MISHEARD ME ORONYM TREE: USING ORONYM TREES TO VALIDATE THE CORRECTNESS OF FREQUENCY DICTIONARIES

A Thesis

Presented to

the Faculty of California Polytechnic State University

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Computer Science

by

Jennifer "Jenee" Gayle Hughes

June 2012

COMMITTEE MEMBERSHIP

TITLE:	Misheard Me Oronym Tree: Using Oronym Trees to Validate the Correctness of Frequency Dictionaries
AUTHOR:	Jennifer "Jenee" Gayle Hughes
DATE SUBMITTED:	June 2012
COMMITTEE CHAIR:	Zoë Wood, Ph.D.
COMMITTEE MEMBER:	Franz Kurfess, Ph.D.
COMMITTEE MEMBER:	John Clements, Ph.D.

Abstract

Misheard Me Oronym Tree: Using Oronym Trees to Validate the Correctness of Frequency Dictionaries

Jennifer "Jenee" Gayle Hughes

In the field of speech recognition, an algorithm must learn to tell the difference between "a nice rock" and "a gneiss rock". These identical-sounding phrases are called oronyms. Word frequency dictionaries are often used by speech recognition systems to help resolve phonetic sequences with more than one possible orthographic phrase interpretation, by looking up which oronym of the root phonetic sequence contains the most common words. However, this approach is highly dependent upon the manner in which the frequency values in a word frequency dictionary are obtained.

Our paper demonstrates a technique used to validate word frequency dictionary values. We use oronym trees to compare phrase frequency values from dictionaries, to the frequency with which our human test subjects heard different variations of the root phrase. We chose to use frequency values from the UNISYN dictionary, which uses tallies each word occurance in a propriatary text corpus (***Glossary term***?).

Given any valid English phrase, herein referred to as the root phrase, our system will first generate all possible correct phonetic sequences for a General American accent. Then, it parses through these phonetic transcriptions depth-first, looking for valid orthographic words for each subsequent phonetic subsequence, generating full and partial phrases from these words. In the event that the entire phonetic sequence branch can be parsed into a valid orthographic phrase, we save this orthographic phrase as an oronym of the root phrase.

We also developed a visual representation of the oronym trees, to allow for visualizing phonetic dead-ends. In the event that a branch's phonetic "tail" is not orthographically interpretable, we visually "dead-end" the branch by drawing a red sphere. A particularly strong orthographic partial phrase before a phonetic dead-end can mislead a listener, causing them to lose track of the words in the rest of the phrase. In the event that the entire phonetic sequence can be parsed into a valid orthographic phrase, we indicate this successfully-found oronym with a green sphere.

Using the oronyms generated from our oronym tree, we then conducted a user study. Our multi-phase user study, incorporated over 851 data points from 208 test subjects. In it, we tested the validity of our oronym generation by having participants record themselves reading an oronym phrase. Then, a second set of subjects transcribed the recordings.

In the first phase, we generated oronym strings for the phrase *"a nice cold hour"*, and had over a dozen people make 72 recordings of the most common oronyms for that phrase. We then compared their pronunciations to the pronunciations we were expecting, and found that in 71 cases, the recorded phrase's phonemics matched our expectation. This indicates that, while not exhaustive, our pronunciation dictionary is a good match for actual American-English pronunciations. In the second phase, we selected 15 of the phase one recordings, and had 30 to 60 different people transcribe each one.

If the frequency dictionary values for our test phrases accurately reflect the real-world expectations of actual listeners, we would expect that the most commonly transcribed phrases in our user study would roughly correspond with our metric for the most likely oronym interpretation of the root phrase.

The best possible use case to show this is the case of *"a nice cold hour"*, whose commonly-misheard **********(find citation)********** oronym is *"an ice cold hour"*. The words "a" and "an" are identical in function, but "an" is only used in the case that the following word starts with a vowel sound. As there are more consonant sounds than vowel sounds, "an" is used far less often than "a" is. The UNISYN dictionary has a frequency value of 7,536,297 for "a", and of 794,169 for "an", for an approximate ratio of 10 to 1, where "a" accounts for 90.46% (p = 0.9047) of the combined count.

In this case, we'd expect that the phrase *"*a *nice cold hour"* would be transcribed nearly ten times as often as *"*an *ice cold hour"*.

In the event that this was not the case, we can conclude that tally-per-occurance frequency dictionary values does not apply well to an average audience's auditory expectations.

During the course of our study, we found that the presence of excessively-common words (such as "the", "is", and "a") threw off our frequency metric when we used per-occurance frequency value. These super-common words have such high per-occurance tallies that it overpowered the effect that any regular word had on a frequency metric. However, when we used document-count frequency values, we found that this effected was mitigated.

The frequency dictionary from the Corpus of Contemporary American English[2] tallies the number of documents that a word is found in, instead of tallying the total number occurances of that over all documents. In this dictionary, "a" has a document-count frequency value of 168619 , and "an" has a frequency value of of 159720 , for a ratio of 1.055 to 1, where "a" accounts for 51.35% (p = 0.5135) of the combined count.

In our user study, we found that 125 people transcribed “a nice cold hour , and 191 people transcribed “an ice cold hour , for a ratio of 0.65 to 1, where “a nice cold hour accounts for 39.56% (p = 0.3956) of the combined count. We did a statistical test with an alpha of .01, and got a value that was so low we can’t find a calculator that has enough decimal places to show it without rounding it to zero. In short, our per-occurance frequency metric predictions don’t even remotely match the projected data.

Our COCAE-derived document-count frequency metric preditions more closely matched our actual findings. We calculated that “a nice cold hour had a COCA-derived frequency metric value of 247719 , and “an ice cold hour had a value of 227405 , giving us a ratio of 1.08 to 1, where “a nice cold hour accounts for 52.17% (p = 0.5217) of the combined count. When compared to our actual results using a one-sample proportion z test, we got a p-value of ***COMPUTE LATER****, which is slightly better, but not great.

We found that using per-occurance frequency values when computing our overall-phrase-frequency metric caused the thrown off by excessively common words, such as “the”, “is”, and “a”. These super-common words have such high per-occurance tallies that it overpowered the effect that any regular word had on a frequency metric. However, when we tally on a document-count basis, instead of a by-occurance basis, we found that this effect was mitigated.

I need help with statistics here. To facilitate comparison, we created two rankings for the actual result phrases: one list ranked by expected frequency, and one ranked by number of actual transcriptions by our test subjects. In our phase two results, we found that out of the 578 transcriptions acquired for 53 unique phrases, only 11 had less than a difference of 5 ranks between the actual and expected occurances. The top 10 unique results, accounting for 88.00% of

total transcriptions, were on average 25 ranks more common than the frequency metric ranking predicted they'd be, with over half of them more than 39 ranks higher (out of 53 total ranks). From this, we can conclude that the frequency dictionary that we used is flawed.

If these frequency dictionary values were correct for the phrase words, we would expect that the most commonly transcribed phrases in our user study would roughly correspond with our metric for the most likely oronym interpretation of the root phrase. In the event that this was not the case, we could conclude that the frequency dictionary values were in error for that phrase. To facilitate comparison, we created two rankings for the actual result phrases: one list ranked by expected frequency, and one ranked by number of actual transcriptions by our test subjects. In our phase two results, we found that out of the 578 transcriptions acquired for 53 unique phrases, only 11 had less than a difference of 5 ranks between the actual and expected occurances. The top 10 unique results, accounting for 88.00% of total transcriptions, were on average 25 ranks more common than the frequency metric ranking predicted they'd be, with over half of them more than 39 ranks higher (out of 53 total ranks). From this, we can conclude that the frequency dictionary that we used is flawed.