# Comparing Priming Effects of Visual and Textual Task Representations

## Texts can Influence Users' Utterances.

Patrick Ehrenbrink and Stefan Hillmann

**Abstract** Textual and visual task representations were compared in an online study to assess the influence of semantic priming on spontaneous, user generated voice commands. Our results indicate that visual representations of tasks lead to a different distribution of used phrases, compared to the textual representation, within the target population. Furthermore, the results suggest that text stimuli can be used to influence user utterances.

## 1 Introduction

With the growing importance of mobile devices and the limited usability of input devices that comes with the mobile domain, spoken language becomes more and more important as an input modality for a variety of mobile applications. For the development of a speech-based interface for a given system, it is important to specify an adequate set of voice commands that the system is supposed to understand and act on. Coming up with those commands can be done with a multitude of methods such as interviews, experiments, corpora analysis and so on. A method that is examined more closely in this paper is to run a small exploratory study in which participants solve a task and are asked to formulate proper and intuitive voice commands. Priming can be an intervening factor in this scenario. Here, priming is a psychological effect that functions as a semantic activation. A priming word can activate semantically related words and increases their probability to be used. Furthermore, it facilitates the response to such words by shortening the response time

Patrick Ehrenbrink
Technische Universität Berlin, Ernst-Reuter-Platz 7, 10587 Berlin e-mail: patrick.ehrenbrink@tu-berlin.de

Stefan Hillmann
Technische Universität Berlin, Ernst-Reuter-Platz 7, 10587 Berlin e-mail: stefan.hillmann@tu-berlin.de

[7]. This means that e.g. if a stimulus like the word “butter” is presented to a person, then this person will be able to respond faster to the word “bread”, because it is semantically associated with the word “butter”.

Data that was collected in an experiment can be biased if the participants are primed by the written or spoken description of a task. Then, this could lead to less diverse answers and reduce the validity of the results. The effect is especially problematic if we wish to study the frequency distribution of used phrases within the population. Such information can be important for the development of heuristics for automatic speech recognition and natural language understanding, e.g. when resolving ambiguous commands. Especially when interacting with Intelligent Personal Assistants in a spoken dialogue, resolving ambiguity is an important part of achieving a natural conversation.

In this paper, we describe an attempt to circumvent the priming of words by using graphical task descriptions instead of textual ones. The two means of task presentation (textual and visual) are compared in the present study in order to determine if they result in a different probability distributions and different absolute quantity of phrases used in the participants’ voice commands.

Hypothesis one

The textual task description primes the participants so that object names and action words from the task description are more likely to be used in the voice commands that they utter. This will result in a higher proportion of the phrases used in the task description compared to other phrases within the collected user commands. Therefore, we expect that object names and action words from the task description are used more frequently in the responses of the textual condition than in the responses of the visual only condition.

Hypothesis two

When using images instead of text in order to describe tasks, there are no words that could prime the participants. Even though priming is likely to have an effect across different modalities, the cross-modal effect is likely to be smaller compared to the uni-modal effect. With the priming bias for a particular phrase reduced, we expect that this will result in a larger amount of unique phrases compared to the textual condition among the collected user commands.

## 2 Related Work

There is some previous work that compared priming effects of textual and graphical representations. Bernsen [1] and Dybkjær [4] compared the influence of textual and

graphical task descriptions. In one task of their study, the participants had to verbalize a specific time. The specific point in time was indicted either in a textual (e.g. "[...] at 7:20." [4]) or a graphical (picture of a clock) scenario description. They found a priming effect in 74.5 % of the answers in the text condition on average. Another study was performed by Möller who used graphical task descriptions in a study about dialogue strategies for spoken dialogue systems [6] for restaurant search. In that study, marked maps were used as task descriptions in order to indicate locations that participants had to search for. Furthermore, shapes of countries were presented to the participant to indicate the nationality of a cuisine (e.g. the shape of Italy was used as an indicator for Italian food). Those graphical task descriptions were used to avoid priming that would possibly have occurred if textual task descriptions had been used. Since no textual task descriptions were used, a comparison between the effects of graphical and textural task descriptions was not performed. Even though those studies present a valid way of representing tasks graphically, the addressed scope in terms of priming is very limited each time. There are not many ways of verbalizing a certain time or indicate a location on a map. Also, the graphics that were used are not (very) ambiguous and leave no room for interpretation. That is, of course, perfectly valid. However, the results are hardly transferable to the task of collecting (a larger number of) suitable utterances that have to be understood by a spoken dialogue system in the addressed domain. From those previous studies it remains unclear if graphical representations are suitable to gain an overview of intuitive verbalizations of an intended object or action. Such an overview includes the most common names for objects and words for actions that can be expected from the target population together with a quantitative distribution of those verbalizations. The study we present in this paper is aimed to investigate the potential of graphical task descriptions in order to collect information about commonly used phrases for the addressed domain and to avoid priming in user studies with spoken language based systems.

## 3 Methods and Participants

For this study, we did not use *abstract* graphical representations to describe the task, but photos of actual objects including the surroundings that are to be addressed. Each task is described in two photos in order to indicate the object to be used and the action to be executed. The latter is indicated by a difference between the two images, e.g. a lamp which is off and on. With this approach, we aim at getting a representative overview of verbal commands that our target population would use to trigger an action in a smart home environment. This overview should include names for the objects and descriptions for the actions.

| | Instruction | |
|---|---|---|
| Task | German | English Translation |
| A | Schalten Sie die Tischlampe ein. | Switch the table lamp on. |
| B | Schalten Sie die Tischlampe aus. | Switch the table lamp off. |
| C | Dimmen Sie die Tischlampe dunkler. | Dim the table lamp brighter. |
| D | Dimmen Sie die Tischlampe heller. | Dim the table lamp darker. |
| E | Schalten Sie den Ventilator ein. | Switch the fan on. |
| F | Schalten Sie den Ventilator aus. | Switch the fan off. |
| G | Öffnen Sie das Fenster. | Open the window. |
| H | Schließen Sie das Fenster. | Close the window. |
| I | Schließen Sie das Rollo. | Close the roller blind. |
| J | Öffnen Sie das Rollo. | Open the roller blind. |
| K | Machen Sie das Rollo halb zu. | Close the roller blind to the half. |
| L | Machen Sie das Rollo halb auf. | Open the roller blind to the half. |
| M | Fahren Sie das Rollo ganz herunter. | Completely close the roller blind. |

**Table 1** English translations of all thirteen German task descriptions.

### *3.1 Study Desgin*

An online-study with between-subjects design was performed using the tool Limesurvey 1.90+[1]. The study included two conditions with 13 tasks each. All used images were photos of the devices (i.e. objects) that the participants were supposed to control via a voice command. In the textual condition, that picture was accompanied by a text that stated the task. An example configuration of the stimulus for Task A is shown in Figure 1. Table 1 shows the descriptions of the 13 tasks that were used in the textual condition and their translation into English.

In the visual only condition, the picture was accompanied by a second picture. That second picture showed the object in a state that was to be induced by a voice command. For example, the first picture showed a lamp that was switched off and the second picture showed a lamp that was switched on. Figure 2 shows the stimulus configuration of Task A in the visual only condition. Even though other objects such as a table were present on the picture, the object or device in question was always in the centre. In the textual condition, the object and action in question was clear to the participant due to the accompanying text. In the visual only condition the object and action in question was clear to the participant because both pictures were identical apart from the target object and its functional state.

The participants were not really able to control the devices on the pictures. However, they were instructed by a text on top of each questionnaire page that they should phrase a verbal command for a spoken dialogue system that would fulfil the task. Furthermore, they were asked to write the phrased command into a text box at the bottom of the page, as the spoken utterance could not be recorded with the used questionnaire system.

[1] `www.limesurvey.org`

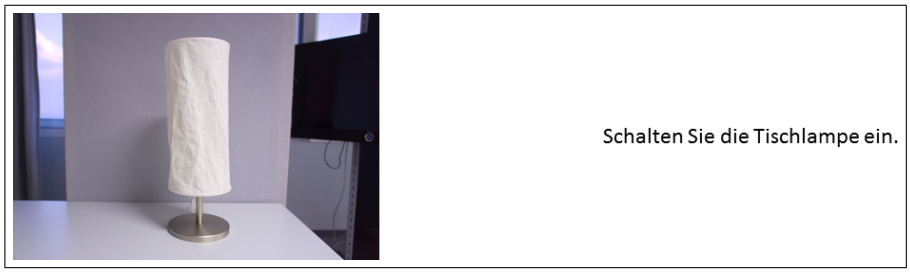


**Fig. 1** Screenshot of the stimulus for task A in the text condition.

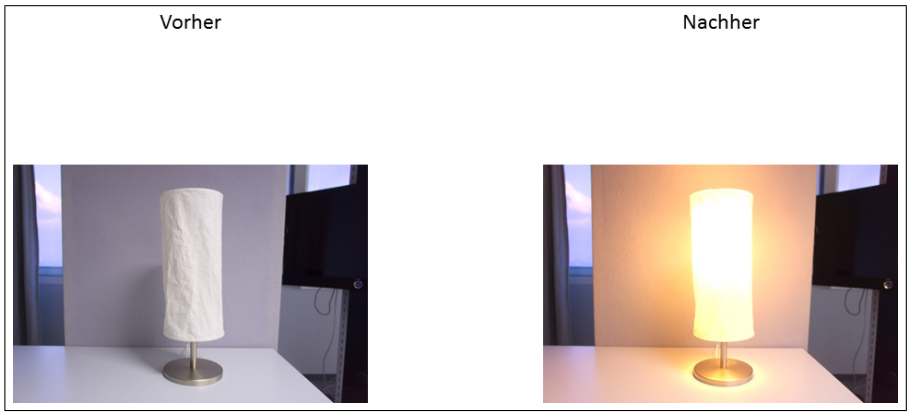


**Fig. 2** Screenshot of the stimulus of task A in the visual only condition. Text and images were shown to the participant simultaneously. Translations: *Vorher*: before, *Nacher*: after.

## *3.2 Participants*

In total, 178 persons took part in the survey. Their average age was 31.48 years and the gender was not taken into account. As a compensation for their participation, three vouchers with a value of 40, 20 and 10 Euro were drawn under all participants. Due to technical limitations of the survey tool, participants could not be assigned randomly to a condition. For that reason, they were assigned to the condition according to their respective age.

The participant's age in years was requested by the questionnaire system at the beginning of the trial. If the age was an even number, the participant were assigned to the visual only condition. In contrast, if the age was an odd number, textual condition was assigned. This procedure resulted in a total number of 92 participants for the visual only condition and 86 participants for the textual condition.

### *3.3 Data Processing*

All responses were examined by three different persons (all were members of our research group). They assessed if the requests were appropriate answers to the stated task. All responses that indicated non-compliance or a misunderstanding of the task were excluded from further analysis. Afterwards, the remaining 166 responses (out of 178) were normalized and processed further. Normalization included the removal of all punctuation marks and typing errors as well as the transformation of upper case characters into lower case. In the subsequent step, words that refer to the name of the object in question (nouns) and words that refer to the actions (verbs) were extracted from each command. This pre-processing ensured that typos or grammatical errors did not result in different word counts when computing the amount of priming.

The amount of priming $p_o$ for the object in a task of the text condition was calculated as shown in Equation 1. $o$ is the name of the object that was stated in the description of respective task in the textual condition, e.g. “table lamp”. Furthermore, $f(o)$ is the relative frequency of $o$ in the set of all phrases that were used to refer to $o$ in the task (see Table 4). The priming of $o$ is the difference of the relative frequency of $o$ in the textual ($t$) and the visual only ($v$) condition.

$$p_o = f(o_t) - f(o_v) \tag{1}$$

Equation 2 shows the computation of the amount of priming for the action ($a$) to be used in a certain task, e.g. “switch on”. The annotation is analogue to Equation 1.

$$p_a = f(a_t) - f(a_v) \tag{2}$$

## 4 Results

Independent sample t-tests were performed on the frequencies of unique phrases for objects and actions in both conditions with IBM SPSS Statistics 22. The frequencies are provided in Table 2.

In Hypothesis one it was stated that we expect that object names and action phrases that are used in the textual descriptions are more likely to be used in the responses from the textual condition, compared to the responses of the visual-only condition. Results show that the proportion of phrases, used by the participants, that appeared in the textual task description was significantly higher in the textual condition than in the visual only condition ($\alpha = 0.05$ and $p < 0.001$). Also, the proportion of uttered action phrases from the textual task description was significantly larger in the textual condition than in the visual only condition ($\alpha = 0.05$ and $p < 0.001$) Table 4 provides the proportions of names and actions from the textual task descriptions as they appeared in each task and condition.

| | Commands | | Object Names | | Actions | |
|---|---|---|---|---|---|---|
| Task | textual | visual | textual | visual | textual | visual |
| A | 26 | 19 | 5 | 5 | 10 | 11 |
| B | 20 | 16 | 5 | 6 | 8 | 7 |
| C | 40 | 35 | 8 | 6 | 23 | 24 |
| D | 37 | 39 | 9 | 6 | 20 | 29 |
| E | 25 | 24 | 7 | 7 | 9 | 10 |
| F | 17 | 24 | 6 | 9 | 7 | 10 |
| G | 17 | 15 | 2 | 4 | 8 | 6 |
| H | 16 | 15 | 1 | 4 | 12 | 7 |
| I | 24 | 39 | 5 | 7 | 13 | 23 |
| J | 22 | 41 | 4 | 8 | 10 | 17 |
| K | 40 | 63 | 3 | 10 | 30 | 41 |
| L | 37 | 57 | 4 | 8 | 32 | 36 |
| M | 29 | 44 | 3 | 8 | 22 | 26 |

**Table 2** Frequency of unique commands (i.e. unique utterances) as well as unique phrases (i.e. names) for the object and the action in each task.

| Condition | Names | Actions |
|---|---|---|
| textual | 4.77 | 15.69 |
| visual | 6.77 | 19.00 |

**Table 3** Average numbers of different Names and Actions for both conditions.

For the comparison of the responses in the two conditions, independent-samples t-test were performed. The average number of unique object names that were retrieved in each task in the visual only condition (6.77) was significantly ($\alpha = 0.05$ and $p = 0.024$) larger than the average number of names that were retrieved in each task of the textual condition (4.77). Furthermore, the average number of actions that were retrieved in each task of the visual only condition (19) was significantly larger ($\alpha = 0.05$ and $p = 0.03$) than the average number of actions that were retrieved in the textual condition (15.69). The average number of object names and actions for each task in both conditions can be seen in Table 2. Table 3 shows the named average frequencies of unique object names and actions.

The amount of priming was calculated according to Equation 1 and 2. Priming increased the probability of a name from the task description to be used in a verbal command by 0.25. For the words that describe actions the average probability increased by 0.08. Beside the amount of priming per task, Table 4 also shows the relative frequencies which were used for the computation of $p_o$ and $p_a$.

| | Textual | | Visual | | Priming | |
|---|---|---|---|---|---|---|
| Task | $f(o_t)$ | $f(a_t)$ | $f(o_v)$ | $f(a_v)$ | $p_o$ | $p_a$ |
| A | 0.202 | 0.179 | 0.026 | 0.052 | 0.176 | 0.127 |
| B | 0.202 | 0.905 | 0.025 | 0.886 | 0.177 | 0.019 |
| C | 0.190 | 0.667 | 0.026 | 0.641 | 0.165 | 0.026 |
| D | 0.179 | 0.690 | 0.026 | 0.487 | 0.153 | 0.203 |
| E | 0.607 | 0.226 | 0.603 | 0.064 | 0.005 | 0.162 |
| F | 0.614 | 0.928 | 0.622 | 0.817 | -0.007 | 0.111 |
| G | 0.843 | 0.265 | 0.789 | 0.184 | 0.054 | 0.081 |
| H | 0.843 | 0.241 | 0.789 | 0.263 | 0.054 | -0.022 |
| I | 0.821 | 0.238 | 0.342 | 0.132 | 0.479 | 0.107 |
| J | 0.833 | 0.238 | 0.368 | 0.158 | 0.465 | 0.080 |
| K | 0.843 | 0.373 | 0.297 | 0.243 | 0.546 | 0.130 |
| L | 0.845 | 0.429 | 0.320 | 0.293 | 0.525 | 0.135 |
| M | 0.857 | 0.369 | 0.324 | 0.378 | 0.533 | -0.009 |

**Table 4** Relative frequency ($f$) of use of phrases from the textual task description, addressing the object ($o$) and action ($a$). Data are shown for the textual ($t$) and visual only ($v$) condition. Priming is the difference between the usage in the textual and visual only condition (see Equation 1 and 2).

## 5 Discussion

*Hypothesis one* can be confirmed. The results show that actions and names for objects that were used in the textual task description appeared significantly more frequently in the responses of the textual condition, compared to the responses of the visual condition. This result can be explained with a priming effect: The task description influenced the likelihood that those words were used by the participants. It can therefore be concluded that a textual task description is not optimal for accessing the proportional distribution of individual phrases in a given population and that visual task descriptions should be used if feasible.

*Hypothesis two* can be confirmed, as well. Our results show that visual task descriptions resulted in the appearance of significantly more names for the objects and actions. It can be concluded, that visual task descriptions are more suitable to collect a wide range of possible utterances than textual descriptions.

From the data it is evident that the amount of priming is relatively low for action words. One possible explanation for this is the fact that the variety of action words in the tasks that were used is larger compared to the variety of names for the objects. However, this is probably not a sufficient explanation. Additionally, a ceiling effect might have occurred. The large quantity of different action words is mostly the result of expressions that appeared only one or two times. For example, in task B, the majority of participants of both conditions used the word “aus” (engl. “off”) (textual: 85%, visual: 84%). Whereas other words such as “ausmachen” (engl. ”turn off”) or “ausschalten” (engl. “switch off”) appeared far less often. The word “aus” was also used in the textual task descriptions. The high usage rate of the word indicates that this is by far the most intuitive action word to be used in that particular situation. Another possible explanation is that the word ”aus” can simply be spoken

relatively fast and might simply be the most efficient way to perform the task of switching something off.

Priming can appear across different modalities [8]. This means that also a visual or auditory stimulus is able to induce verbal priming. The task presentation probably primed the participant in both conditions. Therefore, the priming effect caused by the text was not this powerful here, but we do not consider this as a problem for the validity of the presented study. If it actually has any effect, it would lessen the observed differences between the conditions. Words that are associated with an object are also primed by its visual appearance, so they are primed by the pictures in this study.

It should be kept in mind that priming from cues, regardless of their modality is what designers can benefit from. Priming can help the users to select the appropriate words for their commands. However, if one wants to examine the distribution of different words in the user population, priming becomes a problem. The words that are used in the task description can then bias the results. For instance, this is an issue, if the intuitive usage of a speech based human-machine interface is to be evaluated.

Besides that obvious impact of priming in the tasks that were used for this study, priming is worth to be considered in a variety of other contexts. One of which is persuasive technology. Since semantic priming represents a cognitive bias towards the primed word or object, this effect can be used to influence people's behavior to a small extend. An advantage is that the priming effect takes place rather subliminally [5] and might therefore be an adequate method to avoid negative side effects, such as Psychological Reactance [2, 3], which would otherwise cause the user to be less open to persuasive attempts or even counteract those.

## 6 Conclusion

Results of a survey that uses visual task representation provide a more realistic overview of the variety and quantity of preferred commands among the target population. This means that obtaining verbal commands by using graphical task descriptions also results in more valid data. Priming effects should be considered when collecting supposedly intuitive commands by running an exploratory – study such as the one described in this paper. Apart from the collection of utterances, the effect of textual priming as a factor that influences user behavior could proof beneficial in the persuasive domain, especially for online marketing and shopping advertisements.

**Acknowledgements** This work was supported by the Bundesministerium für Energie und Wirtschaft (Germany) under grant no. 01MG13001G, Universal Home Control Interfaces@Connected Usability (UHCI).

## References


1. Niels Ole Bernsen, Hans Dybkjaer, and Laila Dybkjaer. *Designing Interactive Speech Systems. From First Ideas to User Testing*. Springer, London, 1998.
2. Jack Williams Brehm. *A Theory of Psychological Reactance*. Academic Press, New York, 1966.
3. James Price Dillard and Lijiang Shen. On the nature of reactance and its role in persuasive health communication. *Communication Monographs*, 72(2), 2005.
4. Laila Dybkjaer, Niels Ole Bernsen, and Hans Dybkjaer. Scenario Design for Spoken Language Dialogue Systems Development. In *Proc. of ESCA Workshop on Spoken Dialogue Systems*, pages 93–96, 1994.
5. Johan C Karremans, Wolfgang Stroebe, and Jasper Claus. Beyond vicary's fantasies: The impact of subliminal priming and brand choice. *Journal of Experimental Social Psychology*, 42(6):792–798, 2006.
6. Sebastian Möller. *Quality of Telephone-Based Spoken Dialogue Systems*. Kluwer Academic Publishers, Boston, 2005.
7. Thomas Städtler. *Lexikon der Psychologie*. Kroner, Stuttgart, 2003.
8. David A Swinney, William Onifer, Penny Prather, and Max Hirshkowitz. Semantic facilitation across sensory modalities in the processing of individual words and sentences. *Memory & Cognition*, 7(3):159–165, 1979.