IS THERE A HESITATION BIAS FOR AMBIGUOUS COLOR TERMS?

Simon Betz[1]*, Ricardo Davids*[1]*, Caroline Müller*[1]*, Eva Székély*[2]*, Petra Wagner*[1]*, Maischa Amelie Weber*[1]*, Cassandra Youssef-Baronfeind*[1] *& Sina Zarrieß*[1]

[1]*Bielefeld University, Faculty for Linguistics and Literary Studies,* [2]*Department of Speech, Music and Hearing, KTH Stockholm.*
simon.betz@uni-bielefeld.de

Abstract: Recent endeavors have successfully tried to equip dialogue systems with hesitations [1][2]. Hesitations are useful when it comes to buying dialogue time or to support listeners in comprehension [2]. It has been found that hesitations furthermore create a bias towards complicated concepts [3]. In addition, one particular hesitation, namely lengthening, has gotten comparably little attention in research, despite its subtlety enables it to buy dialogue time without degrading sound quality [4]. In this paper, we investigate whether the hesitation bias can be replicated in click tasks with ambiguous and complex color terms as hard-to-describe entities, that have audio instructions with and without hesitations. We hypothesize that, analogous to the studies by Arnold [3], listeners will infer that the target referent is a complicated rather than a simple color term when the instruction contains a hesitation. While our studies cannot replicate the hesitation bias, we provide valuable insights into the interaction between hesitations and speech processing as well as methodological considerations for future research.

1 Introduction

In this paper, we present basic research on hesitations for dialogue systems. For dialogue systems that have a larger scope than reading out text, it is not only of interest what is being said, but also how it is delivered. In previous studies, we found that hesitations are useful to manage dialogue timing, and to increase participants' performance in interactive tasks (cf. [5] for an overview). Hesitations usually manifest in human speech in one of the following ways: (1) Fillers (*uh, uhm*) (2) Silences (*that...is great*) (3) Lengthening (*the greeeennn tree*). We recently found that lengthening is also useful for conveying attitudes, such as uncertainty, e.g. in systems that describe images [6], which is in line with previous research that found hesitations and temporal patterns to be reliable cues for detecting uncertainty, cf. e.g. [7] [8]. In this study, we further investigate lengthening as a means to resolve reference resolution in image descriptions featuring color combinations. We report a series of experiments in which participants have to click visual referents on a screen based on hesitant and non-hesitant human and synthetic speech input.

1.1 Background

Previous studies have shown that synthesized hesitations can increase task performance of participants in experiments [1, 2]. It could further be shown that lengthenings in particular are elegant ways for dialogue systems to hesitate, as they are hardly perceivable and thus "buy dialogue time for free" [4]. "For free" can be understood as coming without the cost of a loss in sound quality, which is a common problem when synthesizing hesitations [9], which in a way

Figure 1 – Example for a "slider": A color continuum from one color to another which participants can click.

mirrors the bad reputation hesitations sometimes have in human interaction [10]. In this paper we present a series of studies that investigate fundamental temporal and positional properties of hesitation lengthening with the aim to better model hesitations for technical applications. It was found in perception experiments that hesitations create an interpretation bias towards less obvious alternatives [3] and that the position of hesitation lengthening within a property word influences the perception of uncertainty [6]. Our research question is whether the hesitation bias generalizes to dimensions of non-obviousness and ambiguity beyond those examined in [3] and [6]. In [3] the objects were undefinable shapes, in [6] the objects were single color terms; in this study we look at complex color terms and visual items of “easy” and “difficult” colors.

2 Experiments

We first discuss two pretests in which we examine potential direct mappings between hesitations and color terms via perception tests. After that we discuss the main study, in which users have to perform a reference resolution task. In all studies discussed in this paper, color terms are involved. Color terms tend to be a matter of discussion, as people do not always agree on the naming of colors. For creating our visual stimuli, we thus used the color values determined by a large-scale online study[1] to ensure that our understanding of "blue" will be understood as "blue" by the majority of our participants.

2.1 Pretests - Introduction

We first conducted two pretests, henceforth “slider studies 1 & 2”. These pretests are designed to test whether there is a direct mapping between hesitation lengthening and color terms. These come with extremely simplified experimental setups in which there is a clickable 1-dimensional color continuum (“slider”) transitioning from one color to another, cf. Fig.1. Each slider is accompanied by an audio instruction to click at a position that fits the color term. All sliders and audios were further matched in alignment: when the instruction was "blue-green" then blue was on the left hand side and green on the right hand side in the image. When the instruction was "red-blue", red would be left and blue would be right, etc.

The hypothesis for the slider studies is undirected, we simply explore whether lengthening position and extent influence click position. Two scenarios are conceivable. When listeners interpret the lengthening as hesitation, they might infer that the speaker is uncertain about the lengthened color word, cf. section 1, which should result in a click on the other color. Alternatively, it is possible that listeners interpret the lengthening as accentuation, which in terms of duration resembles hesitation lengthening [11], which should lead to clicks on the lengthened color itself. The null hypothesis is that clicks will cluster around the midpoint.

For both studies, the dependent variable is the click position. Under the hood, each click into the sliders is converted into a numerical value between -50 and 50, where -50 is the left border and 50 the right border of the slider. This means that values around 0 support the null hypothesis, values < 0 indicate a bias for the first color term and values > 0 indicate a bias for the second color term. The two pretests were conducted as an online study using PERCY [12].

[1] https://blog.xkcd.com/2010/05/03/color-survey-results/, accessed 01/24/2022

2.2 Pretest - Slider Study 1

2.2.1 Method

In this study, we used the German basic color terms *rot*, *blau* and *gelb* in all six possible combinations and orders for the visual and audio stimuli. The audio stimuli come in three conditions, INI, FIN and NONE, denoting the placement of the lengthening on the initial and final color term, or of no lengthening, respectively. In addition, lengthening is inserted in five different extents, 0, 200, 400, 600 and 800ms duration increase above the NONE condition. All participants got all 60 stimuli, presented in a random order. The stimuli were generated using the open-source synthesis software MaryTTS [13] for maximal control over the duration increase. The distribution of lengthening across the syllable followed the elasticity hypothesis by [14] applied to hesitations as in [15]. Lexically, the stimuli consisted of only the two color words without any carrier sentence.
22 of the participating participants (14 female, 8 male) completed the experiment. Participants were recruited via social media and conducted the experiment at home on their own devices. All participants were volunteers who received no monetary compensation for participation. The entire experiment took about 15 minutes per participant. None of the participant specifications (age, sex, device, mother tongue) had any measurable impact on the results, so they are not regarded further.

2.2.2 Results

We fit two linear mixed effects regression models with click position as the dependent variable, lengthening position / lengthening extent as fixed factors and random slopes for participants as random factors using the R packages lme4 and lmerTest [16]. For this study, we cannot notice any influence of lengthening extent or lengthening position on the resulting click position, rather all clicks cluster around the midpoint, cf. Table 1.

Condition	**Estimate**	**Std.Error**	**t value**
extent0 (intercept)	2.1087	2.2886	0.921
extent200	-0.8007	1.8056	0.6575
extent400	-2.9855	1.8056	0.0985
extent600	-2.4058	1.8056	0.1829
extent800	-2.2826	1.8056	0.2064
INI (intercept)	-2.2409	1.2751	-1.757
FIN	1.1105	2.1027	0.528
NONE	0.9982	1.5617	0.639

Table 1 – Model outputs for slider study 1. Top: model for lengthening extent. Bottom: model for lengthening position.

These results led to the creation of the second slider study. We assume that in this study, too many aspects were over-simplified, so that participants did not know what to do. We identified as potential sources of conflict:

- The color choices were unfortunate - The middle points between two basic colors have names, so referring to e.g. *orange* as *red-yellow* makes no sense.
- Audio instructions might need to be full sentences instead of short phrases.
- Using speech synthesis gives control but degrades sound quality.

- The study might be to long given the simplicity of the input, there is a risk of fatigue.

2.3 Pretest - Slider Study 2

2.3.1 Method

The second pretest uses the same setup as the first, but with several adjustments:
Lexical items. We adapted the sliders to have a basic color on one end and a mix color on the other hand, e.g. "blue-green" instead of "blue-yellow" (where the middle would be green). This way we have an actual color space with blue and green components in the middle. The color combinations used in this study are "blau-grün" (blue-green), "rot-braun" (red-brown) and "grün-gelb" (green-yellow), in all six possible combinations and orders.
Carrier sentences. Presenting phrases consisting only of two color terms appears to be fatiguing and confusing, so for this study we embedded the stimuli into a carrier sentence "click on XY in the color continuum".
Human voice. We opted to replace the speech synthesis voice by recordings of a human voice. One male speaker of German with experience in radio play recordings produced the stimuli.
Lengthening extent. We kept the idea of five different lengthening extents. Using Praat's [17] duration manipulation and resynthesis function, we placed the peak of the lengthening over the vocalic nucleus of the color word in question, raising it for every condition to factors x2, x3, x4, x5 resulting in approximately the same duration increases as in slider study 1 (200ms steps).
Participants. The general participant policy is the same as in the first study, in terms of recruitment and personal data surveyed. In this study, we had 24 participants completing the task that did not take part in the first study. In total, there were 17 female, 5 male and two participants of other gender. None of the speaker-specific factors appeared to influence the results so they are not considered further.

2.3.2 Results

We fit two linear models analogous to slider study 1. As in slider study 1, the extent of the lengthening does not make any difference. The position of the lengthening, however, does. Initial lengthening lowers the value of the click position significantly, which means that participants click more towards the left, i.e. on the lengthened word itself in this condition. For post-hoc analysis, we computed estimated marginal means using the R package emmeans [16]. As can be seen in table 2, INI differs significantly from both FIN and NONE, while there is no significant difference between FIN and NONE.

Condition	**Estimate**	**Std.Error**	**t value**		
extent0 (intercept)	-1.1292	1.5556	-0.726		
extent200	-0.8004	1.7310	-0.462		
extent400	-2.3022	1.7310	-1.330		
extent600	-0.8008	1.7340	-0.462		
extent800	-0.5369	1.7320	-0.310	**Contrast**	**p-value**
INI (intercept)	-5.2420	0.9441	-5.553	INI vs NONE	0.0229
FIN	6.0147	0.9797	6.139	INI vs FIN	<.0001
NONE	4.1125	1.5582	0.639	NONE vs FIN	0.4413

Table 2 – Model outputs for lengthening extent (top) and lengthening position & estimated marginal means contrasts (bottom) for slider study 2.

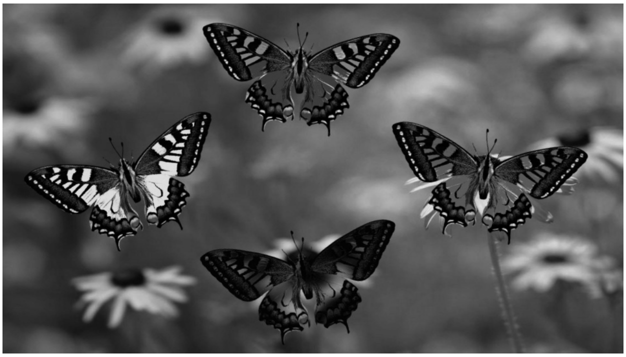

Figure 2 – Example of a visual scene. On top a butterfly with distinct blue and green parts, on the bottom a butterfly colored in a blurry mix of blue and green; left and right random other colors as distractors.

2.4 Main Study: Butterflies

2.4.1 Method

The main study is a reference resolution task in which participants do an online experiment in which they see visual scenes with four potential target objects. Each visual scene comes with an audio instruction to click one of the targets and the instruction may or may not feature hesitation lengthening.

The target objects are butterflies of varying colors on a colored background. In each scene, one of the butterflies is two-colored with parts in color A and parts in color B, cf. Fig.2. Each scene features a competitor butterfly located opposite to the two-colored one. The competitor has the same two colors, but not clearly separated, but rather as a blurry and cloudy mix of both colors across the entire surface. The other two butterflies are of different colors. The visual stimuli are balanced in terms of position, so each butterfly appears equally often at each of the four positions. In order to avoid fatigue, each butterfly is randomly rotated a few degrees to have the scenes appear less static.

These visual scenes are accompanied with one audio stimulus each, telling the participant to click one of the butterflies in the scene, e.g. "Klicke auf den blau-grünen Schmetterling" ("click the blue-green butterfly"). The audio stimuli come in four conditions: (NONE) without hesitation; (INI) with hesitation lengthening on the first color word; (FIN) with hesitation lengthening on the second color word; (FULL) with hesitation lengthening on both color words.

The color combinations investigated are blue-green / green-blue; red-green / green-red; and blue-red / red-blue. Each item occurs twice in the stimulus list, in different positions in the visual scene. In addition, we created 24 distractor stimuli with different colors, yielding 72 stimuli in total. Each participant listened to all 72 stimuli, presented in a random order.

The audio stimuli are produced via voice recordings of one female native speaker of German. The lengthenings are inserted via duration manipulation using Praat [17], yielding duration increases of about 150ms per color word.

In order to make sure that the lengthenings we produced are perceivable, we conducted a small-scale online survey with seven participants before the main study, in which we asked the participants to watch out for hesitation lengthening and feedback on which of the words, if any, they perceive it. This survey revealed that lengthenings can be reliably identified as intended.[2]

The hypothesis is that if the hesitation bias exists for color terms, participants will click on the blurry competitor object more often in hesitation conditions compared to the NONE

[2] Note that lengthenings are notoriously hard to perceive, even by trained disfluency annotators. It thus requires listeners explicitly paying attention to lengthening in certain places to verify the validity of the stimuli.

condition. The main response that is evaluated is thus the click position. In addition, we track the mouse path and the timing of the response. The online experiment is run via PERCY [12].

2.4.2 Results

In the butterfly study, we reject our hypothesis as the presence of hesitations does not create a bias towards the blurry objects. Even more so, 99% of the clicks fall on the non-blurry (the obvious) object. An investigation of the mouse paths also does not reveal a bias of the hesitation towards the blurry object.

What is revealing, though, is the analysis of the timing. We computed task time, which is the interval from the onset of the second color word to the click, which is the interval from the point from which the target can be inferred up to the final decision. It shows that INI comes with the fastest task performance (1.0s) and FULL with the second fastest performance (1.12s), whereas the fluent baseline NONE is second slowest (1.24s) and FIN the slowest condition (1.34s). We fit a linear mixed effects regression model with task time as the dependent variable, lengthening position as fixed factor and random slopes for both user and stimulusID, using the R lmer4 and lmerTest packages [16]. Model comparisons reveal that lengthening position has a significant effect on task time. We then used the R package emmeans [16] to compute estimated marginal means for pairwise comparisons within the factor lengthening position, cf. Table 3. As can be seen, INI differs significantly from FIN and NONE, furthermore there is a borderline significant difference between FULL and FIN.

Condition	Estimate	Std.Error	t value	Contrast	p-value
FULL (intercept)	1.12490	0.08835	12.732	FULL vs INI	0.3430
INI	-0.12727	0.07427	-1.714	FULL vs FIN	0.0584
FIN	0.20222	0.07427	2.723	FULL vs NONE	0.4544
NONE	0.11161	0.07427	1.503	INI vs FIN	0.0013
				INI vs NONE	0.0207
				FIN vs NONE	0.6218

Table 3 – Model output and estimated marginal means contrasts for butterfly study.

3 Discussion

The initial question that sparked this research was whether the hesitation bias observed by [3] exists in a scenario where the hard-to-describe objects are replaced by ambiguous and hard-to-describe properties like color compounds and the hesitations consist exclusively of lengthening in different positions and extents. We cannot confirm this hypothesis, but as we will discuss in this section, this might partly be due to unsuitable methods employed. We can present other findings on the effect of hesitation lengthening, though. As the analyses of task time in the butterfly study reveal, lengthening occurring early in an utterance, i.e. lengthening preceding the relevant information, helps the listener process information. Lengthening occurring late in an utterance may not help or even disturb the listener during processing. These findings are corroborated by the pretests in which only early-occurring lengthening had an effect on the results.

The finding that clicks in the butterfly study ultimately fall on the obvious object in 99% of cases show that a click task is an unsuitable method to investigate hesitation bias. The lexical information, which is ultimately the same, regardless of condition, instructs people to “click a two-colored object”, so people click the obvious two-colored object. However, measuring task time provides valuable insights as they reveal the processing speed during the task. This shows that hesitations do make a difference, just not with regard to the goal of the click task itself, but in the way the goal is reached.

It is further revealing that the conditions that feature lengthening on the first word, INI and FULL, yield the best (i.e. fastest) task performances as opposed to the conditions that lack initial lengthening. We argue that for interpretation of this effect, the lexical structure has to be taken into account: lengthening on the second word means that the hesitation is only perceived when the target is already clear to the participant, thus creating confusion. On the contrary, initial lengthening is always perceived while searching the target and the extra time granted thereby helps to quickly facilitate the task (which is in line with findings by [2]).

With regard to the slider studies, the interpretation is similar. The initial lengthening is used by the participants, at least in the second study, the other lengthening conditions are not. In this extremely simplified setup, the lengthening seems to be interpreted as a prominence-related duration increase “a RED kind of blue”. If the lengthening had been interpreted as a hesitation, we would have expected our participants to not click the lengthened color, based on the findings in [6] which suggests that word-initial lengthening is associated with uncertainty about the word itself. This also suggests that compound-initial lengthening, as used in the studies presented here, works entirely different from word-initial lengthening. Whereas an utterance like "the grrrreen tree", with lengthening before the nucleus of the color word, will be interpreted as "maybe it is a yellow tree", an utterance like "the greeeen-blue tree", with lengthening on the nucleus of the first color word in the color compound seems to evoke different interpretations. Out of context, as in the studies presented here, there seems to be no reason for the listener to interpret it as hesitation, rather the interpretation seems to be related to prominence and emphasis. More research is needed to determine whether there are scenarios where lengthening of this kind is interpreted as hesitation or uncertainty.

4 Conclusion

The studies presented here provide basic research on the function of lengthening in conversation. It has been shown in earlier studies that lengthening is a useful feature not only in human communication, but also in human-machine interaction. Lengthening can buy dialogue time without the listener noticing it and it can help the listener facilitate tasks, which has been identified for hesitations like fillers and silences before, but rarely for lengthening, which is often under the radar. In this paper we cannot replicate the hesitation bias, but we have again found evidence for positive effects of hesitations on cognitive processes and reference resolution.

References

[1] SKANTZE, G. and A. HJALMARSSON: *Towards incremental speech generation in conversational systems. Computer Speech and Language 27*, pp. 243–262, 2013.

[2] BETZ, S., B. CARLMEYER, P. WAGNER, and B. WREDE: *Interactive hesitation synthesis: modelling and evaluation. Multimodal Technologies and Interaction*, 2(1), 2018.

[3] ARNOLD, J. E., C. L. H. KAM, and M. K. TANENHAUS: *If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension. Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(5), p. 914, 2007.

[4] BETZ, S., J. VOSSE, S. ZARRIESS, and P. WAGNER: *Increasing recall of lengthening*

detection via semi-automatic classification. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech 2017, Stockholm)*, pp. 1084–1088. 2017.

[5] BETZ, S.: *Hesitations in Spoken Dialogue Systems*. Ph.D. thesis, 2020. doi:10.4119/unibi/2942254. URL `https://nbn-resolving.org/urn:nbn:de:0070-pub-29422545,https://pub.uni-bielefeld.de/record/2942254`.

[6] BETZ, S., S. ZARRIESS, SZÉKELY, and P. WAGNER: *The greennn tree - lengthening position influences uncertainty perception*. In *Proceedings of Interspeech*, pp. 3990–3994. 2019. URL `https://nbn-resolving.org/urn:nbn:de:0070-pub-29363766, https://pub.uni-bielefeld.de/record/2936376`.

[7] KRAHMER, E. and M. SWERTS: *How children and adults produce and perceive uncertainty in audiovisual speech*. *Language and speech*, 48(1), pp. 29–53, 2005.

[8] PON-BARRY, H. and S. M. SHIEBER: *Recognizing uncertainty in speech*. *EURASIP Journal on Advances in Signal Processing*, 2011, pp. 1–11, 2011.

[9] DALL, R., M. WESTER, and M. CORLEY: *The effect of filled pauses and speaking rate on speech comprehension in natural, vocoded and synthetic speech*. In *Proc. Interspeech*, pp. 56–60. 2014.

[10] FISCHER, K., O. NIEBUHR, E. NOVÁK-TÓT, and L. C. JENSEN: *Strahlt die negative reputation von häsitationsmarkern auf ihre sprecher aus?* In *Proc. 43rd Annual Meeting of the German Acoustical Society (DAGA), Kiel, Germany*, pp. 1450–1453. 2017.

[11] BETZ, S., P. WAGNER, and J. VOSSE: *Deriving a strategy for synthesizing lengthening disfluencies based on spontaneous conversational speech data*. In *Phonetik und Phonologie 12*, pp. 19–23. 2016.

[12] DRAXLER, C.: *Online experiments with the percy software framework - experiences and some early results*. In N. C. C. CHAIR), K. CHOUKRI, T. DECLERCK, H. LOFTSSON, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK, and S. PIPERIDIS (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 235–240. European Language Resources Association (ELRA), Reykjavik, Iceland, 2014.

[13] SCHROEDER, M. and J. TROUVAIN: *The german text-to-speech synthesis system mary: A tool for research, development and teaching*. *International Journal of Speech Technology, 6.*, pp. 365–377, 2003.

[14] CAMPBELL, W. N. and S. D. ISARD: *Segment durations in a syllable frame*. *Journal of Phonetics*, 19(1), pp. 37–47, 1991.

[15] BETZ, S., J. VOSSE, and P. WAGNER: *Phone Elasticity in Disfluent Contexts*. In *Fortschritte der Akustik - DAGA 2017*, pp. 1462–1464. 2017.

[16] R CORE TEAM: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `https://www.R-project.org/`.

[17] BOERSMA, P. and D. WEENINK: *Praat: doing phonetics by computer [computer program]. http://www.praat.org/.* 2014.