A Brief Analysis of Lyric Metrics

Julia Stelman

1/7/2020

Introduction

In this mini analysis of lyric data, I explore a few features of lyric exerpts from 100 songs in 5 different languages. I compare 20 songs written in each language. I want to know if a language spoken across a wider geographical area, in turn, constitutes a language with a broader collection of regional vocabularies, and, in turn, leads to a sample of songs whose joint lyrics utilize a larger set of words than a language spoken across a smaller geographical range does.

In other words, do the songs written by, say, Spanish speaking artists vary more, lyrically, than songs written by, say, Indonesian artists? It would make sense, as Spanish is spoken all over the world, heightening the potential that this language has been split into different dialects by divergent evolution. Meanwhile, Indonesian is only spoken, officially, in Indonesia, a region that, based on the small size of its land area and longitudal range, doesn't look like it provides the Indondesian language as much potential for forming individual dialects.

The following two languages make up the *widely spoken* group:

- English (en)
- Spanish (es)

The following three languages make up the *not widely spoken* group:

- Indonesian (id)
- Dutch (nl)
- Swahili (sw)

The distinction between the first and second groups is decided by whether a language is an official national language on at least three continents.

Summarizing the Data

I've used the natural language processing package `quanteda` to calculate the following metrics for the first few lines of lyrics in each of the 100 sampled songs. (20 songs were randomly sampled from each of the 5 languages.)

- Types –> distinct words
- Tokens –> words
- Type/Token Ratio (TTR)

I summarized these as by-language averages in the table below.

Language	Avg Types per Lyric Excerpt	Avg Tokens per Lyric Excerpt	Avg Type/Token Ratio (TTR)
en	68.10	127	0.544

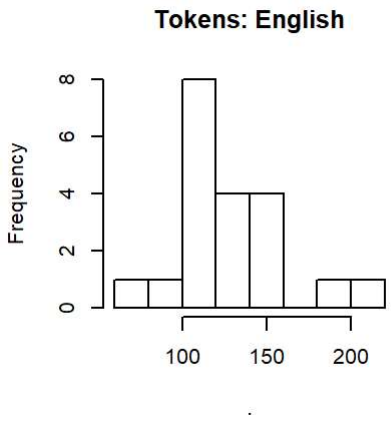
Language	Avg Types per Lyric Excerpt	Avg Tokens per Lyric Excerpt	Avg Type/Token Ratio (TTR)
es	60.00	109	0.560
id	47.70	76	0.625
nl	71.15	128	0.559
sw	61.85	111	0.560

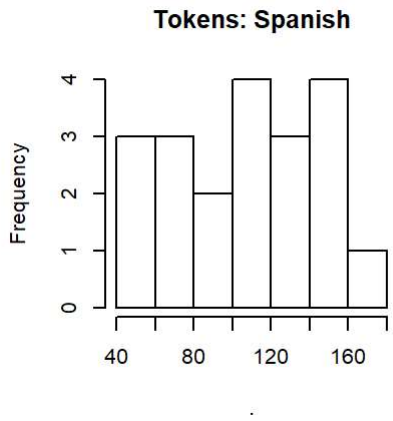
After frankensteining together the 20-song random sample of lyric excerpts from each language, I re-calculated the metrics again treating the combination of all sampled lyrics from one language as if it were from one song. (The code for how I did this is in the Appendix) They are shown in the table below. It's this information that will give me insight on what I'm really looking for.

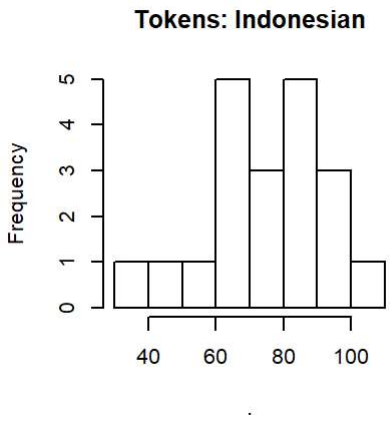
Language	Types	Tokens	TTR
en	588	2543	0.231
es	569	2174	0.262
id	506	1528	0.331
nl	650	2569	0.253
sw	919	2229	0.412

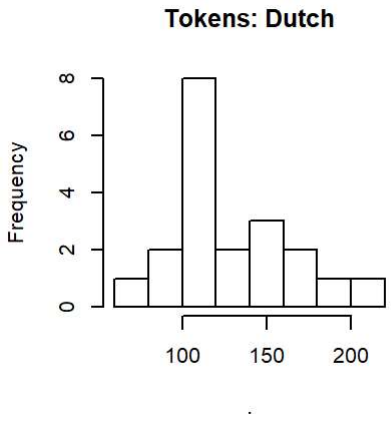
Exploratory Data Analysis

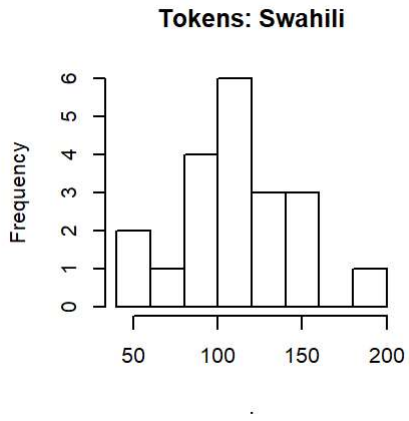
Let's take a look at some histograms of token counts to get a better look.





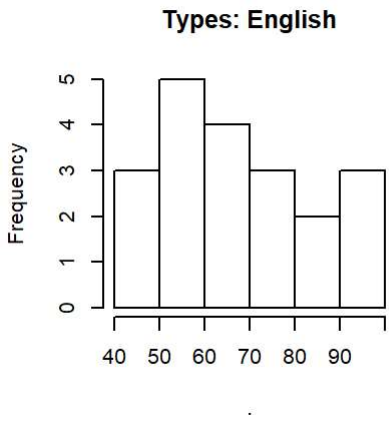


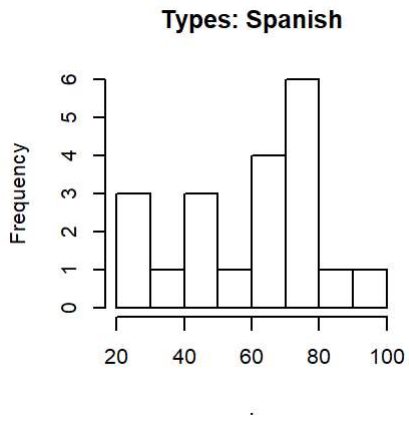


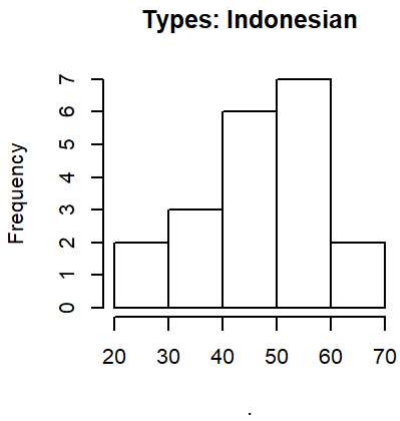


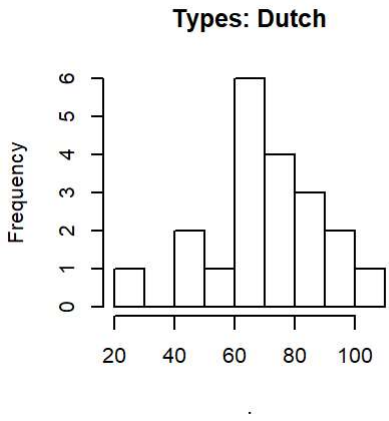
No large gaps and no big outliers. Based on the histogram, it seems like Indonesian song lyric exerpts run small. Indonesian's token count range is about half that of other languages.

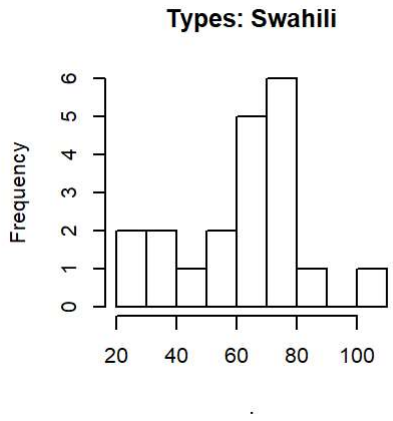
Let's also do the same thing for type counts, just for good measure.



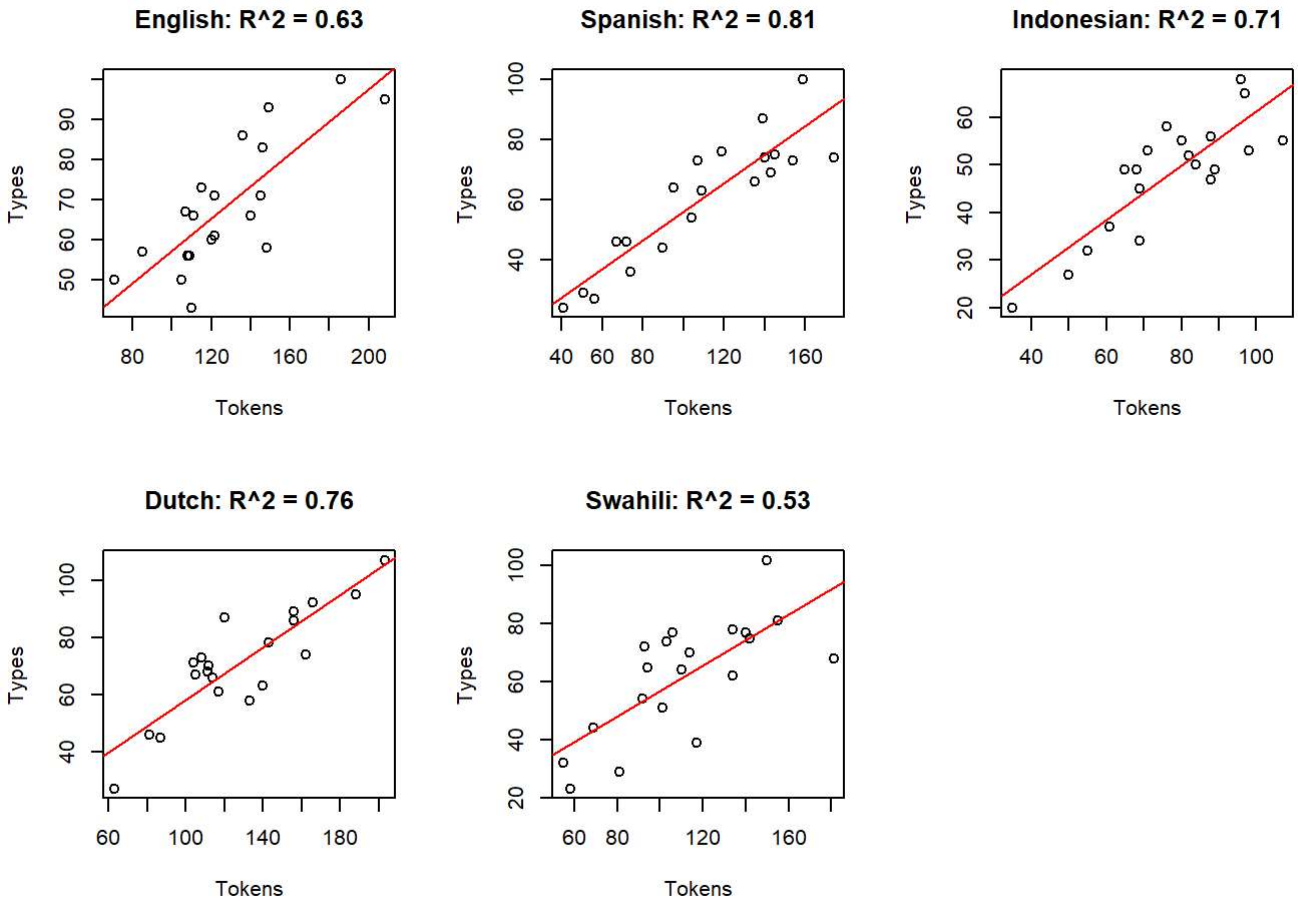








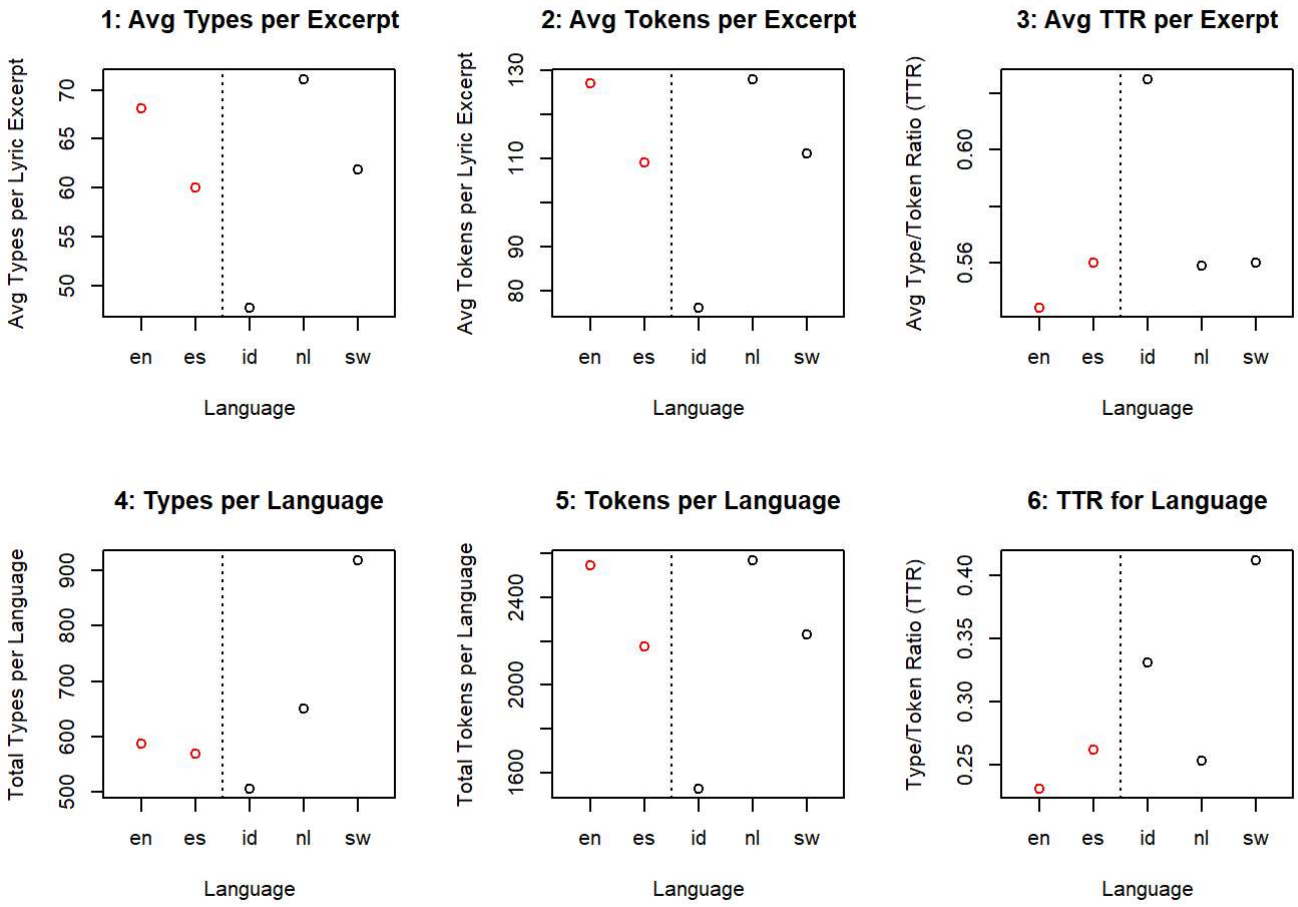
Frankly, all ten of these histograms are far from the normal distribution, but, for the extent of their use in this brief analysis, the data here will work fine. As a final check, let's plot type count against token count in scatter plots.



While these distributions aren't ideal, we just need evidence that there isn't some big bad underlying problem that we should worry is potentially corrupting the patterns that we are seeing. And based on the histograms and scatter plots above, I'm convinced. Let's continue.

Results

In the plot below, the widely spoken languages are drawn in red. The numbers used to make these plots are all taken from the tables in the **Summarizing the Data** section.



Based on Plot 6, the words used in different Swahili songs have the least in common out of any of the five languages– and by a long shot!

Based on Plot 2, it would seem that the Indonesian songs, on average, have rather few words, therefore inflating the Type/Token Ratio of individual Indonesian songs (Plot 3). When the Indonesian lyrics are pooled, the effect on TTR (Plot 6) of low token count (Plot 5) is countered by the effect of low type count (Plot 4).

Notice that, in Plots 5 and 6, Spanish has about 20 fewer types (3% fewer) and around 470 fewer tokens (15% fewer) than English, while Dutch has about 60 more types (10% more) and about 30 more tokens (1% more) than English. Still, in Plot 6, both Spanish and Dutch have a TTR that's about 0.02 to 0.03 units, or 10 to 15 percent, higher than English's TTR. In Plot 6, Spanish's TTR rose above English's TTR purely by leveraging a lower token count in the denominator. Meanwhile, Dutch's TTR had no help from *its* denominator; Dutch's TTR passed English's TTR entirely by type count superiority.

Type/Token Ratio is a measure that is often criticized for how easily it can be manipulated by corpus size, aka token count. It's likely that this quality of TTR is a key cause for Indonesian's differing behaviors in the Plots 3 and 6.

Discussion

It looks like songs written in "not widely spoken" languages might actually vary more, lyrically, than songs written in widely spoken languages, which is the absolute opposite of what I hypothesized earlier. More analysis (and actual statistical tests) would have to be done in order for any real conclusions to be drawn. Still, I've thought of a few reasons for why this pattern might have occured.

- Pop music has a more unified set of common words compared to many other genres. As opposed to artists who write lyrics in English or Spanish, artists who write songs in Swahili have a smaller target market. As a result, they may be more inclined to write song lyrics that they are individually inspired to write rather than song lyrics that will get them into the Top 40.
- The language detector is not perfect. Some songs get misclassified. Being as there was a small pool of Swahili songs to sample from in the first place, a few mistakes would make a bigger difference in Swahili than in English and Spanish.
- My reasoning was incorrect: a larger, and more geographically diverse, set of speakers is *not* affiliated with a larger number of dialects.

The behavior of Swahili was especially interesting to me. To get further insight, I contacted my cousin who lives in Tanzania. He told me that it's a trade language that unifies several countries on the East African Coast, but that, in each one of the countries where Swahili is spoken, it's a little bit different. And furthermore, the Swahili spoken around the coast differs slightly from the Swahili spoken in the interior. So maybe dialectal diversity *is* associated with some of the variation between different Swahili songs.

Acknowledgements

Thank you to Spotify, Musixmatch, Everynoise.com, and lang-detect for the help I got from your libraries, websites, and APIs in the data collection and cleaning phase of this project, which I did in Python.

Also thank you to the R libraries `quanteda`, `tidyverse`, `DescTools`, `caret`, `dplyr`, `stringi`, and `knitr`.

Appendix

Frankensteining the lyric exerpts together for each language

```
# frankenstein songs of the same language together
esinw_df_combined <- esinw_df %>% group_by(lang) %>%
  summarise("text" = paste(text, collapse = " ")) %>%
  select(doc_id = lang, text)

# And create a new corpus object based on this
lyrics_corpus_esinw <- corpus(esinw_df_combined)

# fetch a summary of the corpus composition
lyrics_sum_esinw <- summary(lyrics_corpus_esinw, tolower = T) %>% select(- Sentences)

# Choose the metrics that we want to output in the table
lyrics_sum3ds_esinw <- lyrics_sum_esinw %>%
  mutate(Text = as.character(Text),
         TTR = round(Types/Tokens,3)) %>%
  rename(Language = Text)

# output the table
knitr::kable(lyrics_sum3ds_esinw)
```