Solving The Dorabella Cipher pt2

This part is a summary of the results from my shotgun hill climbing program testing. In other words, I've actually tried to prove that there might be another solution.

Since I've released the plaintext, I've run much longer trials using trigram, digram and monogram scoring. I also downloaded another scoring file. Some of the data is eye opening. Some less so.

While my hill climber has monogram scoring, most of the time it will be useless. Single letter frequencies are not enough to sort out words. When doing analysis by hand, the human brain supplies the needed word recognition and decision making that the computer can't. There was no surprise that it didn't find any significant ciphertext.

A run of 5000 trials using the digram scoring produced results. Throughout the output file are several exact and partial decrypts. Searching for high frequency letter patterns results in finding text with those patterns but little else. In some cases, a few words surface, but nothing that is even close to English text. I was partially surprised by the results because most crypto hill climb references favor trigrams or quadgrams. To get results using digrams is apparently not very common.

Since digram scoring didn't prove anything other than it worked, I did the same with 5000 trials using trigram scoring. The results were virtually identical. No new plaintexts that look like English except for what I already found.

I ran a full 5000 trials using quadgram scoring. Again, nothing new was found and it had a similar frequency of plaintexts as the trigram trials. At this point, all it does is verifies that my program is working as expected. Tests on large approximately 1000 letter texts often find a solution with just one trial. Tests on the smaller texts just show that you usually need more trials to find a solution as the texts get shorter and false positives become more common.

Being a programmable hill climber, it has an alternate swap system that is actually less random, but still designed to be random. Unfortunately there was no change in the outcome. It finds the plaintext about as well as the more random version using digrams, trigrams and quadgrams. As expected monogram scoring doesn't find anything. This still isn't the end. though.

Finding other hill climbing and simulated annealing programs, I adapted a different quadgram scoring table to my own program. I could then test using an alternate scoring method. The results were a surprise. Not all scoring tables are created equal. The best the other scoring table could produce was partial words that were not as good as what I achieved by hand. This may be why some hill climbers have such poor performance on small ciphertexts.

To be fair, I created a “normal” ciphertext with exactly 87 letters. The trigram and quadgram searches found the plaintext with much greater regularity than the Dorabella Cipher. The only letter it didn't guess correctly was the letter X.

All together over 80,000 hill climb iterations have been tested to try and prove that the current plaintext is wrong. So far all that has happened is it has verified that it is right. Alternate transcriptions have failed because of mistakes. Encrypting the good ciphertext with other keys always results in the same plaintext so even the starting point seems to be somewhat arbitrary.

Is there any doubt that the ciphertext is correct? Yes, just a little. The problem lies in the fact that one or more of the lower frequency letters may still be wrong or that there is an error in the transcription. Hill climbing and it's relatives used for cryptographic purposes are not perfect. That can easily be proven by feeding known samples and looking for differences. You always get false positives that look like total garbage with short ciphertexts. The Dorabella Cipher is full of strangeness.

Because it doesn't use normal words, I have to assume that Elgar never really intended for anyone to solve it or it was deliberately made to be difficult to solve. It appears to be constructed to lead the cryptanalyst astray and hides words with alternate spellings. I would say that most attempts would get at least as far as I did by hand. Dora Penny never had a chance.

Below is another arrangement of word divisions. This one makes a little more sense.

```
PBS AFT DALYRENCE MEET B BECO YOUR IDEDTD ALWASE
```
problems after dalliance meet be because your identity always

```
E STUNDER E THINC OLL OR IS IT HIS CH GUISE THNIC
```
e stutter e think ?all? or is it his ?ch? guise think

```
ABU IT ACOA
```
about it ????

And then with a bit more help:

Problems after dalliance meet be because your identity always e stutter e think all or is it his ch guise? Think about it (acoa).

This is perhaps the best idea of what I think it is supposed to say. I think many people have likely come close to the correct solution bu rejected it because it isn't normal. The computer doesn't hold such biases and seems capable of reliably finding the solution.

Links

Hill climb program and source code:
Win32 : HillClmb32-190512.zip
Win64: HillClmb64-190512.zip
source: HillClmbSrc190512.zip

```
Dorabella Cipher text:
ciphertext: dorabella.txt


These command lines create a file that is about 1.53Mb. It shouldn't take
very long even on older hardware. Probably less than a minute on higher end
hardware.

Quadgrams: hillclmb 4rs5000 10000 dorabella.txt out4.txt
Trigrams: hillclmb 3rs5000 10000 dorabella.txt out3.txt
Digrams: hillclmb 2rs5000 10000 dorabella.txt out2.txt
```