

# Dissertation in Progress: An Empirical Analysis of the Costs and Benefits of Naturalness in Spoken Dialog Systems

Ellen Campana

Department of Brain and Cognitive Sciences  
Meliora Hall, RC 220268  
University of Rochester  
Rochester, NY 14607  
United States

## Abstract

In this paper I describe work for my Ph.D. dissertation which is currently in progress. The overarching goal of the work is to develop a methodology for empirically evaluating the effects of different interface design decisions in spoken dialogue systems. The methodology I will use is the dual-task method, borrowed from cognitive psychology, which is advantageous because it provides fine-grained information about the cognitive load of the user while he/she is engaged in interacting with the system. For my dissertation I will focus specifically on the use of definite referring expressions and the question of whether “natural” or “fully-specified” definite referring expressions are easier for users to generate and/or understand. The answers are important because both strategies are used in systems on the market today. More importantly, I hope my work will provide a tool for software developers, and encourage them to carefully weigh the empirically observed costs and benefits of various design decisions.

## Background

An issue that is becoming central to dialog system research is the question of naturalness: Just how closely should Human-Computer Interaction (HCI) approximate Human-Human Interaction (HHI)? The simple (undoubtedly too simple) view is that there are two basic approaches one can take: One possibility is to develop systems that generate and understand a relatively limited set of scripted utterances that follow simple rules, with the user adapting to system limitations. A second possibility is to work toward developing systems that closely approximate human-human communication. Each approach has potential costs and benefits. On the one hand, people seem to be able to adapt their language based on the limitations of a system; however, it may come at the cost of devoting limited capacity mental resources to planning utterances rather than the task at hand. On the other hand, more natural systems require more sophisticated speech and language modules, as well as reasoning capabilities, greatly increasing their complexity; yet if done right they might require less user training and they might be easier to use when users are under stress. Both of these strong views

are championed in the literature though there is very little empirical information available concerning the potential tradeoffs.

The goal of my dissertation work is to develop an empirical methodology that will allow people to quantify and evaluate the costs and benefits of specific design features, thus turning debates into empirical questions. Specifically, I intend to conduct an online evaluation of dialog systems based on the cognitive load imposed on users as they interact verbally with the systems to carry out specific tasks, taking complexity into account. Cognitive load will be measured online using the dual-task paradigm, which is to my knowledge a novel approach in dialog system evaluation. As a test case I will focus specifically on referring expressions, contrasting systems that follow “natural” patterns of reference with those that follow more simplified, less “natural”, sets of rules for reference production and understanding. My hope is that some changes (such as restricting pronoun use unnaturally) will result in increased cognitive load while other changes (such as adding new classes of adjectives) will not. This would suggest that work aimed at making dialog systems more natural should be focused in particular directions—some things may be easy for people to adapt to, and therefore less practically relevant, while other things cause people great difficulty and warrant further research. Thus, the methodology would potentially help researchers decide which problems to focus on.

## Dissertation Outline

### Part I: Methodological Questions

I am extending a classic tool from cognitive psychology, the dual-task paradigm, to dialog system research. This approach is novel, so I will begin by addressing basic methodological issues. Specifically I will attempt to answer the following questions: 1) Is the chosen methodology sensitive to basic differences in cognitive load related to human language understanding? (Answer:

yes! Campana et al. 2004) 2) Is the chosen methodology sensitive to basic differences in cognitive load related to human language generation? 3) Is the chosen methodology sensitive enough to reveal practice effects? And finally, 4) Can the methodology be extended to interactive settings (in which users both speak and listen)?

## Part II: Cognitive Load As Users Understand “Natural” and “Less Natural” Definite References

Once I have demonstrated that the measure is sensitive to cognitive load when users’ are listening, I will demonstrate its usefulness as a measure using the test case of definite referring expressions. I chose this test-case because much is already known about what the “natural” patterns of reference production are within certain contexts, yet implementing these “natural” patterns in real-time dialog systems would greatly increase their complexity for most domains, and often involves domain-specific reasoning. In this part of my dissertation work, I’ll attempt to answer the following questions: 1) Do users experience increased cognitive load when following spoken instructions containing only fully-specified references (see figure 1), compared to the case in which they follow spoken instructions containing only “natural” references? 2) If there is a difference, is it more pronounced in situations where the “natural” reference would be pronoun or a reduced noun phrase? 3) If we observe increased cognitive load in the “less natural” cases, can the effect be reduced as users get more practice with the system? And 4) How long does it take?




## Part III: Cognitive Load As Users Generate “Natural” and “Less Natural” Definite References

After examining cognitive load during human language understanding, I will go on to investigate cognitive load during language generation, answering the questions outlined above for generation as for production. This stage is important because it may be the case that there are some as-yet undiscovered asymmetries between what is difficult for the users to generate, and what is difficult for users to understand. For instance, understanding may be more adaptive than generation. It will be important to discover any such asymmetries before going on to the capstone experiments in the series, described in the next section.

## Part IV: Cognitive Load in Interactive Settings

The final set of experiments in my dissertation will examine cognitive load in interactive situations. That is, situations in which the user both speaks to and listens to the system. I predict that the increased task complexity associated with the interactive setting will increase cognitive load, particularly in conditions that were difficult for users in Part II and Part III. In addition, I predict that effects of practice observed in Part I and Part II will be less pronounced.

**Table 1:** Examples of “natural” and “less natural” references to the leftmost object in each row. Crucially, the reference depends on the context in the “natural” condition, but it is independent of context in the “less natural” condition.

Visual and Discourse Contexts	“Natural” Ref.	Fully-Specified Ref.
 (No previous instruction)	“the circle”	“the big blue circle”
 Previous instruction: “Move the big blue square.”	“it”	“the big blue square”
 Previous instruction: “Move the circle next to the big blue square.”	“the square”	“the big blue square”