Learning to Speak to a Spoken Language System: Vocabulary Convergence in Novice Users

Gina-Anne Levow
University of Chicago
levow@cs.uchicago.edu

Abstract

A key challenge for users and designers of spoken language systems is determining the form of the commands that the system can recognize. Using more than 60 hours of interactions, we quantitatively analyze the acquisition of system vocabulary by novice users. We contrast the longitudinal performance of long-term novice users with both expert system developers and guest users. We find that novice users successfully learn the form of system requests, achieving a significant decrease in ill-formed utterances. However, the working vocabulary on which novice users converge is significantly smaller than that of expert users, and their rate of speech recognition errors remains higher. Finally, we observe that only 50% of each user's small vocabulary is shared with any other, indicating the importance of the flexibility of a conversational interface that allows users to converge to their own preferred vocabulary.

Keywords Spoken Language System; Novice-Expert; Lexical Entrainment

1 Introduction

Most currently deployed interactive spoken language systems employ a restricted vocabulary and syntax for system commands. These constraints provide greater recognition accuracy and faster recognition times from the system's perspective. (Makhoul, 1993) However, they also impose upon the system developer the need to provide a command language that is expressive enough to accomplish the tasks for which the speech system was designed and flexible enough to allow use by a wide variety of users with different levels of experience with the system. In turn, the users must learn the constrained language that is understood by the system. A conversational interface attempts to step away from a rigid command language with, for example, a single form for any command, to provide a set of well-formed inputs that have more varied and natural syntax and admit a range of synonymous terms and constructions. While it has been demonstrated that even with substantial synonymy, users will still choose terms outside the system's vocabulary some percentage of the time (Furnas et al., 1987), it is hoped that the flexibility of a conversational interface will allow some natural variability due to individual speaking styles and potentially ease the task for novice users. A key challenge for the user is thus to produce well-formed input to the system under these restrictions, and for the system designer to provide a set of commands that it is easy for the user to learn. (Brennan, 1998) demonstrate that users adopt the system's terminology, most reliably with explicit correction, but also with implicit correction, similar to the way in which pairs of human speakers converge on a lexical referent. (Walker et al., 1998) observe anecdotally that users learn system vocabulary over time. (Yankelovich, 1996) and (Kamm et al., 1998) explore techniques to guide users to produce well-formed queries, with a variety of strategies and tutorials, respectively. The above studies have focussed on pure novice users within their first few interactions with the system and on the goal of task achievement. Here, we analyze quantitatively the process by which users learn the language understood by the system, by exploring natural interactions during the course of a field trial conducted over a period of months. We analyze not only task completion or command recognition, but also the vocabulary acquired itself.

2 Data Collection

2.1 Speech System Description

The speech system utilized in the field trial is a prototype spoken language system that provides a voice-only interface to a variety of common desktop and information feed services. Specifically, it included e-mail reading and sending, access to one's own and other browsable calendars, weather information, stock quotes, time zone and currency conversions, and a notification system.

Two significant features distinguish this system from other spoken language systems. First, since it was designed primarily for use over the telephone to provide ubiquitous access, it is a voice-only system. With the exception of password entry and escape sequences which use telephone keypad input, all user input is spoken, and all output is through synthesized speech; there are no visual displays for feedback. Speech recognition is performed by BBN's Hark speaker-independent continuous speech recognizer, and synthesis is performed by Centigram's TruVoice text-to-speech system. Locally developed natural language processing and interpretation engines feed the speech recognition results to the appropriate speech application interfaces for each backend system.

Secondly, the spoken language system was designed to provide a ”conversational” interface. A conversational interfaces hopes to provide both ease of use for novice users and efficiency for more experienced users by allowing them to use language which comes naturally for each individual. In addition, it is easy to combine commands or criteria for requests into a single command for more confident and experienced users (e.g. read the third urgent message) or to simply step through the information with a sequence of simple commands for novice users (e.g. ”Go to urgent messages” , ”Next”, ”Next”, ”Next”). All new users are provided with a wallet-sized information card with examples of common commands for each application, but, as we will demonstrate later in this paper, users each rapidly develop their own distinct style and vocabulary.

2.2 Data Collection

Now that we have provided a general overview of the spoken language system, let us turn to a more detailed description of the data collection process. The system was deployed for a field trial to a limited number of participants over an analog telephone connection. All interactions were recorded automatically during the course of the conversation. All speech, both user input and system synthesized responses were digitized and stored at 8kHz sampling rate in 8-bit mu-law encoding on a single channel. Approximately sixty hours of interactions conducted over several months were recorded. In addition to the stored audio, speech recognizer results, natural language analysis results, and the text of all system responses was recorded and time stamped.

2.3 Subjects

The subjects participating in the field trial fell into three distinct classes:

1. Novice Users: Fourteen individuals drawn from the corporation's sales, marketing, and technical staff with no previous experience with this spoken language system.

2. Expert Users: Four long-term members of the system's development staff who had used and worked on the system for at least twelve months.

3. Guest Users: These were one-time users who called in to try out the system from a publicized phone number.

There were three female, two novice and one expert, and fifteen male regular system users, twelve novice and 3 expert. The users engaged in at least ten phone conversations with the system. The distribution of users allows us to examine the development of novice users' interaction style, in terms of vocabulary choice and number of out-of-vocabulary (OOV) utterances. In addition, we can contrast the different recognition accuracy rates and vocabulary distributions of expert and novice users.

2.4 Data Coding

All user utterances were textually transcribed by a paid transcriber. Each transcription of user input was paired with the speech recognizer output for that utterance. Each of these pairs was assigned one of four accuracy codes: Correct, Error minor, Error, or Rejection. The use of the ”Correct” code should be evident. The ”error minor” code assignments generally resulted from a misrecognition of a non-content word (e.g. wrong tense of an auxiliary verb, incorrect article, insertion of ”um” or ”uh”) for which the robust parsing of the natural language component could compensate. The ”error” and ”rejection” codes were assigned in those cases where a user could identify a failure in the interaction. Utterances coded either as Error or Rejection could also receive an additional tag, OOV (Out Of Vocabulary). This tag indicates that either words not in the recognizer's vocabulary or constructions not in the system's grammar were used in the utterances. For simplicity, however, we refer to all these cases as OOV. Two examples appear below:

Unknown Word: Rejection
User Said: Abracadabracadabra
System Heard: <nothing>
Unknown Construction: Misrecognition
User Said: Go to message five eight six
System Heard: Go to message fifty six
Grammar knows: Go to message five hundred eighty six

3 Analysis

In total, there were 7529 recorded user utterances from the field trial. Of these, 4865 were correctly recognized by the speech recognition pass, and 702 contained minor recognition errors, but still resulted in the desired action. There were 1961 complete recognition failures: 1250 of which were rejection errors and 706 of which were substitution misrecognition errors. The remaining errors were due to system crashes or parsing errors. In other words, almost two-thirds of recognition failures were rejections, about twice the number of misrecognitions. 2 Overall, this results in a 25% error rate.

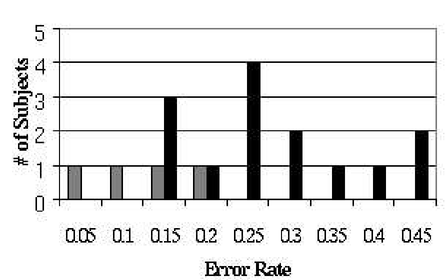

Figure 1: Distributions of Error Rates: Novice (Dark) vs Expert (Light)

3.1 Vocabulary Changes

Now that the basic form of the data has been described we can begin to explore in detail the topic of vocabulary convergence or lexical entrainment in the users in this study. We will look at three specific features of user vocabulary: error and out-of-vocabulary (OOV) rates over time, vocabulary size and rate of new words over time, and degree of vocabulary overlap among users.

3.2 Error and OOV Rates

Let us begin with a longitudinal examination of error and out-of-vocabulary utterance rates. Overall rates are given as averages, and longitudinal rates are in utterances per hundred. Figure 3.2 compares the distributions of overall average error rates for all novice users to that for expert users; Figure 3.2 restricts the comparison to error rates due to out-of-vocabulary or out-of-grammar utterances alone for these two groups. A comparison of novice users with expert users indicates a significantly higher rate of overall recognition (24.86% versus 10.75%) and OOV (7.39% versus 0.76%) errors for novices than for expert users.

The next important question to address is whether these errors rates, especially the higher novice user error rates, change over time, and if so, how and how much. To track these longitudinal changes, or changes over time, we recompute the error and OOV rates from above in terms of the number of errors per hundred utterances for the first, second, and third set of one hundred utterances, and so on.

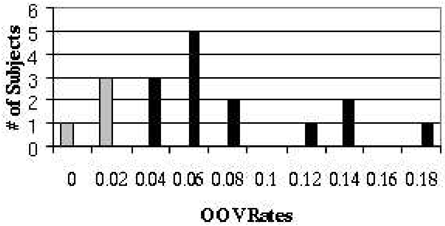

Figure 2: Distributions of OOV Rates: Novice (Dark) vs Expert (Light)

We observe that neither the expert users nor the guest users show any significant change in error rate over time. However, novices show a distinct decrease in errors from the first hundred utterances to the second hundred to a relatively stable and lower error rate. We can quantify this contrast by comparing number of errors in the first hundred utterances to the average number of errors per hundred utterances for the later interactions. This contrast is a significant decrease by t-test, paired, two-tailed. ($p < 0.05$), showing that novice users make fewer errors over time, but still at a much higher rate than expert users.[1]

This observation comes as no surprise; however, we would like to know which features of novice vs. expert user interaction account for this contrast. Specifically, to what degree do out-of-vocabulary utterances or speech acoustics differentially affect the error rates of these two subject groups? Can all contrasts be related to limited knowledge of the system's vocabulary? Experts, naturally, exhibit very few instances of out-of-vocabulary utterances. Here we consider the change in rate of OOV's in novice user utterances over time and contrast it with that of the guest user class. There is a significant decrease in OOV's over time for longer term users in contrast with an almost constant OOV rate for guest users and for expert users. Specifically there is a significant decrease in the number of OOVs between the first 200 utterances and all subsequent interactions. This is clearly a desirable trend, indicating the new users' increasing familiarity with the limited vocabulary understood by the system.

However, by comparing error rates in the first hundred utterances to the average of subsequent hundred utterance sets, we see that when these figures are computed without the errors contributed by OOV-related errors, the decrease in error rates with time is not significant. The decrease in OOV errors is thus the primary contributor to the perceived improvement in recognition rate over time. In addition, even with all OOV errors removed, the error rates of novices are still much higher than those of expert users (18.25% versus 10.25%), indicating that expert use of a spoken language system requires more than just the knowledge of the utterances understood by the system. This knowledge is acquired fairly rapidly as we see by the drop in OOV rates, but the knowledge of proper speaking style is more difficult.

[1]For longitudinal analysis, we consider only those users with more than 200 turns with the system.

1.00	0.30	0.44	0.48	0.41	0.48	0.30	0.37	0.41
0.21	1.00	0.53	0.34	0.26	0.34	0.34	0.42	0.37
0.19	0.32	1.00	0.22	0.24	0.27	0.21	0.32	0.24
0.33	0.33	0.36	1.00	0.26	0.36	0.36	0.28	0.33
0.42	0.38	0.58	0.38	1.00	0.31	0.31	0.35	0.31
0.41	0.41	0.53	0.44	0.25	1.00	0.38	0.38	0.44
0.33	0.54	0.54	0.58	0.33	0.50	1.00	0.33	0.46
0.33	0.53	0.67	0.37	0.30	0.40	0.27	1.00	0.40
0.37	0.47	0.50	0.43	0.27	0.47	0.37	0.40	1.00

Table 1: Proportion of Two Subjects' Vocabulary that is Shared

3.3 Vocabulary Size and Rate of New Word Introduction

The next question to address is how to account for this decrease in OOVs. Does the user simply replace unknown word instances with known words? Does the user's working vocabulary increase, decrease or stay the same?

Here we will use two measures to try to clarify the process of OOV reduction: number of words in working vocabulary (defined as number of discrete words per hundred words spoken) and rate of introduction of new words into the working vocabulary (again in words per hundred). Unsurprisingly, the rate of new word introduction undergoes a significant decrease over time - for all except the guest user category - and, like OOVs, drops dramatically after the first 200-300 words. Analysis of variance of number of new words to point in time is highly significant (F=59.27, df=323, $p < 0.001$)

The trend for the working vocabulary is quite interesting and somewhat unexpected. Again, paralleling the decrease in word introduction, there is a significant decrease in vocabulary size over time. Specifically, there is a significant decrease in the number of unique words per hundred between the first 200-300 words and all later interactions. (F = 8.738, df = 19, $p < 0.01$) Specifically, novice users, after working with the system for an extended period of time, converge on a surprisingly small working vocabulary of an average of 35 distinct words per hundred. This small vocabulary size contrasts strongly with the 50 distinct words per hundred of the expert users [2]. From these analyses, we can see that the decrease in out-of-vocabulary utterances arises from a narrowing of the users' working vocabulary to a fairly small set of words in which the user has high confidence.

3.4 Vocabulary Overlap

What ramifications does this use of a small working vocabulary have for conversational speech user interface design? Is it simply irrelevant since only a small set of words is needed by any user? An analysis of cross-user vocabulary will help to answer these questions. Here we tabulated the percentage of words shared between any pair of users and the percentage of a user's vocabulary that overlaps with any other's. We see that, for any pair of users, between 18 - 57% of vocabulary is held in common, with an average of 21% of the union of the two vocabularies falling in the intersection (Table 1). This translates to each user sharing approximately 50% of their words with any other given user.

This relatively small proportion of overlap between users attests to the value of the conversational interface. While the users individually do not have large vocabularies, the choice of words across users is highly varied. This supports the notion of a flexible vocabulary that allows users to gravitate toward lexical usages that come naturally, and supports wide cross-user utility.

4 Discussion & Conclusion

We observe the significant reduction in recognition errors, largely through a reduction in ill-formed utterances, of novices over their first two to three hundred utterances. This accomplishment supports the anecdotal reports that users learn system vocabulary over time, but most impressively, demonstrates the speed with which users acquire the necessary vocabulary, even in the absence of explicit guidance or correction. Within three hundred utterances, the once-novice users have achieved a rate of out-of-vocabulary errors that rivals that of system developers.

However, these skilled novice users still differ significantly from expert users in two respects: overall recognition accuracy and working vocabulary size. Novice users gradually remove ill-formed utterances from their input to the system. They achieve this result by converging on a small working vocabulary in which they have high confidence. Interestingly, this vocabulary varies substantially among users, suggesting an advantage to the conversational interface that allows users more flexibility in their choice of words and constructions. We still find, though, that even if we exclude all errors resulting from out-of-vocabulary utterances from consideration, novice users suffer from significantly worse speech recognition performance than do the expert system developers. These limitations in overall speech recognition accuracy and restricted vocabulary indicate that additional training that guides users to a suitable speaking style and full exploitation of the system's vocabulary and capabilities is necessary for the competent novice users to become true experts.

Acknowledgments We thank Nicole Yankelovich and Sun Microsystems for access to the field trial data and their assistance in transcription of these materials.

References

S. Brennan, 1998. *The grounding problem in conversations with and through computers*, pages 201–225. Lawrence Erlbaum.

G. Furnas, T. Landauer, L. Gomez, and S. Dumais. 1987. The vocabulary problem in human-system communications. *Communications of the ACM,* 30:964–971.

C. Kamm, D. Litman, and M. Walker. 1998. From novice to expert: the effect of tutorials on user expertise with spoken dialogue systems. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)*, pages 1211–1214.

[2]The expert users do not, in fact, use more of the system applications than novices.

J. Makhoul. 1993. Overview of speech recognition technology. colloquium presentation, human-machine communication by voice. National Academy of Sciences, Irvine, CA.

M. Walker, J. Fromer, G. Di Fabbrizio, C. Mestel, and D. Hindle. 1998. What can i say: Evaluating a spoken language interface to email. In *Proceedings of the Conference on Human Factors in Computing Systems, CHI98*.

N. Yankelovich. 1996. How do users know what to say? *ACM Interactions*, 3(6).