EXPERIMENTAL AND LEXICAL RETRIEVAL IN NATURALISTIC AND MISPERCEPTION

Kevin Tang, Andrew Nevin

1Yale University, Department of Linguistics & University College London, Department of Linguistics, *kevin.tang@yale.edu, a.nevin@ucl.ac.uk*

INTRODUCTION AND DATA

- To understand aspects of the lexical retrieval system, we look into the errors caused by the system when it fails to retrieve the correct intended word.
- To increase the ecological validity of the findings, we examined misperceptions in both naturalistic and experimental settings.
- In this study, we systematically examined two naturalistic datasets collected in the "wild" (of conversational speech as well as song speech) and one mega experimental dataset (of single-word lab experiments).
- We examined the token frequency relationship between the intended word and the perceived word.

Word Confusion Data

- Corpus of Conversational Misperceptions: 2,072 word confusion pairs, 1,231 monosyllabic words and 841 polysyllabic words.
- Mondegreen: 940 word confusion pairs, 5,669 monosyllabic words and 3,040 polysyllabic words.
- Experimental: 1,002 word confusion pairs, 5,085 monosyllabic words and 5,912 polysyllabic words.

HYPOTHESES AND PREDICTIONS

Hypothesis 1 – Correlational

- When the intended word cannot be retrieved, listeners can select a word that the key is or did not correlate with the intended word.
- In a small experiment testing 144 words, Follak, Laherstein, and Deckel (1960) analysed word frequency of the intended word and the perceived words and they did not find a significant correlation.
- In a larger experiment testing 1,428 words, Feltz et al. (2013) found a positive correlation which is significant at a moderate level (r = 0.114, df = 21,842, p = 0.00101) between the intended and perceived words.

Hypothesis 2 – Frequency Differences

- When the intended word cannot be retrieved, listeners can select a word that is significantly less frequent, less frequent or more frequent than the intended word.
- Our work challenges research on single-word perception (Cooke, 2009; Tóth (2015), Chapter 2).
- Graceful degradation in lexical retrieval, we predict that listeners would tend to be more frequent words than expected on the basis of the frequency distribution.

FREQUENCY EXPECTATION OF LYRICS AND EVERYDAY SPEECH

- Naturalistic Misperceptions: exhibited a correlation between the intended word and the perceived word, but the paired T-tests showed that the perceived word is less frequent than the intended word.
- Experimental data exhibited an unusual distribution of the perceived word than the intended word.

CONCLUSIONS AND IMPLICATIONS

- Our frequency findings are stable against potential confounds such as monosyllablity and the amount of word-level segmental differences.
- Our results suggest that listeners are sensitive to their experience and expectation of their phonetic and lexical knowledge during speech perception.

METHODS

Extractions of Word Confusions

- The misperceptions in the Conversational and Misperception corpora are of a sentence-level. Therefore, word confusion data is divided by 10 word-level Levenshtein distance bins as well as monosyllabicity.
- Hypothesis 1: Test (right) Correlational analyses.
- Hypothesis 2: Middle (right)-T tests analyses of frequency differences
- Hypothesis 3: (right-left) T-tests analyses of word durational information.
- The perceptual system uses the information in the context of misperception, the perceptual system uses the information in the context of misperception, the perceptual system uses the information in the context of misperception, the perceptual system uses the information in the context of misperception.

CONCLUSIONS AND IMPLICATIONS

- In the Mondegreen data, listeners would tend to be a more frequent word than the intended word.
- The intended sentences of the Mondegreen data were used as a sample of lyrics (117,326 tokens), while subtle tests were used as a sample of everyday speech (1,032 mil. tokens).
- The frequency distributions of 40,000 word random samples suggest that lyrics contain more frequent words than everyday speech.
- The Vocab Size growth curve analysis (Sawyer, 2001) with 13,000 word random samples showed that the vocabulary size (type/token ratio) is consistently smaller for lyrics than for everyday speech.

DATA AVAILABILITY

- The conversational data used in this paper were extracted from the slips of the EAR (SEAR) Project (www.searproject.org). It contains = 5000 instances of speech perception error of English collected in everyday settings. The archive provides an interactive exploration interface.

SELECTED REFERENCES

- Felty, Allison, et al. (2013). “The misperceptions in the Conversational and Misperception corpora are of a sentence-level. Therefore, word confusion data is divided by 10 word-level Levenshtein distance bins as well as monosyllabicity.
- Felty, Allison, et al. (2013). “The misperceptions in the Conversational and Misperception corpora are of a sentence-level. Therefore, word confusion data is divided by 10 word-level Levenshtein distance bins as well as monosyllabicity.

DISCUSSION

- Our results suggest that listeners are sensitive to their experience and expectation of their phonetic and lexical knowledge during speech perception.
- Firstly, when contextual information (say, word duration) is available, listeners utilize them (as suggested by correlational findings in naturalistic data, and the lack of correlations in the experimental data).
- Secondly, listeners use their expectation and frequency distribution of their lexicons. This is suggested by how the experimental data show that the listeners simply select high frequency words as the perceived words, since this is the best strategy in face of uncertainty due to signal degradation.

- Furthermore, this is highlighted by the paradoxical result from the Mondegreen data of how listeners consistently underestimate the frequency of the intended words.

- Our work challenges research on single-word perception (Cooke, 2009; Tóth (2015), Chapter 2).
- Graceful degradation in lexical retrieval, we predict that listeners would tend to be more frequent words than the intended word.
- This paradoxical result was addressed by the fact that there tends to be less frequent words in lyrics than in everyday speech, therefore the intended word would tend to be a more frequent word than expected on the basis of the frequency distribution of everyday speech.
- This fact also highlights the need for conducting misperception research beyond the word-level as well as the importance of ecological validity.

- The frequency distribution of lyrics can be gene-dependent.
- Mondegreen instances might be more humours as conversational instances, therefore they contain less frequent words.
- Will experimental or naturalistic misperceptions be similar to naturalistic ones?

- The frequency distribution of lyrics can be gene-dependent.
- Mondegreen instances might be more humours as conversational instances, therefore they contain less frequent words.
- Will experimental or naturalistic misperceptions be similar to naturalistic ones?