Lexical Entrainment of Real Users in the Let’s Go Spoken Dialog System

Gabriel Parent, Maxine Eskenazi

Language Technologies Institute, Carnegie Mellon University, Pittsburgh PA USA
gparent@cs.cmu.edu, max@cs.cmu.edu

Abstract

This paper examines the lexical entrainment of real users in the Let’s Go spoken dialog system. First it presents a study of the presence of entrainment in a year of human-transcribed dialogs, by using a linear regression model, and concludes that users adapt their vocabulary to the system’s. This is followed by a study of the effect of changing the system vocabulary on the distribution of words used by the callers. The latter analysis provides strong evidence for the presence of lexical entrainment between users and spoken dialog systems.

Index Terms: spoken dialog systems, lexical entrainment

1. Introduction

The success of a conversation between a user and a spoken dialog system (SDS) is influenced by many of the SDS designer’s choices. Some of those decisions concern the words that the system will use. Users may be confused by ambiguous terms in an SDS, while they also may not know precise but rare terms. Vocabularies constantly change in human dialog. Humans dynamically adapt their choice of words to another to facilitate the flow of the conversation. This process is commonly referred to as lexical entrainment [1]. Wizard of Oz (WOZ) studies [2] have shown that users also tend to adapt to the “system” vocabulary. SDS designers have long known that the users who are successful at using an SDS are those who adapt to the system. Yet, to our knowledge, there is very little literature detailing manners and rates of the user adaptation process [3].

In this paper, we present the results from a study of lexical entrainment of real users of Let’s Go, a bus information SDS [4]. More specifically, we answer two questions: do real users entrain to the SDS’ vocabulary and how do users adapt when new system-spoken primes are introduced. These primes cover different concepts, and include nouns, verbs, adjectives and adverbs. Measures of entrainment (both presence and strength) are calculated for a total of 18981 transcribed dialogs (197,801 turns), spanning a period of over one year of use of the Let’s Go SDS. We employ measures used in the literature, and introduce a new analysis of lexical entrainment based on word frequency before and after prime modification. The large quantity of human-transcribed data allows us to conclude that what others have observed in a WOZ set-up and in offline data studies holds, with strong statistical significance, for real SDS users; the callers do entrain to system primes. Finally, while one short dialog might not be enough to get a seasoned user to say the new system prime, observation over a period of three weeks shows positive evidence of longer-term adaptation.

2. Background

Human-human dialogue research has shown that in a conversation each speaker implicitly gravitates towards the other’s manner of speaking [1] [2]. Levelt [5] proposed that the process was lexically driven, since a lexical item is chosen and then encoded at many levels so that both meaning and form substitutions are possible. Thus the adaptation is both an effort at aligning language and at aligning meaning, aiming at a productive conversation.

A novice SDS user might be more likely to adapt to the system than a frequent user, who should already have adapted. Levow [3] compared novice system users to “experts” to see how quickly the former adapted their speech. Her longitudinal analysis showed that novices can adapt word choice fairly quickly (also reducing their vocabulary). This lowered the out of vocabulary (OOV) rate, but did not raise their speech recognition scores. Again on the lexical level, in WOZ simulations, Gustafson and colleagues [6] detected lexical entrainment and determined that users adapted to a synonym of a word that they had already used if it was spoken by the WOZ, thus proving that a prime could be modified by the system.

This can be useful if the user proposes a prime that is difficult for an automatic system to understand. The system could deal with the difficult word by guiding the user to a longer, more phonetically unique lexical entry. Another good reason to use entrainment strategies when designing an SDS is Nenkova and colleagues’ [7] finding that successful entrainment of high frequency words in human-human dialogues has a high correlation with task success. They observed that where there is more entrainment, there are more overlaps and fewer interruptions, thus producing better dialogue flow and fewer chances for the system to make errors. Matessa [8] shows similar results, where the average message length in a dialog between a lexically “accommodating” agent (the agent picks the same words as the user) and the user is shorter than it is with a “non-accommodating” agent (the agent purposely picks different words from the user).

However, findings from a WOZ setup may not necessarily be replicated with the real users of an SDS. Ai et al [9] note that real users behave differently from paid users. Thus WOZ findings should be verified on an SDS with real users. Stoyanchev and Stent [10] [11] used a subset of Let’s Go data (2184 dialogs) to study the level of entrainment of the user to syntactic primes used by the system. Results were obtained using raw automatic speech recognition (ASR) output. These studies included different lexical entries (2 verbs and 2 prepositions), and give early confirmation that entrainment may be present in real users’ interaction with an SDS. While the present paper first confirms this observation, it goes actively beyond by directly manipulating the system primes. It also exploits a significant amount of data, thus affording observations of user adaptation over time.

3. The Let’s Go vocabulary

This study is based on dialogs between real users and the Let’s Go bus information system [4]. The system is used nightly by Port Authority of Allegheny County bus riders. They call to obtain both bus schedule information and to find out what bus runs in a given neighborhood. Table 1 shows a sample dialog.

In order to study a variety of aspects of users’ lexical entrainment, we first examined all of the existing system prompts and selected a set of key concepts that were likely to...
be expressed by the users and the systems in order for the
dialog to be successful for example “bus immediacy” (e.g.,
“now”, “the next bus”) (Table 2). For each of these concepts,
we selected a word in the existing vocabulary (V1) that was
likely to be entrained by the users. There were 10 words in all.
We built a second vocabulary (V2) which, for each of the 10
words we had selected, contained a corresponding word that
was deemed to contain the same semantic content as the V1
word. For example, “immediately” was chosen as the word
corresponding to “now”. While V1 contains 10 words, V2
contains 8 (2 words in the second vocabulary replaced 4 words
in the first vocabulary in an attempt to narrow V2). The words
cover multiple classes of part-of-speech. It should be noted
that we purposely selected right to be part of V2 even though
it was also in V1 (see Section 5). We verified that all of the
words were synthesized intelligibly, and recognized correctly.
Since all of the dialogs were ultimately hand transcribed, the
quality of the recognizer is not a factor of precision in our
measure of lexical entrainment.

4. Linear regression analysis of lexical
entrainment

4.1. Methodology

To obtain a baseline confirmation that real users of an SDS do
indeed adapt to the system primes, we used the measure pre-
sented in [12]. The idea is to build a linear model of the
prime/target distance and of the frequency of that pair. This
model is built using the data in two orders: normal order, and
random order (the user and system turns are shuffled
randomly). If the models for the two orders are the same, this
means that the order does not affect the relationship between
the two variables. However, our hypothesis is that a user will
reuse a word more often immediately after the system has said
it, and thus the model with the normal order should have a
different slope than the random model. This idea is also ap-
piled in [13] as a way of finding evidence of entrainment on
syntactic features.

We examined one year of Let’s Go real caller data (18,081
dialogs, referred to as D1 hereafter) containing the V1 primes.
These dialogs were transcribed by a crowd of workers through
the Amazon Mechanical Turk (AMT) platform [14].

4.2. Results

To obtain the data for our linear model, we gather data points
by obtaining, for every V1 prime that the system used, the
uptake of that prime by the user within a window of N turns.
Table 3 presents an example of two data points obtained in
this manner way. The first one ([1,1]) is obtained because the
user said the word itinerary one turn after the system did and
the second one ([2,1]) is obtained because the word help
is used by the user two turns after the system used it. If the user
entrained to the system vocabulary, there should be more data
points with a small distance, which the linear model should be
able to capture.

Table 3. Example of data points for the linear regression

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Transcripts</th>
<th>Data points</th>
</tr>
</thead>
<tbody>
<tr>
<td>System:</td>
<td>I am an automated agent that can give you itinerary information about buses. You can ask for help at any time. What can I do for you?</td>
<td>[1,1]</td>
</tr>
<tr>
<td>User:</td>
<td>I’d like the itinerary for the 61C.</td>
<td>[1,1]</td>
</tr>
<tr>
<td>System:</td>
<td>Where are you leaving from?</td>
<td>[2,1]</td>
</tr>
<tr>
<td>User:</td>
<td>Help</td>
<td>[2,1]</td>
</tr>
</tbody>
</table>

Table 4. Slopes and Rs of the linear regression for
different window size. Bold values are statistically
significant (p<0.001)

<table>
<thead>
<tr>
<th>Data</th>
<th>N=5</th>
<th>N=10</th>
<th>N=15</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Slope</td>
<td>R</td>
<td>Slope</td>
</tr>
<tr>
<td>Normal order</td>
<td>-0.73</td>
<td>-0.18</td>
<td>-0.32</td>
</tr>
<tr>
<td>Random order (Baseline)</td>
<td>-0.08</td>
<td>-0.02</td>
<td>-0.01</td>
</tr>
</tbody>
</table>

Table 2. Vocabulary used by Let’s Go for this experiment

<table>
<thead>
<tr>
<th>Category of primes</th>
<th>Words in V1</th>
<th>Words in V2</th>
<th>POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bus immediacy</td>
<td>next</td>
<td>following</td>
<td>ADJ</td>
</tr>
<tr>
<td></td>
<td>previous</td>
<td>preceding</td>
<td>ADJ</td>
</tr>
<tr>
<td></td>
<td>now</td>
<td>immediately</td>
<td>ADV</td>
</tr>
<tr>
<td>Action</td>
<td>leaving</td>
<td>departing</td>
<td>VB</td>
</tr>
<tr>
<td>Domain specific</td>
<td>route</td>
<td>itinerary</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>schedule</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agreement</td>
<td>okay</td>
<td>correct</td>
<td>ADJ</td>
</tr>
<tr>
<td></td>
<td>right</td>
<td></td>
<td></td>
</tr>
<tr>
<td>System interaction</td>
<td>help</td>
<td>assistance</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>query</td>
<td>request</td>
<td>N</td>
</tr>
</tbody>
</table>
rated by a small distance is higher that of pairs at a larger distance.

These results are comparable to [12]; the distance between a prime/target pair has an effect on the frequency of such a pair. However, this approach raises several questions. One is that our measure does not take into account the fact that some words have more synonyms (and thus more chance of fortuitous system/user alignment). Also, as Ward and Litman [12] mention, the size of an optimal window to measure entrainment, and the underlying reason for that window size are still open questions.

More generally, a major issue is that this analysis does not take into account the prior distribution of the words. Section 5 presents another kind of analysis that examines two word distributions: before and after a change in primes. If the system’s vocabulary has an influence on the user’s choice of words, the word distributions should be different.

5. How do users react to a change in primes?

5.1. Methodology

Without informing our callers, we changed the Let’s Go vocabulary from V1 to V2. We took out all 10 V1 words from all system prompts (except for right, as explained in 5.3), so that the results would only reflect the effect of the intended V2 primes. We ran the new set of prompts for 3 weeks (900 dialogs, referred as D2 hereafter), which were all transcribed by an expert transcriber. The analysis is based on the frequency of words in the two vocabularies. If the user does not entrain to the system vocabulary, the distribution of V1 words in the user’s speech should be the same before and after the change. On the other hand, if there is lexical entrainment, we should see more V2 words after the change, since they had been spoken by the system.

5.2. Results

Table 5 shows the frequencies of words in both vocabularies in D1 (one year using V1 primes) and D2 (3 weeks using V2 primes). The results for the dataset D2 are also shown in Figure 1 to allow an easier comparison: the proportion occupied by each word on one line corresponds to its relative frequency in D2. In order to give an order of magnitude of how many times each word was used in D2, we also included the count of the V1 words next to their corresponding label. In order to have an idea of the evolution of user entrainment over time, we calculated for each day the proportion of words used that were from V2 compared to V1. For example, a proportion of 0.4 indicates that out of 10 words used by the caller, 4 were from V2.

5.3. Discussion

From Table 5, we can see that the words we selected to be part of V2 almost never occurred in D1. One explanation is that the seasoned callers are well-aligned with the SDS, which is also corroborated in Section 4 (We should note that we do not have callercaller identification and thus cannot verify who is a repeat caller). The homogeneity of the words used in D1 (almost entirely V1 words) gives further strength to our results. The shift of word frequency toward words in V2, shown in Figure 1, can be tested using a paired t-test. We tested if the frequency shifts from D1 to D2 for words in V1 and V2 were from the same distribution. We obtained a p-value of 0.0011, thus allowing us to reject the null-hypothesis that words in V1 and V2 behave similarly after the prompts change. The Table 5 also makes the difference clear. For example, the word preceding was used 0 times over a period of 1 year prior to our changes and 96 times in the 3 weeks that followed the insertion of that word in the various prompts of the system.

<table>
<thead>
<tr>
<th>Words</th>
<th>D1 Freq. (% rel. Freq)</th>
<th>D2 freq (% rel. Freq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1: next</td>
<td>13204 (99.9%)</td>
<td>492 (82.9%)</td>
</tr>
<tr>
<td>V2: following</td>
<td>3 (0.1%)</td>
<td>101 (17.1%)</td>
</tr>
<tr>
<td>V1: previous</td>
<td>3066 (100%)</td>
<td>78 (44.8%)</td>
</tr>
<tr>
<td>V2: preceding</td>
<td>0 (0%)</td>
<td>96 (55.2%)</td>
</tr>
<tr>
<td>V1: now</td>
<td>6241 (99.8%)</td>
<td>237 (80.1%)</td>
</tr>
<tr>
<td>V2: immediately</td>
<td>10 (0.2%)</td>
<td>59 (19.9%)</td>
</tr>
<tr>
<td>V1: leaving</td>
<td>4843 (98.4%)</td>
<td>165 (70.8%)</td>
</tr>
<tr>
<td>V2: departing</td>
<td>81 (1.6%)</td>
<td>68 (29.2%)</td>
</tr>
<tr>
<td>V1: route/schedule</td>
<td>2189 (99.9%)</td>
<td>174 (94.5%)</td>
</tr>
<tr>
<td>V2: itinerary</td>
<td>2 (0.1%)</td>
<td>10 (5.5%)</td>
</tr>
<tr>
<td>V1: okay/correct</td>
<td>1371 (49.3%)</td>
<td>48 (27.7%)</td>
</tr>
<tr>
<td>V2: right</td>
<td>1409 (50.7%)</td>
<td>125 (72.3%)</td>
</tr>
<tr>
<td>V1: help</td>
<td>2189 (99.9%)</td>
<td>17 (65.3%)</td>
</tr>
<tr>
<td>V2: assistance</td>
<td>1 (0.1%)</td>
<td>9 (34.7%)</td>
</tr>
<tr>
<td>V1: query</td>
<td>6256 (99.9%)</td>
<td>70 (20.4%)</td>
</tr>
<tr>
<td>V2: request</td>
<td>3 (0.1%)</td>
<td>272 (79.6%)</td>
</tr>
</tbody>
</table>

However, it seems that not every word was entrained to the same extent. The prime, itinerary, was the one that callers used the least (from a relative frequency of 0.09% in D1 to 5.75% in D2). This can be explained by the fact that this word is generally less frequent, domain-specific and also ambiguous in the context of a bus information system (it can represent the time when a bus runs, as well as a bus route).

The word request is particularly interesting. Its frequency goes from 0.05% to 79.53%. One possible explanation is that its counterpart in V1, query, is less frequent, thus less known to callers and so probably less natural for them. It should be noted that this prime provides interaction with the system (it’s
mostly used in “start a new request” to restart the dialog). It may be the case that the users entrain more to this type of prime, since it affects the intrinsic behavior of the system, thinking that the system is probably the best reference on how to name a concept like “start a new request”.

The most entrained word (request) and the least entrained one (itinerary) are both nouns. Also, there does not seem to be a pattern in the strength of entrainment for the different parts-of-speech (POS), and thus we do not have enough evidence to conclude that POS could be used to predict entrainment. However, we do note that it seems that more frequent words are more easily entrained. As an example, itinerary, which is much less frequent in day-to-day speech in this domain than schedule, is not chosen by many users. The opposite happens for request and query, being the infrequent term in this case.

The evolution of lexical entrainment between users and the SDS is shown on Figure 2. Although there isn’t a clear upward pattern, there seems to be a positive tendency, especially in the third week. There are also 3 points where the curve goes down, which correspond to the 3 weekend periods during the study. This is likely to be an indication that some of our users only call on the weekend, and thus take more time to adapt to the system vocabulary, although again we cannot confirm this without caller-identification information.

6. Conclusion

This paper presents an experiment aimed at characterizing the lexical entrainment of real users calling the Let’s Go Spoken Dialog System. We first used a convergence measure to determine lexical entrainment of the system users. The results indicate that users tend to adapt to primes that are in system vocabulary, although again we cannot confirm this without caller-identification information.

Future work will examine the contexts in which the primes were used. For example, help and assistance were meant to be used in the same unique context of the user needing help. More complex manipulation could determine if entrainment also holds for prime/target pairs that are not meant to be used in the same context. An example would be if the system, addressed the user with a prompt such as “In the next minutes, I will be asking you some questions” where the word next is used by the system to refer to a time period. By excluding the primes next and following from the normal bus information prompts (“Do you want the next bus?”), only the first greeting would contain the prime. However, such an approach would need some type of compensation for the prior distribution of the two words. This experiment would also be harder to implement in a natural way in a bus information system since the domain is limited. We are beginning to develop an SDS with a wider domain of application that will afford this type of study. That platform will thus allow us to make finer adjustments, and investigate more complex aspects of lexical entrainment. It will also provide “caller id”, which will allow us to individualize the analysis, and eventually the manipulations.

Finally, an effective SDS should not only have the user entrain to its primes, but also detect user primes and modify its own prompts to use them. This is the next step in our effort to endow SDS with entrainment capabilities.

7. Acknowledgements

This work was funded by NSF grant IIS-0914927 “LexE: Using two-part lexical entrainment for more efficient and reliable spoken dialogue systems”. The opinions expressed in this paper do not necessarily reflect those of NSF. We would like to thank Adam Skory and Susi Burger for their help in transcription expertise.

8. References


Figure 2. Proportion of words of V2 in the target words

![Graph showing proportion of words of V2 in the target words over time.](image-url)