The Characteristics of Voice Search: Comparing Spoken with Typed-in Mobile Web Search Queries

IDO GUY, Ben-Gurion University of the Negev

The growing popularity of mobile search and the advancement in voice recognition technologies have opened the door for web search users to speak their queries rather than type them. While this kind of voice search is still in its infancy, it is gradually becoming more widespread. In this article, we report a comprehensive voice search query log analysis of a commercial web search engine's mobile application. We compare voice and text search by various aspects, with special focus on the semantic and syntactic characteristics of the queries. Our analysis suggests that voice queries focus more on audio-visual content and question answering and less on social networking and adult domains. In addition, voice queries are more commonly submitted on the go. We also conduct an empirical evaluation showing that the language of voice queries is closer to natural language than the language of text queries. Our analysis points out further differences between voice and text search. We discuss the implications of these differences for the design of future voice-enabled web search tools.

CCS Concepts: • **Information systems** → *Web search engines*; Search interfaces; Retrieval on mobile devices; *Speech/audio search*;

Additional Key Words and Phrases: Conversational search, mobile search, query log analysis, spoken search, voice search, voice queries

ACM Reference format:
Ido Guy. 2018. The Characteristics of Voice Search: Comparing Spoken with Typed-in Mobile Web Search Queries. *ACM Trans. Inf. Syst.* 36, 3, Article 30 (March 2018), 28 pages.
https://doi.org/10.1145/3182163

1 INTRODUCTION

The popularity of search from mobile devices (*mobile search*) has rapidly increased in recent years (Song et al. 2013). In fact, the number of mobile queries has already exceeded the number of those submitted from desktop devices in the United States and other countries (Shokouhi and Guo 2015). The nature of mobile search has also evolved, with growth in the number of unique queries and shifts in searched topics, from entertainment and adult content to business and commerce, similarly to the shifts of web search topics in its early days (Yi and Maghoul 2011).

A prominent characteristic of the advancement in mobile search is the emergence of *voice search*, allowing users to input queries in a spoken language and then retrieve the relevant entries based on system-generated transcriptions of the voice queries (Jiang et al. 2013). Recent developments

This manuscript is an extended version of Guy (2016).

Author's address: I. Guy, P.O. Box 653 Beer-Sheva 8410501, Israel; email: idoguy@acm.org.

https://doi.org/10.1145/3182163

in speech recognition, backed by high bandwidth coverage and high-quality speech signal acquisition, are enabling higher quality voice search (Chelba and Schalkwyk 2013). Already in 2010, Google presented a case study stating that their goal is to make voice search ubiquitously available and that a level of performance was achieved such that usage is growing, and many users become repeat users (Schalkwyk et al. 2010). Since then, further enhancements to automatic speech recognition (ASR) for web search have been reported (Shan et al. 2010; Zweig and Chang 2011), taking advantage of the large data that started to accumulate on voice search logs (Chelba and Schalkwyk 2013), and applying advanced learning methods (Hinton et al. 2012). The use of voice has also been promoted by the increasing popularity of voice-activated intelligent assistants, such as Google Assistant, Amazon's Alexa, Apple's Siri, and Microsoft's Cortana. These assistants provide context-based query-less personalized advice for mobile users, but also enable web search (Jiang et al. 2015; Kiseleva et al. 2016a). A recent survey of 1400 U.S. smartphone users, executed by Northstar Research and commissioned by Google, found that many teenagers use voice search every day (Awadallah et al. 2015; Shokouhi et al. 2016). It is therefore becoming important for information retrieval researchers and practitioners to understand this new medium of search and its differences from traditional *text search*.

Using voice as a means to search holds various potential advantages. Although typing usability has improved in recent years, querying by voice is still likely to be substantially easier and faster for the vast majority of mobile users. Voice is also the natural way people communicate with one another and express themselves. For users with visual or manual impairment, or with limited literacy skills, voice search may break down the entry barrier into web search. In addition, as searching by voice does not require visual attention or the use of hands, it can be performed in situations such as driving, cooking, or exercising, where typed search might be especially cumbersome, error-prone, and even dangerous. In the aforementioned survey, 78% of the teens who used voice search pointed out its usefulness for multitasking as a key motivating factor.[1]

In spite of its growing popularity, the area of voice search has not received much attention in the information retrieval literature. Early work compared voice and text queries in a laboratory study, however, these did not represent typical web search queries, but rather complex long questions (Crestani and Du 2006). More recent work has mostly focused on voice recognition (Acero et al. 2008; Chelba and Schalkwyk 2013; Moreno-Daniel et al. 2007; Shan et al. 2010; Wang et al. 2008; Zweig and Chang 2011) and query reformulation (Awadallah et al. 2015; Jiang et al. 2013; Shokouhi et al. 2014). A few studies revealed more details about how voice search is performed on commercial search engines (Schalkwyk et al. 2010; Yi and Maghoul 2011); however, we are not aware of a systematic log analysis of voice queries as of yet. Other works conducted laboratory experiments and user studies, which pose their own advantages, but do not enable inspecting behavior "in the wild" and at large scale (Crestani and Du 2006; Jiang et al. 2013).

In this work, we perform a query log analysis of half a million voice queries, issued to the mobile application of the Yahoo commercial web search engine, over a period of 6 months. The log includes English-only queries, from the United States, transcribed from voice to text using high-quality ASR. We compare the voice queries with a similar-size sample of mobile text queries, typed on the same mobile application. Our comparison inspects characteristics of context, clicks, sessions, and, primarily, the query text itself. We examine both semantic and syntactic features and compare them for voice versus text queries. In the final part of our analysis, we directly compare the similarity of the voice and text query language to natural language corpora, which include traditional news articles and the titles of questions in a large community question answering (CQA) website.

[1] https://googleblog.blogspot.co.il/2014/10/omg-mobile-voice-survey-reveals-teens.html.

Our work offers the following key contributions:

—To the best of our knowledge, we present the most comprehensive analysis of a web search engine voice query log.
—We combine a semantic analysis using novel methods, such as analyzing a broad set of triggered cards, with an in-depth syntactic analysis, to shed more light on the commonalities and differences between voice and text queries.
—We provide empirical evidence, based on language modeling, that voice queries are closer to natural language than text queries, yet are still distant from natural question language.

Our findings suggest different ways for search systems to enhance their support and take advantage of the unique characteristics of voice queries. We conclude the article by summarizing the key findings and discussing their implications and future research directions.

2 RELATED WORK

Studies of mobile query log analysis have been published throughout the past decade, ever since mobile devices became ubiquitous. One of the early studies (Kamvar and Baluja 2006) compared search patterns (queries, clicks, time spent on each search phase) on 12-key keypad cellphones, PDAs (with a QWERTY keyboard or a stylus input), and desktop (PC) computers. It found that the diversity of queries on mobile was substantially lower than on desktop and that the most popular query in each of the three device types was different. Baeza-Yates et al. (2007) compared mobile and desktop search queries on Yahoo Japan and found that mobile queries included fewer characters, more queries in the Business category, and fewer in Art. Yi et al. (2008) performed a large-scale query log analysis of the Yahoo OneSearch mobile service and found that mobile query patterns were dynamic, as users were exploring how to use the devices. Pattern use also varied among different geographies and application types.

With the evolution of mobile devices into smartphones, mobile search has also been shown to change. Kamvar et al. (2009) examined search behavior on iPhones and found it was more similar to desktop search than to search on basic mobile phones. Song et al. (2013) performed a broad 3-month log analysis of Bing search on desktop, iPad, and iPhone. They found that both mobile and tablet users issued significantly fewer navigational queries than desktop users, due to the wide availability of mobile apps on these two platforms. Due to the significant differences between user search patterns on the three platforms, they proposed a ranking system that considered platform-specific features. Montanez et al. (2014) focused on the transition of users across device types during the search process by analyzing a search log from desktop computers, smartphones, tablets, and game consoles. In this work, we focus on mobile devices for the comparison between voice and text queries.

With the advancement of speech recognition technologies, studies of mobile search using voice started to emerge. Many of the studies focused on voice recognition challenges. Wang et al. (2008) defined voice search as *"the technology underlying many spoken dialog systems that provide users with the information they request with a spoken query"* and reviewed key challenges, such as environmental noise, pronunciation variance, and linguistic issues. Acero et al. (2008) described the architecture of the speech recognition interface of Microsoft's "Live Search for Mobile". The key challenge they pointed out was the loss of signal-to-noise ratio caused by the fact that users often speak at arms length while looking at the screen or use the application in inherently noisy environments, such as in cars or on the street. Moreno-Daniel et al. (2007) discussed the interleaving of ASR with information retrieval (IR) systems and suggested to combine acoustic and semantic models to enhance performance.

In recent years, alongside the enhancement of ASR technologies with deep learning (Hinton et al. 2012), various studies suggested advanced methods for voice search ASR and reported further performance enhancements. Chelba and Schalkwyk (2013) leveraged the data on Google's voice search logs to enhance language modeling and achieved *"small but significant"* gains in speech recognition performance. Their follow-up study showed that augmenting the language model with geo-location information further improved the results (Chelba et al. 2015). Shan et al. (2010) described a system for Mandarin Chinese voice search and reported *"excellent performance on typical spoken search queries under a variety of accents and acoustic conditions."* Zweig and Chang (2011) found that the use of Model M (exponential *n*-gram language model) with personalization features improved the speech recognition performance on Bing voice search. Chan et al. (2016) developed a neural network architecture that subsumes the acoustic, pronunciation, and language models to support conversational speech recognition. Shokouhi et al. (2016) showed that using web search clicks and reformulations as training data for validating or correcting queries can substantially reduce the recognition error rate in voice search. In this work, we take advantage of the advancement in speech recognition, to explore a high-quality transcribed query log, but do not delve into speech recognition aspects.

Another body of research has focused on voice query reformulation, showing that users sometimes respond to voice recognition errors by different reformulation patterns, such as repeating a query or refining it (Jiang et al. 2013). Classifiers were built to predict and categorize voice reformulations, extending text-based approaches with features such as voice recognition time and confidence (Awadallah et al. 2015; Levitan and Elson 2014). Researchers also found that users do not tend to switch between voice and text when reformulating queries (Shokouhi et al. 2014). While our study does not focus on query reformulation, we report related statistics for voice versus text queries in our session analysis.

Most closely related to our research are three studies that directly referred to the comparison between voice and text queries. The first described a case study of the development of "Google Search by Voice" (Schalkwyk et al. 2010). It stated the system "*achieved a level of performance such that usage is growing rapidly, and many users become repeat users.*" While most of the report is focused on describing the technology and the evaluation of the voice recognition component, a section is dedicated to evaluating the user experience based on a 4-week query log analysis. It was found that the query categories "food & drink" and "local" (e.g., place names or business listings) were more popular with voice searches. Also, short queries, in particular one- and two-word queries, were relatively more frequent in voice searches than in typed searches, while longer queries (five+ words) were far rarer. A poster by Yi and Maghoul (2011) inspected the change in mobile search on Yahoo from 2007 to 2010 and provided a short comparison of 79K voice queries to typed mobile and desktop queries, which examined query length and categories. The most comprehensive comparison between voice and text queries was performed in a lab study from over a decade ago (Crestani and Du 2006). The 12 participants were students from the local research lab, as voice search was in its infancy and required IR experience. They were asked to formulate 10 TREC topics as queries in a process that took 1 to 3 minutes per query. The resulting queries did not reflect typical web queries and were complex and long (23.1 words for an average voice query, 9.5 for text).[2] Moreover, the study did not involve a search system and participants were not exposed to search results. In addition to query characteristics, such as length and typing duration, the retrieval effectiveness of typed

[2]An example for a spoken query was: "*I want to find document about Grass Roots Campaign by Right Wing Christian Fundamentalist to enter the political process to further their religious agenda in the U.S. I am especially interested in threats to civil liberties, government stability and the U.S. Constitution, and I'd like to find feature articles, editorial comments, news items, and letters to the editor*" Its textual counterpart was: "*Right wing Christian fundamentalism, grass roots, civil liberties, US Constitution.*".

versus spoken queries was evaluated. In our analysis, when relevant, we tie to the results reported in these three studies and discuss the commonalities and differences with our extended findings.

3 RESEARCH SETTINGS

Our analysis is based on a random sample of 500,000 queries from the Yahoo mobile search application, performed by over 50,000 unique users of the voice interface along a period of exactly 6 months (April–October 2015) in the United States. The mobile search application transcribes a voice query into a text query using state-of-the-art ASR, and from this point onward treats it as a text query. In other words, the multi-modal interface allows inputting queries by voice, but returns results using the same information retrieval techniques and the standard mobile search user interface. For comparison, we collected an identical number of queries performed using the "regular" keyboard-based interface of the same mobile application. We refer to the former set of queries as *voice queries* and to the latter as *text queries*. The text queries were collected along the same period of 6 months for a similar number of users. Moreover, we sampled an identical number of voice and text queries in each day of the experimental period. When inspecting day-of-week distribution and session statistics, we compared all queries from all users in our voice sample with all queries from all users in our text sample, during 2 months of the experimental period, to allow suitable analysis.

Each query in the log, either voice or text, included, in addition to the query itself, a timestamp (adapted to the time zone in which it was performed), a location in the form of city and state, the device type (desktop or smartphone), and, for logged-in users, the user's age and gender. In addition, for each query we had information about its associated clicks, if any were performed, including the corresponding URLs and ranks within the search engine results page (SERP).

Our analysis is organized as follows. Section 4 compares basic characteristics of voice and text queries, including context, query length, and session characteristics. Section 5 examines the query semantics by inspecting different categories as well as specific queries and terms. Section 6 looks into click behavior and distribution of clicked domains, reflecting on the findings in the query semantics analysis. Section 7 examines query syntax as reflected in characteristics of parsing and distribution of part-of-speech tags. The semantic and syntactic analyses reveal various differences between voice and text queries, of which many indicate that voice queries are phrased closer to natural language than text queries. The final part of our evaluation therefore explicitly compares the similarity of both types of queries to natural language corpora, and is described in Section 8.

4 BASIC CHARACTERISTICS

In this section, we compare basic characteristics of voice and text queries, including context aspects, basic query features, and session characteristics.

4.1 Context

We found similar contextual characteristics for voice and text queries in terms of searcher's age and geography (cities and states). There was a slight tendency toward male searchers in the voice log compared to the text log (up 3%).

The distribution across day-of-week was similar for voice and text queries: in both, there was a slight peak on weekends compared to weekdays (5% more searches on average). In contrast, there was a noticeable difference between voice and text queries with regard to time-of-day, as depicted in Figure 1. Voice queries were more frequent during day hours (from 8am to 8pm), while higher portions of the text queries (relative to voice) were performed during evening, night, and early morning hours (8pm to 8am). These differences were consistent during weekdays and weekends. Previous work found that mobile search in general is more common during evening hours, while desktop search is more common during work hours (Song et al. 2013). Our results indicate that currently, voice search is closer to desktop search use w.r.t. time-of-day.

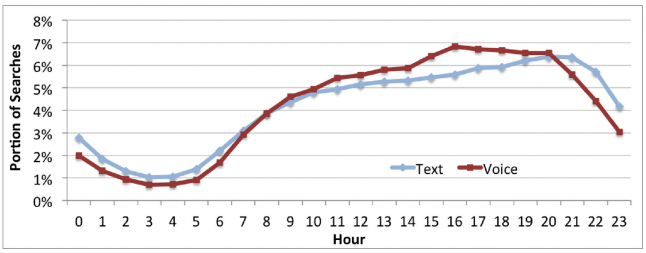

Fig. 1. Query distribution by hour of the day.

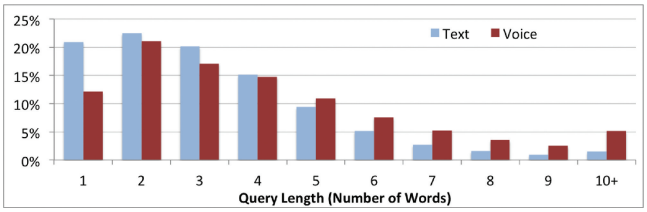

Fig. 2. Query length distribution by number of words.

In our analysis, we inspected the results while controlling for factors that were found to be different between voice and text queries, including time-of-day and gender. When relevant, we report the influence of these factors on the results.

4.2 Queries

The average query length was significantly higher for voice queries at 4.2 (std: 2.96, median: 4, max: 109) versus 3.2 for text (std: 2.38, median: 3, max: 308). Query length was measured by the number of words, using white-space tokenization. The substantially higher maximum value in text queries is likely due to the use of the copy-paste feature, which does not exist for voice. Figure 2 shows the detailed distribution of voice versus text queries by length. It can be seen that one-word queries were particularly rarer on voice (12.2% vs. 21% for text), perhaps implying a lower portion of navigational queries. Voice queries were more common starting at queries of five words, which have been recently referred to as "verbose" queries (Gupta and Bendersky 2015). Overall, 34.5% of the voice queries were of five words or more, compared to only 21.2% of the text queries. The length difference between voice and text queries has a major effect on query syntax, which we examine more closely in Section 7.

Previous work was somewhat inconsistent with regard to voice versus text query length. While a Google case study found that voice queries tend to be shorter than typed queries (2.5 versus 2.9 on average, respectively) (Schalkwyk et al. 2010), other studies found voice queries to be longer (e.g., 3.4 versus 2.2 on a Yahoo study) (Crestani and Du 2006; Yi and Maghoul 2011). Jiang et al. (2013) pointed out this discrepancy and stated that further studies are needed to identify the characteristics of queries in voice search.[3] Our findings evidently support voice queries being substantially longer than text queries (while also indicating a general trend of queries becoming longer on mobile search).

[3]The authors state in their paper: "*Schalkwyk et al. found that voice search queries were tend to be shorter than in conventional searches, whereas Crestani et al. found that voice queries tend to longer and more similar to natural language. Since we did not conduct conventional search experiments for comparison, we cannot come to an answer to this disputable issue. We suggest that further studies are needed to identify the characteristics of queries in voice search.*"

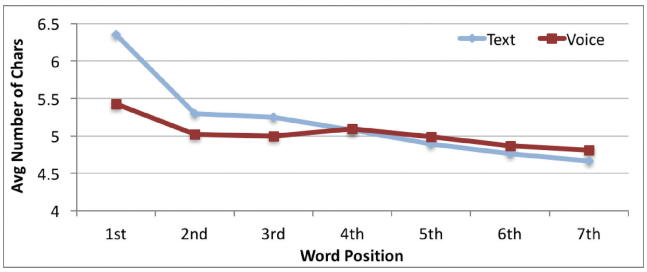

Fig. 3. Query length distribution by number of characters.

Table 1. Session Characteristics

	Text	Voice
Portion of one-query sessions	65.1%	67.1%
Avg. (std.) session length	1.74 (1.61)	1.77 (2.03)
Avg. (std.) idle in seconds	174.4 (207.9)	121.8 (182.7)
Median idle in seconds	84	63.5
Portion of identical queries	13.7%	15.1%
Portion of refining queries	10%	7.6%

The portion of unique queries for voice was 73.9%, compared to 77.6% for text. This stands somewhat in contrast to voice queries being longer, as we would expect more repetition for shorter queries. We believe this finding stems from two main reasons. First, the lower query diversity for voice search characterizes search in its earlier stages, as has been the case for web and mobile search (Yi and Maghoul 2011). Second, the use of abbreviations, spelling variants, and punctuation marks, makes text queries more diverse. We will further demonstrate this in Section 5.

The length of voice queries compared to text is also reflected by the number of characters per query: 23.9 for voice (std: 15.9, median: 22) versus 20 for text (std: 18.4, median: 18). Figure 3 depicts the average of number of characters per word for the ith word in a query, for $i = 1, 2, \ldots, 10$ (for queries with i words or more). A sharper fatigue effect can be observed for text queries: the length of the word decreases more intensely as its position in the query increases, likely reflecting the higher effort required for typing words versus saying them. First words in text queries are especially long. We conjecture that this is due to one-word queries, which often contain site names that tend to be longer than the average word. As we have seen in Figure 2, one-word queries are substantially more common in text queries.

4.3 Sessions

A session is a series of queries issued by an individual user in close succession, often with all queries being related to the same topic. Using a common approach for defining sessions (Teevan et al. 2011), we considered queries that occur in a sequence without 15 minutes of inactivity as part of the same session. Table 1 shows session statistics. The average session length on voice and text queries was very similar, with higher standard deviation for voice. About two-thirds of the sessions, on both voice and text, included only one query. Average and median idle times were shorter on voice queries, likely as a result of the faster inputting enabled by voice. This gap may also reflect the fact that voice queries often focus on topics that require little interaction with the results, as our analysis will later demonstrate. Finally, we inspected the relation of a query in a session to its previous query (when such exists): the bottom of Table 1 shows the portion of

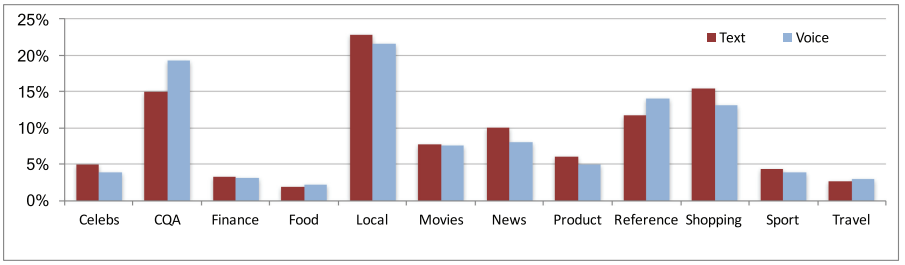

Fig. 4. Query distribution by category.

identical and refining queries (a query q2 refines a query q1 if q1 is a prefix of q2, but not identical to q2). Overall, the numbers are similar, with somewhat higher portions of identical queries on voice and refining queries on text. It could be that the slightly higher portion of identical queries on voice is due to retrying (without success) to change the query's transcription (Levitan and Elson 2014). Other than that, however, we did not find evidence for a substantially higher rate of query reformulation on voice compared to text.

5 QUERY SEMANTICS

In this section, we compare query semantics between voice and text queries. We first examine higher level query categories both using a query classification tool and by inspecting the triggering of vertical cards. We then examine the queries themselves more closely, by comparing the most popular queries, the most distinctive query terms, and characterizing language patterns.

5.1 Query Categories

To get a high-level sense of query categories, we used an in-house query classification tool, which is based on named entity recognition and supervised machine learning trained on a huge corpus of queries, categorized according to their clicked websites. We removed two broad categories ("Web" and "Search"), which left us with 12 categories. We only considered the category with highest confidence for each query; in case the overall confidence was too low (below 50%), we discarded the query in our analysis. Overall, 54.7% of the voice queries and 54.9% of the text queries could be classified according to this scheme. Their distribution across the categories is shown in Figure 4.

The categories that emerge as more common for voice queries are CQA (with a ratio of 1.29 between the portion of voice queries and text queries) and Reference (encyclopedias, dictionaries, museums, and similar, with a ratio of 1.2). Two smaller categories more popular for voice are Food (1.16) and Travel (1.11). On the other hand, categories that are more common for text queries include Celebrities (0.77), News (0.81), Product (0.82), Shopping (0.87), and Sport (0.9). For Local (0.94), Finance (0.95), and Movies (0.99), the differences were minor. Overall, we see that voice queries are more popular for question answering and knowledge seeking, while text queries are more popular for news and business.

5.2 Triggered Cards

The recent evolution of web search has introduced richer experience of the SERP, with cards (also referred to as "oneboxes" or "direct displays") showing results of verticals such as weather, direct factual answers, or live sport scores (Shokouhi and Guo 2015). These cards extend organic search results ("ten blue links") by addressing a user's specific information need directly on the search page (Chilton and Teevan 2011), often sparing the need for a click (Lagun et al. 2014). Commercial

Table 2. Voice(V)/text(T) Card Triggering Ratio

Frequent Cards		Infrequent Cards	
Card Name	V/T Ratio	Card Name	V/T Ratio
Music Videos	1.32	Time	5.65
CQA	1.3	Countries	1.52
Recipe	1.3	Dictionary	1.43
Maps	1.27	Weather	1.19
Movies	1.13	Cars	0.95
Finance	1.02	TV	0.9
People	0.75	Lottery	0.56
Sports	0.73	Horoscope	0.36

search engines trigger cards upon identification of a relevant user's intent. The card triggering technology largely relies on query pattern matching that provides high precision, since the presentation of a wrong card might substantially degrade the user experience and is therefore highly undesirable. In the following analysis, we examine the distribution of presented cards for voice versus text queries, relying on the search engine's technology for card triggering to shed light on semantic differences between the two types of queries. The card triggering technology allows us to build on the ability to capture clear and specific user intent based on different patterns of language. For example, the dictionary card is triggered by queries such as "what is <term>," "<term> definition," or "meaning of <term>."

Overall, the portion of queries for which at least one card was triggered was 43.3% for voice queries and 40.6% for text. The higher portion for voice may be associated with two related aspects: (1) as cards often spare the need for a click, they reflect queries that may require less interaction from the user, which is suitable for the voice scenario; (2) some cards reflect a very narrow user intent, with a focused answer (e.g., a dictionary definition, weather, or time), which may be more common on voice search.

Table 2 presents the "triggering ratio" for 16 different cards (i.e., the ratio between the portion of voice queries for which the card was presented and the portion of text queries for which it was presented). We excluded very broad cards (news, shopping, and digital magazines) and cards that reflect a vague or ambiguous intent (e.g., company, which also includes websites such as Facebook or Amazon). The left side of the table shows the ratio for frequent cards, triggered for at least 0.5% of both voice and text queries, while the right side focuses on narrower less common cards that still appear for at least 0.1% of either voice or text queries. Among the frequent cards, music videos, which are triggered by queries such as song names and singer names, were the most common for voice compared to text (highest triggering ratio). CQA cards, which include direct (inline) answers from community question answering sites, recipe cards, and map cards, were also more common for voice. On the other hand, sports and people cards were considerably more common for text queries. The latter is a particularly common card, triggered for celebrity names and sometimes a refining keyword such as age, spouse, or height. As we have already seen in the query classification analysis, the "celebs" category is particularly more common on text compared to voice.

Reviewing the less frequent cards, on the right side of Table 2, the time card was largely more common on voice queries, with a triggering ratio of over 5.5. This card is typically triggered by queries that ask for the time in a specific location (e.g., "savanna time" or "what is the hour in chicago"). Countries (country names, sometimes with refinements such as "capital of" or "population"), dictionary, and weather, were also more popular on voice, whereas lottery and even more

Table 3. Most Popular Queries

Text	Voice
facebook	youtube
youtube	yahoo mail
pornhub	facebbok
google	google
yahoo mail	hello
craiglist	craigslist
porn	yahoo
xnxx	walmart
yahoo	amazon
amazon	home depot
redtube	yahoo.com
xvideos	amazon.com
facebook login	news

so horoscope were more popular on text. Interestingly, TV cards have a substantially lower ratio than movies, even though the corresponding queries are similar in nature, with the difference being the artifact—a TV show versus a movie. This reflects, perhaps, a stronger use of voice search on the go compared to use at home. Overall, we see that many of the infrequent cards are more commonly triggered for voice. Additionally, it appears that cards with concise answers (time, definition, weather) are more commonly triggered for voice queries, while cards that require higher user engagement, as they present richer content or are more likely to require interaction (horoscope, lottery), are more commonly triggered for text queries.

Some of the findings in this section coincide with the Google case study (Schalkwyk et al. 2010), which identified "food & drink" and "local" as the more popular categories for voice (out of a total of eight), corresponding with our findings w.r.t. the "recipe" and "maps" cards, respectively. That study also found "online communities" and "adult" categories less frequent for voice, to which we show support in the next section.

5.3 Popular Queries

After inspecting different types of query categories, we now look more closely into specific queries and terms. We first compare popular queries, then examine distinctive terms for voice versus text queries and vice versa, and finally inspect characterizing language. Table 3 shows the most popular queries for text versus voice (popularity is measured by the number of unique users who issued the query at least once). The top query in each list is already different: "facebook" for text versus "youtube" for voice. The difference between the two is substantial: on text queries, "facebook" was issued by a number of users larger by a factor of 1.7 than the number of users who queried for "youtube," while on voice this ratio was only 0.44. It can also be seen that adult site queries ("porn," "xnxx," "redtube," "xvideos") appear only on the top text list. On the other hand, the voice list includes more retail brands ("walmart," "home depot") and the query "hello," which is likely used for experimentation with the voice system. In addition, the voice list includes the use of the suffix ".com" for popular navigational queries, for example, both "yahoo" and "yahoo.com" are on the top list for voice (and similarly for "amazon"). Across all queries, however, the ".com" suffix was less common on voice than on text, as we will show later in this section.

Table 4. Top Text and Voice Distinctive Query Terms

Unigrams		Bigrams		Trigrams	
Text	Voice	Text	Voice	Text	Voice
pornhub	the	yahoo mail	is the	facebook sign up	what is the
2015	is	facebook sign	what is	credit card login	how do you
xnxx	a	sign up	how do	fargo bank login	phone number for
tumblr	what	you tube	do you	bobbi kristina brown	do you spell
facebook	in	dear abby	in the	dicks sporting goods	how old is
redtube	you	online login	number for	online login site	how do i
tx	to	big tits	of the	drudge report 2015	what time is
ca	of	near me	phone number	yahoo mail inbox	where is the
st	how	mlb scores	north carolina	chase online login	i need the
login	end	schedule 2015	pictures of	wireless my account	time is it
vs	i	crossword clue	new york	verizon wireless my	what is a
nc	for	yahoo news	for the	scrabble word finder	what are the
craigslist	on	card login	what's the	online banking login	i want to
ny	do	horoscope 2015	where is	craigslist los angeles	take me to
dr	number	season 2	show me	toys r us	i'm looking for

5.4 Distinctive Query Terms

To further inspect semantic differences, we set out to explore which terms mostly characterize voice versus text queries. To this end, we used Kullback-Leibler (KL) divergence, which is a non-symmetric distance measure between two given distributions (Berger and Lafferty 1999; Carmel et al. 2012). Specifically, we calculated the terms that contribute the most to the KL divergence between the voice and text query language models, for unigrams, bigrams, and trigrams.[4] Table 4 reports the terms with the highest KL divergence for text queries (w.r.t. voice queries) and for voice queries (w.r.t. text). Inspecting the unigrams, the terms on the voice list mostly include common function words (determiners, prepositions), question words, and pronouns. The only two nouns on the list are "end" and "number." For "end," closer inspection verified that this is due to the ASR often confusing it with the more common "and" (e.g., "can i eat onions end garlic while breast-feeding" or "the preacher end the bear song"). The text unigram list, on the other hand, includes nouns such as site names (especially social media and adult sites) and common abbreviations (state or city names: "ny," "tx," "nc"; "st"; "dr"; "vs"), which are hardly ever used on voice.

For bigrams and trigrams, the text lists include website and entity names, sometimes with extending keywords such as "sign up," "login," "online," "2015," "news," or "scores." On the other hand, the voice list includes many common parts of natural language ("what is," "in the"), requests phrased in natural language ("show me," "take me to," "i'm looking for," and "phone number for"), and also a few state names that appear in their standard form ("new york," "north carolina").

Finally, we also used the KL analysis to examine distinctive unigrams positioned at the beginning and at the end of a query. The most distinctive words to open a voice query were the question words "what" and "how" and the most distinctive for text queries were the site names "pornhub" and "facebook." The most common word to terminate a voice query was "please," again indicating the use of natural language, while for text it was "2015," perhaps as it is easy to write but relatively long to pronounce.

[4]We elaborate on the smoothing method in Section 8.

Table 5. Language Differences Between Voice and Text Queries

Pattern	%v	v/t	Pattern	%v	v/t	Pattern	%v	v/t
^wh-question	9.9	2.67	'	4.6	2.65	texas	1.23	2.5
wh-question	11.5	2.68	,	0.67	1.12	california	0.76	3.53
^question	11.9	2.55	.$	0.2	1.42	new york	0.75	3.57
^how	3.6	2.25	"	0.04	0.41	ohio	0.74	1.81
^what	3.5	3.06	?$	0.01	0.05	florida	0.65	2.32
^where	0.8	3.07	's	3.5	2.46	illinois	0.38	3.22
^who	0.7	3.2	'm	0.26	6.29	pennsylvania	0.24	6.53
^when	0.5	2.49	don't	0.21	5.89	chicago	0.19	0.92
^why	0.4	1.83	'll	0.07	5.03	houston	0.18	0.92
^which	0.15	4.55	.com	3.1	0.6	phoenix	0.08	1.18
^is	0.59	2.16	www	0.1	0.07	los angeles	0.06	0.71
^can	0.48	2.02	http	0	0.02	philadelphia	0.05	0.82
^does	0.22	1.81	kim kardashian	0.05	0.63	walmart	0.42	1.17
^do	0.21	1.87	bruce jenner	0.04	0.97	amazon	0.26	0.69
^are	0.18	2.55	lebron james	0.03	0.85	home depot	0.2	1.49
^what's	0.32	14.39	donald trump	0.03	0.83	ebay	0.18	0.72
^who's	0.04	6.42	^kim kardashian$	0.02	0.48	lowe's	0.15	1.82
^how to	1.3	1.51	^bruce jenner$	0.01	0.47	mcdonald's	0.09	6.43
^how do you	0.44	7.43	^lebron james$	0.01	0.57	walgreen	0.08	1.32
^how do i	0.19	4.81	^donald trump$	0.01	0.42	best buy	0.07	1.02

5.5 Additional Language Characteristics

Table 5 includes additional differences that stood out in our analysis between voice and text queries. The table shows the statistics of queries that match different patterns, presented as regular expressions (where "^" marks the beginning of a query, "$" marks the end of a query, and "*" marks a sequence of zero or more characters). The "%v" column shows the portion of voice queries that matched the pattern and the "v/t" column shows the ratio between the portion of voice queries that matched the pattern and the portion of text queries that matched it. On the leftmost column of the table, we focus on the use of question words, which exhibits one of the most prominent differences between voice and text queries. A recent paper studied this form of "question queries," which *"take the form of natural language"* (White et al. 2015). We used a similar methodology to identify this type of queries for voice and text. Overall, 9.9% of the voice queries begin with a wh-question word (one of the 5W1H), 2.67 times higher than the portion of text queries. Adding yes/no questions (start with "is," "can," "does," etc.), the portion further grows to 11.9%, with a ratio of 2.55 to text queries (see the entry for "^question" in the table).

The two most popular question words to open a query, by a large margin, were "how" (3.6% of all voice queries) and "what" (3.5%). The voice/text ratio is substantially higher for "what" than for "how." Also with high ratio are "which" (although the most uncommon in general), "who," and "where." "Why" queries have the lowest voice/text ratio at 1.8, probably as these are more

open-ended questions, often characterized by longer answers that require more exploration on the part of the user (Verberne 2007).

The next section in the table focuses on the five most common words to open a yes/no question. In general, the ratios for these are lower than for wh-question words. Aside from these five, "would" and "could" had a particularly high ratio between voice and text, probably as they represent courteous language more characteristic of speech.

Queries that begin with "what's" and "who's" were extremely more common on voice. This is partially since in text it is also possible to see a version without the apostrophe (e.g., "whats"), which never appears on voice.

The final section on the left of the table focuses on three specific types of "how" queries: while the "how to" form is only ×1.5 more common on voice, the ratio for "how do i" and even more so "how do you," which reflect a more natural language, is very high in favor of voice (albeit less common in general).

The middle column of Table 5 starts with various punctuation marks. Generally, punctuation marks are uncommon on mobile queries and are even more uncommon on voice compared to text queries. Most punctuation marks on voice queries are inherently added by the speech-to-text system (e.g., toys "r" us contains quotation marks by definition), rather than explicitly pronounced by the user. Therefore, very few voice queries end with a question mark compared to text queries, where it is also rare, but still ×20 more common than voice. Queries ending with a period are actually more common on voice. A closer inspection reveals this is due to abbreviations that explicitly use a period and appear at the end of a query, such as "st.," "ave.," or "inc." Commas are also slightly more common on voice, due to patterns such as large numbers (e.g., "60,000"), addresses ("east michigan avenue, jackson, michigan"), and dates ("may 29, 1948"). Double quotes are less commonly used on voice than on text (used on voice mostly as part of entity names, such as toys "r" us and unit marks, such as 3"). Finally, apostrophes are substantially more common on voice, due to inherent use of possession, and to a smaller extent contractions (e.g., "I'm," "don't," or "we'll").

The third section in the middle column of the table examines common patterns of URLs, all less frequently used on voice. The ratios for "www" and "http" are especially low, suggesting users much more rarely say them than type them.

The bottom section of the middle column shows a few examples of the most commonly searched celebrities during our experiment's period. As we have already seen, celebrities are searched more often on text. The voice/text ratio further decreases when inspecting queries that include the person's name only, without any extensions, such as "how old is," "spouse," or "leaked photos." Apparently, when the information need is more focused, the gap in celebrity-related queries between text and voice queries narrows. Another thing to note is that on text queries, we can find more diverse name forms. For example, for "catlyn jenner" we see both "catlyn" and "catlin" on text, while only the former on voice. For NBA player Stephen Curry, the formal name version is rarely used on voice, while the short version "steph curry" and incorrect "stefan curry" more often used on voice. The use of both Stefan and Stephan on voice is likely due to different pronunciations, indicating that morphology may also increase the entropy over entity names in voice search.

The rightmost column of Table 5 examines more types of named entities. The first section focuses on the full names of the seven most populated states in the United States. Evidently, such full names are much more common on voice queries. On the other hand, as we have already seen, the short two-letter abbreviations are more commonly used on text queries. It can also be seen that the voice/text ratio increases as the state's full name becomes longer and harder to type, until surpassing 6.5 for "pennsylvania." Inspecting the "%v" column, it is also noticeable that states with longer names are less frequently spoken: "california" is the most populated while "new york" is

Table 6. Example Voice and Text Queries that "Landed" on the Same CQA Page

Voice Query	Text Query
looking for a restaurant that serves oysters in san francisco	oysters restaurant sf
how many minutes are played in women's soccer	women soccer duration
what restaurant did colin farrell and vince vaughn have dinner at in brooklyn	colin farrell vince vaughn joint dinner
need to see old sites i visited	view browsing history
is priority shipping and standard shipping the same thing	priority vs standard shipping
if you're 65 years old do you need a fishing license	fishing license senior citizen
i need the phone number for walmart in canton connecticut please	walmart canton connecticut phone

both a state and a city, yet both are preceded by "texas"; "ohio" is the least populated of the seven but is fourth in terms of frequency on voice queries. This trend disappears when it comes to city names, as can be seen in the next section of the table, which includes the names of the five most populated U.S. cities, aside from New York, which is also a state name. The voice/text ratio is close to 1 and even below it for most cities. We conjecture that the reason for these findings is that for states, users tend more strongly toward using the abbreviated version on text queries as the state's name becomes more difficult to type.

We finally inspect shopping brand names, particularly of big retailers in the United States. One trend that can be observed is that online shopping websites, such as Amazon and eBay, are more popular with text queries, while retailers that are prominently active in physical locations (Walmart, The Home Depot) are more popular with voice queries, perhaps since more queries are issued on the go. Inspecting the example of Lowe's, the number of queries is similar between voice and text queries. For voice, however, all of them appear with the standard form "lowe's," while for text they are equally split between "lowe's" and "lowes." Similarly, more than half of the text queries that refer to "mcdonald's" do not use the apostrophe.

Our analysis repeatedly implies that voice queries are more commonly used on the go (e.g., the map card ratio, movie versus TV cards, physical versus online shopping brands). To further explore this aspect, we examined queries for which both the search activity's zipcode and the user's zipcode were available (24.9% of the voice queries and 26.2% of the text queries). We found that while for 61.7% of the text queries the two zipcodes were identical, for voice queries they were identical for only 45.2%, giving another indication that voice queries are more often performed away from home. Previous work has already shown that mobile users issue queries from a more diverse set of locations, and particularly away from their home, as compared to both desktop and tablet users (Kravi et al. 2015; Song et al. 2013). Our findings suggest that this trend becomes even more substantial for voice search users.

Thus far, we have seen many quantitative characteristics by which voice queries differ from text queries. Next, we show a few anecdotal examples of voice and text queries used to perform a semantically similar search. To this end, we inspected queries in our voice sample that landed (i.e., resulted in a click) on CQA pages, as these often reflect a specific information need. For such queries, we matched text queries that landed on the same CQA page during a period of 1 week (our text sample did not contain such matches, thus we had to inspect a larger log). Table 6 presents seven examples of voice and text query pairs that landed on the same CQA page and express a similar information need. These examples nicely demonstrate some of the findings pointed out during this section. We note, however, that we cherry-picked examples where the voice query was especially different from the corresponding text query. In other cases, voice queries were similarly

Table 7. Click Statistics

	Voice/Text Ratio
Click-through rate (CTR)	0.78
Avg. # of clicks	0.83
Multi-click queries	0.73
MRR for all queries	0.78
MRR for clicked queries	0.97
% Blue links	0.94
% Rank 1	0.85
% Unique domains (hosts)	0.71
% Top domains	1.1

phrased to text queries. For instance, inspecting our original samples of voice and text queries, 13.1% of the voice queries were completely identical to a query in the text sample.

6 CLICKS

In this section, we compare the click behavior for voice versus text queries and inspect the distribution of clicked domains. Table 7 shows the ratio for various click characteristics between voice and text queries.[5] It can be seen that the click-through rate (CTR; the portion of queries for which at least one click was performed) and average number of clicks per query were substantially lower for voice queries. We conjecture that voice queries are often conducted in a situation that allows less interaction with the device, including clicking on search results. Indeed, the portion of multi-click queries (Kravi et al. 2016), that is, queries for which the user has performed two or more clicks, is even further lower for voice queries, as compared to text queries. Multi-click queries were found to often express a complex information need, with exploratory nature (Kravi et al. 2016).

The mean reciprocal rank (MRR) across all queries is also considerably lower for voice queries. The MRR across clicked queries only is similar for voice and text queries, indicating that the difference in the general MRR is mostly due to the lower CTR of voice queries.

The portion of clicks on organic search results ("blue links") out of all clicks on the SERP is lower for voice queries. This means that higher portions of the clicks on voice queries are performed on ads and cards. We have already seen that cards more often appear on voice SERPs in Section 5.2. Out of the blue-link clicks, the portion of clicks on the top-ranked blue link was lower for voice queries. This could indicate a lower portion of navigational queries, which are usually characterized by a click on the top result (Chakrabarti et al. 2009).

Inspecting the clicked domains (a domain is determined by the "host" part of the clicked URL), the portion of unique domains out of all clicks is substantially lower for voice queries, indicating lower diversity. Further analysis indicated that a higher portion of the voice clicks are performed on top domains, determined using a list of the 100 most commonly clicked domains during our experiment's period. Our analysis inspected the entire set of voice versus text queries; future research may compare the click characteristics of voice and text over the same queries.

Table 8 shows the top clicked domains for text and voice queries. The rightmost column shows, for each of the top voice domains, the "click ratio that is, the ratio between its number of clicks on the voice query sample and number of clicks on the text query sample. Differences emerge between the two lists from their top. While the most clicked domains for text queries are Wikipedia and Facebook, they are only 2nd and 5th, respectively, on the voice click list, with low click ratios,

[5] Actual values are not disclosed due to business sensitivity.

Table 8. Most Clicked Domains on Text Queries and Voice Queries

	Text	Voice	
	Domain	Domain	V/T Ratio
1	en.wikipedia.org	video.search.yahoo.com	2.05
2	facebook.com	en.wikipedia.org	0.81
3	pornhub.com	youtube.com	1.26
4	video.search.yahoo.com	answers.yahoo.com	1.26
5	youtube.com	facebook.com	0.51
6	answers.yahoo.com	pornhub.com	0.61
7	xvideos.com	local.yahoo.com	1.18
8	local.yahoo.com	maps.yahoo.com	1.99
9	xnxx.com	yellowpages.com	1.75
10	amazon.com	answers.com	1.76
11	redtube.com	amazon.com	0.76
12	imdb.com	xvideos.com	0.54

The "V/T Ratio" column presents the click ratio between voice queries and text queries for the most clicked voice domains.

especially for the social network, at 0.51. Instead, the top of the voice query list is dominated by video domains, video.search.yahoo.com, with more than double the clicks as on text queries, and the popular video sharing site Youtube. Also higher on the voice list are CQA sites, such as Yahoo Answers at 4th (6th on text) and Answers.com at 10th (not among the top 12 for text). On the top voice list only, with high click ratios, are also maps.yahoo.com and Yellowpages, reflecting more specific information needs (we have already seen "phone number" is a distinctive term for voice queries). Another evident difference is with adult sites: while four are among the top text domains (Pornhub, Xvideos, Xnxx, and Redtube), only two make the top voice domains, with low click ratios. As we saw, text queries are more popular during night hours, which could explain the difference. Inspecting the portions of adult site clicks by the hour of the day, we indeed observed a sharp increase during night hours compared to day hours; however, the gap between text and voice clicks persists throughout all hours of the day.

Further inspecting the lists of top clicked domains revealed more differences between voice and text searches. Table 9 presents additional commonly clicked domains with their voice/text click ratio. It can be seen that clicks on health-related sites (WebMD, Drugs.com, NIH; as well as the appearances of "health" within all clicked domain strings) were more common on text queries (click ratio smaller than 1). Additionally, text queries more commonly led to clicks on shopping, news, finance, celebrities, and social network sites (Twitter, LinkedIn), with the latter having a particularly low ratio between voice and text clicks. Also, despite the differences in favor of voice clicks for audio and video results, there was no such difference for photo sites, such as Photobucket and Pinterest. On the other hand, voice queries had more clicks on CQA sites, maps, weather, dictionary, recipes, video streaming, and music sites.

7 QUERY SYNTAX

In Section 4.2, we found that voice queries tend to be longer than text queries. In this section, we delve deeper into syntactic analysis of voice versus text queries. Our analysis includes two parts: we first inspect the characteristics of syntactic parsing of both types of queries and then examine the distribution of key part-of-speech (POS) tags. In our analysis, we used two corpora for additional comparison: the first is the Wall Street Journal (WSJ) corpus (Sections 2–23) (Marcus

Table 9. Additional Clicked Domains with Their Voice/Text Click Ratio

Domain	V/T Ratio	Domain	V/T Ratio
webmd.com	0.88	wikihow.com	1.3
drugs.com	0.77	ehow.com	1.22
nih.gov	0.64	ask.com	1.23
"health"	0.91	stackoverflow.com	2.1
ebay.com	0.92	"answers"	1.67
craigslist.org	0.54	"how.com"	1.28
"shopping"	0.84	maps.google.com	1.83
news.yahoo.com	0.82	mapqest.com	1.75
"news"	0.7	"maps"	1.74
sports.yahoo.com	0.72	weather.com	1.35
espn.go.com	0.8	accuweather.com	2.22
"sport"	0.83	"weather "	1.48
finance.yahoo.com	0.48	dictionary.reference.com	1.43
"finance"	0.56	thefreedictionary.com	1.75
"bank"	0.89	"dictionary"	1.35
celebs.yahoo.com	0.63	all-recipes.com	1.58
"celeb"	0.81	recipe.com	1.45
twitter.com	0.17	"recipe"	1.53
linkedin.com	0.37	screen.yahoo.com	1.97
photobucket.com	0.97	"screen"	1.77
pinterest.com	0.65	"itunes"	1.88
"photo"	0.87	"music"	1.33

The double-quoted strings represent all domains that contain the respective string.

et al. 1993), which primarily includes business and financial news articles, broken into sentences (a total of 42,248 sentences). The second is a collection of 500,000 question titles in English, randomly sampled from the Yahoo Answers CQA website. We only considered question titles of one sentence (over 90% of all titles on the site).

7.1 Parsing Characteristics

For this analysis, individual sentences from each of the four corpora—WSJ, question titles, voice queries, and text queries—were tokenized, pos-tagged, and syntactically parsed using the Stanford parser.[6] The parser first generates an unlexicalized PCFG parse (Klein and Manning 2003) and then produces typed dependencies by matching patterns on CFG trees (De Marneffe et al. 2006).

Table 10 reports four measures of syntactic complexity (four middle columns) for each of the four corpora. The first column shows the median and average number of tokens per parsed item.[7]

[6] http://nlp.stanford.edu/software/lex-parser.shtml.

[7] The number of tokens is slightly higher than reported in Section 4.2, due to the use of the Stanford tokenizer instead of white-space tokenization.

Table 10. Syntactic Properties of Four Corpora

Corpus	Median (mean) token count	Median (mean) tree depth	$root \rightarrow NN*$ edges (%)	$root \rightarrow VB*$ edges (%)	Median (mean) parse score
WSJ	20 (20.41)	6 (6.5)	9.2	84.1	−6.2 (−6.3)
CQA question titles	10 (10.6)	4 (4.3)	16.7	75.0	−9.3 (−10.4)
Voice queries	4 (4.3)	3 (2.9)	55.7	37.1	−13.7 (−13.9)
Text queries	3 (3.3)	2 (2.4)	66.5	28.6	−15.3 (−15.6)

The four leftmost properties are proxies of syntactic complexity. The fifth property (parser score) is a proxy of parsing Difficulty. The $root \rightarrow NN*$ and $root \rightarrow VB*$ columns present the fraction of edges from the root of the parse tree that go to words POS-tagged as nouns and verbs, respectively.

Table 11. Part-of-Speech Distribution

	Text	Voice	Titles	WSJ
% Nouns (NN)	64.3	52.4	30.6	34.5
% Adjectives (JJ)	9.9	9.6	6.8	8.0
% Verbs (VB)	8.7	12.1	21.6	16.2
% Prepositions (IN)	5.5	7.6	8.9	11.8
% Determiners (DT)	2.0	4.5	7.5	9.8
% Pronouns (PR)	1.7	3.6	9.9	3.4
% Adverbs (RB)	2.2	3.5	6.4	4.6

The second column depicts the median and mean dependency tree depth, defined as the number of edges in the longest path from the root node to a leaf in the tree. The third and fourth columns present the fraction of dependency tree root edges that go to tokens POS-tagged as nouns or verbs, respectively. We use these two measures as proxies to the syntactic category of the input text, with noun roots often indicating simple noun phrases and verb roots often indicating more complex syntactic forms (Pinter et al. 2016). Finally, the rightmost column of Table 10 presents the median and average length-normalized log probability score of the PCFG parse, which serves as a proxy for grammaticality (a more negative score reflects a lower probability of the parse).

These results indicate substantial differences between voice and text queries: voice queries have more tokens, higher tree depth, higher portion of root nodes that govern a verb, lower portion of root nodes that govern a noun, and higher parse score. All of these differences make voice queries more similar to question titles (which are, in turn, closer to news articles), relative to text queries. This analysis suggests that on the scale where text queries are at one extreme (shorter, less grammatical) and natural-language news articles are at the other (longer, better-formed), voice queries are positioned somewhere in-between text queries and question titles.

7.2 POS Tags

The second part of the query syntax evaluation focuses on the distribution of part-of-speech tags using the Stanford POS tagger[8] (Toutanova et al. 2003). In our analysis, we removed all punctuation tokens. Mobile queries in general include very few punctuation marks, as already mentioned in Section 5: only 0.7% of the text tokens and 0.3% of the voice tokens were punctuation marks, compared to 12.4% and 12.8% for question titles and WSJ, respectively (largely due to the use of question marks at the end of question titles and periods at the end of WSJ sentences). We worked with a lowercase version of all four corpora, as all the voice queries and the vast majority of text queries were lowercase in their original form.

Table 11 displays the portion of primary POS tags (as the portion out of all tokens in the corpus) for the four corpora. The first row refers to nouns, which are prevalent in queries, as previously

[8] http://nlp.stanford.edu/software/tagger.shtml.

Table 12. Part-of-Speech Distribution for Text (T) vs. Voice (V) by Query Length (Number of Tokens)

	2		3		4		5		6		7	
	T	V	T	V	T	V	T	V	T	V	T	V
% Nouns (NN)	80.6	73.6	72.1	65.2	66.6	60.1	60.5	54.5	53.9	50.1	47.7	45.1
% Adjectives (JJ)	9.5	10.8	12.5	12.6	11.5	10.9	10.7	10.2	10.0	9.5	9.3	9.1
% Verbs (VB)	5.8	7.4	6.3	8.3	7.5	9.8	8.7	11.2	10.5	12.4	12.8	13.5
% Prepositions (IN)	0.2	1.0	2.4	3.7	5.2	6.5	7.5	8.1	9.1	9.3	10.2	10.3
% Determiners (DT)	0.3	1.1	0.8	1.9	1.3	2.6	2.0	3.6	2.9	4.6	3.6	5.6
% Pronouns (PR)	0.5	1.3	0.6	1.8	1.1	2.5	1.6	3.0	2.1	3.5	3.2	4.2
% Adverbs (RB)	1.1	1.7	1.0	1.7	1.2	2.2	2.1	3.2	3.2	3.7	4.2	4.5

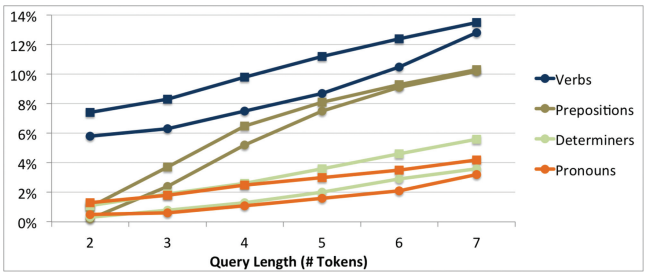

Fig. 5. POS distribution for voice (square markers) vs. text (round markers) queries.

shown by Barr et al. (2008). Yet, for voice queries the portion of nouns is substantially lower than for text queries (52.4% vs. 64.3%), although still considerably higher than for WSJ (34.5%) and question titles (30.6%). Adjectives are also somewhat more common for queries, with similar portions for text (9.9%) and voice (9.6%). For the other parts of speech, it can be seen that the portions for voice queries are higher than text queries, and closer to the portions for titles and WSJ, although not quite as high. These differences are consistent across all five lower POS types in the table, and especially salient, with more than a double ratio between voice and text, for determiners and pronouns (the latter are also especially common in question titles).

The richer language used in voice queries is largely due to their length: as we saw, voice queries are a token longer on average than text queries. Yet, differences also emerge when comparing queries of the same length. Table 12 shows the POS tag portions for voice versus text across queries of 2–7 tokens. There is a clear general trend in POS distribution by query length: the portion of nouns decreases as the number of tokens increases, the portion of adjectives remains stable, while the portion of other POS types increases with the length of the query. For queries with a fixed length, differences can still be observed between voice and text in the portion of nouns, on the one hand, which is higher for text queries, and verbs, prepositions, determiners, pronouns, and adverbs, which are higher for voice queries. These differences somewhat diminish as query length grows, yet such queries are quite rare, especially for text (e.g., only 2.7% of the text queries are of 7 tokens). The differences for determiners and pronouns remain solid even for queries of 7 tokens. Overall, we see that even for queries of the same length, there is a difference in POS distribution between voice and text queries, with more diverse language used in voice.

Figure 5 visualizes these results for four POS types—verbs, prepositions, determiners, and pronouns—to further demonstrate the differences between voice and text queries for fixed query lengths.

8 SIMILARITY TO NATURAL LANGUAGE

Our analysis so far has revealed semantic and syntactic differences between voice and text queries. In this section, we set out to validate that the language of voice queries is indeed closer than the language of text queries to natural language. To this end, we built two natural language models (LMs). The first is based on the WSJ corpus (sections 2–23, 42,248 sentences). The second, marked QT, was built based on a random sample of 50M question titles from the Yahoo Answers CQA site, posted between 2006 and 2015. Yahoo answers is a large and diverse website, acting not only as a medium for sharing technical knowledge, but as a place where one can seek advice, gather opinions, and satisfy curiosity about a wide variety of topics (Adamic et al. 2008). We opted to use these two corpora since one (WSJ) represents classic formal language, commonly used in natural-language processing research, while the other (QT) represents a more up-to-date web language contributed by the "crowd" in order to ask questions.

8.1 Lexical Language Model

We built unigram, bigram, and trigram LMs, with Jelinek-Mercer smoothing (Zhai and Lafferty 2001), $\lambda = 0.8$, as learned throughout our experiments.[9] Smoothing unknown unigrams was done using the standard ϵ of $1/|v|$, where $|v|$ is the size of the vocabulary.

For measuring the similarity to a natural language model we used perplexity, which is perhaps the most commonly used measure of model quality in speech, natural-language processing, and information retrieval research (Rosenfield 2000). Perplexity quantifies the error between the predicted probability of an event proposed by a language model, compared to the empirical probability of the event. In our case, we used perplexity to measure the quality of a natural-language model (WSJ or QT) with regard to a corpus of queries (i.e., the events) from either voice or text. In other words, we measured how likely the set of voice versus text queries is to originate from the given language model.

More formally, given a language model LM and a set of observed probabilities P, the perplexity of LM is defined as $2^{H(P;LM)}$ where $H(P;LM)$ is the cross entropy of the probability model LM with respect to the observed probabilities P, summed over all events in P. The closer the estimated probabilities for each event to the actual probabilities, the lower the perplexity. In our case, each event is a query q in a corpus Q (either voice or text), with an observed probability $1/|Q|$. Therefore, the cross entropy can be calculated as

$$H(Q;LM) = -\sum_{q \in Q} P(q) \cdot \log_2 LM(q) = -\frac{1}{|Q|} \sum_{q \in Q} \log_2 LM(q),$$

where $\log_2 LM(q)$ is the length-normalized log probability of the query q based on the language model LM. The perplexity itself is calculated by using the cross entropy as the exponent—the lower its value, the better is the language model for "predicting" the generation of the given query corpus.

The first two rows of Table 13 show the perplexity results for the WSJ and QT LMs. We refer to it as *lexical perplexity*, since the corresponding LMs consider the lexical words of the text. Later in this section, we will also discuss LMs that consider part-of-speech tags. It can be seen that the perplexity for voice queries is considerably lower than for text queries, for unigrams, bigrams, and trigrams. The "V/T" columns explicitly show the proportion between the two. It can be seen that the ratio decreases, that is, the gap between voice and text queries grows, when moving from unigrams to bigrams and then to trigrams, indicating that the difference is largely due to the structure of the voice query language, rather than merely due to its vocabulary.

[9] In our experiments, we worked with $\lambda = 0.5, 0.6, \ldots, 0.9$. We only report the results with $\lambda = 0.8$ for clarity of presentation, but note that the respective outcomes for other values of λ were very similar.

Table 13. Lexical Perplexity of Text (T) and Voice (V) Queries w.r.t. Natural Language Models: Wall Street Journal (WSJ) and Question Titles (QT)

	Unigrams			Bigrams			Trigrams		
Corpus	Text	Voice	V/T	Text	Voice	V/T	Text	Voice	V/T
WSJ	30,463	13,563	0.445	11.9*M*	792,952	0.066	17.3*B*	133*M*	0.008
QT	35,843	12,865	0.359	70,064	11,764	0.168	402,837	21,127	0.067
QT 42*K*	12,667	6,665	0.526	466,428	83,924	0.18	49.5*M*	2.89*M*	0.058

The "QT 42K" line represents a question title corpus with an identical number of sentences to WSJ.

Table 14. Lexical Perplexity by Query Length for Text (T) and Voice (V)

		Unigrams			Bigrams			Trigrams		
	Length	T	V	V/T	T	V	V/T	T	V	V/T
WSJ	3	33.1*K*	17.4*K*	0.53	1.5*M*	519*K*	0.35	139*M*	33*M*	0.24
	5	16.2*K*	9.2*K*	0.57	169*K*	63*K*	0.37	4*M*	1*M*	0.26
	7	8.4*K*	5.4*K*	0.65	42.9*K*	19.1*K*	0.45	544*K*	181*K*	0.33
QT	3	43.4*K*	19.1*K*	0.44	40.8*K*	15.8*K*	0.39	112*K*	32.4*K*	0.29
	5	17.3*K*	9.5*K*	0.55	7.2*K*	3.7*K*	0.51	12.5*K*	5.5*K*	0.44
	7	6.7*K*	4.6*K*	0.69	2.3*K*	1.4*K*	0.6	3*K*	1.6*K*	0.55

The general perplexity values for bigrams and trigrams are substantially higher for WSJ than for QT. This can be either due to the WSJ language being less similar to query language than QT, or due to the large volume of training data for QT, which produced a better language model. To further explore this, we randomly sampled 42,248 question titles from the 50M QT corpus—identical to the number of sentences in the WSJ corpus—and trained unigram, bigram, and trigram LMs based on this smaller corpus. The bottom row of Table 13 shows the respective results. It can be seen that the bigram and trigram perplexity values are higher than for the massively trained QT model, but are still lower by roughly an order of magnitude for bigrams and two orders of magnitude for trigrams compared to the WSJ LM.[10] This gives a stronger indication that the queries are more likely to be generated from the QT LM than the WSJ LM. The differences between voice and text queries remain—the perplexity ratios are similar to the full QT LM.

While being longer is an inherent characteristic of voice queries, we set out to explore whether for queries of the same length, there are still differences in the perplexity between text and voice queries. Table 14 shows the results for both WSJ and QT, as measured for text and voice queries of length 3, 5, and 7 tokens. Generally, the perplexity values indeed decrease as query length increases, indicating, as conjectured, that longer queries are closer to natural language. For both WSJ and QT, the perplexity across all *n*-grams is still noticeably lower for voice versus text queries of the same length. While the ratio is not as sharp as for the general query population, the difference is still clear and the trend of decreasing ratio from unigrams to bigrams and from bigrams to trigrams persists. This indicates that when we take a voice query and a text query of the same length, the former will still have a language closer to natural. We also see that as query length grows, the

[10]The unigram perplexity is substantially lower for the smaller training set. We suspect that the large training set adds many rare unigrams to the vocabulary that do not at all appear in the query sets, but reduce the probability of other unigrams that do appear in the queries.

Table 15. POS Perplexity for Text (T) and Voice (V) Queries

	Bigrams			Trigrams			4-grams			5-grams		
Corpus	Text	Voice	V/T	Text	Vice	V/T	Text	Voice	V/T	Text	Voice	V/T
WSJ	32.5	29.3	0.9	57.1	48.8	0.8	71.9	57.1	0.79	114.1	90.2	0.79
QT	17.4	16.6	0.96	23.6	19.3	0.82	31.4	23.2	0.74	39.0	26.8	0.69
QT 42K	17.4	16.6	0.96	23.5	19.4	0.82	31.9	24.4	0.77	40.2	30.5	0.76

Table 16. POS Perplexity by Query Length for Text (T) and Voice (V) Queries

		Bigrams			Trigrams			4-grams			5-grams		
	Length	T	V	V/T	T	V	V/T	T	V	V/T	T	V	V/T
WSJ	3	25.7	25.4	0.99	28.6	28.7	1	31.3	32.5	1.04	49.2	51.7	1.05
	5	19.6	20.1	1.02	21.6	21.8	1.01	27.2	27.8	1.02	45.0	46.0	1.02
	7	18.5	18.5	1	19.9	19.7	0.99	26	25.4	0.98	45.7	43.3	0.95
QT	3	15.2	15.2	1	15.6	14.5	0.93	16.6	14.1	0.85	21.4	16.4	0.77
	5	13.3	12.8	0.96	13.6	12.2	0.9	15	12.6	0.84	17.5	13.5	0.77
	7	12.8	12.2	0.95	12.5	11.4	0.91	13.5	11.8	0.88	15.2	12.8	0.84

perplexity gap between text and voice query is somewhat shrinking, but it is still evident and the trend of decreasing ratio from unigrams to bigrams and from bigrams to trigrams holds.

8.2 Part-of-Speech Language Model

As already mentioned, we also examined language models based on the POS tags rather than the lexical words. To this end, we trained models based on the WSJ and QT corpora, where each token was replaced by its corresponding POS tag. We then calculated the perplexity as before, where each query was represented by a sequence of POS tags corresponding to its tokens. Since the vocabulary is much smaller in this case (includes 36 POS tags in total), we also examined n-grams with higher values of n, specifically, bigram, trigram, 4-gram, and 5-gram LMs. Table 15 presents the results of this *POS perplexity*. It can be seen that the perplexity for the voice POS-tag language is lower than for text queries, both for WSJ and for QT, and the ratio decreases as n grows, reflecting a closer structure of voice queries to a natural POS-tag language. This trend also persists for the smaller QT corpus that contains a similar number of sentences to the WSJ corpus.

Table 16 shows the POS perplexity results for queries of length 3, 5, and 7 tokens. In this case, it can be seen that the POS perplexity for voice and text queries is similar w.r.t. the WSJ corpus, indicating that the POS language for queries of fixed length resembles that of WSJ for both voice and text. For the QT corpus, however, the ratio is smaller than 1, indicating a stronger POS-tag language similarity of the voice queries than the text queries, even for fixed lengths. The ratio further decreases as n grows.

8.3 Lexical and Part-of-Speech Language Model

We finally set out to examine the combination of the lexical words and POS tags. This enabled us to inspect a finer-grain level of the language, where a word is characterized by its syntactic role in addition to its text (e.g., the word "address" will have representations both as a noun and as a verb). To this end, we trained models based on the WSJ and QT corpora, with each token represented by the concatenation of its corresponding POS tag to its text (e.g., "addressNN"). We then calculated the perplexity as before, where each query was represented by the sequence of its tokens concatenated to their POS tags. Due to the richer vocabulary in this case, we only

Table 17. Lexical+POS Perplexity for Text (T) and Voice (V) Queries w.r.t. Question Titles

	Unigrams			Bigrams			Trigrams		
Length	T	V	V/T	T	V	V/T	T	V	V/T
all	95,264	26,545	0.28	713,719	41,930	0.06	22.3*M*	214,018	0.01
3	91,766	37,849	0.41	106,057	41,137	0.39	471,854	138,633	0.29
5	37,216	18,080	0.486	17,563	7,625	0.434	43,636	15,859	0.363
7	13,016	7,857	0.604	5,085	2,559	0.503	9,064	3,961	0.437

The "all" row presents the results across all queries, regardless of their length, while the other rows present the results for specific query lengths.

report the results over the larger QT corpus, which contains 50M sentences. As can be seen in Table 17, the perplexity values generally increase compared to the lexical perplexity. Yet, the ratio between the perplexity for voice queries and text queries is even lower, and further decreases from unigrams to bigrams and from bigrams to trigrams, reflecting again the closeness of voice queries to natural language, compared to text queries. The lower ratio of the lexical+POS perplexity compared to lexical perplexity persists when inspecting fixed query lengths: 3, 5, and 7 tokens.

9 DISCUSSION AND IMPLICATIONS

Our study disclosed various differences between voice and text search. In this section, we summarize the key findings, discuss their implications, with regard to both the information retrieval process and the user experience, and suggest directions for future work.

Query Categories. Both our query semantics and click analysis revealed that voice queries are more focused on audio-video content, such as from music channels or video sharing websites. It seems that voice search is more often used when the result is also expected to include voice. In addition, we saw that higher portions of the voice queries triggered the "direct answer" card and yielded clicks on popular CQA sites. This may be a result of the fact that higher portions of the voice queries were phrased as questions: White et al. (2015) found that such "question queries" typically have informational intent and often result in visits to CQA sites. Such queries have also been identified as a primary type of verbose queries, typically using rich language to express a narrow information need (Gupta and Bendersky 2015). We also found evidence for higher portions of recipe-related queries and clicks, implying that voice search is used while cooking. On the other hand, lower portions of voice queries referred to social networking and adult sites, which may represent more sensitive or personal content (Moorthy and Vu 2015). These noticeable differences in query categories suggest that search services that build on query classification, such as vertical selection, card triggering, ad targeting, query expansion, and even result ranking, may need to be adapted when used for voice search. For example, as voice queries are often phrased as questions, the identification of CQA queries (e.g., for presenting a CQA vertical on the SERP) may need to change (Tsur et al. 2016).

While we observed a gap along all hours of the day for adult search, we saw that people do search for adult sites using voice queries, even though it requires them to explicitly pronounce a site's name and/or relevant content terms. In addition, voice queries provide a clear evidence for the unique identity of the searcher. While these are naturally entry barriers, reflected by the lower ratio of adult queries, they do not completely hold back users from searching for adult content by voice.

Device Interaction. Voice search tends to focus on topics that require less interaction with the device's touchscreen. This was reflected by a substantially lower number of clicks, higher portions

of queries that triggered cards , and more queries that expressed a narrow information need (time, dictionary, weather, or phone numbers). We also saw this trend on celebrity queries, where the more open-style queries that only include the person's name, were relatively less common than queries that refer to a refined aspect, such as age or spouse. On the other hand, queries for research topics (e.g., health) and, as already mentioned, social network sites, which require a higher level of engagement and interaction (Guy et al. 2015, 2016), were less frequent on voice. We also saw that question queries that start with "why" were the least frequent on voice relative to text (Verberne 2007).

These results have two key implications. First, they suggest that voice search should enable voice-based result presentation, to support a complete hand-free interaction with the user. While short voice answers have started to emerge on commercial web search engines, we believe these capabilities should be further extended, to support interaction with more result types, exploration of different search results, query suggestion , and ultimately a complete dialogue with the user, as already done by intelligent assistants for personal advice (Jiang et al. 2015; Kiseleva et al. 2016a). Another idea from the world of personal assistants is supporting full activation by voice, for example, "Ok Google" or "Hey Cortana," to spare an extra click on the voice search button (Jiang et al. 2015).

The second implication relates to the IR evaluation process. Recent studies have argued that with modern search, especially on mobile devices, the merit of clicks as a primary evaluation measure decreases (Lagun et al. 2014). Other measurements, such as "good abandonment," have been proposed (Li et al. 2009; Williams et al. 2016). Our findings show that voice search clicks, as a form of interaction, are even rarer than text search clicks on mobile devices. Thus, new metrics for evaluating user satisfaction of voice queries should be developed.

Search on the Go. We found various signals that voice queries are more common on the go than text queries. These include higher portions of queries and clicks that relate to maps and phone numbers; higher portions of queries that refer to physical rather than online shopping brands; more frequent triggering of movies than TV cards; higher portion of the searches conducted during day hours (8am to 8pm); and, last but not least, lower portion of searches from a zip code identical to the user's home zip code. This suggests that contextualization techniques may help benefit voice search and help satisfy search needs when users are away from their home. Future research should look into more advanced models for detecting search on the go (Kravi et al. 2015) and examine more closely the tie between voice search and the user's location.

Pronouncing vs. Writing. In our analysis, a variety of issues were observed that relate to the difference between speaking queries and typing them. We saw more frequent use of words that are easy to pronounce but hard to write (e.g., long state names), and on the other hand, less frequent use of abbreviations or calendar years (2015), which are easy to type but harder to pronounce. Related to this is the absence of a copy-paste feature, reflected by the rare use of URLs in voice queries. Voice alleviates typing issues, such as correctly spelling named entities, but may present other challenges, such as pronunciation. Assuming the latter are less frequent (future work might want to prove this), named entities can sometimes be searched more easily by voice. In addition, some typing styles are "standardized" by the transcription process, for example, the use of an apostrophe for possession (on text, "s" is commonly used both with and without the preceding apostrophe) or the use of diacritical marks, such as in *beyoncé*. Finally, the use of punctuation marks and uppercase letters, which is infrequent on mobile search, is even rarer on voice search.

All of these differentiating properties may influence users' choice of a voice versus a text search system, as well as the retrieval process itself. For example, it is harder to query about an error message just received on the screen by voice, but easier to query about a person or place whose

name were just heard over the phone. Also, the ranking algorithm for voice queries may consider the "phonetic difficulty" of a word in order to weight it.

Voice Query Language. Our syntactic analysis, based on parsing and POS tagging, showed that voice queries are not only longer, by a token on average, than text queries, but also use richer language. Our semantic analysis demonstrated this with the use of natural language phrases such as "i'm looking for," "take me to," and "please." The perplexity-based analysis showed that the language of voice queries is indeed closer to natural language, even when controlling for query length. While it has long been claimed that voice is a richer, more expressive media than written text (Chalfonte et al. 1991; Crestani and Du 2006), our study demonstrates it for the first time over web search queries.

Having said that, the language of voice queries was found to still be far from natural-language questions and even farther from news articles. We also saw that voice queries may often be as short and identically phrased as text queries. These findings suggest that voice queries pose their own type of language, in-between traditional text queries and natural-language questions.

One question that stands, which we did not explore in this work, is how the length and richer language of voice queries can help improve the search process. Previous studies found that longer queries, closer to natural language, do not necessarily improve the retrieval effectiveness compared to typical keyword-based queries (Crestani and Du 2006; White et al. 2015). On the other hand, taking advantage of the general growth of query length on web search, recent studies proposed various methods for applying linguistic analysis on long queries, including part-of-speech tagging, dependency parsing, and entity and relation extraction, in order to enhance search performance. Linguistic analysis can also be applied on the document side and matched against the query's analysis, to resolve ambiguity and further enhance query-document match calculation (Carmel et al. 2014). A recent book on "verbose" queries (five words or more; 34.5% of all voice queries) summarizes this body of research and explains that for such queries, not only is it more feasible to apply linguistic analysis, but it is also essential for the understanding of the specific intent and the relevance of the returned results (Gupta and Bendersky 2015).

Linguistic means are essential because in verbose queries there is more noise, such as extraneous terms that users believe are important to conveying their information needs, but in fact are confusing to automatic systems (Gupta and Bendersky 2015). The performance of textbook information retrieval techniques for verbose queries is not as good as that for their shorter counterparts and search engines mostly focus on short queries as these account for the most popular (head) searches, while long queries constitute the long tail. Technologies such as query parsing or entity extraction are often avoided since search engines focus on head queries and as these technologies are still considered computationally expensive to apply at runtime. We have seen that the number of unique queries for voice is lower than for text, but we expect this to change over time, as voice search evolves. And since voice queries are longer in nature, the number of head queries is likely to reduce, while the long tail is destined to grow.

Previous work has tried to automatically transform queries into questions, by adding missing functional words or using question templates (Dror et al. 2013; Lin 2008), motivated by a variety of reasons, such as helping with intent disambiguation, improving search over CQA archives , enhancing query expansion, and allowing to post a query directly as a question on a CQA site when the answer cannot be found. Since voice search queries are more often phrased as questions, they may directly enable these benefits.

Reflections on Conversational Search. Conversational search is a new paradigm that has recently emerged alongside the rise of intelligent personal assistants (Jiang et al. 2015; Sano et al. 2016), chat bots (Yan et al. 2016), and similar technologies (Burtsev et al. 2017). As opposed to traditional

search, where the user inputs a query, receives a page with results (SERP), and may either interact with these results or leave the page, in conversational search the system conducts, typically using a voice interface, a dialogue with the user, in order to refine their information needs and reach the desired outcome (Radlinski and Craswell 2017; Shiga et al. 2017). This paradigm shift may lead to a radical change in future search and information access systems (Kiseleva et al. 2016b).

In this work, we focused on traditional web search when comparing voice and text queries. We did not perform any analysis of conversational search. Yet, our work's tie to conversational search is twofold. First, our analysis reveals a broad set of differences in the way users phrase their queries verbally, even when the search paradigm remains identical. This is an important first step toward understanding how voice-based conversational search systems should satisfy user needs (Kiseleva et al. 2016a, 2016b). In addition, as already mentioned, our results indicate that users of voice search currently refrain, to a certain extent, from searches that require more interaction. Conversational search may fundamentally change this by supporting a natural-language voice-based interaction between the search system and the user.

10 CONCLUSION AND FUTURE DIRECTIONS

This work presents a broad study of voice search as reflected in the log of a commercial web search engine. Voice queries are compared to text queries by a variety of aspects and are shown to pose many unique properties. Prominently, the language of voice queries is different than the text query language and is closer to natural language. In addition, the topics within the focus of voice search spread differently than in text search, and user behavior, such as time of use and clicks, are also different. All of these differences need to be accounted for as new search systems that desire to place more emphasis on voice interfaces are developed.

There is plenty of room (and need) for further research. Voice queries contain phonetic characteristics that are not included in text queries, such the speaker's speed, intonation, and stress, which can help enhance the retrieval process and personalization. Also, our research settings in this work is based on a large-scale log analysis. Conducting qualitative user studies can complement our findings and help gain in-depth understanding of the voice search process.

Our analysis does not include user-segmented data, as we were only allowed to expose data aggregated over the entire set of users in our query log sample. In their lab study, Crestani et al. (Crestani and Du 2006) found that some users phrased their voice queries very similarly to their text queries, while others phrased them entirely differently. Future research should further characterize the use of voice search by different types of users.

Our analysis suggests that voice search is still in its early stages, as reflected by the smaller portion of unique queries and the lower diversity of clicked domains. Future work should follow the dynamics of voice search as it is poised to become ubiquitous and further evolve.

REFERENCES

A. Acero, N. Bernstein, R. Chambers, Y. C. Ju, X. Li, J. Odell, P. Nguyen, O. Scholz, and G. Zweig. 2008. Live search for mobile: Web services by voice on the cellphone. In *Proc. ICASSP*. 5256–5259.

Lada A. Adamic, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. 2008. Knowledge sharing and Yahoo answers: Everyone knows something. In *Proc. WWW*. 665–674.

Ahmed Hassan Awadallah, Ranjitha Gurunath Kulkarni, Umut Ozertem, and Rosie Jones. 2015. Characterizing and predicting voice query reformulation. In *Proc. CIKM*. 543–552.

Ricardo Baeza-Yates, Georges Dupret, and Javier Velasco. 2007. A study of mobile search queries in Japan. In *Query Log Analysis (WWW'07 Workshop)*.

Cory Barr, Rosie Jones, and Moira Regelson. 2008. The linguistic structure of English web-search queries. In *Proc. EMNLP*. 1021–1030.

Adam Berger and John Lafferty. 1999. Information retrieval as statistical translation. In *Proc. SIGIR*. 222–229.

Mikhail Burtsev, Aleksandr Chuklin, Julia Kiseleva, and Alexey Borisov. 2017. Search-oriented conversational AI (SCAI). In *Proc. ICTIR*. 333–334.

David Carmel, Avihai Mejer, Yuval Pinter, and Idan Szpektor. 2014. Improving term weighting for community question answering search using syntactic analysis. In *Proc. CIKM*. 351–360.

David Carmel, Erel Uziel, Ido Guy, Yosi Mass, and Haggai Roitman. 2012. Folksonomy-based term extraction for word cloud generation. *ACM Transactions on Intelligent Systems and Technology* 3, 4 (2012), 60:1–60:20.

Deepayan Chakrabarti, Ravi Kumar, and Kunal Punera. 2009. Quicklink selection for navigational query results. In *Proc. WWW*. 391–400.

Barbara L. Chalfonte, Robert S. Fish, and Robert E. Kraut. 1991. Expressive richness: A comparison of speech and text as media for revision. In *Proc. CHI*. 21–26.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Proc. ICASSP*. IEEE, 4960–4964.

Ciprian Chelba and Johan Schalkwyk. 2013. *Empirical Exploration of Language Modeling for the google.com Query Stream as Applied to Mobile Voice Search*. Springer Science+Business Media, New York. 197–229.

Ciprian Chelba, Xuedong Zhang, and Keith Hall. 2015. Geo-location for voice search language modeling. In *Proc. INTERSPEECH*, 1438–1442.

Lydia B. Chilton and Jaime Teevan. 2011. Addressing people's information needs directly in a web search result page. In *Proc. WWW*. 27–36.

Fabio Crestani and Heather Du. 2006. Written versus spoken queries: A qualitative and quantitative comparative analysis. *JASIST* 57, 7 (2006), 881–890.

Marie-Catherine De Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proc. LREC*. 449–454.

Gideon Dror, Yoelle Maarek, Avihai Mejer, and Idan Szpektor. 2013. From query to question in one click: Suggesting synthetic questions to searchers. In *Proc. WWW*. 391–402.

Manish Gupta and Michael Bendersky. 2015. Information retrieval with verbose queries. *Foundations and Trends in Information Retrieval* 9, 3–4 (2015), 209–354.

Ido Guy, Roy Levin, Tal Daniel, and Ella Bolshinsky. 2015. Islands in the stream: A study of item recommendation within an enterprise social stream. In *Proc. of SIGIR*. 665–674.

Ido Guy, Inbal Ronen, Naama Zwerdling, Irena Zuyev-Grabovitch, and Michal Jacovi. 2016. What is your organization 'like'?: A study of liking activity in the enterprise. In *Proc. of CHI*. 3025–3037.

G. Hinton, Li Deng, Dong Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine* 29, 6 (2012), 82–97.

Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. 2015. Automatic online evaluation of intelligent assistants. In *Proc. WWW*. 506–516.

Jiepu Jiang, Wei Jeng, and Daqing He. 2013. How do users respond to voice input errors? Lexical and phonetic query reformulation in voice search. In *Proc. SIGIR*. 143–152.

Maryam Kamvar and Shumeet Baluja. 2006. A large scale study of wireless search behavior: Google mobile search. In *Proc. CHI*. 701–709.

Maryam Kamvar, Melanie Kellar, Rajan Patel, and Ya Xu. 2009. Computers and iphones and mobile phones, oh my!: A logs-based comparison of search users on different devices. In *Proc. WWW*. 801–810.

Julia Kiseleva, Kyle Williams, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016a. Predicting user satisfaction with intelligent assistants. In *Proc. SIGIR*. 45–54.

Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016b. Understanding user satisfaction with intelligent assistants. In *Proc. CHIIR*. 121–130.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proc. ACL*. 423–430.

Elad Kravi, Eugene Agichtein, Ido Guy, Yaron Kanza, Avihai Mejer, and Dan Pelleg. 2015. Searcher in a strange land: Understanding web search from familiar and unfamiliar locations. In *Proc. SIGIR*. 855–858.

Elad Kravi, Ido Guy, Avihai Mejer, David Carmel, Yoelle Maarek, Dan Pelleg, and Gilad Tsur. 2016. One query, many clicks: Analysis of queries with multiple clicks by the same user. In *Proc. CIKM*. 1423–1432.

Dmitry Lagun, Chih-Hung Hsieh, Dale Webster, and Vidhya Navalpakkam. 2014. Towards better measurement of attention and satisfaction in mobile search. In *Proc. SIGIR (SIGIR'14)*. 113–122.

Rivka Levitan and David Elson. 2014. Detecting retries of voice search queries. In *Proc. ACL*. 230–235.

Jane Li, Scott Huffman, and Akihito Tokuda. 2009. Good abandonment in mobile and PC internet search. In *Proc. SIGIR*. 43–50.

Chin-Yew Lin. 2008. Automatic question generation from queries. In *Proc. Workshop on the Question Generation Shared Task*. 156–164.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The penn treebank. *Computational Linguistics* 19, 2 (1993), 313–330.

George D. Montanez, Ryen W. White, and Xiao Huang. 2014. Cross-device search. In *Proc. CIKM*. 1669–1678.

Aarthi Easwara Moorthy and Kim-Phuong L. Vu. 2015. Privacy concerns for use of voice activated personal assistant in the public space. *International Journal of Human–Computer Interaction* 31, 4 (2015), 307–335.

A. Moreno-Daniel, S. Parthasarathy, B. H. Juang, and J. G. Wilpon. 2007. Spoken query processing for information retrieval. In *Proc. ICASSP*, Vol. 4. IV-121–IV-124.

Yuval Pinter, Roi Reichart, and Idan Szpektor. 2016. Syntactic parsing of web queries with question intent: A distant supervision approach. In *Proc. NAACL*. 670–680.

Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proc. 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR'17)*. 117–126.

Roni Rosenfield. 2000. Two decades of statistical language modeling: Where do we go from here? *Proceedings of the IEEE* 88, 8 (2000), 1270–1278.

Shumpei Sano, Nobuhiro Kaji, and Manabu Sassano. 2016. Prediction of prospective user engagement with intelligent assistants. In *Proc. ACL*. 1203–1212.

Johan Schalkwyk, Doug Beeferman, Françoise Beaufays, Bill Byrne, Ciprian Chelba, Mike Cohen, Maryam Kamvar, and Brian Strope. 2010. Your word is my command: Google search by voice: A case study. In *Advances in Speech Recognition*, Amy Neustein (Ed.). Springer US, 61–90.

Jiulong Shan, Genqing Wu, Zhihong Hu, Xiliu Tang, Martin Jansche, and Pedro J. Moreno. 2010. Search by voice in Mandarin Chinese. In *Proc. INTERSPEECH*. 354–357.

Sosuke Shiga, Hideo Joho, Roi Blanco, Johanne R. Trippas, and Mark Sanderson. 2017. Modelling information needs in collaborative search conversations. In *Proc. SIGIR*. 715–724.

Milad Shokouhi and Qi Guo. 2015. From queries to cards: Re-ranking proactive card recommendations based on reactive search history. In *Proc. SIGIR*. 695–704.

Milad Shokouhi, Rosie Jones, Umut Ozertem, Karthik Raghunathan, and Fernando Diaz. 2014. Mobile query reformulations. In *Proc. SIGIR*. 1011–1014.

Milad Shokouhi, Umut Ozertem, and Nick Craswell. 2016. Did you say u2 or youtube?: Inferring implicit transcripts from voice search logs. In *Proc. WWW*. 1215–1224.

Yang Song, Hao Ma, Hongning Wang, and Kuansan Wang. 2013. Exploring and exploiting user search behavior on mobile and tablet devices to improve search relevance. In *Proc. WWW*. 1201–1212.

Jaime Teevan, Daniel Ramage, and Merredith Ringel Morris. 2011. #TwitterSearch: A comparison of microblog search and web search. In *Proc. WSDM*. 35–44.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. NAACL*. 173–180.

Gilad Tsur, Yuval Pinter, Idan Szpektor, and David Carmel. 2016. Identifying web queries with question intent. In *Proc. WWW*. 783–793.

Suzan Verberne. 2007. Paragraph retrieval for why-question answering. In *Proc. SIGIR*. 922.

Ye-Yi Wang, Dong Yu, Yun-Cheng Ju, and A. Acero. 2008. An introduction to voice search. *IEEE Signal Processing Magazine,* 25, 3 (2008), 28–38.

Ryen W. White, Matthew Richardson, and Wen-tau Yih. 2015. Questions vs. queries in informational search tasks. In *Proc. WWW*. 135–136.

Kyle Williams, Julia Kiseleva, Aidan C. Crook, Imed Zitouni, Ahmed Hassan Awadallah, and Madian Khabsa. 2016. Detecting good abandonment in mobile search. In *Proc. WWW*. 495–505.

Zhao Yan, Nan Duan, Jun-Wei Bao, Peng Chen, Ming Zhou, Zhoujun Li, and Jianshe Zhou. 2016. DocChat: An information retrieval approach for chatbot engines using unstructured documents.. In *Proc. ACL*. 516–525.

Jeonghe Yi and Farzin Maghoul. 2011. Mobile search pattern evolution: The trend and the impact of voice queries. In *Proc. WWW*. 165–166.

Jeonghee Yi, Farzin Maghoul, and Jan Pedersen. 2008. Deciphering mobile search patterns: A study of yahoo! mobile search queries. In *Proc. WWW*. 257–266.

Chengxiang Zhai and John Lafferty. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proc. SIGIR (SIGIR'01)*. 334–342.

Geoffrey Zweig and Shuangyu Chang. 2011. Personalizing model M for voice-search. In *Proc. INTERSPEECH*. 609–612.

Received April 2017; revised December 2017; accepted January 2018