Movie Gender Dialogue Analysis

By Mack Campbell

Background Information

- Current LGS, variationist, and discourse analytic frameworks approach gender as performative and nonbinary (Sauntson 2020).
- For the sake of a quantitative approach with large amounts of data, some of these considerations are harder to apply.
- Pop culture helps perpetuate the stereotype that men and women speak differently, but that is disproven by research (Sauntson 2020).
  - To that end I will be comparing the same metrics across all gender markers and highlighting similarities and differences.

Cornell Movie Dialogue Corpus

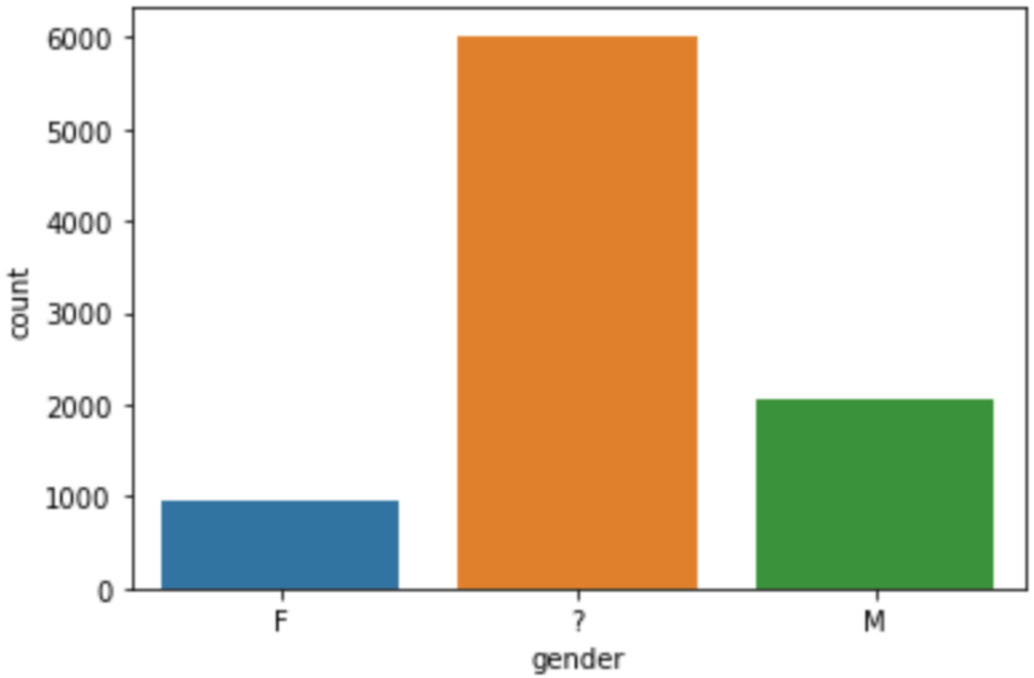
- 617 movies from 1927 – 2010
- 9,033 characters
- 304,713 utterances
- 4,181,442 tokens

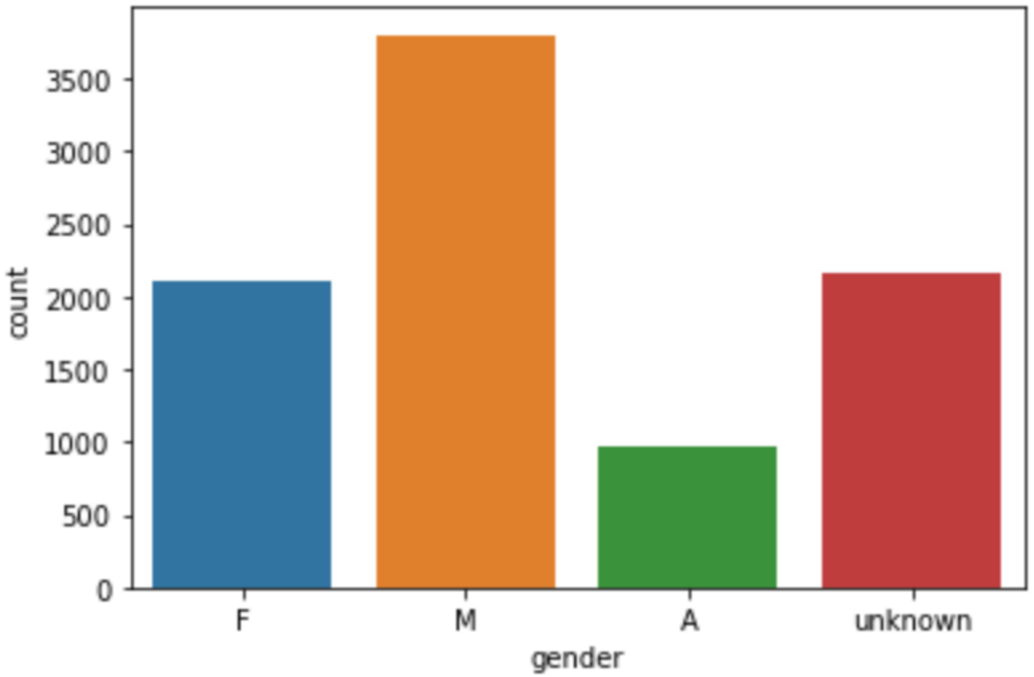
Research Questions

- How are gender roles represented across the time span of the corpus?
  - Do number and length of turns change over time?
- How do hedges and parts of speech usage differ across gender?

Generating Gender

- Out of the 9,033 characters, 6,018 of them were missing gender when I got the data set.
- By using NLTK’s names lists and looking for specific strings I was able to add gender markers for 3,859 (64%).





Names and Strings

NLTK Names

- **Male:** Aamir, Aaron, Abbey, Abbie, Abbot, Abbott, Abby, Abdel, Abdul, Abdulkarim
- **Female:** Abagael, Abagail, Abbe, Abbey, Abbi, Abbie, Abby, Abigael, Abigail, Abigale
- **Ambiguous:** Abbey, Abbie, Abby, Addie, Adrian, Adrien, Ajay, Alex, Alexis, Alfie

Strings

- **Male:** family terms (dad, son, brother), address terms (sir, Mr., Herr), positions (lord, priest, emperor), random (dude, guy)
- **Female:** family terms (mom, daughter, sister), address terms (Ms., Mrs.), positions (waitress, hostess, nun), random (chick)
- **Ambiguous:** positions (Dr., nurse, student, customer, lawyer, pilot)

Caveat About Gender

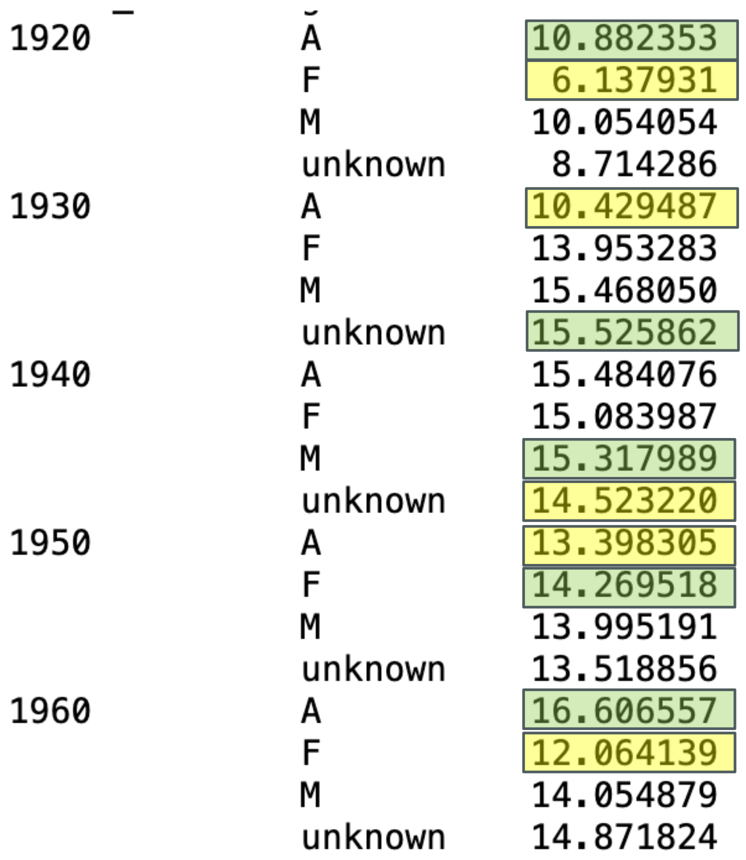
- Gender markers in the corpus do not indicate anything about the actor's or the character's actual gender presentation/identity.
- With ambiguous names and positions there is an element of a nonbinary approach to gender, but work could be improved on this area in the future.
- Most of the movies in this corpus take place before the shift in how we perceive gender reached mainstream.

Turns

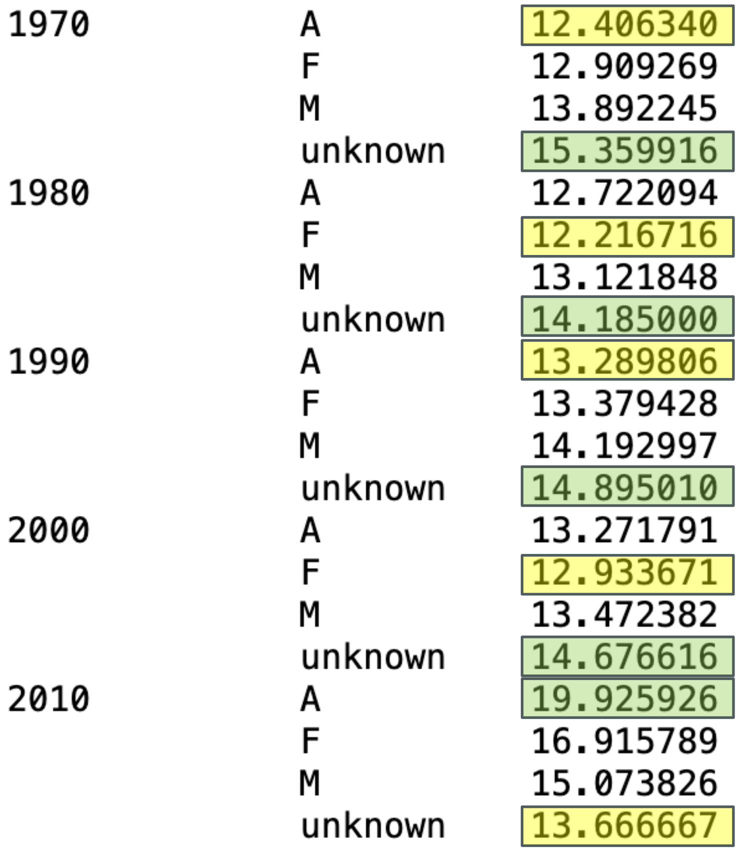
- Turns are the building blocks of conversations and are the focus of conversation analysis/discourse analysis.
- Using Halmari 1999 as a reference, longer turns can be equated to more situational power.
- I measured turns in token count, sentence count, and sentence length.
- By token count, who takes longer turns? How long is the average sentence?

Turn Data by Decade

- Average token count per turn is 13.73
- Who takes longer turns – by average token per turn



```
1920         A         10.882353
             F          6.137931
             M         10.054054
             unknown    8.714286
1930         A         10.429487
             F         13.953283
             M         15.468050
             unknown   15.525862
1940         A         15.484076
             F         15.083987
             M         15.317989
             unknown   14.523220
1950         A         13.398305
             F         14.269518
             M         13.995191
             unknown   13.518856
1960         A         16.606557
             F         12.064139
             M         14.054879
             unknown   14.871824
```



```
1970         A         12.406340
             F         12.909269
             M         13.892245
             unknown   15.359916
1980         A         12.722094
             F         12.216716
             M         13.121848
             unknown   14.185000
1990         A         13.289806
             F         13.379428
             M         14.192997
             unknown   14.895010
2000         A         13.271791
             F         12.933671
             M         13.472382
             unknown   14.676616
2010         A         19.925926
             F         16.915789
             M         15.073826
             unknown   13.666667
```

Part of Speech

- Tagged using NLTK’s POS tagger (moving to SpaCy)
- Adverbs (RB, RBR, RBS)
- Adjectives (JJ, JJR, JJS)
- Coordinating conjunction (CC)
- Interjection (UH)

Part of Speech Data

- Jupyter Notebook demonstration

Hedges

- Used to be less assertive or less certain in an utterance.
- Hedge list:
  - I guess
  - I think
  - Maybe
  - Might
  - Perhaps
  - Possibly

Next Steps

- Get SpaCy POS tagger up and running
- Run analysis on hedging
- Linear regression on turn length over time

Thanks!

Questions, comments, concerns?

References

- Halmari, H. (1999). Power relationships and register variation in Väinö Linna’s *Here under the Northern Star*. *Journal of Finnish Studies*, *3*(2), 36-49.
- Saunston, H. (2020b). *Researching language, gender and sexuality: A student guide*. Routledge.