G B

ology@github

Search this website .

Kirk vs Spock Linguistics Head-to-Head

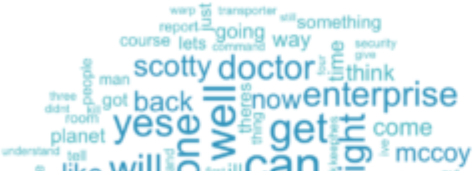
2017-08-12 BY GENE

Today I harvested the three seasons of Star Trek TOS episode scripts from a friendly website hosting the annotated text. Then I extracted just the lines spoken by Captain Kirk and just the lines of Mr Spock.

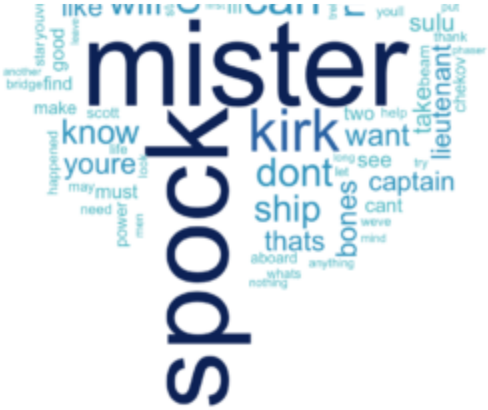
So how does Kirk match up to Spock linguistically?

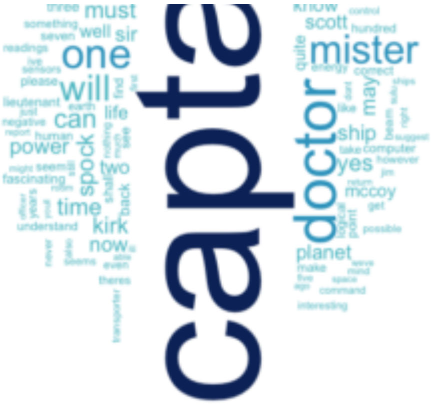
First of all, the Lingua::EN::Fathom module that computes text readability, says that the mean FOG for the entire series is 6.97. This measure can be roughly interpreted as the 7th grade level. Kirk speaks at just below the average at 6.41 FOG, and Mr Spock weighs in at 8.53 FOG. Makes sense. Also, Kirk has a lot more to say than Spock. He speaks a whopping 115,119 words compared to Spock's 62,616. This also makes sense.

Who says what the most? A bit of handy word cloud code generates these charts:



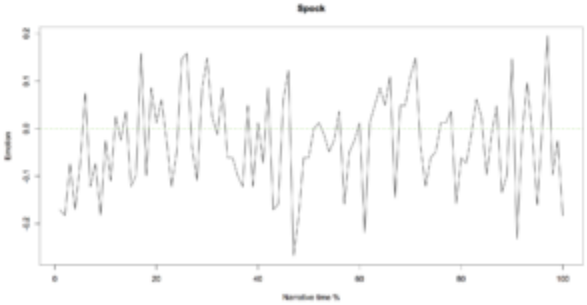
captain
phaser
say
enterprise
indicate
perhaps
four
course
way
believe
moment
know

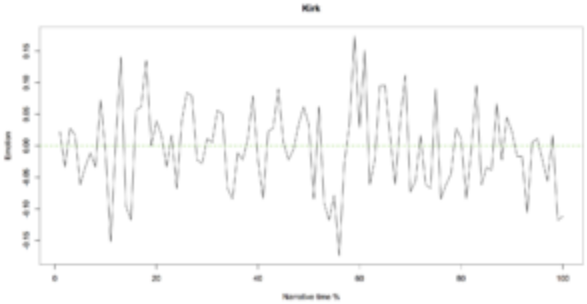




I’m sure you can tell who is who!

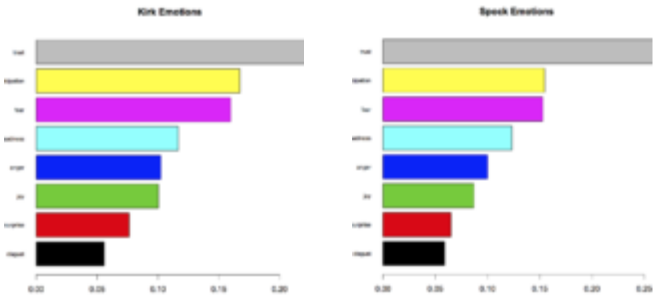
Since emotional sentiment can be easily plotted with the syuzhet package, what does that tell us?





It appears that Spock actually has a *slightly* greater emotional range than Kirk!

And what about the actual emotions expressed by each?



To my surprise, each expresses exactly the same emotions but in differing amounts.

As for lexical diversity (how “rich is the vocabulary” = unique words divided by all words), Kirk scores 0.123 but Spock scores a bit more with 0.194.

What are the top 5, 4-word phrases that are repeated the most for each? (This is done with the ngram package.) First is Captain Kirk:

```
                      ngrams freq         prop
1        this is the captain.    32 2.820825e-04
2          are you all right?    21 1.851166e-04
3               i want you to    20 1.763015e-04
4              i want to know    20 1.763015e-04
5            get out of here.    16 1.410412e-04
```

Next up is Mr Spock:

```
                      ngrams freq         prop
1          would seem to be    14 2.241112e-04
2      i fail to understand     6 9.604764e-05
3         there seems to be     6 9.604764e-05
4 on the planet's surface,      6 9.604764e-05
5          might be able to     6 9.604764e-05
```

It is obvious from the frequency that Mr Spock repeats himself a lot less often.

Here are the top 5 of Kirk’s repeated 5-word phrases:

```
                         ngrams freq         prop
1 kirk to enterprise. come in.    12 1.057819e-04
2    captain james kirk of the    12 1.057819e-04
3       let's get out of here.    10 8.815155e-05
4          lay in a course for     9 7.933640e-05
5       kirk here. what is it?     9 7.933640e-05
```

And here are Spock's:

```
                              ngrams freq         prop
1                it would seem to be    5 8.004098e-05
2                   i do not wish to    4 6.403278e-05
3 galileo to enterprise. galileo to    4 6.403278e-05
4      nine hundred and ninety point    4 6.403278e-05
5     hundred and ninety point seven    4 6.403278e-05
```

In the context of the whole series (considering everything spoken in every episode), what are the most unique bigrams ("two word phrases") for Kirk and Spock? This is the TF-IDF measure and was done with code of my own making, instead of an R package.

```
kirk.txt
        1. doctor coleman = 0.0010379422
        2. mister lurry = 0.0008303538
        3. cestus three = 0.0008303538
        4. janice lester = 0.0008303538
        5. mister president = 0.0007265595
        6. doctor janice = 0.0006227653
        7. storage compartments = 0.0006227653
        8. one distress = 0.0004151769
        9. mister lemli = 0.0003113827
        10. first punch = 0.0003113827

spock.txt
        1. mister boma = 0.0023780836
        2. mister gaetano = 0.0011097724
        3. mister flint = 0.0008477113
        4. berthold rays = 0.0007926945
        5. time portal = 0.0007064260
        6. gamma hydra = 0.0007064260
        7. doctor sevrin = 0.0007064260
        8. commodore stocker = 0.0007064260
```

```
9. edith keeler = 0.0007064260
10. hydra four = 0.0007064260
```

“Fascinating.”

FILED UNDER: DATA
TAGGED WITH: LINGUISTICS, R, STAR TREK

^ **Top**