Speech Graffiti habitability: What do users really say?

Stefanie Tomko and Roni Rosenfeld
Language Technologies Institute, School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh PA 15213
{stef, roni}@cs.cmu.edu

Abstract

The Speech Graffiti interface is designed to be a portable, transparent interface for spoken language interaction with simple machines and information servers. Because it is a subset language, users must learn and adhere to the constraints of the language. We conducted a user study to determine habitability and found that more than 80% of utterances were Speech Graffiti-grammatical, suggesting that the language is acceptably learnable and usable for most users. We also analyzed deviations from grammaticality and found that natural language input accounted for the most deviations from Speech Graffiti. The results will suggest changes to the interface and can also inform design choices in other speech interfaces.

1 Introduction

Speech Graffiti (a.k.a. the Universal Speech Interface project) is an attempt to create a standardized, speech-based interface for interacting with simple machines and information servers. Such standardization offers several benefits, including domain portability, lower speech recognition error rates and increased system transparency for users (Rosenfeld et al., 2001). This is realized via a subset language that must be learned and remembered by users.

This study was designed to assess habitability by measuring Speech Graffiti-grammaticality: how often do users actually speak within the subset grammar when interacting with a Speech Graffiti system? A high level of grammaticality would suggest that the language is reasonably habitable, while low grammaticality would indicate that the language design requires substantial changes.

1.1 Speech Graffiti basics

Speech Graffiti interfaces comprise a small set of standard rules and keywords that can be used in all Speech Graffiti applications. The rules are principles governing the regularities in the interaction, such as *"input is always provided in phrases with the syntax* 'slot `is` value*'"* and *"the system will tersely paraphrase whatever part of the input it understood."* The keywords are designed to provide regular mechanisms for performing interaction universals such as help, orientation, navigation and error correction.

By standardizing user input, Speech Graffiti aims to reduce the negative effects of variability on system complexity and recognition performance. At the same time, we hope that introducing a universal structure that is intended to be used with many different applications will mitigate any negative effects that might be otherwise associated with learning an application-specific command language.

1.2 Related work

Although several studies have previously explored the usage of constrained or subset languages (for example, Hendler & Michaelis, 1983; Guindon & Shuldberg, 1987; Ringle & Halstead-Nussloch, 1989; Sidner & Forlines, 2002), they have generally been concerned with performance effects such as task completion rates. Sidner & Forlines (2002) reported a "correct utterance" rate of approximately 60-80% for their user studies, although this was not a main focus of their work. While we understand the focus on such performance measures, we believe that it is also important to understand how habitable the constrained language is for users, in what ways users deviate from it, and what impact habitability has on user satisfaction.

2 Method

Our data was generated from a user study in which participants were asked to complete tasks using both a Speech Graffiti interface to a telephone-based movie information system (MovieLine) and a natural language interface to the same data. Tasks were designed to have the participants explore a variety of the functions of the systems (*e.g.* "list what's playing at the Squirrel Hill Theater" and "find out & write down what the ratings are for the movies showing at the Oaks Theater").

After interacting with each system, each participant completed a user satisfaction questionnaire rating 34 subjective-response items on a 7-point Likert scale. This questionnaire was based on the Subjective Assessment

of Speech System Interfaces (SASSI) project (Hone & Graham, 2001) and included statements such as "I always knew what to say to the system" and "the system makes few errors." An overall user satisfaction rating was calculated for each user by averaging that user's scores for each of the 34 response items. Users were also asked a few comparison questions, including system preference. In this analysis we were only concerned with results from the Speech Graffiti MovieLine interactions and not the natural language MovieLine interactions (see Tomko & Rosenfeld, 2004). System presentation order was balanced and had no significant effect on grammaticality measures.

2.1 Participants

Twenty-three participants (12 female, 11 male) accessed the systems via telephone in our lab. Most were undergraduate students from Carnegie Mellon University and all were native speakers of American English. We also asked users whether they considered themselves "computer science or engineering people" (CSE) and how often they did computer programming; the distributions of these categories were roughly equal.

2.2 Training

The Speech Graffiti approach requires users to learn the system prior to using it via a brief tutorial session. 15 participants received *unsupervised* Speech Graffiti training consisting of a self-directed, web-based tutorial that presented sample dialog excerpts (in text) and proposed example tasks to the user. The other eight participants received *supervised* Speech Graffiti training. This training used the same web-based foundation as the unsupervised version, but participants were encouraged to ask the experimenter questions if they were unsure of anything during the training session.

Both supervised and unsupervised training sessions were balanced between web-based tutorials that used examples from the MovieLine and from a FlightLine system that provided simulated flight arrival, departure, and gate information. This enabled us to make an initial assessment of the effects of in-domain training.

2.3 Analysis

The user study generated 4062 Speech Graffiti MovieLine utterances, where an utterance is defined as one chunk of speech input sent to our Sphinx II speech recognizer (Huang et al., 1993). We removed all utterances containing non-task-related or unintelligible speech, or excessive noise or feed, resulting in a cleaned set of 3626 utterances (89% of the total). We defined an utterance to be grammatical if the Phoenix parser (Ward, 1990) used by the system returns a complete parse with no extraneous words.

3 Results

82% (2987) of the utterances from the cleaned set were fully Speech Graffiti-grammatical. For individual users, grammaticality ranged from 41.1% to 98.6%, with a mean of 80.5% and a median of 87.4%. These averages are quite high, indicating that most users were able to learn and use Speech Graffiti reasonably well.

The lowest individual grammaticality scores belonged to four of the six participants who preferred the natural language MovieLine interface to the Speech Graffiti one, which suggests that proficiency with the language is very important for its acceptance. Indeed, we found a moderate, significant correlation between grammaticality and user satisfaction, as shown in Fig. 1 ($r = 0.60$, $p < 0.01$). We found no similar correlation for the natural language interface, using a strict definition of grammaticality.

Users' grammaticality tended to increase over time. For each subject, we compared the grammaticality of utterances from the first half of their session with that of utterances in the second half. All but four participants increased their grammaticality in the second half of their Speech Graffiti session, with an average relative improvement of 12.4%. A REML analysis showed this difference to be significant, $F = 7.54$, $p < 0.02$. Only one of the users who exhibited a decrease in grammaticality over time was from the group that preferred the natural language interface. However, although members of that that group did tend to increase their grammaticality later in their interactions, none of their second-half grammaticality scores were above 80%.

Summary by training and system preference. No significant effects on Speech Graffiti-grammaticality were found due to differences in CSE background, programming experience, training supervision or training domain. This last point suggests that it may not be necessary to design in-domain Speech Graffiti tutorials;

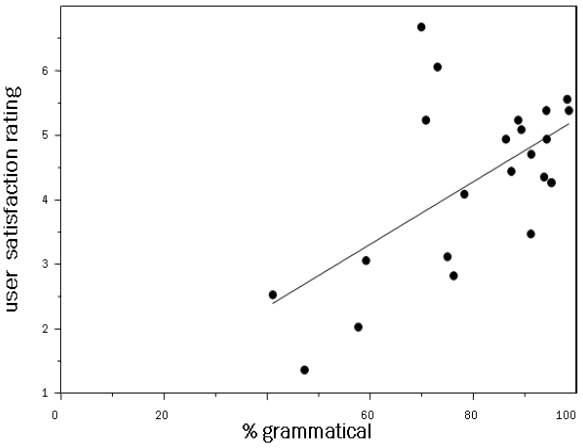

Figure 1. Grammaticality and user satisfaction for Speech Graffiti MovieLine.

instead, a single training application could be developed. The six users who preferred the natural language MovieLine generated 45.4% of the ungrammaticalities, further supporting the idea of language proficiency as a major factor for system acceptance.

3.1 Deviations from grammar

To help determine how users can be encouraged to speak within the grammar, we analyzed the ways in which they deviated from it in this experiment. We identified 14 general types of deviations from the grammar; Fig. 2 shows the distribution of each type. Four trivial deviation types (lighter bars in Fig. 2) that resulted from unintentional holes in our grammar coverage comprised about 20% of the ungrammaticalities. When these trivial deviations are counted as grammatical, mean grammaticality rises to 85.5% and the median to 91.3%. However, we have not removed the trivial ungrammaticalities from our overall analysis since they are likely to have resulted in errors that may have affected user satisfaction. Each of the ten other deviation types is discussed in further detail in the sections below.

General natural language syntax, 20.6%: Speech Graffiti requires input to have a *slot is value* phrase syntax for specifying and querying information. The most common type of deviation in the Speech Graffiti utterances involved a natural language (NL) deviation from this standard phrase syntax. For example, a correctly constructed Speech Graffiti query to find movie times at a theater might be *theater is Galleria, title is* Sweet Home Alabama*, what are the show times?* For errors in this category, users would instead make more NL-style queries, like *when is* Austin Powers *playing at Showcase West?*

Slot only, 14.6%: In these cases, users stated a slot name without an accompanying value or query words. For example, a user might attempt to ask about a slot without using *what*, as in *title is Abandon, show times*. In about a third of slot-only instances, the ungrammatical input appeared to be an artifact of the *options* function, which lists slots that users can talk about at any given point; users would just repeat back a slot name without adding a value, confirming Brennan's (1996) findings of lexical entrainment.

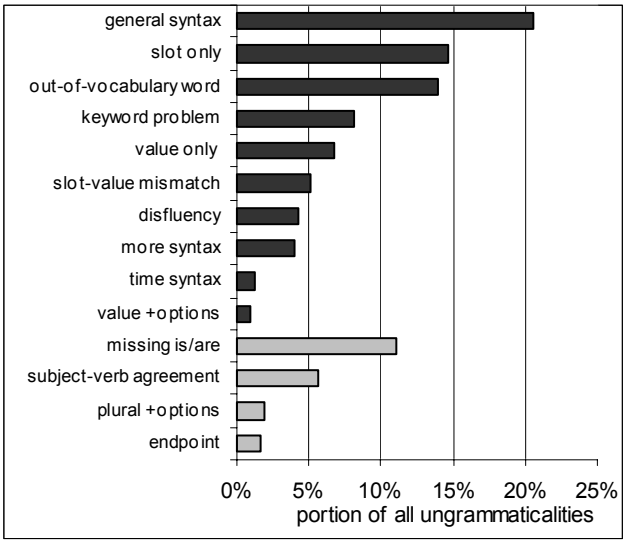

Figure 2. Distribution of ungrammatical utterances by type.

Out-of-vocabulary word, 14.0%: These were often movie titles that were not included in the database or synonyms for Speech Graffiti-grammatical concepts (*e.g. category* instead of *genre*).

Keyword problem, 8.1%: Participants used a keyword that was not part of the system (*e.g. clear*) or they used an existing keyword incorrectly.

Value only, 6.7%: Users specified a value (*e.g. comedy*) without an accompanying slot name.

Slot-value mismatch, 5.1%: Users paired slots and values that did not belong together. This often occurred when participants were working on tasks involving locating movies in a certain neighborhood. For instance, instead of stating *area is Monroeville*, users would say *theater is Monroeville*. Since the input is actually in the correct *slot is value* format, this type of ungrammaticality could perhaps be considered more of a disfluency than a true habitability problem.

Disfluency, 4.3%: This category includes utterances where the parser failed because of disfluent speech, usually repeated words. 81% of the utterances in this category were indeed grammatical when stripped of their disfluencies, but we prefer to leave this category as a component of the non-trivial deviations in order to account for the unpredictable disfluencies that will always occur in interactions.

***More* syntax, 4.0%:** This is are a special case of a keyword problem in which participants misused the keyword *more* by pairing it with a slot name (*e.g. theater, more*) rather than using it to navigate through a list.

Time syntax, 1.3%: In this special case of natural language syntax ungrammaticality, users created time queries that were initially well-formed but which had time modifiers appended to the end, as in *what are show times after seven o'clock?*

Value + *options*, 1.1%: In grammatical usage, the keyword *options* can be used either independently (to get a list of all available slots) or paired with a slot (to get a list of all appropriate values for that slot). In a few cases, users instead used *options* with a value, as in *Squirrel Hill options*.

4 Discussion

We have shown a significant correlation between grammaticality and user satisfaction for the Speech Graffiti system. Grammaticality scores were generally high and tended to increase over time, demonstrating that the system is acceptably habitable.

Based on the data shown in Fig.1, it appears that 80% is a good target for Speech Graffiti grammaticality. Nearly all participants with grammaticality scores over 80% gave positive (*i.e.* > 4) user satisfaction scores, and more than half of our users achieved this level. Furthermore, users with grammaticality above 80% completed an average of 6.9 tasks, while users with grammaticality below 80% completed an average of only 3.5 tasks. A fundamental question for our future work is "what can we do to help everyone speak within the bounds of the system at the 80% level?"

Several possible refinements are immediately apparent beyond fixing our trivial grammar problems. System responses to *options* should be reworked to reduce incorrect lexical entrainment and alleviate slot-only deviations. The out-of-vocabulary instances can be analyzed to decide whether certain synonyms should be added to the current system, although this will generate only domain-specific improvements. Many ungrammaticality types can also be addressed through refinements to Speech Graffiti's help and tutorial functions.

Addressing the general NL syntax category poses the biggest problem. Although it is responsible for the largest portion of ungrammaticalities, simply changing the grammar to accommodate these variations would likely lead to increased system complexity. A main concern of our work is domain portability, and Speech Graffiti's standard structure currently allows for fairly rapid creation of new interfaces (Toth et al., 2002). Any natural language expansion of Speech Graffiti grammar will have to be balanced with the ability to port such a grammar to all domains. We are currently analyzing the ungrammatical utterances in this and the time syntax categories to determine whether any Speech Graffiti-consistent modifications could be made to the interface. However, most of the improvement in this area will likely have to be generated via better help and training.

An important additional finding from this work is the scope of general NL syntax deviations. Considering items like movie and theater names as equivalence class members, the NL utterances used by participants in the Speech Graffiti system reduced to 94 patterns. In comparison, the NL utterances used by participants in the natural language MovieLine reduced to about 580 patterns. One of the main differences between the NL patterns in the two systems was the lack of conversational phrases like "can you give me…" and "I would like to hear about…" in the Speech Graffiti system. Thus the knowledge that they are interacting with a restricted language system seems to be enough to make users speak more simply, matching results from Ringle & Halstead-Nussloch (1989).

Although many of our ungrammaticality types may appear to be specific to Speech Graffiti, they reinforce lessons applicable to most speech interfaces. The slot-only issue demonstrates that lexical entrainment truly is a factor in spoken language interfaces and its effects should not be underestimated. Out-of-vocabulary words are a persistent problem, and keywords should be chosen with care to ensure that they are task-appropriate and that their functions are as intuitive as possible.

Overall, this study has provided us with a target level for Speech Graffiti-grammaticality, suggested changes to the language and provided insight about what aspects of the system might need greater support through help and tutorial functions. We plan to implement changes based on these results and re-evaluate the system through further user testing.

References

Brennan, S.E. 1996. Lexical entrainment in spontaneous dialog. In *Proceedings of the International Symposium on Spoken Dialogue*, pp. 41-44.

Guindon, R. & Shuldberg, K. 1987. Grammatical and ungrammatical structures in user-adviser dialogues: evidence for sufficiency of restricted languages in natural language interfaces to advisory systems. In *Proc. of the Annual Meeting of the ACL*, pp. 41-44.

Hendler, J. A. & Michaelis, P. R. 1983. The Effects of Limited Grammar On Interactive Natural Language. In *Proceedings of CHI*, pp. 190-192.

Hone, K. & Graham, R. 2001. Subjective Assessment of Speech-System Interface Usability. In *Proceedings of Eurospeech,* Aalborg, Denmark.

Huang, X., Alleva, F., Hon, H.W., Hwang, M.Y., Lee, K.F. & Rosenfeld, R. 1993. The Sphinx-II Speech Recognition System: An Overview. *Computer, Speech and Language,* 7(2):137-148.

Ringle, M.D. & Halstead-Nussloch, R. 1989. Shaping user input: a strategy for natural language design. *Interacting with Computers* 1(3):227-244

Rosenfeld, R., Olsen, D. & Rudnicky, A. 2001. Universal Speech Interfaces. *Interactions,* 8(6):34-44.

Sidner, C. & Forlines, C. 2002. Subset Languages for Conversing with Collaborative Interface Agents. In *Proc. of ICSLP*, Denver CO, pp. 281-284.

Tomko, S. & Rosenfeld, R. 2004. Speech Graffiti vs. Natural Language: Assessing the User Experience. To be published in *Proc. of HLT/NAACL.*

Toth, A., Harris, T., Sanders, J., Shriver, S. & Rosenfeld, R. 2002. Towards Every-Citizen's Speech Interface: An Application Generator for Speech Interfaces to Databases. In *Proc. of ICSLP*, Denver, CO.

Ward, W. 1990. The CMU Air Travel Information Service: Understanding Spontaneous Speech. In *Proc. of the DARPA Speech and Language Workshop.*