HOW MAY I INTERRUPT? LINGUISTIC DESIGN GUIDELINES FOR PROACTIVE IN-CAR VOICE ASSISTANTS

Anna-Maria Meck

BMW Group, Institute of Phonetics and Speech Processing, Ludwig Maximilian University of Munich
anna-maria.meck@bmw.de

Abstract: Proactivity is a sought-after feature in Voice Assistants (VA) and considered the next sensible step in growing from merely reactive to more conversational interfaces. Creating proactive use cases thereby poses a challenging design task with a multitude of potential pitfalls. Proactive interactions possibly interrupt users who are already conducting important, demanding, and potentially even security-relevant primary tasks, e.g., driving a car. So far, research has identified precise timing and careful consideration of primary task engagement as key to successful proactive interactions. While there is substantial literature on proactive VAs in terms of when to interrupt users, how to interrupt them has received less attention. Previous research has shown that VA users are susceptible to differing formulations of VA system outputs (prompts) though and that they prefer some formulations over others. Syntactical, grammatical, and lexical nuances play a role in how a VA is being perceived. To close the gap of to date insufficient linguistic-driven design guidelines for proactive VAs, a crowdsourcing study was conducted to examine users' formulation preferences for proactive prompts in an automotive setting. Our findings show concrete syntactical best practices for formulating proactive in-car prompts, thereby allowing for the compilation of hands-on design guidelines.

1 Introduction

Researchers and industry alike agree on proactivity as the next sensible step forward for voice assistants (VAs), making a leap from merely reactive to proactive assistants [1; 2]. Nothdurft defines proactivity as "an autonomous, anticipatory system-initiated behavior, with the purpose to act in advance of a future situation, rather than only reacting to it" [1]. Contrary to this definition, current VAs are primarily designed to respond to their users' wishes with little or no proactive behavior, leading to a "pull paradigm" [3] with interactions originating from the users' side. Although this practice prevents unexpected VA behavior and distractions – crucial for environments where distractions are potentially critical like e.g., automotive settings – proactivity is a most interesting use case for VAs and popular among users. Studies have shown that users wish for proactivity in VAs, with subjects indicating proactive behavior as being rather or even extremely important for them [4; 5]. While these studies show the importance of proactive features, it needs mentioning that designing these features is an ambitious design task. Proactive interactions are possibly disrupting people who are already conducting important, demanding, and potentially even security-relevant primary tasks. Precise timing of proactive interactions at moments which are opportune for users constitutes a major design concern which should be addressed through context-aware and context-sensitive design strategies. Besides awaiting opportune moments for engaging in proactive behavior, VA designers need a legitimate and beneficial reason for interrupting and weigh potential costs and benefits of proactive interactions thoroughly.

There is substantial literature on proactive voice assistants in terms of interruptibility, intrusiveness, and primary task engagement. This research answers the question of *when* to interrupt users. Best practices for formulating proactive VA responses, so *how* to interrupt users, has

to the best of our knowledge not been studied in detail yet. Previous studies have shown an impact and an effect of syntax, grammar, and wording on users' evaluations of prompts though [6; 7]. By means of a crowdsourcing study, this paper examines user preferences regarding the formulation of proactive prompts in an automotive setting. Prompts were modified regarding syntactical, grammatical, and lexical parameters to identify linguistically supported best practices when designing proactive VA prompts. This work thereby closes the gap of to date insufficient design guidelines for formulating proactive prompts by providing concrete linguistic design guidance for proactive in-car interactions.

2 Related Work

2.1 Proactivity

According to Kraus et al., proactivity fosters a greater sense of team and team performance as well as trust towards a VA compared to merely reactive VAs. In Kraus' studies, different levels of proactivity had a significant impact on users' trust towards a VA. Low- and medium-level proactive system actions (meaning assistants providing notifications and suggestions) were trusted more than their strictly reactive baseline counterparts. Kraus et al. conclude that proactivity can relieve stress in challenging situations by confirming or reinforcing decision making processes during task execution. Especially for novice users, proactivity was found to be a meaningful feature [8].

While proactivity is oftentimes perceived as useful and helpful, there are users who judge autonomous speech outputs to be unwelcome intrusions. As Reicherts et al. put it, "proactive VAs need to strike the right balance between being helpful and being intrusive" [2]. Research around proactivity is therefore oftentimes concerned with users' "interruptibility", meaning timing and circumstances of opportune moments for proactive interactions. Interruptions at opportune moments can facilitate the engagement of users with the recommended proactive content as "timing and relevance of proactive services are critical to the user experience" [9]. In general, interruptibility requires auditory and verbal channel availability [9], but there is a diverse set of factors playing into interruptibility for proactive interactions. Comparing agent-initiated and user-initiated interactions, Reicherts et al. found proactivity to be rated best when not interrupting already ongoing interactions, if a user was alone as opposed to in a group setting, and if a message was suggestive rather than imposed [2]. Iqbal et al. studied effects of low and high mental workload on resumption lag, annoyance, and social attribution and found moments with low workload from a primary task to be so-called "best moments" for interrupting users. Current involvement in high workload primary tasks was accordingly labelled a "worst moment" for interrupting. Interruptions in "best moments" led to less resumption lag and annoyance and fostered a higher degree of social attribution than interruptions in "worst moments". The suitability of disruptions through proactive behavior is hence influenced by the timing of the proactive interrupting task relative to the primary task with the momentane mental workload being an effective measure for opportune interruptibility [10]. These results are backed by Cha et al. [9] who also found concentration, engagement, urgency, and busyness to be critical factors for interruptibility. In their study, users were most susceptible to interruptions when occupied with a highly engaging but not challenging task.

In automotive settings, interruptibility needs to be considered with extra care as users are preoccupied with a challenging and highly security-relevant primary task: driving. Here too, proactive behavior may not be intrusive and overload drivers [4]. Kim et al. found drivers to engage in compensatory behavior like speed reductions and micromanagement of steering wheel position when presented with a secondary task [11]. In another study, they showed that drivers lowered their secondary task engagement when feeling overloaded, resulting in not engaging with proactive behavior at all [12]. This means that primary and secondary tasks are

mutually dependent in automotive settings. A demanding primary task will shut down secondary task engagement while a demanding secondary task will result in poorer performance of the primary task. Still, proactivity in in-car settings is an appealing feature. Use cases can range from driving-related use cases like e.g., refueling or offering to find a parking space to non-driving-related use cases such as reading the news. Schmidt et al. found driving-related proactive scenarios to achieve higher acceptance ratings as well as higher additional value scores compared to non-driving-related use cases [4]. Furthermore, they found that drivers reacted significantly faster to a personal assistant's proactive actions than to its non-proactive counterpart [4]. Assessing proactive and non-proactive use cases with the DALI questionnaire, the researchers found no significant differences between both conditions which they interpret as proactivity being no more demanding than reactive dialogs. SASSI ratings were also found to be similar between proactive and non-proactive behavior, with items "fun" and "useful" being rated "relatively high" [4] for the proactive condition.

2.2 Linguistic Design of VA Responses

While research in the area of linguistic-centered prompt design is still expandable, studies have shown that VA users have preferences for certain formulations of VA responses regarding syntax, grammar, and wording [6; 7]. Stier et al. examined syntactic complexity and driving performance and compared paratactical and hypotactical prompts. They found syntactically less complex parataxes to be preferred over more complex hypotaxes in terms of naturalness and comprehensibility. Furthermore, hypotactical sentences had a detrimental effect on driving performance [13]. Meck et al. broadened the understanding of linguistic preferences further by comparing 28 syntactical, grammatical, and lexical parameters with an impact on the evaluation of VA prompts. In their study, the researchers categorized prompts according to different conversation types, namely prompts in functional conversations (VA responses after being asked to carry out a function), prompts in informational conversations (VA responses after being asked for information), and chit chat prompts (VA responses relating to small talk). Formulation preferences partly differed depending on the type of conversation, suggesting that best practices in designing proactive conversations may also differ from previous design guidelines. Lastly, the research team identified three superordinate user needs regarding VA prompts: a suitable level of (in)formality, a suitable level of complexity/simplicity, and a suitable level of (im)mediacy [6]. These user needs can be catered to by adhering to syntactical, grammatical, and lexical best practices.

Proactivity needs to enhance user experience without being intrusive and putting high information processing demands on users. Previous work by Meck et al. revealed 28 linguistic parameters with an impact on the evaluation of a prompt [6]. Six of these 28 parameters can be related to intrusiveness and information processing on a linguistic level.

Sentence structure: an intricate sentence structure plays directly into high processing demand as the more complex a prompt, the more complex its processing is too [14]. Straightforward prompts with a paratactical sentence structure can therefore aid in reducing information processing demand, while hypotactical sentences add to it.

Sentence length: because of its fleeting nature, processing speech requires high attentional and memory skills [15] when building the situation model of comprehension [16]. The longer the prompt, the higher the concomitant processing demand.

Position of sub-clauses: sub-clauses can be categorized into prepositive and postpositive sub-clauses: Prepositive sub-clauses precede a main clause; postpositive sub-clauses succeed it. Prepositive sub-clauses directly indicate the reason for a proactive interruption through the conditional, temporal, or causal conjunction they are introduced with. Prompts leading with sub-clauses therefore aid in explaining a proactive "intrusion" to a given user.

Form of address: VAs are generally seen as service-oriented assistants, making self-referencing of a VA with "I" (e.g., "I can help you with that") a valid design choice. Still, as discussed in Limerick et al. [17], the usage of VAs is linked to a diminished sense of agency. Addressing users with "you" could aid in the matter of intrusiveness by providing users with an elevated sense of agency by verbally giving them control.

Politeness: studies on politeness in HCI point to an overuse of it leading users into the uncanny valley [18]. Still, opting for politeness could counteract the felt intrusiveness of a proactive VA prompt.

Voice: the usage of an active grammatical voice puts an agent in the foreground whereas the usage of a passive voice focuses more on an action itself [19]. A previous study found active voice to be the preferred grammatical voice for VA prompts [7]. Still, an agent could feel less intrusive when using the passive voice, thereby focusing on the purpose and the message of the interaction rather than on the messenger.

3 Method

3.1 Research Question & Hypotheses

The present work wants to extend design guidelines around proactivity by linguistically informed recommendations for formulating proactive prompts. To explore the occurrence of potential best practices, a within-subjects study was designed to answer the following research question: are there best practices for the formulation of proactive prompts on syntactical, grammatical, and lexical levels?

The following hypotheses emerge:

H1: parataxes are preferred over hypotaxes in proactive prompts

H2: short sentences are preferred over long sentences in proactive prompts

H3: 2nd ps. sg. ("you") is preferred over 1st ps. sg. ("I") in proactive prompts

H4: prepositive sub-clauses (SCs) are preferred over postpositive SCs in proactive prompts

H5: politeness is preferred over no politeness in proactive prompts

H6: passive voice is preferred over active voice in proactive prompts.

3.2 Crowdsourcing Study

3.2.1 Use Cases and Study Prompts

Six proactive voice use cases served as basis for study prompts in the crowdsourcing study: 1) availability of a faster route, 2) parking suggestions, 3) intelligent destination proposals, 4) customizing the navigation map, 5) offering to activate a relaxing mode, and 6) information on the remaining fuel range.

Study prompts were modified regarding sentence structure, sentence length, form of address, position of sub-clauses, politeness, and voice. For each of the parameters, two comparison prompts were designed. These prompts differed in only one of the above-mentioned parameters, thereby enabling the direct comparison of e.g., different sentence structures. To ensure comparability within prompt pairs, all prompts were examined regarding their complexity in terms of comprehensibility by means of the so-called readability index, or LIX [20]. The LIX calculates a sentence's complexity by considering its number of words, number of clauses, the average clause length, and the number of long words (words with more than 6 characters). All prompt pairs were examined regarding their readability and only qualified as study prompts if

they reached the same level of complexity. Furthermore, they were kept consistent in terms of number of sentences. Lastly, the varied parameter was always positioned either at the beginning or at the end of a respective prompt to make use of primacy and recency effects.

3.2.2 Design and Conduct of the Crowdsourcing Study

The crowdsourcing study was conducted online in form of an A/B testing. After giving informed consent, subjects answered questions on age, gender, and experience with in-car VAs. Study participants were instructed to imagine driving a vehicle equipped with a proactive VA. They were then presented with two comparison prompts differing in only one syntactical, grammatical, or lexical parameter and asked to choose the prompt they intuitively liked better. Each prompt pair was accompanied by an introductory text, explaining the respective in-car use case to embed the proactive interaction in a concrete scenario. In total, each participant obtained 12 prompt pairs in randomized order to counteract sequence effects.

Prompts were presented in written form and not auditorily. The decision for this approach is driven by three reasons. Firstly, written prompts control for the potential influence of a synthetic TTS (text-to-speech) voice. Secondly, a study by Stier et al. found that comparing prompts differing in e.g., syntactical structures was impractical via speech. In their study, participants were not able to detect syntactic differences between prompts [13] when they were presented to them via speech. Thirdly, and most importantly, a study by Meck et al. found no differences in the evaluation of prompts between a crowdsourcing study where prompts were presented in text form and a driving simulator study where prompts were delivered via a TTS voice [21]. In this study, the evaluation of prompts did not differ between the two testing conditions, making crowdsourcing studies with text-based prompts a valid alternative to more resource-intensive audio studies in a driving simulator setting.

Due to the ordinal nature of the A/B data, the sample size for the crowdsourcing study was computed for a two-tailed McNemar's Test. An a priori power analysis conducted in G*Power [22] with α=.05 and β=.95 set the sample size to N=99 subjects. The effect size was estimated according to a similar study conducted by Stier et al., who compared different syntactical structures and reported effect sizes of r=.23 to r=.32 [13]. Conservatively, the effect size for the present study was set to r=.23.

3.2.3 Study participants

N=100 participants were invited to take part in the study. 62% of study participants were between 18-34 years old, followed by 31% in the age group between 35-60, and 7% of over 60-year-old subjects. With 54% of subjects identifying as male, 43% identifying as female, and 3% identifying as diverse, the gender balance within the sample was acceptable. Subjects were asked for their usage of in-car VAs, with most participants indicating to use their VAs multiple times a week (43%) or multiple times a month (20%).

4 Results

Due to the dichotomous nature of the A/B data, McNemar's tests were conducted in R [23]. Regarding sentence structure, parataxes were found to be the preferred sentence structure for prompts over hypotaxes with p=0.016. Findings for sentence length showed that short prompts were preferred over long prompts with p=.006. H1 and H2 can therefore be supported. Regarding form of address, 57% of subjects preferred prompts which reference themselves with "you" over self-referencing of the system with "I", but this result was not significant with p=.16. H3 can therefore not be supported. The results for sub-clauses, politeness, and voice did not prove to be significant. Prepositive sub-clauses were slightly preferred over postpositive sub-clauses with 56% (p=.23). Polite prompts were only marginally preferred over their counterparts

without lexical politeness with 55% (p=.32). Lastly, active voice was preferred over passive voice with 56% but this effect too was not significant (p=.23). Hence, H4 to H6 need to be rejected.

5 Discussion

For sentence structure, subjects found paratactical sentences more suitable than hypotactical sentences for proactive interactions. In paratactical sentences, information is split into small and distinct processing units, thereby aiding users in efficiently processing a prompt. This is especially important for proactive settings and proactivity in the car. Proactivity is unsolicitedly interrupting users who are potentially carrying out a demanding primary task. To counteract overloading users, formulating prompts in a straightforward and easily comprehensible manner is advised. As expected for sentence length, short prompts with a word count of 20 words were preferred over long prompts (30 words). As a higher word count means more processing capacities, shorter prompts put less burden on users when processing prompts. Again, not overloading users is key and can be accomplished by choosing shorter rather than longer prompts. Regarding form of address, results were not significant. While a study found significant preferences regarding form of address for other types of conversations with VAs (e.g., conversations focusing on conveying information or small talk) [6], the same does not hold true for proactive interactions. Voice control as a modality is commonly associated with a diminished sense of agency compared to e.g., touch [17]. Due to their more conversational nature (compared to strictly reactive systems), proactive agents are found to foster trust. This elevated trust may lead to reduced strain in terms of agency and can explain the non-significant result for form of address.

Prepositive subordinate clauses were slightly preferred over postpositive sub-clauses. Subclauses are introduced by conditional, temporal, or causal conjunctions. Hence, prepositively put sub-clauses directly indicate the reason for an interruption, thereby helping users to quickly understand their purpose. The use of politeness was preferred over less polite prompts although results were not significant. Politeness is a debated concept in HCI as politeness itself is an inherently human concept, possibly leading into the uncanny valley in HCI [18]. Results for voice did not reach critical levels of significance. Active voice means putting a respective speaker in the foreground (the VA in our case), while passive voice focuses more on a proposed action (the proactive suggestion). In the A/B study, subjects preferred active voice. Although this can only be viewed as a tendency, the result mirrors findings from previous studies where active voice was preferred over passive voice in VA prompts [6].

Compared to other types of prompts (e.g., prompts conveying information or small talk) in Meck et al.'s study [6], proactive prompts show fewer concrete best practices when it comes to syntax, grammar, and wording. Compared to the above-mentioned types of prompts and conversations, proactive interactions are not triggered by the user. On the contrary, proactive interactions are potentially interrupting already ongoing primary tasks. Therefore, the question of *when* to interrupt users may overshadow *how* to interrupt users, in that the formulation of a proactive prompt is secondary compared to its occurrence.

5.1 Limitations and Future Work

Although this paper considered a bandwidth of syntactical, grammatical, and lexical parameters, it cannot raise a claim to completeness. Furthermore, proactive use cases outside the driving environment could reveal different best practices as the car is a unique environment with a demanding primary task, compared to e.g., the smart home. Moreover, it needs mentioning that the study was conducted in German and results may very well be language dependent.

The focus of this study lay in the formulation of prompts and did not take considerations around timing of proactivity into account. Future work could combine both topics and compare differently formulated prompts presented in concrete driving scenarios at opportune moments.

6 Conclusion

Ample research has been published on interruptibility and intrusiveness of proactive VAs and proactive in-car conversations. Work on how exactly a proactive prompt should be formulated once the right circumstances for proactive interactions appear, has received less attention. This paper presents findings from a study shedding light on concrete linguistic guidelines for formulating proactive VA prompts. An A/B study was conducted via crowdsourcing to obtain an overview over linguistic preferences for proactive in-car conversations. We found that study participants indeed preferred certain syntactical structures and prompt lengths over others. This paper shows that the existing framework for proactivity needs to be expanded to include linguistic considerations. It thereby closes the gap of insufficient prompt design guidelines for proactive prompts by providing concrete linguistic design guidance on syntactical levels for proactive in-car interactions.

7 References

[1] NOTHDURFT, F., ULTES, S. AND MINKER, W.: *Finding Appropriate Interaction Strategies for Proactive Dialogue Systems—An Open Quest.* **In** *Proceedings of The 2nd European and the 5th Nordic Symposium on Multimodal Communication*, 73–80. 2014.

[2] REICHERTS, L., ZARGHAM, N., BONFERT, M., ROGERS, Y. AND MALAKA, R.: *May I Interrupt? Diverging Opinions on Proactive Smart Speakers.* **In** *CUI 2021 - 3rd Conference on Conversational User Interfaces*, 1-10. 2021.

[3] SEMMENS, R., MARTELARO, N., KAVETI, P., STENT, S. AND JU, W.: *Is Now A Good Time?: An Empirical Study of Vehicle-Driver Communication Timing.* **In** *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 12 pages. 2019.

[4] SCHMIDT, M., MINKER, W. AND WERNER, S.: *How Users React to Proactive Voice Assistant Behavior While Driving.* **In** *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 485–490. 2020.

[5] SCHMIDT, M., STIER, D., WERNER, S. AND MINKER, W.: *Exploration and assessment of proactive use cases for an in-car voice assistant.* **In** *ESSV Konferenz Elektronische Sprachsignalverarbeitung*, 148–155. 2019.

[6] MECK, A.M. AND PRECHT, L.: *How to Design the Perfect Prompt: A Linguistic Approach to Prompt Design in Automotive Voice Assistants – An Exploratory Study.* **In** *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 237–246. 2021.

[7] STIER, D. AND SIGLOCH, E.: *Linguistic Design of In-Vehicle Prompts in Adaptive Dialog Systems: An Analysis of Potential Factors Involved in the Perception of Naturalness.* **In** *UMAP '19: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, 191–195. 2019.

[8] KRAUS, M., WAGNER, N., CALLEJAS, Z. AND MINKER, W.: *The Role of Trust in Proactive Conversational Assistants.* **In** *IEEE Access, vol. 9*, 112821–112836. 2021.

[9] CHA, N., KIM, A., PARK, C.Y., KANG, S., PARK, M., LEE, J.-G., LEE, S. AND LEE, U.: *"Hello There! Is Now a Good Time to Talk?": Opportune Moments for Proactive Interactions with*

Smart Speakers. **In** *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, Volume 4, Issue 3*, 1–28. 2020.

[10] IQBAL, S.T. AND BAILEY, B.T.: *Investigating the Effectiveness of Mental Workload as a Predictor of Opportune Moments for Interruption.* **In** *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, 1489–1492. 2005.

[11] KIM, A., CHOI, W., PARK, J., KIM, K. AND LEE, U.: *Interrupting Drivers for Interactions: Predicting Opportune Moments for In-vehicle Proactive Auditory-verbal Tasks.* **In** *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, Volume 2, Issue 4*, 28 pages. 2018.

[12] KIM, A., PARK, J.-M. AND LEE, U.: *Interruptibility for In-vehicle Multitasking: Influence of Voice Task Demands and Adaptive Behaviors.* **In** *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, Volume 4, Issue 1*, 22 pages. 2020.

[13] STIER, D., HEID, U., KITTEL, P., SCHMIDT, M. AND MINKER, W.: *The Influence of Syntax on the Perception of In-Vehicle Prompts and Driving Performance* **In** *Conversational Dialogue Systems for the Next Decade*, 349–362. 2021.

[14] CHAFE, W. AND DANIELEWICZ, J.: *Properties of spoken and written language.* **In** *Comprehending oral and written language*, 83–113. 1987.

[15] WOLF, M.C., MUIJSELAAR, M.M.L., BOONSTRA, A.M. AND DE BREE, E.: *The relationship between reading and listening comprehension: shared and modality-specific components.* **In** *Reading and Writing 32*, 1747–1767. 2018.

[16] KINTSCH, W. AND VAN DIJK, T.A.: *Toward a model of text comprehension and production.* **In** *Psychological Review, 85(5)*, 363–394. 1978.

[17] LIMERICK, H., MOORE, J.W. AND COYLE, D.: *Empirical Evidence for a Diminished Sense of Agency in Speech Interfaces.* **In** *CHI '15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 3967–3970. 2015.

[18] CLARK, L.: *Social boundaries of appropriate speech in HCI: A politeness perspective.* **In** *Proceedings of the 32nd International BCS Human Computer Interaction Conference*, 1–5. 2018.

[19] SCHMIDT, M., BHANDARE, O., PRABHUNE, A., MINKER, W. AND WERNER, S.: *Classifying Cognitive Load for a Proactive In-Car Voice Assistant.* **In** *2020 IEEE Sixth International Conference on Big Data Computing Service and Applications*, 9–16. 2020.

[20] LENHARD, W. AND LENHARD, A.: *Berechnung des Lesbarkeitsindex LIX nach Björnson.* 2011.

[21] MECK, A.-M., DRAXLER, C. AND VOGT, T.: *A Question of Fidelity: Comparing Different User Testing Methods for Evaluating In-Car Prompts.* **In** *CUI '22: Proceedings of the 4th Conference on Conversational User Interfaces*, 1–5. 2022.

[22] FAUL, F., ERDFELDER, E., BUCHNER, A. AND LANG, A.-G.: *G * Power 3.1 manual.* 2021.

[23] RSTUDIO TEAM: *RStudio: Integrated Development Environment for R.* Boston, MA. 2022.