UNIVERSITY OF PITTSBURGH www.ischool.pitt.edu

School of Information Sciences

# How Do Users Respond to Voice Input Errors?

## Lexical and Phonetic Query Reformulation in Voice Search

**Jiepu Jiang, Wei Jeng, Daqing He**

School of Information Sciences,
University of Pittsburgh

# EXAMPLE

- I am a big fan of the famous Irish rock band U2. Are they going to have a concert in Dublin recently? Maybe I can go to a concert after SIGIR.

- Then, I take out my smartphone ….

# EXAMPLE: VOICE INPUT ERROR

- ***Voice Input Error***
  - The query received by the search system is different from what the user meant to use.
- **Speech recognition error**

| User's Actual Query | System's Transcription |
|---|---|
| *U2* | *Youtube* |

- **Improper system interruption**
  - The user is interrupted before finishing speaking all of the query terms.

# EXAMPLE: QUERY REFORMULATION

- **Lexical changes**

| Original Query | Reformulation |
|---|---|
| U2 | ***Irish rock band*** U2 |

- **Phonetic changes**
  - Overstate “U2” at speaking

- **Probably related to the voice input errors**

# RESEARCH QUESTIONS

1. How do voice input errors affect the effectiveness of voice search?

2. How do users reformulate queries in voice search?

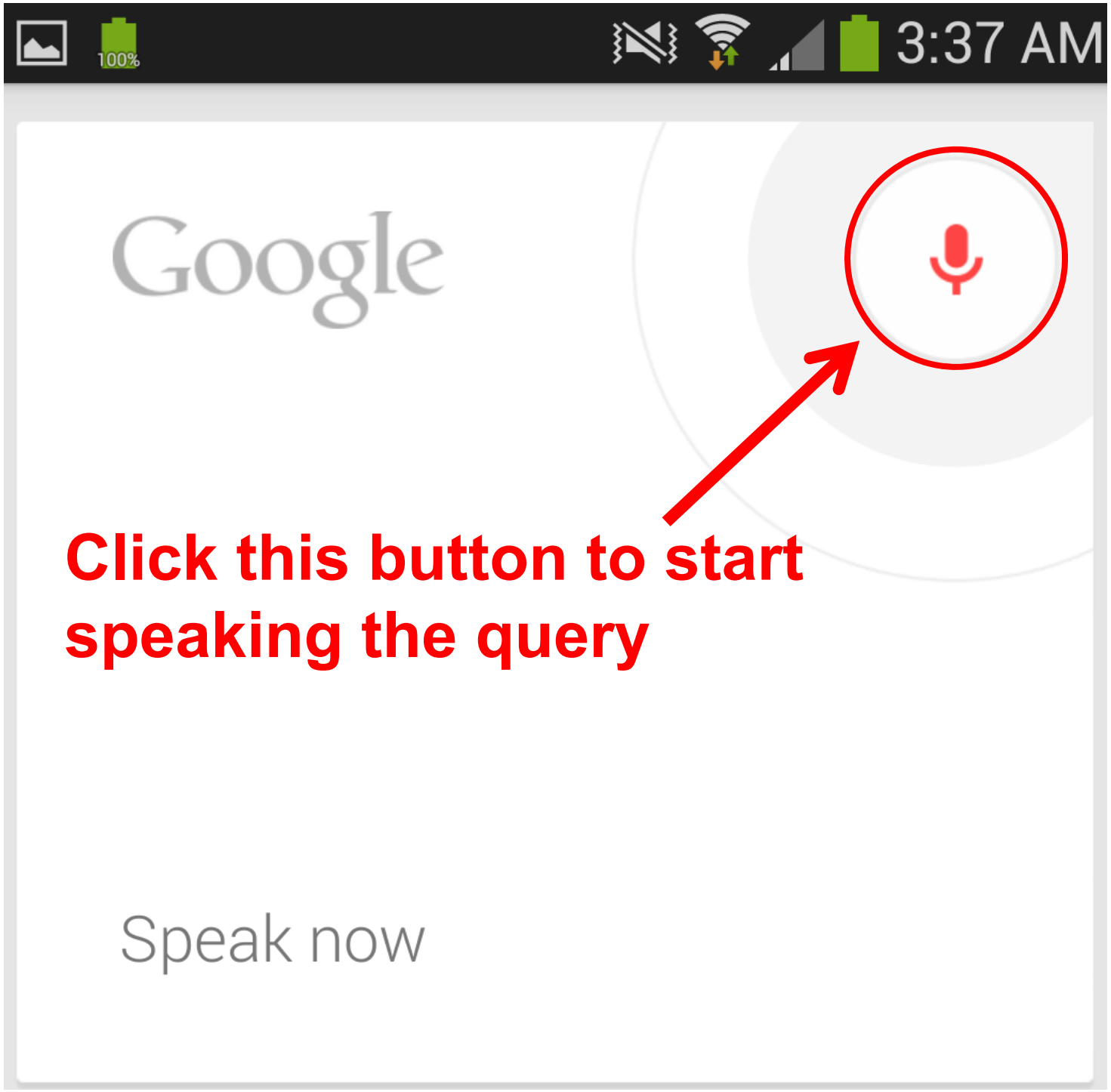
3. Are users' query reformulations related to voice input errors? If yes, do they help the solve the voice input errors?

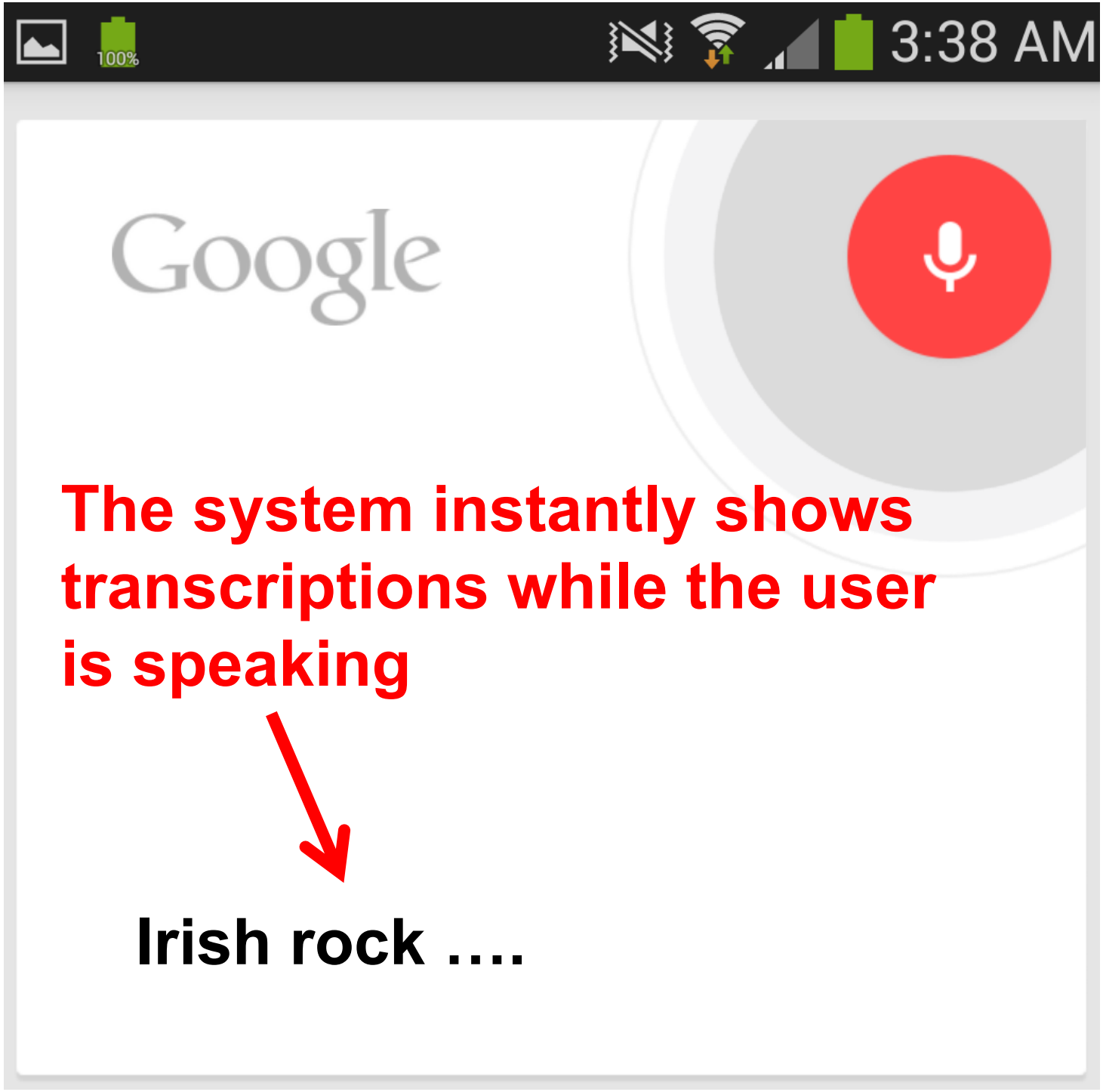
# OUTLINE

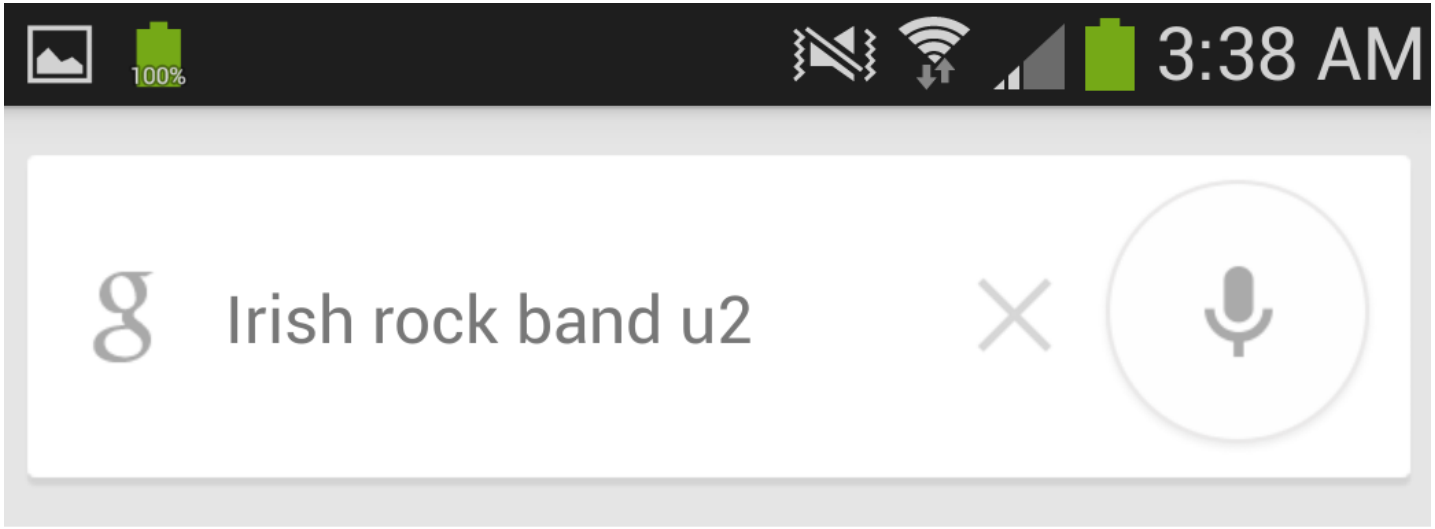
- Objectives
- *Experiment Design*
- Data
- Voice Input Errors
- Query Reformulations

# EXPERIMENT DESIGN

- **Objective**
  - To collect users' natural responses to voice input errors
- **System**
  - Google voice search app on iPad


100%
3:37 AM
Google
Click this button to start speaking the query
Speak now


100%
3:38 AM
Google
The system instantly shows transcriptions while the user is speaking
Irish rock ….




**U2 - Wikipedia, the free encyclopedia**

en.wikipedia.org/wiki/U2

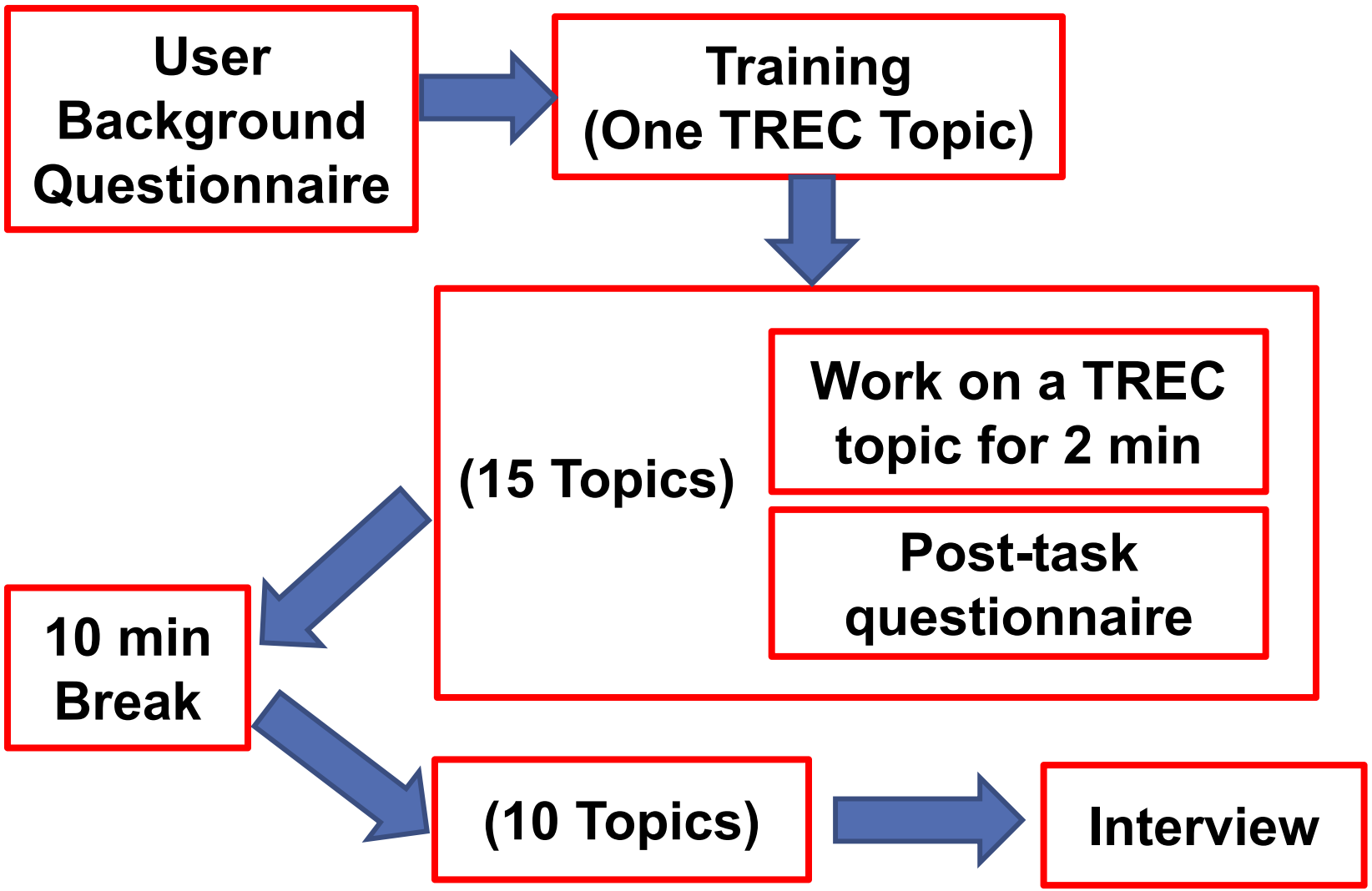
Jump to "Reapplying for the job of the best band in the world" (2000–06) - The **band** ... In 2005, Bruce Springsteen inducted **U2** into the **Rock** and ... The move was criticised in the **Irish** ...

**Finally, the system retrieves results according to its transcriptions**

# SEARCH TASKS

- **Work on TREC topics**
  - 30 from robust track, 20 from web track
- **Search session (2 minutes)**
- **Users can**
  - Reformulate queries
  - Use Google's query suggestions
  - Browse and click results
- **Users cannot**
  - Type on the iPad to input queries

# EXPERIMENT PROCEDURE (90 MIN)



# LIMITATIONS OF THE DESIGN

- **Lack of contexts of using voice search**
  - Topics
  - Experiment environment
- **Query Input**
  - Our experiment: voice only
  - Practical cases: voice + typing on iPad
- **Influence on our results & conclusions**
  - Details in the paper

# OUTLINE

- Objectives
- Experiment Design
- *Data*
- Voice Input Errors
- Query Reformulations

# OVERVIEW OF THE DATA

- **20 English native speaker participants**
- **500 search sessions (20 participants × 25 topics)**
- **1,650 queries formulated by participants themselves**
  - 3.3 voice query per user session
- **32 cases of using query suggestions**
- **1.41 (SD=1.14) clicked results per user session.**

# QUERY TRANSCRIPTION

- **$q_v$ (a voice query's actual content)**
  - manually transcribed from the recording
  - two authors had an agreement of 100%, except on casing, plurals, and prepositions
- **$q_{tr}$ (the system's transcription of a voice query)**
  - available from the log

# EVALUATION OF EFFECTIVENESS

- **No Explicit Relevance judgments**

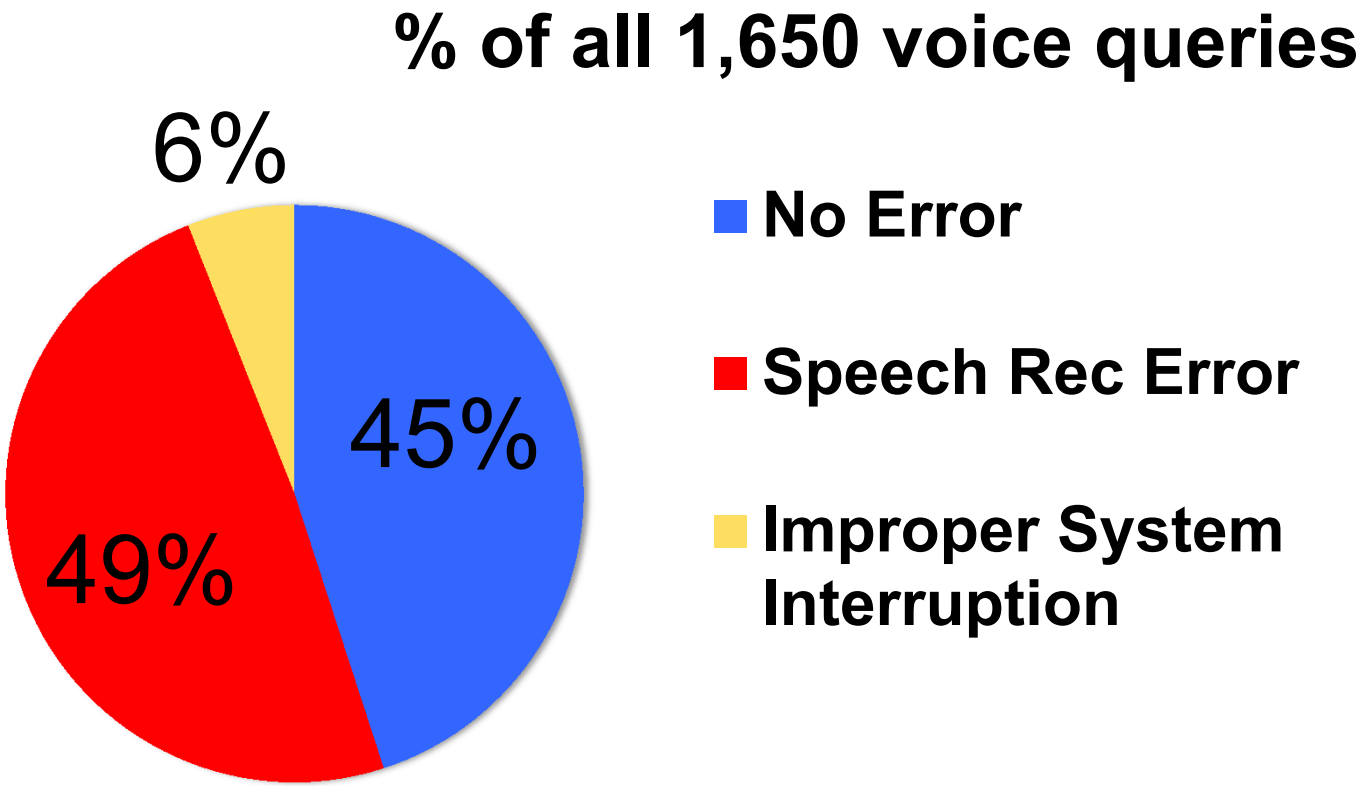
- **For each topic, we aggregate all users' clicked results on this topic as its relevant documents**
  - 9.76 (SD=3.11) unique clicked results per topic
  - For each clicked result, relevance score = 1

# OUTLINE

- Objectives
- Experiment Design
- Data
- **Voice Input Errors**
  - ***Individual Queries***
  - Search Sessions
- **Query Reformulations**

# INDIVIDUAL QUERIES

- **908 queries have voice input errors (55% of 1,650)**
  - 810 by speech recognition error
  - 98 by improper system interruption

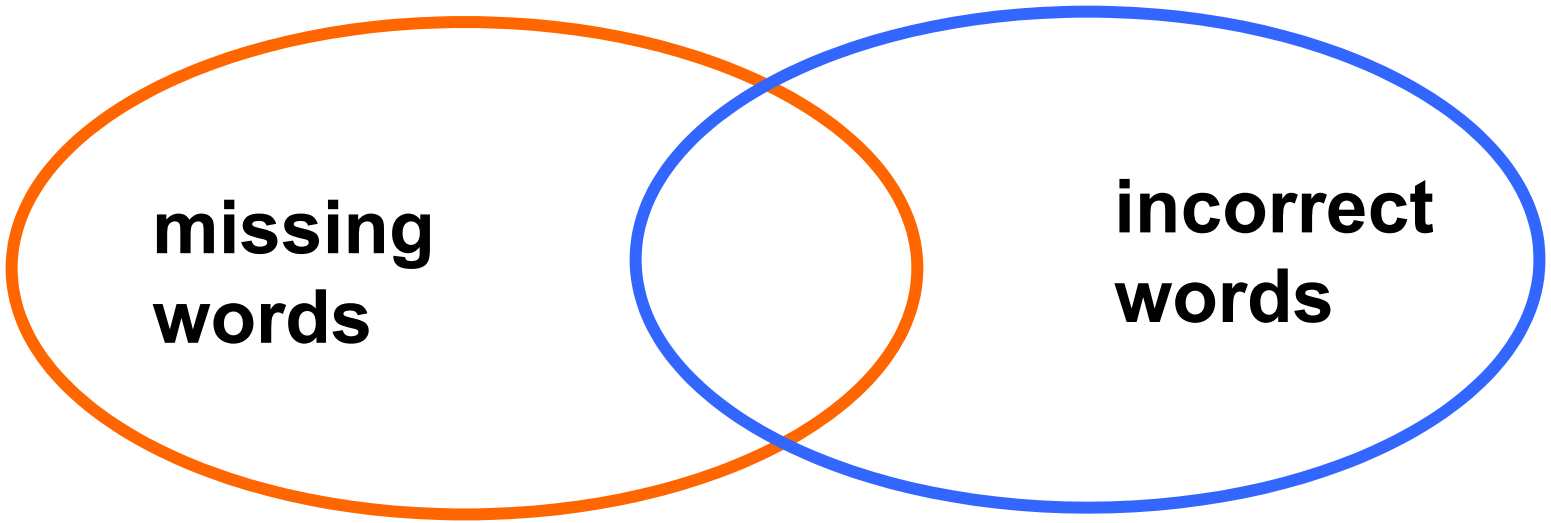


# INDIVIDUAL QUERIES: WORDS

- ***Missing words***: words in $q_v$ but not in $q_{tr}$
- ***Incorrect words***: words in $q_{tr}$ but not in $q_v$

$q_v$: a voice query's actual content

$q_{tr}$: the system's transcription



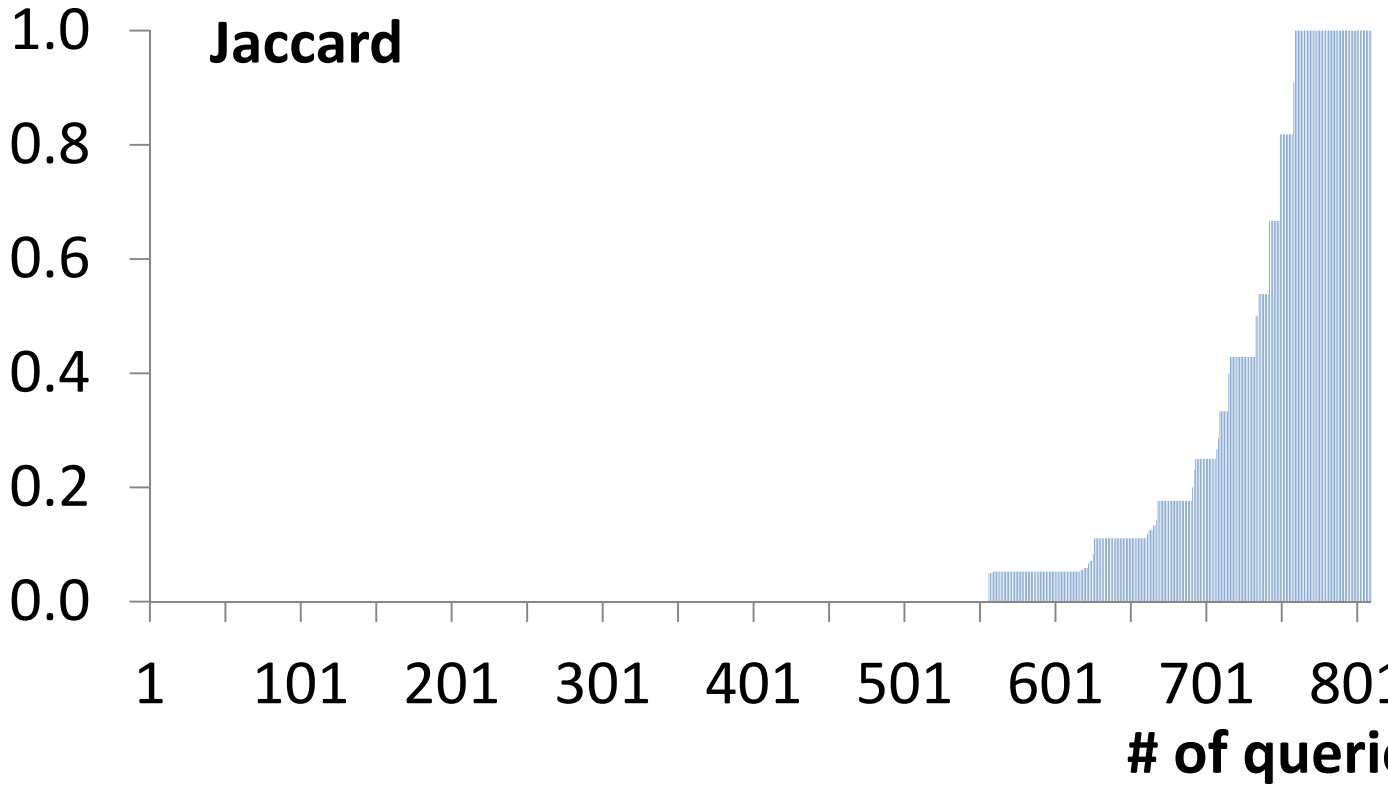
# INDIVIDUAL QUERIES: WORDS

- About half of the query words have errors

| | Speech Rec Errors 810 Queries | |
|---|---|---|
| | *mean* | *SD* |
| Length of $q_v$ | 4.14 | 1.99 |
| Length of $q_{tr}$ | 4.21 | 2.31 |
| # missing words in $q_v$ | 1.77 | 1.09 |
| # incorrect words in $q_{tr}$ | 1.84 | 1.44 |
| % missing words in $q_v$ | 49.7% | 29% |
| % incorrect words in $q_{tr}$ | 49.3% | 31% |

# INDIVIDUAL QUERIES: RESULTS

- **For 810 queries with speech recognition errors**
  - Very low overlap between the results of $q_v$ and $q_{tr}$
  - Jaccard similarity of top 10 results = **0.118**



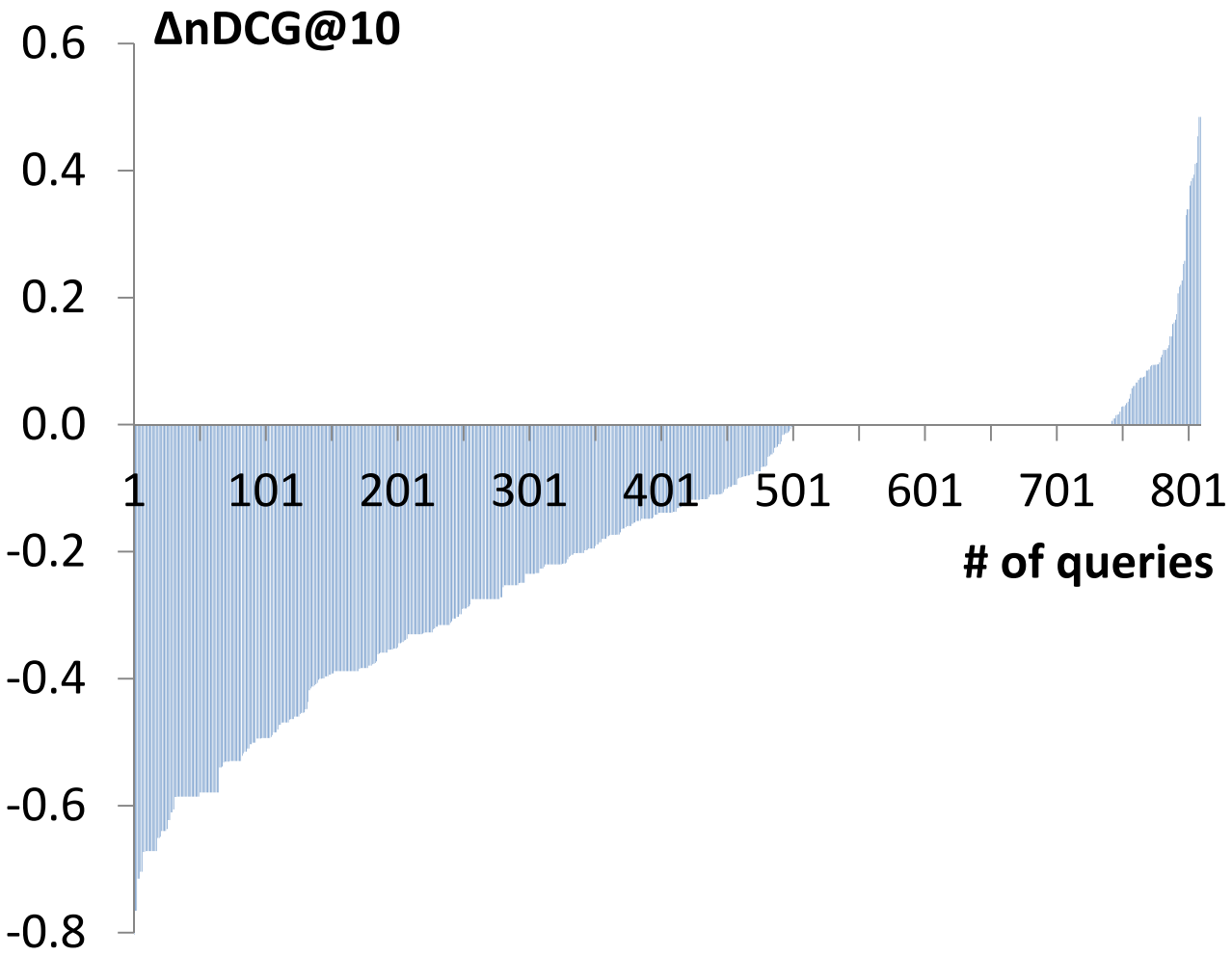
# INDIVIDUAL QUERIES: PERFORMANCE

- **Significant decline of search performance (nDCG@10)**

| | No Errors 742 Queries | | Speech Rec Errors 810 Queries | |
|---|---|---|---|---|
| | ***mean*** | ***SD*** | ***mean*** | ***SD*** |
| **nDCG@10 of $q_v$** | 0.275 | 0.20 | 0.264 | 0.22 |
| **nDCG@10 of $q_{tr}$** | 0.275 | 0.20 | 0.083 ↓ | 0.16 |
| **ΔnDCG@10** | - | - | -0.182 | 0.23 |

# INDIVIDUAL QUERIES: PERFORMANCE

- **Significant decline of search performance (nDCG@10)**



# INDIVIDUAL QUERIES: PERFORMANCE

- **Improper system interruption**
  - The worst search performance

| | No Errors 742 Queries | | Speech Rec Errors 810 Queries | | Improper System Interruptions 98 Queries | |
|---|---|---|---|---|---|---|
| | mean | SD | mean | SD | mean | SD |
| **nDCG@10 of $q_v$** | 0.275 | 0.20 | 0.264 | 0.22 | - | - |
| **nDCG@10 of $q_{tr}$** | 0.275 | 0.20 | 0.083 ↓ | 0.16 | 0.061 ↓ | 0.14 |

# OUTLINE

- Objectives
- Experiment Design
- Data
- **Voice Input Errors**
  - **Individual Queries**
    - ***Half of the words have errors***
    - ***Very different search results***
    - ***Significant decline of search performance***
  - **Search Sessions**
- **Query Reformulations**

# OUTLINE

- Objectives
- Experiment Design
- Data
- **Voice Input Errors**
  - Individual Queries
  - ***Search Sessions***
- **Query Reformulations**

# SEARCH SESSION

- **Significantly more voice queries were issued**
  - Increased efforts of users
  - 2/3 queries have voice input errors

| | 187 Sessions w/o Voice Input Errors | | 313 Sessions w/ Voice Input Errors | |
|---|---|---|---|---|
| | mean | SD | mean | SD |
| **# queries** | 1.44 | 0.82 | 4.41 ↑ | 2.51 |
| **# unique queries** | 1.44 | 0.82 | 3.30 ↑ | 1.87 |
| **# queries w/o voice input errors** | 1.44 | 0.82 | 1.51 | 1.36 |

# SEARCH SESSION

- **Slightly less (4%) unique relevant results retrieved in the session, although about 3 times of total results were returned**
  - more results were retrieved, probably increased efforts of users for judging results

| | 187 Sessions w/o Voice Input Errors | | 313 Sessions w/ Voice Input Errors | |
|---|---|---|---|---|
| | mean | SD | mean | SD |
| **# unique relevant results by $q_{tr}$** | 2.90 | 1.56 | 2.78 | 1.71 |
| **# unique results by $q_{tr}$** | 13.38 | 6.66 | 37.95 ↑ | 21.00 |

# SEARCH SESSION

- **In sessions with voice input errors**
  - Slightly less clicked results over the session
  - 15% more likelihood with no clicked results

| | 187 Sessions w/o Voice Input Errors | | 313 Sessions w/ Voice Input Errors | |
|---|---|---|---|---|
| | mean | SD | mean | SD |
| **# clicked results in the session** | 1.39 | 1.01 | 1.34 | 1.23 |
| **% sessions user clicked results** | 84.49% | - | 69.97% | - |

# OUTLINE

- Objectives
- Experiment Design
- Data
- **Voice Input Errors**
  - Individual Queries
  - **Search Sessions**
    - **Users made extra efforts to compensate**
    - **Overall slightly worse performance over session**
- **Query Reformulations**

# OUTLINE

- Objectives
- Experiment Design
- Data
- Voice Input Errors
- **Query Reformulations**
  - ***Patterns***
  - **Performance**
  - **Correcting Error Words**

# TEXTUAL PATTERNS

- **Query Term Addition (ADD)**

| | Voice Query | Transcribed Query | ADD words |
|---|---|---|---|
| $q_1$ | the sun | the son | |
| $q_2$ | the sun **solar system** | the sun solar system | solar system |

- **Query Term Substitution (SUB)**
  - SUB word pairs are manually coded (93% agreement)

| | Voice Query | Transcribed Query | SUB words |
|---|---|---|---|
| $q_1$ | art theft | test | |
| $q_2$ | art embezzlement | are in Dublin | theft → embezzlement |
| $q_3$ | stolen artwork | stolen artwork | embezzlement → stolen<br>art → artwork |

# TEXTUAL PATTERNS

- **Query Term Removal (RMV)**

| | Voice Query | Transcribed Query |
|---|---|---|
| $q_1$ | advantages of same sex schools | andy just open it goes |
| $q_2$ | same sex schools | same sex schools |

- **Query Term Reordering (ORD)**

| | Voice Query | Transcribed Query |
|---|---|---|
| $q_1$ | interruptions to ireland peace talk | is directions to ireland peace talks |
| $q_2$ | ireland peace talk interruptions | ireland peace talks interruptions |

# PHONETIC PATTERNS

- **Partial Emphasis (PE)**
  - Overstate a specific part of a query

| PE Type | Example | Explanation |
|---|---|---|
| **Stressing (STR)** | ***rap*** and crime | put stress on “rap” |
| **Slow down (SLW)** | rap and c-r-i-m-e | slow down at “crime” |
| **Spelling (SPL)** | P·u·e·r·t·o Rico | spell out each letter in “Puerto” |
| **Different Pronunciation (DIF)** | Puerto Rico | pronounce “Puerto” differently |

# PHONETIC PATTERNS

- **Whole Emphasis (WE)**
  - Overstate the whole query at speaking

- **2 authors manually coded the phonetic patterns**
  - agreement 87.6%
  - 5 Labels
    - STR/SLW
    - SPL
    - DIF
    - WE
    - REP (repeat without observable patterns)

# USE OF DIFFERENT PATTERNS

- **When previous query has voice input error**
  - Increased use of SUB & ORD
  - Less use of ADD & RMV

| Patterns | Prev Q Error | Prev Q No Error | Overall |
|---|---|---|---|
| ADD | 90.50% | 32.98% ↓ | 53.82% |
| SUB | 15.04% | 16.34% ↑ | 14.87% |
| RMV | 66.75% | 37.93% ↓ | 48.37% |
| ORD | 33.51% | 43.03% ↑ | 39.58% |
| (All Lexical) | 99.74% | 77.36% ↓ | 85.47% |

# USE OF DIFFERENT PATTERNS

- **Use of phonetic patterns are nearly always associated with previous voice input errors**

| Patterns | Prev Q Error | Prev Q No Error | Overall |
|---|---|---|---|
| STR/SLW | 0% | 14.84% ↑ | 9.46% |
| SPL | 0% | 0.60% ↑ | 0.39% |
| DIF | 0% | 0.90% ↑ | 0.57% |
| WE | 0.26% | 9.30% ↑ | 6.02% |
| (All Phonetic) | 0.26% | 25.64% ↑ | 16.44% |
| Repeat | 0% | 20.54% ↑ | 13.58% |

# OUTLINE

- Objectives
- Experiment Design
- Data
- Voice Input Errors
- **Query Reformulations**
  - **Patterns**
    - ***Lexical + Phonetic; related to voice input errors***
  - ***Search Performance***
  - **Correcting Error Words**

# REFORMULATION: PERFORMANCE

- **Overall slightly improvement (10% in nDCG@10)**
- **But highly depends on whether or not voice input error happened after query reformulation**
- **Did not reduce the likelihood of voice input errors**

| The reformulated query has / is | nDCG@10 (before → after) | # of cases |
|---|---|---|
| No Error | 0.150 → 0.233 ↑ | 474 (40%) |
| Speech Rec Error | 0.104 → 0.079 ↓ | 597 (51%) |
| Interruption | 0.156 → 0.056 ↓ | 79 (6.7%) |
| Query Suggestion | 0.201 → 0.223 ↑ | 32 (2.7%) |
| Overall | 0.129 → 0.143 ↑ | 1,182 |

# OUTLINE

- Objectives
- Experiment Design
- Data
- Voice Input Errors
- **Query Reformulations**
  - **Patterns**
  - **Search Performance**
  - ***Correcting Error Words***

# REFORMULATION: CORRECTING ERRORS

- **Do query reformulation help correct error words?**
  - no substantial difference in terms of the # of error words (if speech recognition error happened after reformulation)

| The reformulated query has | # missing words | # incorrect words |
|---|---|---|
| | before → after | before → after |
| **No Errors** | 1.75 → 0.00 | 1.81 → 0.00 |
| **Speech Rec Errors** | 1.89 → 1.74 ↓ | 1.72 → 1.78 |

# REFORMULATION: CORRECTING ERRORS

- **Does query reformulation help correct error words?**
  - Yes, it indeed corrected parts of the error words
  - But new error words come out

| The reformulated query has | # missing words corrected after reformulation | # missing Words removed after reformulation | # new missing words |
|---|---|---|---|
| **No Errors** | 1.13 | 0.61 | 0.00 |
| **Rec Errors** | 0.52 | 0.34 | 0.72 |

# SUCCESS RATE OF CORRECTING ERRORS

- **SUB & ORD as the most effective patterns**
- **PE and WE: not much higher than simply repeat**

| | Success rate of correcting missing words | nDCG@10 before → after |
|---|---|---|
| **ADD** | 40.73 % | 0.085 → 0.119 |
| **SUB** | 73.53 % | 0.052 → 0.156 ↑ |
| **RMV** | - | 0.077 → 0.111 |
| **ORD** | 69.14 % | 0.062 → 0.147 ↑ |
| **PE** | 62.50 % | 0.022 → 0.150 ↑ |
| **WE** | 60.94 % | 0.028 → 0.110 ↑ |
| **Repeat** | 59.73 % | 0.051 → 0.142 ↑ |
| **Overall** | 47.45 % | 0.058 → 0.132 ↑ |

# OUTLINE

- Objectives
- Experiment Design
- Data
- Voice Input Errors
- **Query Reformulations**
  - **Use of reformulation related to voice input errors**
  - **Some are effective for correcting error words**
  - **Did not reduce the likelihood of voice input errors**
  - **Overall not much improvement of search performance**

# WRAP UP

- **Voice input errors**
  - largely affect search performance and users' efforts

- **Voice Query Reformulation**
  - New patterns
  - Lexical reformulation for correcting voice input errors
  - Currently query reformulation is not much effective
  - Overall lack of support for query reformulation
    - Users have to speak the whole query again rather than correcting individual words
    - Query suggestion were seldom used

# LIMITATION

- **What may not be generalizable (due to TREC topics)**
  - The frequency of voice input errors
  - The frequency that different patterns were used
- **What may be generalizable**
  - The limited effectiveness of query reformulation
  - The comparative effectiveness of different patterns
- **Experiment environment (e.g. noise, interruption)**
  - The effectiveness of query reformulation could be even worse

# Thank you

# ACKNOWLEDGEMENTS


- **Google Voice Search**
  - Absolutely the best ever voice search system we found
- **Supports**
  - SIGIR student travel grant (Jiepu Jiang)
  - Google travel grant for women (Wei Jeng)
  - Student travel grant, School of Information Sciences, University of Pittsburgh (Jiepu Jiang & Wei Jeng)
- **People**
  - Participants of the study
  - Shuguang Han
  - Kelly Shaffer
  - Jessica Benner
  - Usability Lab (ULAB), Information Science, University of Pittsburgh