Language Pattern Analysis for Automotive Natural Language Speech Applications

Ute Winter
GM Advanced Technical Center
HaManofim Street 11
Herzeliya 46725, Israel
+972-9-9720622
ute.winter@gm.com

Tim J Grost
GM HMI User Experience
30100 Mound Rd
Warren, MI 48090
586-986-0075
timothy.grost@gm.com

Omer Tsimhoni
GM Advanced Technical Center
HaManofim Street 11
Herzeliya 46725, Israel
+972-9-9720618
omer.tsimhoni@gm.com

ABSTRACT

Natural language speech user interfaces offer a compelling choice of user interaction for the automotive market. With the increasing number of domains in which speech applications are applied, drivers must currently memorize many command words to control traditional speech interfaces. In contrast, natural language interfaces demand only a basic understanding of the system model instead of memorizing keywords and predefined patterns. To utilize natural language interfaces optimally, designers need to better comprehend how people utter their requests to express their intentions. In this study, we collected a corpus of utterances from users who interacted freely with an automotive natural language speech application. We analyzed the corpus by employing a corpus linguistic technique. As a result, natural language utterances can be classified into three components: information data, context relevant words, and non context relevant vocabulary. Applying this classification, users tended to repeat similar utterance patterns composed from a very limited set of different words. Most of the vocabulary in longer utterances was found to be non context restrictive providing no information. Moreover, users could be distinguished by their language patterns. Finally, this information can be used for the development of natural language speech applications. Some initial ideas are discussed in the paper.

Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: Natural language, Voice

General Terms

Human Factors, Performance, Languages

Keywords

Automotive Speech Interfaces, Natural Language Speech Applications, Language Patterns, Speech Corpora

1. INTRODUCTION

Speech applications have been deployed in vehicles for years [2]. The intent is generally to simplify the user interaction with infotainment systems and to provide an additional method of user interaction while driving. Increasingly complex vehicle systems put pressure on user interaction designers to leverage speech input in their designs. The inclusion of personal devices such as mobile phones, MP3 players, Personal Navigation Devices, and others, in the driving environment adds to the complexity and pressure to support speech controls for these devices [5].

The design, development and testing of traditional speech applications have focused on the improvement of recognition accuracy of the driver's utterances, with the expectation that the higher the recognition accuracy, the better the interface and the more satisfying the user experience. Dialog design often focuses on the selection of reasonably intuitive command words, which offer the speech recognition engine diverse phonetic content for improved disambiguation, thus better recognition accuracy. For example the driver has to say “play artist” and add the artist name to enable a music device. This concentrated focus on raw recognition performance has produced vehicle speech user interfaces which can boast incredibly low word error rates.

What is often not given enough consideration is the requirement that the user memorizes the command words needed to control the system. If the user fails to produce an utterance with the command structured exactly the way it is understood by the system, e.g. saying “listen to artist” and the artist name, the recognition attempt is likely to fail. It is considered to be one type of miscommunication and categorized as an error of the speaker [1]. The user may have known exactly what he wanted to do, he may have even recalled using the feature in the past, but simply not remembered exactly the word or combination of words necessary to execute the task. This is a common flaw in speech user interface design. Overlooked is the rate at which users produce perfectly ‘reasonable’ requests which are not covered in the system's recognition grammar, and thus rejected or substituted for another request by the system. Repeated incidents of this failure may lead many users to the conclusion that the interface is not usable.

By increasing the number of features, traditional command and control speech user interfaces require the user to remember an ever increasing list of commands. By adopting a more natural and flexible set of possibilities of requests, the system designer may increase the usability of the system. This reduces the responsibility of the user from having to memorize commands, to simply having a basic understanding of the system model and the features supported by speech.

Collecting and analyzing a corpus of naturally spoken utterances from users offering their most likely phrases for their requests,

AutomotiveUI'10, November 11-12, 2010, Pittsburgh, Pennsylvania
ACM 978-1-4503-0437-5

enables the designer to understand the distribution of likely utterances for each supported feature. The inclusion of an appropriate amount of this data and the modeling of the data's characteristics into the speech application may greatly improve the performance and usability of the system interface. We can conclude that it is vital to understand the data's characteristics. One aspect of user utterances, while interacting with automotive speech applications, is their typical vocabulary, language patterns and complexity. Only if we understand the user language, we can transform the discovered knowledge into features, which the application can utilize.

This paper studies this aspect and analyzes vocabulary and language patterns of a corpus of audio data collected from users, while enabling a natural language speech user interface in a vehicle environment. We introduce a classification of the vocabulary using context information as qualitative measurement for classification. This classification helps quantify the language complexity and variety of utterances in the corpus, between users and across domains. Finally, we conclude how our findings may impact the design of natural language speech user interfaces, and may help improve task success and dialog quality of the speech system, prerequisites for a natural and intuitive user experience in interaction with such interfaces.

2. SPEECH CORPUS CREATION

2.1 Participants

The study collected user data from 33 participants. Participant age was distributed across three categories, with the largest representative group (18 participants) between the ages of 31 and 45 years. The remaining participants were fairly evenly split between those aged 21-30 years (8) and 46-60 years (7). Within each of these three age groups there were equal number of men and women, except only three women in the older age group.

2.2 Application Environment

The speech recognition system used in this study was embedded in a 2009 Buick Enclave vehicle so the user could experience the use cases and associated interaction while actually driving. The user activated a dialog session by pressing a button which was located on the steering wheel. Voice prompts and audio feedback were provided by way of the vehicle's audio system. The speech system used the embedded microphone located in the vehicle's overhead console.

The speech recognition system allowed for the control of several features over four domains. The feature support was concentrated on the features which are generally accepted to be more time consuming, difficult or confusing to perform by way of a visual manual interface. Four functional domains were supported by this study.

Radio Tuning – Among the primary radio features supported in the study, the user was allowed to utter requests, which switched between broadcast radio bands, tuned the radio to specific AM and FM frequencies, and tuned the XM radio to specific XM channels by including the channel name or number in his request. Various lesser features such as seeking up and down for stations, XM category selection, were also available.

Music Selection – Included in the primary music selection features supported by the system were the ability to identify music selections by any combination of the following attributes of the music content: artist or group name, album name, and song title. Also supported was the ability for the user to ask for feedback to identify available music content, e.g. available artists/groups, available songs and albums by a given artist. Again, lesser features were supported, such as skipping to the next/previous tracks.

Phone Dialing – The study made use of communication features for phone calling, which included the ability to speak commands to dial phone numbers and to place calls to entries in the phonebook. Each contact stored in the phonebook had multiple numbers available for home, mobile, and work phone.

Navigation Destination Entry – The use cases included for destination entry allowed the user to either identify a place of interest (POI) category to display on the map, or to enter an address as a specific destination for route guidance.

In an effort to maximize task completion rate and minimize task completion time, nearly all tasks were supported as 'one-shot' commands. This meant that all the information needed for the system to complete the task was allowed to be delivered in a single user utterance. If the speech recognizer could not complete the task after the first utterance, it initiated a directed dialog to extract missing information and thus, achieve task success. The focus of this study is on the first utterance.

The command structure was left as flexible as possible, allowing the user to format requests at their discretion. For example, the commands "Brown-eyed Girl by Van Morrison", "Play Brown-eyed Girl by Van Morrison" and "I wanna listen to Van Morrison, Brown-eyed Girl" were considered to be equivalent.

2.3 Task Procedure

The data analyzed in this study were raw audio data recorded in the test vehicle. Each utterance was recorded as it was being recognized by the speech recognition engine. Recognition results, confidence scores, and semantic interpretation were logged.

Each subject sat in the driver's seat and received a few moments of introduction to the system. The administrator sat in the passenger seat and the introduction was delivered while the vehicle was running at idle in park. The administrator provided only a high level description of the speech recognition system which was developed to recognize requests spoken naturally by the user to control a variety of in-vehicle systems. The functional domains were briefly explained to give the user an idea of the scope of the speech user interface.

The individual features of each domain which would be offered as use cases for the session were mentioned but were not discussed in detail. For example, the administrator explained that there was an internal hard drive loaded with 1000 songs by 100 artists and that they would be asked to select music for playback from that device. The user was also told that there was a simulated phone book complete with contact entries which they would be accessing to establish phone connections. The user was given a list of the artists represented on the hard drive, as well as the contact names in the simulated phonebook. This was done to allow them an opportunity to familiarize themselves with the system to some degree along the lines of how it would be familiar to them if the music content and phonebook were theirs coming from their personal devices. It was also mentioned that the navigation database for this study was limited to the local geographical area, and that when the use cases for entering a destination address were performed, the destination address would have to come from that area.

The process of initiating the speech session was explained and then demonstrated. The administrator pointed out the Speech Recognition button on the steering wheel, and asked the user to press it. The administrator adjusted the vehicle audio system volume to a level to which the user was comfortable. The administrator made sure that the user heard the prompt and the audible ‘beep’ which signified the start of the listening window. The user was instructed that they would need to wait for the beep before they began delivering their request utterance.

The study administrator did not provide an example of a spoken request nor did he allow the user to deliver a test request to the system at this time. The test procedure was set up as described to give the user a basic understanding of the application and its possibilities, but not to influence the participant’s choice of language in provided utterances.

2.4 Corpus Description

The corpus includes only the first utterance of each dialog. The main reason is that in continuing dialogs the speech application decides to take the initiative and directs the driver through the process of task completion by allowing only a limited set of answers without the option to express the intention naturally.

The corpus contains a total of 2067 utterances, which were distributed over all four domains as shown in Table 1. Each participant made, on average, 63 utterances though not all speakers provided the same number of utterances

Table 1. Domain Distribution of Utterances

Music	Radio	Navigatio n	Phone	Total
599	484	664	320	2067

The corpus was analyzed to include a textual representation of all audio requests. To create these transcriptions, the raw recognition results were first used and then each utterance was manually transcribed to ensure the veracity of the data. The transcription exactly matched the spoken utterances, including filled pauses such as *uh*, *ah*, *er*, *um*. Because a speech application would normally treat these interjections as words which have to be distinguished from other words, they were treated alike in the transcription.

3. ANALYSIS METHOD

3.1 Empirical Method

Our approach aims to discover knowledge in the corpus regarding the use of language and its complexity, attempting to trace a path from collected data to knowledge about language patterns and vocabulary properties, which can be used to improve performance of automotive speech user interfaces [4].

Our guideline is the method of “3A” perspective (Annotation, Abstraction and Analysis) in corpus linguistics, first introduced by Wallis and Nelson [8]. A key point is the abstraction, which maps the language present in the corpus to an abstract model serving the research goals. Without the step of abstraction, knowledge discovery would not be effectively dedicated to a research topic. The annotation metadata has to be defined in line with the abstract model. Finally the analysis of the corpus evolves from exploring the annotated corpus and discovering features or recurring patterns in line with the abstract model.

The abstraction is conducted by modeling the recognition task, which a natural language speech application undertakes, with respect to vocabulary use and its meaning for the recognition result. In a second step, the corpus is annotated regarding properties which the model has identified to be essential for the interpretation of language during the recognition task. Because the subsequent analysis evolves from the results of the first two steps (abstraction and annotation), in this paper, we describe the analysis methods in parallel to the presentation of the results in section 4 and 5.

For the abstraction we need to define language complexity. With Rescher [7] we understand it “as a characteristic feature of the real,” which consists of compositional, structural and functional elements. When referring to language and our research goal, it contains investigating characteristics such as the number of constituent elements in an utterance, the variety of those constituent elements, and the different possible ways of arranging the constituents with consideration of their interrelationships, among others.

3.2 Recognition Task, Action and Data

The expected first utterance for the speech application conveys a user’s request for a system action. The user may formulate this primary act in a secondary form, such as a question (“Can you tune the radio to XM 8?”) or as a statement (“I am hungry,” when looking for restaurants). In any case the user expects the system to perform the requested action and, in some cases, to provide spoken feedback on its understanding.

Consequently, the speech application needs to extract the user’s intention from his utterance by performing two sub tasks. It has to recognize the provided informational data and to interpret the type of action. Table 2 shows some examples of user requests.

Table 2. Examples of User Requests

Utterance	Requested Action	Provided Data
“Find Shopping Mall”	Navigate to Place of Interest	“Shopping Mall”
“Play Doobie Brothers”	Play Music, Artist	“Doobie Brothers”
“List Songs by James Taylor”	List Music, Songs	“James Taylor”

Automotive natural language speech applications typically handle the task in two steps [3]. An Automatic Speech Recognizer (ASR) takes the acoustic signal and transforms it into a raw textual utterance. This engine is either based on Statistical Language Models with slots for data or on sophisticated Finite State Grammars, modeling a variety of typical natural utterances. A semantic interpreter (often referred to as Natural Language Understanding module or NLU) post-processes the text to extract the user’s intention by understanding the provided data and requested action. A variety of technologies are possibly used, e.g. name entity extraction, chart parsing, or Statistical Language Models, all in combination with semantic knowledge bases.

The speech application can perform the recognition and interpretation task if the user provides two pieces of information. First, the user needs to explicitly mention the data. For each request there is a sufficient number and type of data for unambiguous recognition, though the user can provide more data than needed. An example is “Kryptonite by Three Doors Down”, where the artist is mentioned in addition to the song title, although in most cases the song title is sufficient for system recognition.

Second, the user should provide explicit or implicit information about the action, if it deviates from the speech application’s default action. If a user requests an artist name, e.g. “Van Morrison”, then the speech application decides to play songs of the artist, assuming there is no ambiguous possibility to recognize this name. If the intention is different, e.g. to list all songs of this artist in a prompt or on the display, then the utterance has to indicate this as seen in the above example “List songs by James Taylor”.

3.3 Ambiguity and Context

If the user makes wrong assumptions about the default action or does not give sufficient explicit or implicit information about type of requested action, the utterance may be ambiguous. In this case the speech application needs to continue the dialog and request more information of the described type. Examples are the utterances “Chicago” or “Find Chicago”, which can refer to an artist, a street name or an XM radio station. Instead, when uttering “Play Chicago”, the context is narrowed down to an artist or radio station. Utterances such as “Let me hear the artist Chicago” or “Tune the radio to Chicago” are considered unambiguous, because the context points to one possible action.

The examples above show that ambiguity is a major issue for which speech applications have to provide solutions. They also show that incorporating information into the recognition process about the context may help resolve ambiguities. Traditional command and control speech interfaces address this problem by demanding predefined commands from the user, which restrict the domains or indicate the action.

When users are allowed to speak their intention naturally they may provide context-restrictive information in their utterances. Although this may happen with new users unconsciously, over time all users are likely to learn that well-chosen additional information increases their success in interacting with the speech application.

Consequently, it is important for speech application designers to learn from corpora of natural language utterances about typical use cases in a vehicle environment. Such corpora can be used to learn how users choose their vocabulary when speaking to a natural language speech system in order to provide what they perceive as needed additional information and in order to restrict context.

3.4 Data Annotation

The annotation was done on a textual representation of all of the spoken first utterances. We can derive from the preceding discussion, what metadata is necessary for the annotation of the corpus:

1. Distinguish between the user’s provided data and the additional vocabulary in the natural utterance. This facilitates the possibility to analyze and classify the remaining words other than data regarding quantity and context restrictive properties. Moreover, it enables us to view all utterances as a set of patterns, where the data is exchangeable.
2. Provide information about the domain and sub domain of the user’s request. Not only do we want to analyze the corpus across users and domains, the annotation allows us to find levels of context restriction.
3. Augment each utterance in the corpus with a speaker ID and audio wave ID for the desired quantification of language complexity among users.

Table 3 shows an example list of utterances and their annotation.

Table 3. Examples of Annotation

Utterance	Corpus Annotation	Domain / Sub Domain
“I need directions”	I need directions	Navigation / -
“Find gas station”	Find <place_of_interest>	Navigation / POI
“I need directions to 608 Whitcomb”	I need directions to <house_no> <street>	Navigation / Address
“Tune to XM 20”	Tune to XM <radio_station>	Radio / XM
“Play Best of You by Foo Fighters”	Play <song> by <artist>	Music / Song
“Please list Rascal Flatts Songs”	Please list <artist> songs	Music / Artist
“Call Bill Smith”	Call <name>	Phone / Name

4. CORPUS LANGUAGE ANALYSIS

4.1 Vocabulary Analysis

A vocabulary analysis of all utterances by breaking them into single words and excluding the data pieces revealed that the corpus contained 223 different words. This number includes separately filled pauses and plural forms, which also occur in singular. Without plural variations and filled pauses the total number of words would be less than 200.

How can such a small number of words build the entire corpus? Many of the utterances can be clustered because they are very similar and can be distinguished by one or two words and the word order. Table 4 shows an example of utterances with the word “dial” in the phone calling domain. The entire corpus contains 53 utterances with the listed 8 patterns, while only seven different words combine them.

Table 4. Examples of Similar Utterance Patterns

Utterance Pattern	Words
dial <phone_no; name>	dial
dial <name> number	dial, number
dial <name> number from my phone	dial, number, phone, from, my
dial a phone number	dial, number, phone, a
dial phone number	dial, number, phone,
dial phone number <phone_no>	dial, number, phone,
dial the number <phone_no>	dial, number, the
dial the phone number	dial, number, phone, the

This implies that a lot of words are frequently recurring in the utterances. 94 words out of 223 have at least 5 occurrences in the corpus. Table 5 lists the twenty most frequent words, including the number of occurrences in the corpus and the number of domains, in which they occurred. Note that the number of occurrences in the Table is not weighted by the unequal number of utterances collected for each domain.

Table 5. Frequent Vocabulary

Word	# of Domains	Occur-ences	Word	# of Domains	Occur-ences
to	4	313	by	1	121
play	2	267	a	4	100
find	3	220	wanna	4	95
I	4	199	tune	2	78
XM	1	181	need	2	77
me	4	139	AM	1	75
the	4	137	nearest	1	71
directions	1	136	dial	1	71
hear	2	128	FM	1	66
call	1	124	get	4	60

4.2 Vocabulary Classification

As pointed out earlier, we aim for a classification of the vocabulary that matches the level and quality of information usable for the recognition task. This can be viewed as a determination of level and quality of context restrictiveness on the overall possibilities for recognition and interpretation. In a task where the space of possibilities can be very large and ambiguous because of an increasing number of domains and data, such as all street names and places of interest in a country, narrowing down the context can significantly restrict the search space.

Consequently, and also observable in Table 5, the extent to which a word reduces the recognition task to certain domains or sub domains, can be used to classify the vocabulary. For this corpus with requests in four domains we propose to categorize all words with at least five occurrences in the corpus into

- **non domain restrictive -** words which are used in three or four domains and where we can’t claim with confidence, that the meaning of the word restricts the context
- **2-domain restrictive** - words which occur in only two domains
- **domain restrictive** - words which are restricted to one domain
- **sub domain restrictive** - words which significantly reduce the domain space to one or more sub domains.

The classification is presented in the order of the level of restriction from wide to narrow context.

The 94 words, which occur at least five times in the corpus, can be classified as follows: 40 words are non domain restrictive, and 54 words restrict the recognition task to domains or sub domains. 14 words are 2-domain restrictive, mostly because they point to one of the music enabling tasks in a radio or MP3 player, which are very similar. Words such as “play”, “tune”, “hear”, “listen”, “next” belong to this category. Twenty eight words, such as “directions”, “call”, “dial”, “song”, “navigate”, and “station,” are domain restrictive, and expectedly so. There are, however, some frequently used words, which are restrictive but are less expected. One example is the word “by”, which is only used for the patterns “<song> by <artist>” and “<music/songs/albums> by <artist>”. Finally there are 12 words, which are only used in the context of sub domains, e.g. “track”, “artist”, “street”, “XM”.

In non domain restrictive vocabulary there are words that are distributed similarly over all domains, such as “me”, “a”, “wanna”, “to”, and “please”. But some words show a preference for one domain, although they are used in all four domains, e.g. “find” (mostly navigation), or “list” (mostly music selection).

A second kind of context restrictive class, which evolves from the corpus, is **action restrictiveness**, where action is used as defined in 3.2. Words such as “call”, “play”, “list”, “tune”, or “previous” are used only for a specific request for action. The level of domain restrictiveness for these words can vary, e.g. “call” is domain restrictive for phone dialing, whereas “list” is used across all domains, but always requests the action of giving the user a list of their address book, available artists, or places of interests such as restaurants. This means that both classes of context restrictive words are not mutually dependent.

We can expect that an extension of an in-vehicle speech user interface to more domains will lead to more ambiguity and increase the vocabulary to new domain and action restrictive words. Moreover, part of the words, which can be classified as context restrictive for the examined four domains, may become less restrictive, because they will be used for new domains, as well. We can conclude that although the classification is derived from the recognition task and independent from the corpus, the list of words for each class can vary depending on the configuration of domains and possible actions of the application.

4.3 Patterns and Context in Utterances

The corpus contains 693 different utterance patterns Table 6 shows the ten most frequent patterns (on the left) with their number of occurrences in the corpus and randomly chosen infrequent patterns (on the right), which occur between 2 to 5 times in the corpus.

Table 6 clearly shows that the most frequent patterns in the corpus are short and are focused on the data. Sparsely added words are mostly context restrictive. It can also be observed that infrequent patterns are substantially longer than the frequent patterns.

Table 6. Examples of Utterance Patterns

Frequent Patterns (Number of Occurrences in Corpus)	Infrequent Patterns (Randomly Chosen, 2-5 Occurrences in Corpus)
<data> (166)	…
call <data> (164)	I wanna call <data>
play <data> (98)	play a song by <data> >
XM <data> (63)	let me hear XM <data>
find <data> (56)	give me directions to <data>
dial <data> (45)	I wanna dial a number
<data> <data> (35)	I wanna hear <data> <data>
get directions (26)	how do I get to the <data>
AM <data> (23)	please tune to AM <data>
FM <data> (22)	change the radio to <data>

What is the distribution of context restrictive vocabulary in all utterances? First, we study its quantitative distribution. The corpus includes only two patterns that contain data without additional words. All other patterns have between one to six words complementing the data, very rarely there are more than 6 words. About 90% of the utterance patterns include at least one context restrictive word. If the data is complemented by one single word, it is already context restrictive in 80% of the utterances. The number increases to 95% for all utterance patterns with 5 additional words. Hence, recognizing context restrictive words with high confidence can be used to restrict the search space for the data and to interpret the desired action.

Calculating the average number of context restrictive words per utterance pattern shows that it is not significantly increasing compared to the length of the utterance. While an utterance pattern with two additional words include 1.2 context restrictive words, an utterance pattern with 6 words includes 1.9 context restrictive words. The gain of information about the user's intention does not grow when utterances become longer. What is mainly increasing is the general vocabulary, which cannot be used to indicate the intention of the user in any way.

Table 7 shows a statistical summary of the observations:

Table 7. Examples of Utterance Patterns

# of Words in Utterances	# of Utterance Patterns	% of Utterances with Context Vocabulary	Average # of Context Words
0 words	2	0 %	0 words
1 word	76	80 %	0.8 words
2 words	164	89 %	1.2 words
3 words	190	92 %	1.5 words
4 words	127	92 %	1.7 words
5 words	82	95%	1.8 words
6 words	37	92 %	1.9 words

Traditional speech user interfaces require command words to appear in predefined patterns at the beginning of an utterance before the requested data. Therefore, an interesting question for a corpus of naturally built utterances is: how often do users intuitively use context restrictive words as a first word in an utterance? In 45% of all patterns, which composed 53% of all utterances, context restrictive words appeared as the first word in the utterance. Therefore, users would have naturally started a request according to the required patterns of traditional speech interfaces only about half of the time.

Another noteworthy point is the tendency of users to position context vocabulary before the data. In this corpus, users added context vocabulary after the data or after the first data piece only in 4% of all utterance patterns, which constitute only 3% of all utterances in the corpus.

Finally, it is interesting to examine the position of context restrictive words relative to the position of data in the utterance. For this investigation we examined only utterances which contained both data and context restrictive words. About 35% of all utterances were excluded because they contained either only data or only context vocabulary, e.g. saying only data when requesting directions or when making a phone call. For the remaining utterances we measured the distance between context vocabulary and data as the smallest number of words between all data and context word locations. For instance the pattern "I wanna listen to <song> by <artist>" was tagged as "I wanna <context> to <data> <context> <data>". In this case the distance was zero, because the closest <context>/<data> pair was not divided by a word. Table 8 summarizes the distance between context and data.

Table 8. Distance between Context and Data

Distance (# of Words)	% of Utterance Patterns	% of Utterances
0	73%	70%
1	89%	95%
2	97%	99%
3 and 4	100%	100%

Clearly, context vocabulary tends to be positioned close to the data and not necessarily in the beginning of an utterance.

5. USER ANALYSIS

5.1 Differences among Users

Previous results on the overall quality of the corpus vocabulary and the development of its classification regarding context restrictiveness enable us to find differences between users and their preferred vocabulary and patterns. Table 9 shows ten randomly chosen music selection request patterns from two distinct users in the order they were requested:

Table 9. Examples of Utterance Patterns

User 1	User 2
Play <artist>	Play <artist>
Play <artist>	Let's hear <song>
Play <artist> <song>	Listen to <song>
Play <artist> <song>	Let's hear some <artist>
Play <song> by <artist>	Do we have any <artist> songs
Play <song> by <artist>	Do we have any <artist> music
Play <song> by <artist>	Want to listen to <artist>
Play <artist>	I wanna hear some <artist>
List songs by <artist>	MP3 player play

There appears to be a difference between these two users in the choice and quantity of words, which constitutes the utterance patterns, and which influences the variety of their utterances. Less easy to estimate at first glance is the amount of context information, that each user provides additional to the data.

5.2 Utterance Complexity among Users

The complexity of user utterances can be described by the average **vocabulary quantity** that a user needs to form his utterances. As an utterance is constituted by data and additional words, which are combined to the overall utterance length, these are the three values we take into account to understand differences among users.

Figure 1 shows all three average values per user ordered by utterance length. The mean utterance length (averaged across all the utterances of each user) varies across users from less than two words to more than four words (including data). Users can therefore clearly be differentiated by their vocabulary quantity. Overall, the increase in length is a result of an increase in additional vocabulary rather than an increase in data. This is because the natural choice of a user for data is limited to 0 to 2 data pieces. Still, the user's choice in providing non mandatory data, e.g. the artist name in addition to a song title, does not show a tendency to increase with the utterance length. The increase of utterance length is primarily caused by adding non-data words.

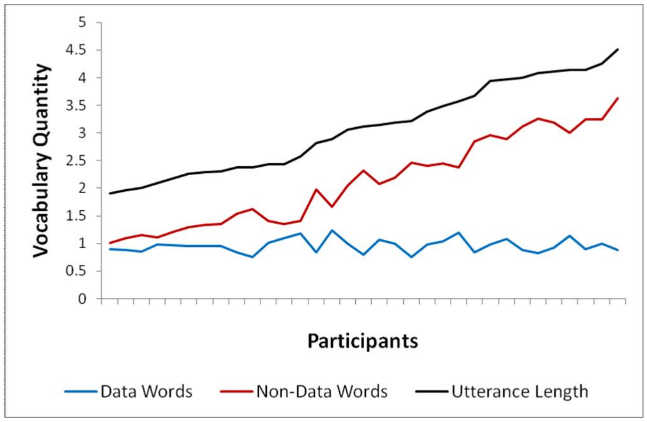

Figure 1. Vocabulary Quantity per User

Another interesting factor in utterance complexity is **language variety**, which measures the number of different words in utterances and the possible ways to arrange words into utterances. For this purpose we measure variety by the number of utterance patterns relative to the total number of utterances per user. Additionally we measure the number of different words relative to the number of all words in the user's utterances. Both variety values are measured as the percentage of different words and patterns relative to the total number of words and patterns per user.

It is evident that both measurements are necessary, when studying utterances from different users. A user can decide to choose different words for each utterance pattern as was shown for user 2. According to table 4, even a small number of words can be arranged into a variety of utterance patterns, and there are users who do so. Figure 2 shows the distribution of language variety among users.

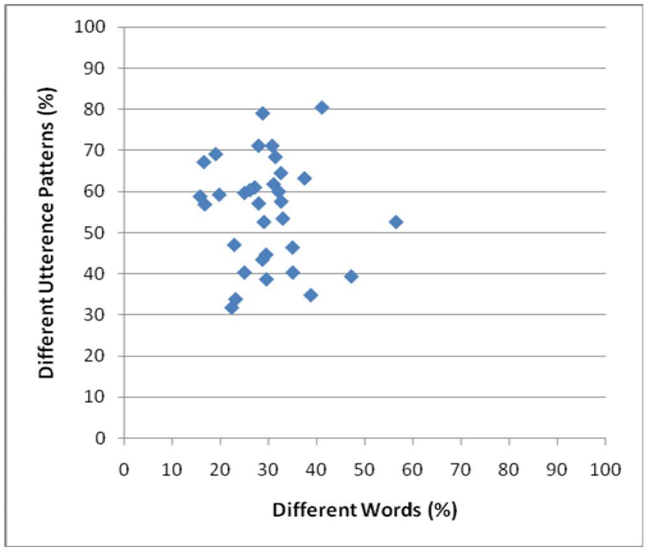

Figure 2. Language Variety per User

In general, language variety differs among users, but only to a certain extent and without the clear tendency that was found for vocabulary quantity. Users mostly choose a limited word variety (15-35% of all words) and a slightly higher utterance pattern variety (40-70% of all utterances), where part of the patterns can be very similar. It is therefore evident that users make use of only part of the range of variety provided by speech interfaces and language in general. Users express their intentions only via a limited subset of preferred possibilities.

When comparing user behavior across application domains it can be observed that utterances for phone call requests were different from requests of all other tested domains. The average utterance length for phone call requests was more than one word shorter than the other domains across all users. The reduction was observed in general vocabulary rather than in context informative vocabulary along the lines of the overall behavior of the corpus for vocabulary quantity, see Table 7 above.

A possible explanation is that the phone dialing domain offers a smaller variety of options for requests than do other domains. Alternatively, it is possible that users adapted their utterances for phone call requests because they were familiar with command and control phone speech interfaces.

5.3 Context Relevance of Users

The majority of utterances contained context restrictive words (as shown in 4.3). Table 7 shows that context information was not significantly higher in long utterances than in short utterances. Figure 3 now explores **context relevance** per user by measuring the portions of data, context restrictive words, and non context restrictive words per user. Users who tended to provide longer utterances did not provide substantially more information in those utterances. The amount of general vocabulary, which does not include information about the user's intention, varied from 0.25 to 2.25 words and explained most of the increase in utterance length (r^2=0.94). The change in amount of data did not explain any of the change in utterance length (r^2=0.01). The amount of context restrictive vocabulary increased moderately, explaining only a small part of the increase in utterance length (r^2=0.56). If one viewed context relevance as the relation between the number of context restrictive words and number of general words, context relevance decreased as length of utterances grew.

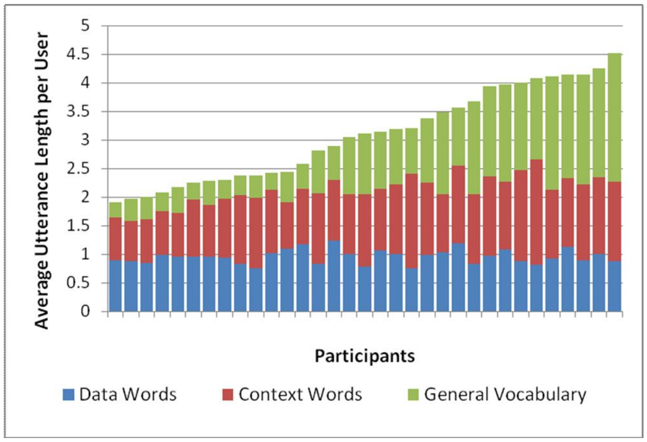

Figure 3. Language Variety per User

6. CONSEQUENCES FOR SPEECH INTERFACES

The analysis points to several significant characteristics of speech utterances that can be used to improve accuracy and overall dialog quality of speech applications.

Users tend to repeat similar utterance patterns composed from a very limited set of different words. Statistical Language Models of the ASR engine should reflect this behavior. Optimally, the training data for Statistical Language Models should be natural language audio data in authentic use conditions. But the models are still often trained based on textual data collections, which may lead to less reliability regarding authenticity. Our results can help to explore, in how far the utterances in textual data collections reflect the actual language usage discovered in this study. This knowledge may lead to an improvement of training material and Statistical Language Models reproducing the variety of user utterances.

Except for recognition of data, the engine should be designed to pay special attention to high accuracy for context restrictive vocabulary, which is found in 90% of all utterances. If it is action restrictive, its recognition is even necessary to complete a task. If the action vocabulary is recognized but the data is not recognized, the system can continue dialog and ask specifically for missing data and efficiently complete the task. Therefore knowing the requested action may also improve dialog quality. If the vocabulary is domain restrictive, the engine can restrict the search space for the data. This may be even possible during the acoustic recognition process, as the corpus shows that users tend to provide context restrictive vocabulary before data in all tested tasks. When post-processing the raw textual utterance, the NLU engine can make use of these words, only if the context restrictive vocabulary is recognized. Here it helps to resolve ambiguities in data and to understand the requested action.

The study has only analyzed context relevance for words which are exclusively used in certain domains. But we have found that words are used in multiple domains with clear preference for a certain domain. In statistically based machines, such as ASR and NLU engines, these words can be still considered as context restrictive and used for improvement in the above described way.

It is evident that most of the vocabulary is likely to be non-context restrictive in longer utterances and does not provide information. This is especially true for vocabulary which is unknown to the recognition engine, because it is not part of the Statistical Language Models or Finite State Grammar. Therefore an ASR engine may be trained to invest less into the recognition of general vocabulary and may be able to overlook unknown vocabulary. What has not been investigated in this study though is the possibility that non context restrictive words can be combined to context restrictive phrases. This corpus does not provide enough material for an investigation, as only 10% of the utterances do not contain context restrictive words and could be used for such an analysis. It is clear though, that the list of context restrictive vocabulary needs to be enhanced by such phrases, if we want to make use of context relevance.

The study also proves that users differ in language patterns and complexity. In the automotive environment, where the number of users in a vehicle is limited, we can use the knowledge to improve the interaction between machine and driver by adapting the system to the user's preferences. The speech application can learn from the user's history of interaction with the speech interface not only about preferred data requests, but also about typical utterance patterns and length, vocabulary variety, and choice and distribution of context relevant words. The system, which is originally built for a large variety of users, can reduce its language complexity to the characteristics of a small number of specific users and increase accuracy for this target group.

There is evidence that users are adapting themselves to their dialog partners, even if it is a machine [6]. A speech interface may be designed to make use of this regarding one of the main observed differences among users, the vocabulary quantity, respectively the amount of non informative vocabulary. A carefully designed system can try to shape user utterances, e.g. via TTS prompts [9], to desired patterns with focus on data and context restrictive vocabulary and reduced general vocabulary. As example the user interface could emphasize or reinforce the inclusion of explicit action restrictive words, if users tend toward implicit action requests. In such way, the user will develop a limited pattern set over time, which will be easier to recognize.

Our next steps in research will be to take our findings and examine our future in-vehicle speech applications. We may for instance analyze and improve the Statistical Language Model training data in the same way as it is recommended here and prove that it will lead to an overall performance improvement for the system. In summary, we will focus on training data and configuration of authentic language patterns for all parts of the application, the dialog management, ASR, TTS and NLU engine.

7. REFERENCES

[1] Danieli, M. 2004. Designing error recovery dialogs. In *Practical Spoken Dialog Systems*, D. Dahl, 85-104. Text, Speech and Language Technology 26, Kluwer Academic Publishers, Dordrecht

[2] Höge, H. et al. 2008. Automotive speech recognition. In *Automatic Speech Recognition on Mobile Devices and over Communication Networks,* Z.-H. Tan, B. Lindberg, 347-373, Advances in Pattern Recognition IV, London.

[3] Jurafsky, D., Martin, J.H. 2008. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. 2nd ed. Pearson Education, Upper Saddle River, NJ

[4] McEnery, T., Wilson, A. 2001. Corpus linguistics: an introduction. Edinburgh University Press, Edinburgh

[5] Minker, W., Bühler, D., Dybkjær, L (Ed.). 2005. Spoken multimodal human-computer dialogue in mobile environments. Springer, Dordrecht

[6] Nass, C., Brave, S., 2005. Wired for Speech: How voice activates and advances the Human-Computer relationship. MIT Press, Cambridge, MA.

[7] Rescher, N. 1998. Complexity: A Philosophical Overview. Transaction Publishers, New Brunswick, NJ

[8] Wallis, S., Nelson, G. 2001. Knowledge discovery in grammatically analysed corpora. *Data Mining and Knowledge Discovery* 5, 4 (Oct. 2001), 305-335.

[9] Zoltan-Ford, E. 1991. How to get people to say and type what computers can understand. *International Journal of Man-Machine Studies* 34, 4 (1991), 527-547.