Hesitation Lengthening Elicitation and Detection via Target Words in a Card Game Study

Simon Betz
Universität Bielefeld
simon.betz@uni-bielefeld.de

Abstract: We have created a strategic card game as a framework to elicit hesitations under varying degrees of cognitive load. In this study we investigate whether there is a special role of property words with regard to hesitations. We have shown that our framework elicits unusually large amounts of hesitations, now we investigate whether property words, specifically, color words exhibit a "magnetism" for hesitations. We assume this because our game setup dictates that colors are the main objects that participants have to consider, so it could be that hesitations gravitate towards the concepts that speakers think about. Another aspect that we regard in this study is the question whether it is possible to build a simple classifier that can detect hesitation lengthening on said color words using read samples as a baseline. To foreshadow the results, we do not observe a "hesitation magnetism" of property words, and the simple classification method exhibits high recall but low precision.

1 Introduction and Background

1.1 Hesitations

We have developed a card game for the elicitation of spontaneous speech phenomena related to cognitive load [1]. The phenomena of interest for us are hesitations. We use the term *hesitations* as an umbrella term for the following three phenomena: silences, fillers and lengthenings. Our working definitions for the three are:

- Silences are any intervals without speech by the active speaker.
- Fillers are vocalizations of a central vowel and optionally a nasal *(uh, uhm)*.
- Lengthenings are stretches of markedly elongated segments or syllables.

In general, hesitations temporally extend the speech signal, thus allowing extra time for speakers to formulate and for listeners to comprehend. This usefulness in dialogue stands in contrast to their bad reputation [2]. For thorough overviews on the topics of hesitation and disfluency, which is sometimes used as a synonym, see [3, 4, 5].

1.2 Cognitive Load

Cognitive load can generally be understood as the amount of working memory dedicated to an ongoing task. It has been thoroughly studied with regard to speech production [6] as well as in connection with hesitation and pausing behavior. [7] report that the frequency of fillers and silences are effective predictors of cognitive load. [8] observes a hesitation bias in perception experiments - in the presence of hesitations, listeners assume that the more complicated of two referents is the target. In our previous study [1], we found that hesitation frequency increases linearly as a function of cognitive load, administered by game complexity, cf. Sec.1.3.

1.3 The Game

We designed a game as a general framework to elicit ecologically valid hesitations while controlling cognitive load. Results reported in [1] indicate that the method works well, yielding very high hesitation rates (14.9 per minute) and revealing clear connections between cognitive load and spoken as well as gestural hesitations.

The game is a strategic card game inspired by established card and board games like Magic, Hearthstone, Terraforming Mars or Wingspan. All of these games have in common that they contain cards that consist of cost and effect. Players need to accumulate certain resources in order to pay the costs of cards, which then allows them to utilize their effect. Fig.1 shows example cards.

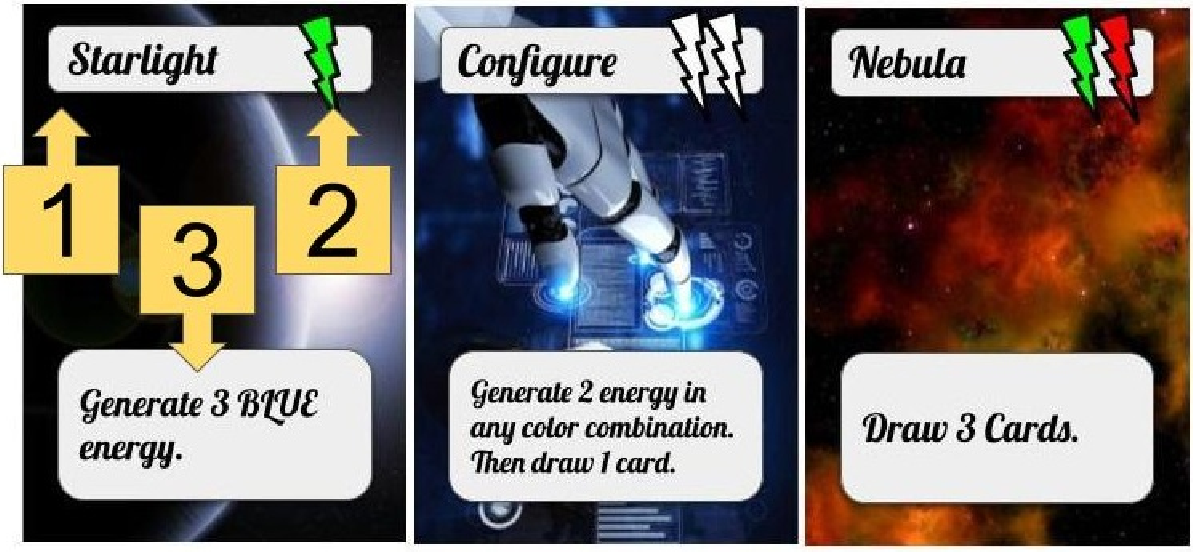


Figure 1 – Example cards. (1) name, (2) cost, (3) effect. Card costs explained: "Starlight" costs 1 green energy, "Configure" costs two energy of any color, "Nebula" costs 1 green and 1 red energy.

Our game features three types of card effects which are illustrated in the example cards in Fig.1: energy generation (left card), card drawing (right card), and energy type filtering (middle card). The game is played solitaire and the goal is to draw the entire deck of 21 cards. While doing so, the player is asked to communicate every game move to the experimenter. Game moves are playing cards and executing their effects, which may further involve moving tokens to count resources across the table. The crucial part of the game is to generate and filter energy to acquire the right amount and colors of energy to play cards. Fig.2 shows the game setup and a participant playing the game.

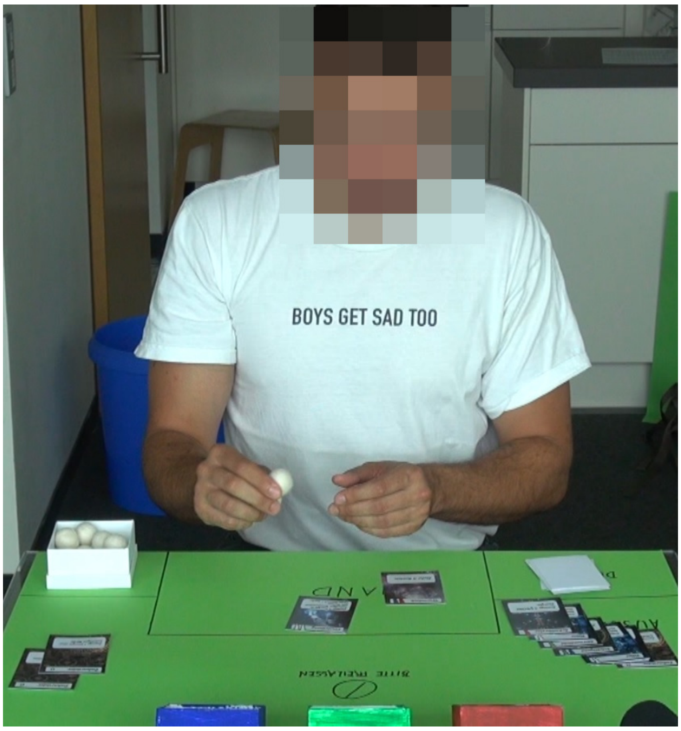


Figure 2 – Example of a participant playing the game.

Cognitive load within the game is modeled as a function of card complexity. Card complexity is operationalized as the number of choices a card offers. As an example, the cards in Fig.1 would have the following complexity values: the left and right cards would have 0 complexity as no choice is involved to play them. The costs and effects are clearly defined. In contrast, the middle card would have 4 complexity, as the cost and the effect offer 2 choices each, i.e. 2x which type of energy to spend and 2x which type to generate.

1.4 The Lengthening Problem

To our knowledge, our study [1] was the first attempt to investigate the effects of cognitive load on hesitation lengthening. The present study continues the research and addresses methodical questions. The aim of this study is to determine whether it is possible to make research and analysis of hesitation lengthening more effective. At the moment, there is the following problematic situation: hesitation lengthening has been shown to be a very versatile tool, both in human and in human-machine interaction [5, 9, 10]. It can generate extra dialogue time, which helps both speaker and listener, while at the same time it is not perceived as obtrusive as other hesitations, especially fillers, often are. This is most likely due to the fact that lengthening is a very subtle and elusive acoustic event that often passes unnoticed, even if clearly present in the signal [11]. This in turn means that even trained annotators frequently miss lengthenings [11], which in turn leads to a data sparsity problem. Generally, lengthening is regarded as a frequent hesitation phenomenon, but not as frequent as silences or fillers. In our previous study [1], however, lengthening occurred more often than fillers.

1.5 A Novel Approach to Lengthening Analysis

For this study, we explore new ways for lengthening analysis. In a previous study, we artificially inserted lengthening into perception test stimuli and were able to show that listeners resolve the referents differently based on the position of lengthening within a property word [12]. In this study, we approach lengthening from the other side: our game scenario prompts users to utter the property (color) words red, green and blue frequently and during increased cognitive load. We hypothesize that this method not only gives rise to a large amount of hesitation, but also that there will be hesitation lengthening predominantly on color words, as the colors are the main matter that participants of this experiment have to think about. If it is possible to so limit the loci of hesitation lengthening, it could provide a shortcut for lengthening annotation and analysis. Instead of annotating an entire corpus, it could be sufficient to only annotate the target words to create a data set for lengthening analysis. This aspect will be addressed as our first research question, cf. Sec.1.6. The classification of a word as hesitant or non-hesitant is usually done based on annotators' perception. For a target word-based analysis, a small baseline corpus of read out instances of the target words can be created to perform the classification automatically, or semi-automatically as an aid for the annotator. We hypothesize that the read-out words resemble non-lengthened instances of the same words in the in-game situation and that in-game instances of these words that are markedly longer than the read-out ones can be classified as lengthening; this will be investigated as our second research question for this study, cf. Sec.1.6.

1.6 Research Questions and Hypotheses

The research questions addressed in the present study are: (1) Does hesitation lengthening occur more often on color words than on other words? (2) Can we use a sample of read speech to build a simple classifier for hesitation lengthening? For (1) we hypothesize that this will indeed be the case. Question (2) is exploratory, we do not have a hypothesis for its outcome.

2 Methods

The main corpus data used in this study was recorded for the study described in [1]. It contains data of 19 participants (7 female, 12 male; median age = 30). The corpus has a total length of 110 minutes and contains 1648 hesitations, which amounts to 14.9 hesitations per minute.

In order to answer research question (1), we perform a simple exploration and determine the percentage of lengthening on a target color word compared to the total amount of lengthening.

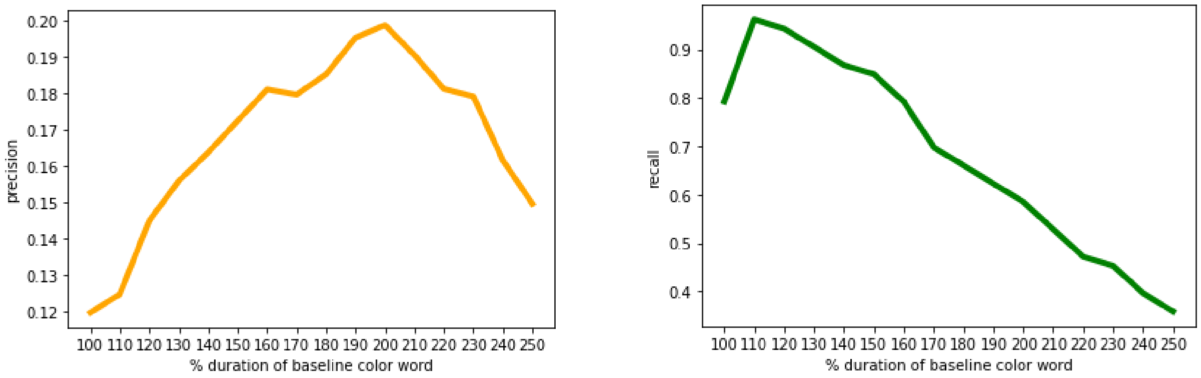
For research question (2) we build a simple classifier that uses read speech as a baseline. Precisely, we explore how much precision and recall we could achieve by comparing target word duration from read text to their counterparts in spontaneous speech in varying levels of tolerance. In addition to the corpus of game recordings used in [1], we created a corpus of read speech, recorded immediately after the game recording of each participant. Each participant read a short text of 11 statements containing a total of 12 target color words (4 each) in both phrase-medial and phrase-final position. The read-out texts were annotated automatically using the BAS Webservices [13]. One annotator then corrected the annotations of the target words to ensure that the right durations could be extracted. In addition, the root morpheme of each instance of a target word was annotated on a different tier. This is necessary as the recordings were carried out in German, which adds inflectional endings to some words, see examples 1 & 2:

ex.1.EN: *"This energy is blue."*
ex.1.DE: *"Diese Energie ist* ***blau****."*
ex.2.EN: *"I generate one blue energy"*
ex.2.DE: *"Ich erzeuge eine* ***blaue*** *Energie."*

Most instances of color words in the main corpus (495 of 606 or 81.7%) come as the root morpheme form or as the inflected form with {-e} suffixed. Any other realization is not regarded further in our analysis, as we lack appropriate baseline forms. For the remaining 495 instances of color words, we then tried to build a classifier that uses the duration of read color words to determine whether a given spontaneously uttered color word is hesitant or not. We calculated precision and recall for 16 subsequent levels of tolerance. The highest level of tolerance means that any color word from the main corpus that is longer than the baseline is classified as hesitation. We then lowered the tolerance step-wise by 10%. For the next tolerance level, a word from the main corpus must be longer than 110% of the baseline duration. The lowest tolerance level tested is 250% of the baseline duration. The baseline duration is the average duration of a word that matches in meaning (red, blue, green), in word form (with or without suffix {-e}) and in speaker. Doing so, we hope to find a "sweet spot" configuration for future analyses, i.e. a tolerance level that yields the best balance of precision and recall.

3 Results

The first question, whether lengthening predominantly falls on color words in our setup can be answered clearly with "no". Of a total of 498 lengthenings observed in our corpus, only 51 or 10.2% fall on color words (22 red, 15 blue, 14 green).



(a) Precision

(b) Recall

Figure 3 – Precision & recall as function of stepwise tolerance decrease

Turning to the second question the answer is a bit more complex. As can be seen in Fig.3a, precision in general remains below 0.2. It peaks at 200% duration. Recall, as can be seen in Fig.3b, is much higher, up to 0.96, peaking at 110% duration and then steadily declining. Plotting both precision and recall together (Fig.4), confirms that there is no "sweet spot" value to set the tolerance to, as the peaks of precision and recall are many steps apart.

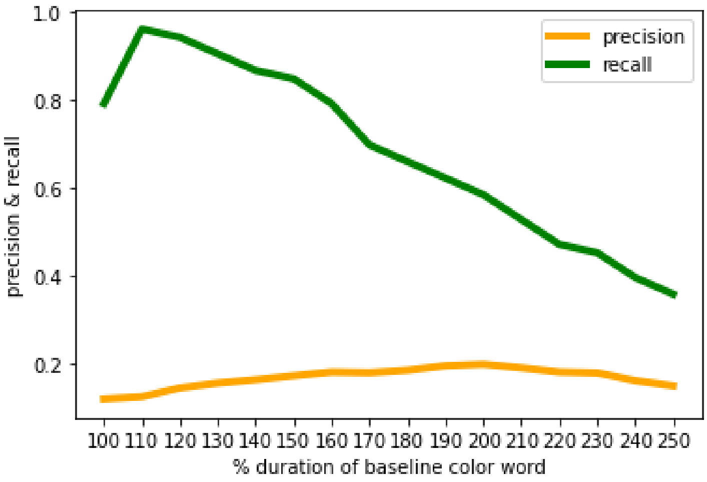


Figure 4 – Precision & recall combined

4 Discussion

In order to interpret the results, we need some reference values. With regard to our first research question, we observed that only 10% of hesitation lengthening in the corpus falls on color words. For comparison, 7% of the words in the corpus are color words. In total we observe 51 lengthened color words in 110 minutes of corpus data, which means that short-cutting the analysis and annotating only color words for lengthening would yield a rate of 0.46 hesitation lengthenings per minute. Interestingly, this rate would still be higher than the rate observed in the DUEL corpus (0.42) [14], but lower than the rate found in the GECO corpus (0.57) [15]; rates taken from [11]. This, however, might be due to the fact that our game setup yields a very high hesitation per minute rate in general (14.9) [1], so we cannot conclude that short-cutting the analysis and limiting it to color words is a good idea because of the large amount of instances that would be missed. Further, we have to reject our hypothesis that lengthening favors property words. From the clear interpretability of the function of lengthening on color words observed in a perception test [12], we would have expected that a production study that demands uttering large amounts of color words would produce large amounts of lengthened color words as well, but this is not the case. This reveals that cognitive load gives rise to large amounts of hesitations, but that the hesitations do not necessarily manifest themselves on the very concepts that speakers think about. This might be a downside of our scenario - there is a large amount of cognitive load in a relatively short span of time, which renders it difficult to map speech phenomena exactly to their underlying reasons.

Despite the low amount of color word lengthening, we investigated whether a simple classification method is viable. Results regarding this second question are mixed. The precision between 0.1 and 0.2 is very low, meaning that 80-90% of the hits are actually false positives. Recall, however, is very high, especially when the tolerance level is at 110%. These values incidentally resemble our previous attempts at semi-automatic lengthening classification based on duration: in [11] we attested 0.29 precision and 0.81 recall. The former study required full phonemic annotation of the corpus data but was able to account for any lengthening on any word that occurred. The present study only required certain target words, but as a result only accounted for lengthening on these very words. It is interesting to note, however, that regardless of the level of analysis, whether it be phonemes or words, the duration-based classification of lengthening seems to work similarly. In order to utilize this method further, using the classifier as an annotation aid would be conceivable. The classifier ensures that the human annotator does not miss instances due to the perceptual elusiveness of lengthening, the human annotator can filter out the false positives emitted by the machine.

5 Conclusion

In general we conclude, as observed in [1], that our card game scenario is a very promising environment for studying hesitation phenomena. In this study we investigated a side aspect of that scenario, namely the role of property words. The game is designed in a way that demands participants to utter large amounts of specific property words, in this case the color words red, green and blue. We expected that these property words would exhibit a "hesitation magnetism" and specifically attract hesitation lengthening, but this is not the case. We assume the reason for this is the high hesitation density which blurs the exact mapping of speech onto cognition. Consequently, this setup does not appear suitable for target word-based analyses. However, the exploration of the classifier suggests that if target words are available, a classifier can be a valuable assistant for the human annotator.

References

[1] BETZ, S., N. BRYHADYR, O. TÜRK, and P. WAGNER: *Cognitive load increases spoken and gestural hesitation frequency*. *Language*, Submitted.

[2] 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.

[3] EKLUND, R.: *Disfluency in Swedish human–human and human–machine travel booking dialogues*. Ph.D. thesis, Linköping University Electronic Press, 2004.

[4] LICKLEY, R. J.: *Fluency and Disfluency*, pp. 445–474. John Wiley & Sons, Inc, 2015. doi:10.1002/9781118584156.ch20. URL `http://dx.doi.org/10.1002/9781118584156.ch20`.

[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] LIVELY, S. E., D. B. PISONI, W. VAN SUMMERS, and R. H. BERNACKI: *Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences*. *The Journal of the Acoustical Society of America*, 93(5), pp. 2962–2973, 1993. doi:10.1121/1.405815. URL `https://doi.org/10.1121/1.405815`. `https://doi.org/10.1121/1.405815`.

[7] MONTACIÉ, C. and M.-J. CARATY: *High-level speech event analysis for cognitive load classification*. In *Proc. Interspeech 2014*, pp. 731–735. 2014. doi:10.21437/Interspeech.2014-110.

[8] 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.

[9] Betz, S., S. Zarriess, and P. Wagner: *Synthesized lengthening of function words - The fuzzy boundary between fluency and disfluency.* In L. Degand (ed.), *Proceedings of the International Conference Fluency and Disfluency*, pp. 15–19. 2017.

[10] Betz, S., B. Carlmeyer, P. Wagner, and B. Wrede: *Interactive hesitation synthesis: modelling and evaluation. Multimodal Technologies and Interaction*, 2(1), 2018.

[11] 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.

[12] 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`.

[13] Kisler, T., U. Reichel, and F. Schiel: *Multilingual processing of speech via web services.* *Computer Speech & Language*, 45, pp. 326–347, 2017. doi:10.1016/j.csl.2017.01.005.

[14] Hough, J., Y. Tian, L. de Ruiter, S. Betz, D. Schlangen, and J. Ginzburg: *DUEL: A Multi-lingual Multimodal Dialogue Corpus for Disfluency, Exclamations and Laughter.* In *10th edition of the Language Resources and Evaluation Conference*, pp. 1784–1788. 2016.

[15] Schweitzer, A. and N. Lewandowski: *Convergence of articulation rate in spontaneous speech.* In *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013, Lyon)*, pp. 525–529. 2013.