Searching by Talking: Analysis of Voice Queries on Mobile Web Search

Ido Guy
Information Systems Engineering Department
Ben Gurion University of the Negev, Beer Sheva, Israel
Yahoo Research, Haifa, Israel
idoguy@acm.org

ABSTRACT

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 paper, we examine the logs of a commercial search engine's mobile interface, and compare the spoken queries to the typed-in queries. We place special emphasis on the semantic and syntactic characteristics of the two types of queries. We also conduct an empirical evaluation showing that the language of voice queries is closer to natural language than typed queries. Our analysis reveals further differences between voice and text search, which have implications for the design of future voice-enabled search tools.

Keywords: mobile search; query log analysis; voice search

1. INTRODUCTION

The popularity of search from mobile devices (*mobile search*) has rapidly increased in recent years [34]. 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 [32]. 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 [19]. Recent developments in speech recognition, backed by high bandwidth coverage and high-quality speech signal acquisition, are enabling higher quality voice search [8]. 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 [30]. Since then, further enhancements to automatic speech recognition (ASR) for web search have been reported [31, 43], taking advantage of the large data that started to accumulate on voice search logs [8], and applying advanced learning methods [17]. The use of voice has also been promoted by the increasing popularity of voice-activated intelligent assistants, such as Siri, Google Now, and Cortana. These assistants provide context-based query-less personalized advice for mobile users, but also enable web search [18]. A recent survey of 1400 U.S. smartphone users found that 55% of the teenagers use voice search every day [3, 14]. 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. 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 [14].

In spite of its growing popularity, the area of voice search has not received much attention in the IR 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 [10]. More recent work has mostly focused on voice recognition [1, 8, 27, 31, 38, 43] and query reformulation [3, 19, 33]. A few studies revealed more details about how voice search is performed on commercial search engines [30, 40], however we are not aware of a systematic log analysis of voice queries as of yet.

In this work, we perform a query log analysis of half a million voice queries, issued to the mobile application of a commercial web search engine, over a period of six 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.

SIGIR '16, July 17-21, 2016, Pisa, Italy
© 2016 ACM. ISBN 978-1-4503-4069-4/16/07…$15.00
DOI: http://dx.doi.org/10.1145/2911451.2911525

Our work has 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 common and different 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 paper 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 [20] compared search patterns on 12-key keypad cellphones, PDAs, 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. [4] 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. [41] 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. With the evolution of mobile devices into smartphones, mobile search has also been shown to change. Kamvar et al. [21] examined search behavior on iPhones and found it was more similar to desktop search than to search on basic mobile phones. Song et al. [34] performed a broad 3-month log analysis of Bing search on desktop, iPad, and iPhone. Due to the significant differences between user search patterns on the three platforms, they proposed a ranking system that considered platform-specific features.

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. [38] 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. [1] described the architecture of the speech recognition interface of "Live Search for Mobile". Moreno-Daniel et al. [27] discussed the interleaving of ASR with 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 [17], various studies suggested advanced methods for voice search ASR and reported further performance enhancements. Chelba et al. [8] leveraged the data on Google's voice search logs to enhance language modeling and achieved *"small but significant"* gains in speech recognition performance. Shan et al. [31] 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 [43] found that the use of Model M (exponential n-gram language model) with personalization features improved the speech recognition performance on Bing 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.

Some of the recent work 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 [19]. Classifiers were built to predict and categorize voice reformulations, extending text-based approaches with features such as voice recognition time and confidence [3]. Researchers also found that users do not tend to switch between voice and text when reformulating queries [33]. 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" [30]. 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 1- and 2-word queries, were relatively more frequent in voice searches than in typed searches, while longer queries (5+ words) were far rarer. A poster by Yi and Maghoul [40] 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 a decade ago (pre-smartphone era) [10]. 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). 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 versus spoken queries was evaluated. In our analysis, when relevant, we tie to the results reported in these three studies and discuss the common and different with our own 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 six 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 six 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 two 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 timezone in which it was performed), a location in the form of city and state, 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 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 towards 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 regards 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.

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

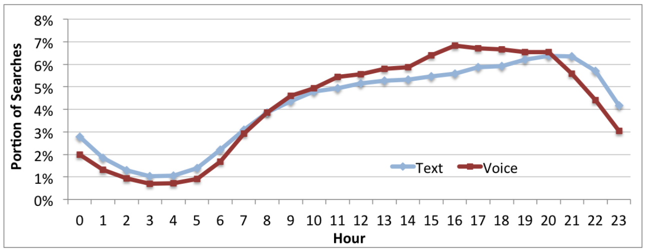

Figure 1: Query distribution by hour of the day.

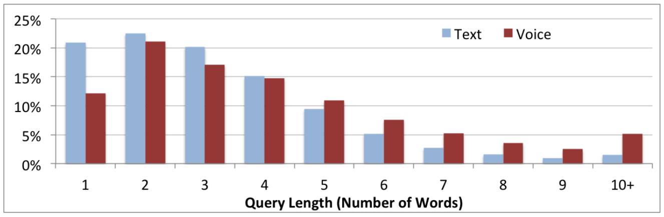

Figure 2: Query length distribution.

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 5 words, which have been recently referred to as "verbose" queries [15]. Overall, 34.5% of the voice queries were of 5 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 regards 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) [30], other studies found voice queries to be longer (e.g., 3.4 versus 2.2 on a Yahoo study) [10, 40]. Jiang et al. [19] pointed out this discrepancy and stated that *"further studies are needed to identify the characteristics of queries in voice search."*. 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).

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 [40]. Second, the use of abbreviations, spelling variants, and punctuation marks, makes text queries more diverse. We will further demonstrate this in Section 5.

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 [35], we considered queries that occur in a sequence without 15 minutes of inactivity as part of the same session.

	Text	Voice
% 1-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
% identical queries	13.7%	15.1%
% refining queries	10%	7.6%

Table 1: Session characteristics.

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 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.

5. QUERY SEMANTICS

In this section, we compare voice and text query semantics. We first examine higher level query categories by inspecting the triggering of vertical cards. We then examine the query themselves more closely, by comparing the most popular queries and the most distinctive query terms.

5.1 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 [32]. These cards extend organic search results ('ten blue links') by addressing a user's specific information need directly on the SERP [9], often sparing the need for a click [23]. Commercial search engines trigger cards upon the 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 reflect a more frequent use for information needs that can be directly satisfied on the SERP and require less interaction from the user.

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 appeared 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 or height.

Inspecting specific queries that refer to popular celebrities, we indeed observed a ratio smaller than 1 between voice and text, which becomes particularly low for queries that only include the celebrity's name, without any refinements. For example, the voice/text ratio for queries that included "lebron james" was 0.85, but this ratio decreases to 0.57 when only considering the exact query "lebron james". Similarly, for "bruce jenner" these ratios were 0.97 and 0.47, respectively; for "kim kardashian" 0.63 and 0.48; and for "donald trump", 0.83 and 0.42.

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 so horoscope were more popular on text. 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 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 [30], which identified "food & drink" and "local" as the more popular categories for voice (out of a total of 8), corresponding with our findings w.r.t the "recipe" and "maps" cards. That study also found the "online communities" and "adult" categories were less frequent for voice, to which we show support in the next section.

5.2 Popular Queries

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 it was issued by less than half (0.44). It can also be seen that adult site queries ("porn", "xnxx", "redtube") appear only on the top text list.

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

Table 2: Voice(V)/text(T) card triggering ratio.

Text		Voice	
1. facebook	7. porn	1. youtube	7. yahoo
2. youtube	8. xnxx	2. yahoo mail	8. walmart
3. pornhub	9. yahoo	3. facebook	9. amazon
4. google	10. amazon	4. google	10. home depot
5. yahoo mail	11. redtube	5. hello	11. yahoo.com
6. craiglist	12. facebook login	6. craiglist	12. amazon.com

Table 3: Most popular queries.

On the other hand, the voice list includes more retail brands ("walmart", "home depot") and the query "hello", which is likely used for experimenting with the voice system. In addition, the voice list includes the use of the suffix '.com' for popular navigational queries, e.g., 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 (ratio of 0.6), likely due to substantially sparser use of full URLs on voice queries: the prefixes 'www' and 'http' appeared much more commonly on text queries, with a voice/text ratio of 0.07 and 0.01, respectively.

5.3 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 an asymmetric distance measure between two given distributions [6]. 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[1]. 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 site names (especially social media and adult) and common abbreviations (states, e.g., "tx", "ca", "nc"; and also "st", "vs"; "dr"), 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", "phone number for"), and also a few state names that appear in their standard form ("new york", "north carolina").

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

Table 4: Most distinctive query terms.

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.

The difference in use of question words is one of the most prominent between voice and text queries. A recent study examined this form of "question queries", which *"take the form of natural language"* [39]. 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-word (one of the 5W1H), compared to only 3.7% of the text queries (ratio 2.67). Adding yes/no questions (start with 'does', 'did', 'can', etc.), the portions grow to 11.9% and 4.7%, respectively (ratio 2.55). The two most popular question words, by a large margin, were "how" (3.6% of all voice queries) and "what" (3.5%); while for "what" the voice/text ratio was 3.1, for "how" it was lower at 2.3. The lowest ratio among the 5W1H was for "why" queries at 1.8 (these queries account for 0.4% of all voice queries), probably as these are more open-ended questions, often characterized by longer answers that require more exploration on the part of the user [37]. On the other hand, common prefixes for factoid questions were substantially more common on voice, e.g., "how old is" (ratio 5.74) or "who is the" (5.36) [16].

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 one week (our text sample did not contain such matches, thus we had to inspect a larger log). Table 5 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 phrased to text queries. For instance,

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

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

Table 5: Example voice and text queries that landed on the same CQA page.

	Voice / Text Ratio
Click-through rate	0.78
Avg #clicks	0.83
MRR for all queries	0.78
MRR for clicked queries	0.97
% Unique domains (hosts)	0.71
% Top domain clicks	1.1

Table 6: Click statistics.

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

6. CLICKS

Table 6 shows the ratio for various click characteristics between voice and text queries[2]. It can be seen that the click-through rate (CTR; the portion of queries for which at least one click was made) 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. The mean reciprocal rank (MRR) across all queries is also substantially 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.

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.

Table 7 shows the top clicked domains for text and voice queries. The rightmost column shows, for each of the top voice domains, the "click ratio", i.e., 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, 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 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 (Pornhub, Xvideos, Xnxx, and Redtube) are on the top text domains, 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.

	Text	**Voice**	
	Domain	Domain	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

Table 7: Most clicked domains.

Further inspecting the lists of top clicked domains revealed differences towards voice queries in other CQA sites (e.g. wikiHow with a click ratio of 1.3), maps (MapQuest 1.75; appearances of "maps" within all clicked domain strings with a ratio of 1.74), weather (AccuWeather 2.22; "weather" string 1.48), dictionary (dictionary.reference.com 1.43; thefreedictionary.com 1.75, "dictionary" 1.35), recipes (allrecipes.com 1.58, "recipe" 1.53), video streaming (screen.yahoo.com 1.97; "screen" 1.77), and music (iTunes 1.88). On the other hand, more dominant on text queries were shopping sites (eBay 0.92, Craiglist 0.54, "shopping" 0.84), health (WebMD 0.88, drugs.com 0.77, nih.gov 0.64), news (news.yahoo.com 0.82, "news" 0.7), sports (sports.yahoo.com 0.72, espn.go.com 0.8, "sport" 0.83), finance (finance.yahoo.com 0.48, "finance" 0.56, "bank" 0.89), celebs (celebs.yahoo.com 0.63, "celeb" 0.81), and social network sites with a particularly low ratio between voice and text clicks (Twitter 0.17, Linkedin 0.37). Despite the differences in favor of voice clicks for audio and video results, there was no such difference for photo sites (Photobucket 0.97, Pinterest 0.65, "photo" 0.87).

7. QUERY SYNTAX

In Section 4.2, we saw 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) [26], which primarily includes business and financial news articles, broken into sentences (a total of 42,248 sen-

[2] We cannot disclose actual values due to business sensitivity.

tences). 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[3]. The parser first generates an unlexicalized PCFG parse [22] and then produces typed dependencies by matching patterns on CFG trees [11].

Table 8 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[4]. 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 [28]. Finally, the rightmost column of Table 8 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 Tagging

The second part of the query syntax evaluation focuses on the distribution of part-of-speech tags using the Stanford POS tagger[5] [36]. In our analysis, we removed all punctuation tokens. Mobile queries in general include very few punctuation marks: 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 lower-case version of all four corpora, as all the voice queries and the vast majority of text queries were lower case in their original form.

Table 9 displays the portion of primary POS tags (as the portion out of of all tokens in the corpus) for the four corpora. The first row refers to nouns, which are prevalent in queries, as previously shown [5]. 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 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 10 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 number 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.

8. NATURAL LANGUAGE RESEMBLANCE

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 to natural language than text queries. 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 [2]. 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.

We built unigram, bigram, and trigram LMs, with Jelinek-Mercer smoothing [42], λ=0.8, as learned throughout our experiments[6]. Smoothing unknown unigrams was done using the standard ϵ of 1 over the vocabulary size.

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 [29]. Perplexity quantifies the error between the predicted probability of

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

[4] 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.

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

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

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	4.3 (4)	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)

Table 8: Syntactic properties of four corpora.

	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

Table 9: Part-of-speech distribution.

	2		3		4		5		6		7	
	T	V	T	V	T	V	T	V	T	V	T	V
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
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
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
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
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
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
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

Table 10: Part-of-speech distribution for text (T) vs. voice (V) by query length (number of tokens).

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 regards 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 $\frac{1}{|Q|}$. 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 11 show the perplexity results for the WSJ and QT LMs. 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, which reflects a ratio between two probabilities. It can also be seen that the ratio decreases, i.e., 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.

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 lowest row of Table 11 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[7] . This gives a stronger indication that the queries are more likely to be generated from the QT LM than the WSJ. 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 12 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 ratio decrease 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 have a language closer to natural.

9. DISCUSSION AND IMPLICATIONS

Our study disclosed various differences between voice and text search. In this section, we summarize the key findings, discuss implications, 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 ques-

[7] Interestingly, 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.

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

Table 11: Perplexity of text (T) and voice (V) queries w.r.t natural language models.

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

Table 12: Perplexity by query length for text and voice.

tions: White et al. [39] found that such "question queries" typically have informational intent and often result in visits to CQA sites. We also found evidence for higher portions of recipe-related queries and clicks, perhaps 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 [13]. 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.

•**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, and 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 higher level of engagement, were less frequent on voice. 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 [18]. 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 [23]. Other measurements, such as "good abandonment", have been proposed [24]. 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.

•**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. In addition, some typing styles are "standardized" by the transcription process, e.g., the use of 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 upper case letters, which is infrequent on mobile search, is even rarer on voice search. Voice search systems need to take into account this set of unique qualities characterizing voice queries.

•**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 [7, 10], our study demonstrates it for the domain of 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 follows, 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 [10, 39]. 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. A recent book on "verbose" queries (5 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 to the understanding of the specific intent and the relevance of the returned results [15].

Previous work has tried to automatically transform queries into questions, by adding missing functional words or using question templates [12, 25], 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.

There is plenty of room (and need) for further research, such as exploring the use of voice queries' phonetic characteristics, e.g., the speaker's stress, speed, and intonation for search personalization, or conducting user studies to gain in-depth understanding of the voice search process. 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 research should follow the dynamics of voice search as it is poised to become ubiquitous and further evolve.

.

10. ACKNOWLEDGMENTS

My thanks to Alex Nus, Dan Pelleg, Yuval Pinter, and Gilad Tsur for their useful comments. Special thanks to Avihai Mejer for his useful advice throughout this research.

11. REFERENCES

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

[2] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge sharing and yahoo answers: Everyone knows something. In *Proc. WWW*, pages 665–674, 2008.

[3] A. H. Awadallah, R. Gurunath Kulkarni, U. Ozertem, and R. Jones. Characterizing and predicting voice query reformulation. In *Proc. CIKM*, pages 543–552, 2015.

[4] R. Baeza-Yates, G. Dupret, and J. Velasco. A study of mobile search queries in japan. In *Query Log Analysis (WWW workshop)*, 2007.

[5] C. Barr, R. Jones, and M. Regelson. The linguistic structure of english web-search queries. In *Proc. EMNLP*, pages 1021–1030, 2008.

[6] A. Berger and J. Lafferty. Information retrieval as statistical translation. In *Proc. SIGIR*, pages 222–229, 1999.

[7] B. L. Chalfonte, R. S. Fish, and R. E. Kraut. Expressive richness: A comparison of speech and text as media for revision. In *Proc. CHI*, pages 21–26, 1991.

[8] C. Chelba and J. Schalkwyk. Empirical exploration of language modeling for the google.com query stream as applied to mobile voice search. In *Mobile Speech and Advanced Natural Language Solutions*, pages 197–229. 2013.

[9] L. B. Chilton and J. Teevan. Addressing people's information needs directly in a web search result page. In *Proc. WWW*, pages 27–36, 2011.

[10] F. Crestani and H. Du. Written versus spoken queries: A qualitative and quantitative comparative analysis. *JASIST*, 57(7):881–890, 2006.

[11] M.-C. De Marneffe, B. MacCartney, and C. D. Manning. Generating typed dependency parses from phrase structure parses. In *Proc. LREC*, pages 449–454, 2006.

[12] G. Dror, Y. Maarek, A. Mejer, and I. Szpektor. From query to question in one click: Suggesting synthetic questions to searchers. In *Proc. WWW*, pages 391–402, 2013.

[13] A. Easwara Moorthy and K.-P. L. Vu. Privacy concerns for use of voice activated personal assistant in the public space. *International Journal of Human-Computer Interaction*, 31(4):307–335, 2015.

[14] Google official blog. http://googleblog.blogspot.co.il/2014/10/omg-mobile-voice-survey-reveals-teens.html. [Accessed 2016-01-10].

[15] M. Gupta and M. Bendersky. Information retrieval with verbose queries. *Foundations and Trends in Information Retrieval*, 9(3-4):209–354, 2015.

[16] I. Guy and D. Pelleg. The factoid queries collection. In *PROC. SIGIR*, 2016.

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

[18] J. Jiang, A. Hassan Awadallah, R. Jones, U. Ozertem, I. Zitouni, R. Gurunath Kulkarni, and O. Z. Khan. Automatic online evaluation of intelligent assistants. In *Proc. WWW*, pages 506–516, 2015.

[19] J. Jiang, W. Jeng, and D. He. How do users respond to voice input errors? lexical and phonetic query reformulation in voice search. In *Proc. SIGIR*, pages 143–152, 2013.

[20] M. Kamvar and S. Baluja. A large scale study of wireless search behavior: Google mobile search. In *CHI*, pages 701–709, 2006.

[21] M. Kamvar, M. Kellar, R. Patel, and Y. Xu. Computers and iphones and mobile phones, oh my!: A logs-based comparison of search users on different devices. In *Proc. WWW*, pages 801–810, 2009.

[22] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *Proc. ACL*, pages 423–430, 2003.

[23] D. Lagun, C.-H. Hsieh, D. Webster, and V. Navalpakkam. Towards better measurement of attention and satisfaction in mobile search. In *Proc. SIGIR*, pages 113–122, 2014.

[24] J. Li, S. Huffman, and A. Tokuda. Good abandonment in mobile and pc internet search. In *SIGIR*, pages 43–50, 2009.

[25] C. Y. Lin. Automatic question generation from queries. In *Workshop on the Question Generation Shared Task*, pages 156–164, 2008.

[26] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 1993.

[27] A. Moreno-Daniel, S. Parthasarathy, B. Juang, and J. Wilpon. Spoken query processing for information retrieval. In *Proc. ICASSP*, volume 4, pages IV–121–IV–124, 2007.

[28] Y. Pinter, R. Reichart, and I. Szpektor. Syntactic parsing of web queries with question intent: A distant supervision approach, 2016. Under Submission to NAACL 2016.

[29] R. Rosenfield. Two decades of statistical language modeling: Where do we go from here? *Proceedings of the IEEE*, 2000.

[30] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope. Your word is my command: Google search by voice: A case study. In *Advances in Speech Recognition*, pages 61–90. 2010.

[31] J. Shan, G. Wu, Z. Hu, X. Tang, M. Jansche, and P. J. Moreno. Search by voice in mandarin chinese. In *Proc. INTERSPEECH*, pages 354–357, 2010.

[32] M. Shokouhi and Q. Guo. From queries to cards: Re-ranking proactive card recommendations based on reactive search history. In *Proc. SIGIR*, pages 695–704, 2015.

[33] M. Shokouhi, R. Jones, U. Ozertem, K. Raghunathan, and F. Diaz. Mobile query reformulations. In *Proc. SIGIR*, pages 1011–1014, 2014.

[34] Y. Song, H. Ma, H. Wang, and K. Wang. Exploring and exploiting user search behavior on mobile and tablet devices to improve search relevance. In *WWW*, pages 1201–1212, 2013.

[35] J. Teevan, D. Ramage, and M. R. Morris. #twittersearch: A comparison of microblog search and web search. In *Proc. WSDM*, pages 35–44, 2011.

[36] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. NAACL*, pages 173–180, 2003.

[37] S. Verberne. Paragraph retrieval for why-question answering. In *Proc. SIGIR*, pages 922–922, 2007.

[38] Y. Y. Wang, D. Yu, Y.-C. Ju, and A. Acero. An introduction to voice search. *Signal Processing Magazine*, 25(3):28–38, 2008.

[39] R. W. White, M. Richardson, and W. Yih. Questions vs. queries in informational search tasks. In *Proc. WWW*, pages 135–136, 2015.

[40] J. Yi and F. Maghoul. Mobile search pattern evolution: The trend and the impact of voice queries. In *Proc. WWW*, pages 165–166, 2011.

[41] J. Yi, F. Maghoul, and J. Pedersen. Deciphering mobile search patterns: A study of yahoo! mobile search queries. In *Proc. WWW*, pages 257–266, 2008.

[42] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proc. SIGIR*, pages 334–342, 2001.

[43] G. Zweig and S. Chang. Personalizing model m for voice-search. In *Proc. INTERSPEECH*, pages 609–612, 2011.