present the outcomes in Table 2. We measure statistical significance using standard one-way repeated measures analysis of variance (ANOVA) tests, accepting comparisons with $p < 0.05$ as statistically significant. We observe statistically significant differences using the following metrics: *disfluency*, *WC*, *terms*, and *TTR*.

The ANOVA test for *disfluency* resulted in $F2,282 = 16.39$, $F_{critical} = 3.02$, $p < .00001$. Post-hoc Tukey tests (Tukey, 1949) produced $p = 0.906$ (N1, N2), $p < 0.00001$ (N2, N3), and $p < 0.00001$ (N1, N3), illustrated in the box plot in Figure 1. Interestingly, although we anticipated that subjects would use more words and have a higher ratio for words per sentence as noise level increased, only one of these hypotheses held true. While *WC* correlated with noise level ($F2,282 = 6.33$, $F_{critical} = 3.02$, $p < 0.01$), measures of *words per sentence* were too close to draw significant conclusions. As a contrast to the earlier plot, we examine this finding further in Figure 2. Post-hoc tests on *WC* revealed significant differences between (N2, N3) and (N1, N3), both with $p < 0.01$, but not between (N1, N2) with $p = 0.98$.

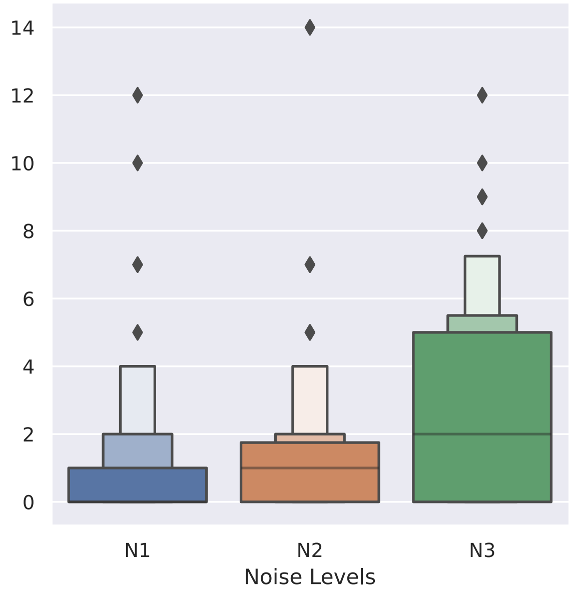

Figure 1: *Disfluency* for each noise level.

Our ANOVA test for *TTR* ($F2,282 = 7.9$, $F_{critical} = 3.02$, $p < 0.001$) followed a similar pattern to *disfluency*, indicating differences between conditions. Tukey's post-hoc test showed that with the exception of (N1, N2) with $p = 0.291$, the cognitive loads of the other two pairs, (N1, N3) and (N2, N3) both with $p < 0.01$, differed significantly, revealing an inverse correlation with noise level. Surprisingly, analyses performed on the lexical complexity indices did not follow our expectations. With $p = 0.41$, $p = 0.50$, and $p = 0.09$, the *Flesch Kincaid Grade*, *SMOG Index*, and *Gunning Fog Index* were not found to significantly increase with increased noise levels. Additionally, the scores for these three indices (often also associated with education level) suggest that regardless of noise level, understanding recorded conversations with IVAs requires minimal education.

Discussion

The findings from our analyses convey an interesting and consistent portrait of noise conditions and their associated influence on workplace use of IVAs. We find significant differences between two or more noise conditions as measured by *WC*, *terms*, *TTR*, and *disfluency*, with significant effects between (N2, N3) and (N1, N3) confirmed for all four measures in a post-hoc Tukey test. On the other hand, we did not observe significant differences between any noise conditions using the *Flesch Kincaid Grade*, *SMOG Index*, or *Gunning Fog Index*. This contradicts previous studies (F. Chen et al., 2016), leading us to suspect that there is still much to explore among the unique research opportunities provided by this dataset.

The clear divide between measures of verbosity and verbal redundancy (*WC*, *terms*, and *TTR*) and lexical complexity

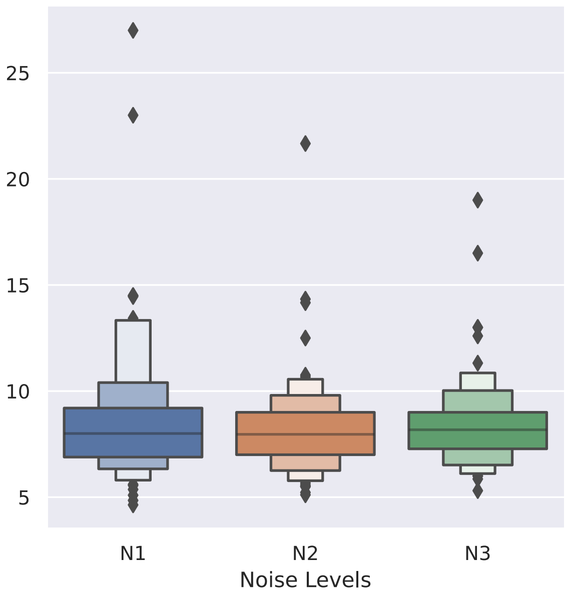

Figure 2: *Words per sentence* (WPS) for each noise level.

(*FKG*, *SMOG*, and *GFI*) reveals a strong relationship between noise level and required (or perceived requirement of) verbal load, without necessitating changes to word choice itself. That is, although individuals may adapt their syntactic structure or conciseness based on environment or their perception of the IVA's ability to understand, they do not feel a complementary need to adapt their lexicon in terms of length or complexity. The difference observed under the *disfluency* measure sheds further light on this. In many cases, the observed increase in *WC* and *terms* and decrease in *TTR* as noise level increases may result from an associated increase in verbal *disfluency* at greater noise levels, as individuals are forced to repair and repeat their utterances. For example, consider the utterance spoken in a quiet environment (e.g., N1):

> *Can you set a timer for twenty minutes?*

This may be compared to a counterpart in a noisy environment (e.g., N3) that interferes with spoken communication:

> *Can you set can you can you set a timer for twenty um twenty minutes?*

Thus, the outcomes of our analyses raise an important question: *What is necessary to promote the increased use of IVAs in workplace environments, such that their potential benefits can be realized?* Our findings reveal several recommendations. First, speech recognition quality is important, particularly in noisy environments. Under-performing speech recognizers may force users to repeat, repair, and otherwise edit their dialogue, reducing IVA efficiency and user experience. Second, workplace IVAs should anticipate and incorporate effective natural language understanding methods to interpret and leverage disfluent speech. Although preventing users from producing disfluencies through successful automated speech recognition may be the first form of defense against poor user experience, users may still be likelier to produce disfluencies in noisy environments due to increased situational distraction. Finally, IVAs should avoid anticipating simpler terms following disfluency detection. Despite an observed relationship between verbosity and noise level as measured by *WC*, *terms*, and *TTR*, there is not a corresponding association between noise level and lexical selection (as quantified by *FKG*, *SMOG*, or *GFI*).

Limitations

This work has three main limitations. First, the experiments do not provide comprehensive coverage of all potential workplace tasks. Although we developed task scenarios that are common across many office environments, it is unclear whether the identified relationships between noise level and linguistic measures of verbal load and language complexity would generalize to other tasks. Second, our analyses do not examine all possible linguistic measures of verbal behavior. Although we selected well-known measures that have historically demonstrated interesting correlations with cognitive load (Chotlos, 1944; Flesch, 1948; Laughlin, 1969; Gunning, 1952), we may be missing connections that are best demonstrated using other measures. Finally, our dataset is small relative to general-domain dialogue datasets and is limited to English. Collecting follow-up data allowing us to study generalizability across task and language is an intriguing avenue for future work.

Conclusion

In this work, we introduce *OfficeDial*, a new multimodal dataset comprising conversations between virtual assistants and human participants as they perform everyday office tasks captured at varying noise levels. The dataset includes manually-corrected transcripts, physical measures of cognitive load (i.e., pupil dilation, saccades, and index of cognitive activity), and computed linguistic behavioral measures. We make this dataset available for public access through Zenodo (Arvan et al., 2023), offering a novel resource for the study of dialogue between humans and virtual assistants and furthermore its intersection with cognitive load.

Through statistical analyses, we reveal interesting associations between verbal behavior and ambient noise level. Specifically, we find increases in *WC*, *terms*, and *disfluencies* as noise level increases, as well as an inverse relationship with *TTR*. We find no significant effects using measures of lexical complexity. This suggests that individuals adapt their overall verbal load to handle challenges that arise at greater noise levels when conversing with IVAs in office environments, but they do not correspondingly simplify their word choice. It is our hope that *OfficeDial* and our included analyses will inspire other researchers to further extend our collective understanding of the relationship between verbal behavior, noise level, and cognitive load during conversations with intelligent virtual assistants in office environments.

Acknowledgments

This research was supported by a research grant from the Office Ergonomics Research Committee (OERC). We thank the anonymous reviewers for their helpful suggestions, which were incorporated into the final version of this manuscript.

References

Arvan, M., Valizadeh, M., Haghighat, P., Nguyen, T., Jeong, H., & Parde, N. (2023, May). *Officedial dataset.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.7922480 doi: 10.5281/zenodo.7922480

Babiloni, F. (2019). Mental workload monitoring: New perspectives from neuroscience. In L. Longo & M. C. Leva (Eds.), *Human mental workload: Models and applications - third international symposium, H-WORKLOAD 2019, rome, italy, november 14-15, 2019, proceedings* (Vol. 1107, pp. 3–19). Springer. Retrieved from https://doi.org/10.1007/978-3-030-32423-0_1 doi: 10.1007/978-3-030-32423-0_1

Beaver, I., Freeman, C., & Mueen, A. (2020). Towards awareness of human relational strategies in virtual agents. In *Thirty-fourth aaai conference on artificial intelligence.*

Berthold, A., & Jameson, A. (1999). Interpreting symptoms of cognitive load in speech input. In *Um99 user modeling* (pp. 235–244). Springer.

Budzianowski, P., Wen, T., Tseng, B., Casanueva, I., Ultes, S., Ramadan, O., & Gasic, M. (2018). Multiwoz - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 conference on empirical methods in natural language processing, brussels, belgium, october 31 - november 4, 2018* (pp. 5016–5026). Association for Computational Linguistics. Retrieved from https://aclanthology.org/D18-1547/

Byrne, B., Krishnamoorthi, K., Sankar, C., Neelakantan, A., Duckworth, D., Yavuz, S., … Kim, K. (2019). Taskmaster-1: Toward a realistic and diverse dialog dataset. *CoRR, abs/1909.05358*. Retrieved from http://arxiv.org/abs/1909.05358

Chen, F., Zhou, J., Wang, Y., Yu, K., Arshad, S. Z., Khawaji, A., & Conway, D. (2016). *Robust multimodal cognitive load measurement*. Springer. Retrieved from https://doi.org/10.1007/978-3-319-31700-7 doi: 10.1007/978-3-319-31700-7

Chen, H., Liu, X., Yin, D., & Tang, J. (2017, November). A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explor. Newsl.*, *19*(2), 25–35. Retrieved from https://doi.org/10.1145/3166054.3166058 doi: 10.1145/3166054.3166058

Cheng, J., Agrawal, D., Alonso, H. M., Bhargava, S., Driesen, J., Flego, F., … Johannsen, A. (2020). Conversational semantic parsing for dialog state tracking. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing, EMNLP 2020, online, november 16-20, 2020* (pp. 8107–8117). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/2020.emnlp-main.651 doi: 10.18653/v1/2020.emnlp-main.651

Chotlos, J. W. (1944). Iv. a statistical and comparative analysis of individual written language samples. *Psychological Monographs*, *56*(2), 75.

Eberhart, Z., Bansal, A., & Mcmillan, C. (2020). A wizard of oz study simulating api usage dialogues with a virtual assistant. *IEEE Transactions on Software Engineering*, 1-1. doi: 10.1109/TSE.2020.3040935

Eric, M., & Manning, C. D. (2017). Key-value retrieval networks for task-oriented dialogue. *CoRR*, *abs/1705.05414*. Retrieved from http://arxiv.org/abs/1705.05414

Farzana, S., Deshpande, A., & Parde, N. (2022, May). How you say it matters: Measuring the impact of verbal disfluency tags on automated dementia detection. In *Proceedings of the 21st workshop on biomedical language processing* (pp. 37–48). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.bionlp-1.4

Farzana, S., Valizadeh, M., & Parde, N. (2020, May). Modeling dialogue in conversational cognitive health screening interviews. In *Proceedings of the 12th language resources and evaluation conference* (pp. 1167–1177). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.lrec-1.147

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, *32*(3), 221.

Gao, J., Galley, M., & Li, L. (2019). Neural approaches to conversational ai. *Foundations and Trends® in Information Retrieval*, *13*(2-3), 127-298. Retrieved from http://dx.doi.org/10.1561/1500000074 doi: 10.1561/1500000074

Giorgi, A., Ronca, V., Vozzi, A., Sciaraffa, N., Florio, A. D., Tamborra, L., … Borghini, G. (2021). Wearable technologies for mental workload, stress, and emotional state assessment during working-like tasks: A comparison with laboratory technologies. *Sensors*, *21*(7), 2332. Retrieved from https://doi.org/10.3390/s21072332 doi: 10.3390/s21072332

Grover, T., Rowan, K., Suh, J., McDuff, D., & Czerwinski, M. (2020). Design and evaluation of intelligent agent prototypes for assistance with focus and productivity at work. In *Proceedings of the 25th international conference on intelligent user interfaces* (p. 390–400). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3377325.3377507 doi: 10.1145/3377325.3377507

Gunning, R. (1952). *Technique of clear writing*. McGraw-Hill.

Haghighat, P., Nguyen, T. V., Valizadeh, M., Arvan, M., Parde, N., Kim, M., & Jeong, H. (2023). Effects of an intelligent virtual assistant on office task performance and workload in a noisy environment. *Applied ergonomics*, *109*, 103969.

Herms, R., Wirzberger, M., Eibl, M., & Rey, G. D. (2018).

Coloss: Cognitive load corpus with speech and performance data from a symbol-digit dual-task. In N. Calzolari et al. (Eds.), *Proceedings of the eleventh international conference on language resources and evaluation, LREC 2018, miyazaki, japan, may 7-12, 2018.* European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2018/summaries/1008.html

Hollenstein, N., Barrett, M., & Björnsdóttir, M. (2022). The copenhagen corpus of eye tracking recordings from natural reading of danish texts. *arXiv preprint arXiv:2204.13311.*

Hollenstein, N., Rotsztejn, J., Troendle, M., Pedroni, A., Zhang, C., & Langer, N. (2018). Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, *5*(1), 1–13.

Hollenstein, N., Troendle, M., Zhang, C., & Langer, N. (2020). Zuco 2.0: A dataset of physiological recordings during natural reading and annotation. In N. Calzolari et al. (Eds.), *Proceedings of the 12th language resources and evaluation conference, LREC 2020, marseille, france, may 11-16, 2020* (pp. 138–146). European Language Resources Association. Retrieved from https://aclanthology.org/2020.lrec-1.18/

Hornung, O., & Smolnik, S. (2021). Ai invading the workplace: negative emotions towards the organizational use of personal virtual assistants. *Electronic Markets*. Retrieved from https://doi.org/10.1007/s12525-021-00493-0 doi: 10.1007/s12525-021-00493-0

Kaelin, V. C., Valizadeh, M., Salgado, Z., Parde, N., & Khetani, M. A. (2021, Nov 4). Artificial intelligence in rehabilitation targeting the participation of children and youth with disabilities: Scoping review. *J Med Internet Res*, *23*(11), e25745. Retrieved from https://doi.org/10.2196/25745 doi: 10.2196/25745

Khawaja, M. A., Chen, F., & Marcus, N. (2014). Measuring cognitive load using linguistic features: Implications for usability evaluation and adaptive interaction design. *Int. J. Hum. Comput. Interact.*, *30*(5), 343–368. Retrieved from https://doi.org/10.1080/10447318.2013.860579 doi: 10.1080/10447318.2013.860579

Kim, K., de Melo, C. M., Norouzi, N., Bruder, G., & Welch, G. F. (2020). Reducing task load with an embodied intelligent virtual assistant for improved performance in collaborative decision making. In *2020 ieee conference on virtual reality and 3d user interfaces (vr)* (p. 529-538). doi: 10.1109/VR46266.2020.00074

Laughlin, G. H. M. (1969). Smog grading-a new readability formula. *Journal of Reading*, *12*(8), 639–646. Retrieved 2022-05-03, from http://www.jstor.org/stable/40011226

Li, C., & Yang, H. (2021, April 19). Bot-x: An ai-based virtual assistant for intelligent manufacturing. *Multiagent and Grid Systems*, *17*(1), 1–14. doi: 10.3233/MGS-210340

Luger, E., & Sellen, A. (2016). "Like having a really bad PA": The gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 chi conference on human factors in computing systems* (p. 5286–5297). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2858036.2858288 doi: 10.1145/2858036.2858288

Marshall, S. (2002). The index of cognitive activity: measuring cognitive workload. In *Proceedings of the ieee 7th conference on human factors and power plants* (p. 7-7). doi: 10.1109/HFPP.2002.1042860

McTear, M., Callejas, Z., & Griol, D. (2016). *The conversational interface: Talking to smart devices* (1st ed.). Springer Publishing Company, Incorporated.

Merdivan, E., Singh, D., Hanke, S., Kropf, J., Holzinger, A., & Geist, M. (2020, February). Human Annotated Dialogues Dataset for Natural Conversational Agents. *Applied Sciences*, *10*(3), 762. Retrieved from https://hal.archives-ouvertes.fr/hal-03081727 doi: 10.3390/app10030762

Müller, C., Großmann-Hutter, B., Jameson, A., Rummer, R., et al. (2001). Recognizing time pressure and cognitive load on the basis of speech: An experimental study. In *International conference on user modeling* (pp. 24–33).

Peskov, D., Clarke, N., Krone, J., Fodor, B., Zhang, Y., Youssef, A., & Diab, M. (2019, November). Multi-domain goal-oriented dialogues (MultiDoGO): Strategies toward curating and annotating large scale dialogue data. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 4526–4536). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1460 doi: 10.18653/v1/D19-1460

Planke, L., Lim, Y., Gardi, A., Sabatini, R., Kistan, T., & Ezer, N. (2020). A cyber-physical-human system for one-to-many UAS operations: Cognitive load analysis. *Sensors*, *20*(19), 5467. Retrieved from https://doi.org/10.3390/s20195467 doi: 10.3390/s20195467

Radlinski, F., Balog, K., Byrne, B., & Krishnamoorthi, K. (2019). Coached conversational preference elicitation: A case study in understanding movie preferences. In *Proceedings of the annual meeting of the special interest group on discourse and dialogue (SIGDIAL).*

Rastogi, A., Zang, X., Sunkara, S., Gupta, R., & Khaitan, P. (2020, Apr.). Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 8689-8696. Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/6394 doi: 10.1609/aaai.v34i05.6394

Saha, A., Khapra, M. M., & Sankaranarayanan, K. (2018). Towards building large scale multimodal domain-aware conversation systems. In S. A. McIlraith & K. Q. Weinberger (Eds.), *Proceedings of the thirty-second AAAI conference on artificial intelligence, (aaai-18), the 30th innovative ap-*

plications of artificial intelligence (iaai-18), and the 8th AAAI symposium on educational advances in artificial intelligence (eaai-18), new orleans, louisiana, usa, february 2-7, 2018 (pp. 696–704). AAAI Press. Retrieved from https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17104

Sarikaya, R. (2017). The technology behind personal digital assistants: An overview of the system architecture and key components. *IEEE Signal Processing Magazine*, *34*(1), 67-81. doi: 10.1109/MSP.2016.2617341

Schwarz, J., & Fuchs, S. (2017). Multidimensional real-time assessment of user state and performance to trigger dynamic system adaptation. In D. D. Schmorrow & C. M. Fidopiastis (Eds.), *Augmented cognition. neurocognition and machine learning - 11th international conference, AC 2017, held as part of HCI international 2017, vancouver, bc, canada, july 9-14, 2017, proceedings, part I* (Vol. 10284, pp. 383–398). Springer. Retrieved from https://doi.org/10.1007/978-3-319-58628-1_30 doi: 10.1007/978-3-319-58628-1_30

Sevcenko, N., Ninaus, M., Wortha, F., Moeller, K., & Gerjets, P. (2021). Measuring cognitive load using in-game metrics of a serious simulation game. *Frontiers in Psychology*, *12*, 906.

Sexton, J. B., & Helmreich, R. L. (2000). Analyzing cockpit communications: the links between language, performance, error, and workload. *Human Performance in Extreme Environments*, *5*(1), 63–68.

Sundstrom, E., Town, J. P., Rice, R. W., Osborn, D. P., & Brill, M. (1994). Office noise, satisfaction, and performance. *Environment and Behavior*, *26*(2), 195-222. Retrieved from https://doi.org/10.1177/001391659402600204 doi: 10.1177/001391659402600204

Sweller, J., van Merriënboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, *31*(2), 261–292.

Templin, M. C. (1957). *Certain language skills in children: Their development and interrelationships* (Vol. 10). JSTOR.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, *5*(2), 99–114. Retrieved 2022-05-17, from http://www.jstor.org/stable/3001913

Umair, M., Mertens, J., Albert, S., & de Ruiter, J. P. (2022). Gailbot: An automatic transcription system for conversation analysis. *Dialogue & Discourse*, *13*(1). doi: 10.5210/dad.2022.103

Ure, J. (1971). Lexical density and register differentiation. *Applications of linguistics*, *23*(7), 443–452.

Valizadeh, M., Ranjbar-Noiey, P., Caragea, C., & Parde, N. (2021, June). Identifying medical self-disclosure in online communities. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 4398–4408). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.naacl-main.347 doi: 10.18653/v1/2021.naacl-main.347

Veltman, J., & Jansen, C. (2005). *The role of operator state assessment in adaptive automation* (Tech. Rep.). TNO DEFENCE SECURITY AND SAFETY SOESTERBERG (NETHERLANDS).