

Quantitative Corpus Analysis

...

Linguistic Differences in HC Andersen and Grimm Fairy Tales

Overview

- Background
- Research Questions
- Data pre-processing
- Data processing
- Analysis & Results
- Conclusion

Background: fairy tales of H. C. Andersen

- Danish author most known for his literary fairy tales
- 156 stories, 9 volumes
 - individualistic, original stories
- reflective and introspective quality, societal commentary
- motifs: transformative power of love, friendship, empathy
 - (O. Althobaiti, 2023)
 - focus on character development

Background: fairy tales of Jacob and Wilhelm Grimm

- German linguists and storytellers
- Preservation of traditional folk tales from Germany
 - literary aspirations but also deep sense of national pride
- collection: orig. “Children’s and Household Tales”
- motifs: transformative power of goodness, triumph of virtue, resilience
- romanticism

(O. Althobaiti, 2023)

→ moral teachings to reinforce societal virtues and values

Background: examples of fairy tales

Hans Christian Andersen

The Brave Tin Soldier

The Ugly Duckling

The Little Match-Seller

The Loveliest Rose in the World

The Philosopher's Stone

The Toad

Jacob and Wilhelm Grimm

The Three Spinners

Rumpelstiltskin

Hans in Luck

The Devil's Sooty Brother

The Devil and His Grandmother

The Wolf and the Man

Research Questions

Linguistic complexity analysis instead of thematic comparison

Research question 1

How does the language of two of the most influential European authors of fairy tales differ in linguistic complexity?

Research question 2

How does the sentiment of the language of the fairy tales by Hans Christian Andersen and by the Grimm brothers reflect their varying motifs?

Hypothesis:

1. Due to their more individualistic and imaginative nature, the stories of H.C. Andersen should be linguistically more complex than those of the Grimm Brothers.
2. Due to the emphasis on traditional values and moral excellence, the stories of the Grimm brothers should display more dominance than those of H.C. Andersen. Andersen due to his focus on love should show a more positive sentiment.

Data pre-processing

H.C. Andersen fairy tales

- 142 tales (1835 - 1873)
- translated by Susanna Mary Paull 1872; Horace Elisha Scudder (14 tales)
- Almost all works included
- automatic text scraping with python script
- **subcorpus size:**
389469 tokens
101086 types

Grimm fairy tales

- 209 tales (1812 - 1850)
- translated by Margaret Hunt 1884
- whole collection
- automatic text scraping with python script
- **subcorpus size:**
322853 tokens
83070 types

Data Processing

- Analysis conducted with R, programming language for statistical computing.
- Packages: `quanteda` (text analysis and statistics), `spacy` (tagging and parsing), and `koRpus` (readability and lexical metrics).
- Texts were turned into a data frame with the number of types, tokens and sentences of each work as variables (example in image)

```
18 ##### reading grimm
19 inputpath = paste(getwd(), "/data", sep = "")
20 df_grimm = read.xlsx(paste(inputpath, "/Corpus_grimm_with_text.xlsx", sep=""))
21
22 text = c()
23 i = 1
24 for (i in 1:length(df_grimm$text)){
25   text[i] = df_grimm$text[i]
26 }
27
28 grimm_corpus = corpus(text)
29 summary(grimm_corpus)
30
31 docvars(grimm_corpus, 'title') = df_grimm$title
32 docvars(grimm_corpus, 'author') = 'Grimm'
33
34 df_grimm = summary(grimm_corpus, n=Inf)
```

Data Processing

quanteda:

- Sum of total number of tokens
- MSTTR (Mean Segmental Type-Token Ratio) → not sensitive to lengths of fairy tales
- Flesch-Kincaid readability score

spacy:

- text lemmatized and tagged with its part-of-speech
- removed part-of-speech that might skew measurements: punctuation, numerals, symbols and spaces.

R functions:

- hapax (words that only occur once)
- lexical & grammatical density (percentage of the sum of nouns, adjectives, proper nouns, adverbs and verbs)
- syntactic complexity with Fichtner's C (the number of verbs, times the number of words, divided by the number of sentences squared)

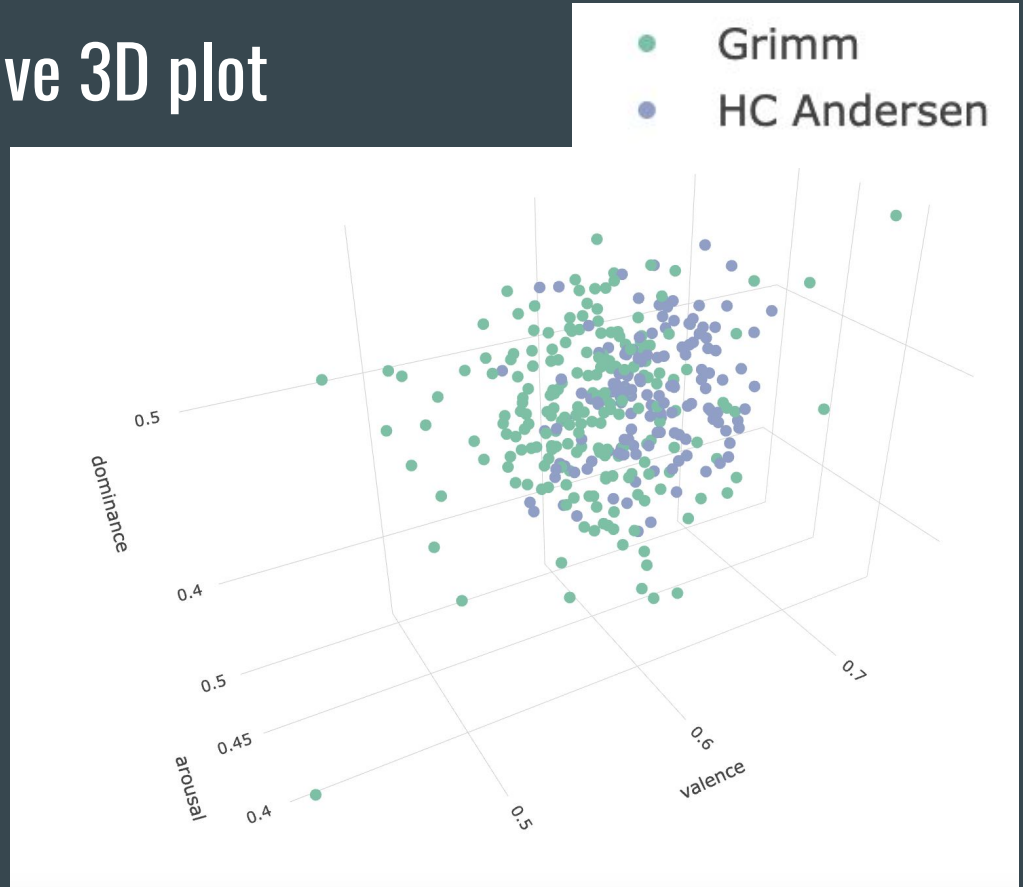
Data Processing

Sentiment analysis:

- NRC-VAD lexicon for English (Mohammad, 2018)
- lexicon contains 20,000 words tagged with values for three dimensions to model human emotions - valence, arousal and dominance (VAD)
- **Valence:** how positive or negative
- **Arousal:** how calm or exciting
- **Dominance:** how submissive or dominating
- for each text, each emotional dimension value was summed up for all tokens, divided by the total number of lemmas for each text
- The averages for each of the three emotional dimensions were then added as variables to the data frame

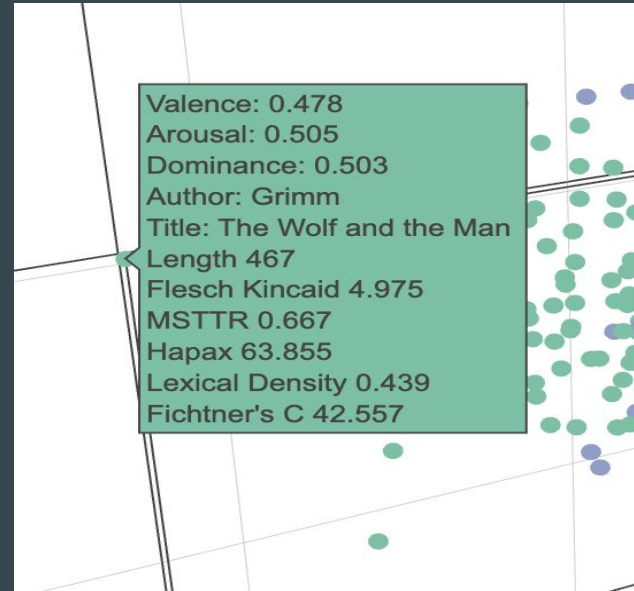
Analysis & Results - Interactive 3D plot

- 3d-space with emotions as axis
- Every text tagged with average complexity and emotion score
- Fairy tales readable on click



Example - The Wolf and the Man

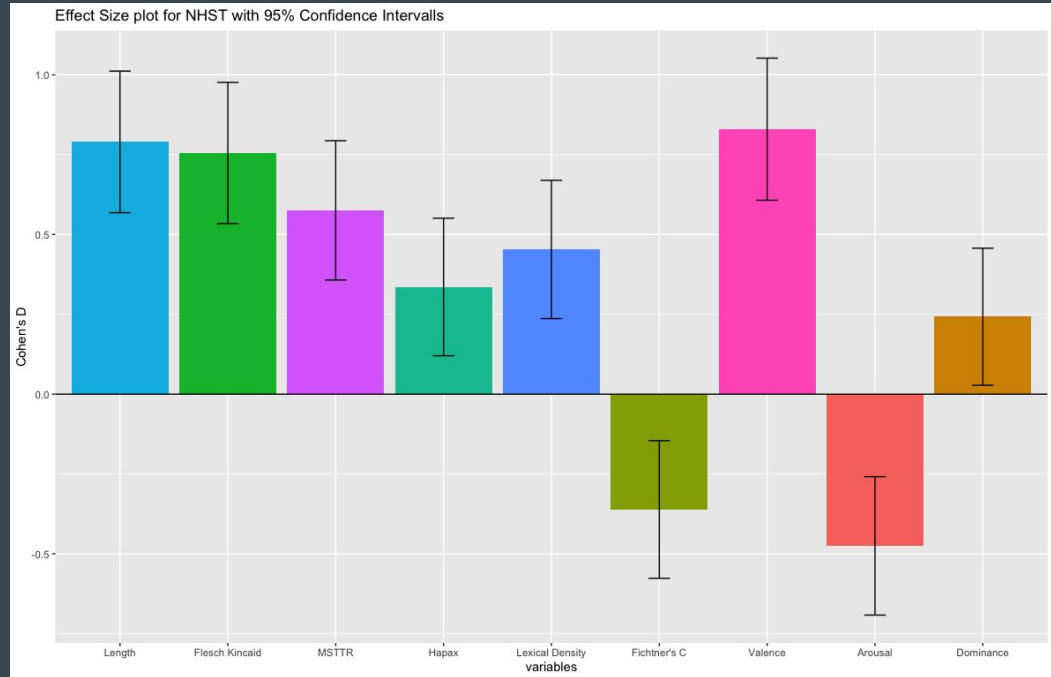
- low valence, high arousal and mid-dominance



- “When the huntsman saw him he said, it is a pity that I have not loaded with a bullet, aimed, and fired his small shot in his face. The wolf pulled a very wry grimace, but did not let himself be frightened, and attacked him again, on which the huntsman gave him the second barrel”

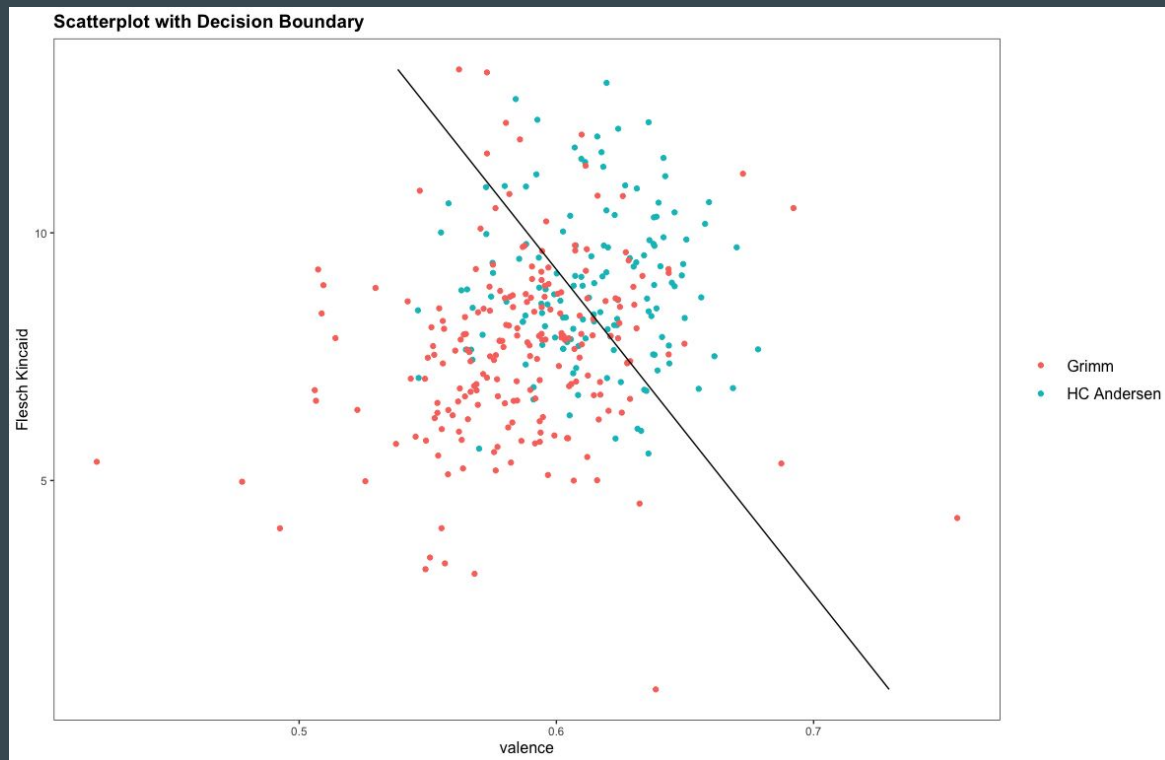
Analysis & Results - Null Hypothesis Significance Testing

- Effect Size: Cohen's D
- Valence largest effect
- Andersen writes complexer
- Grimm scores higher in syntactical complexity and Arousal



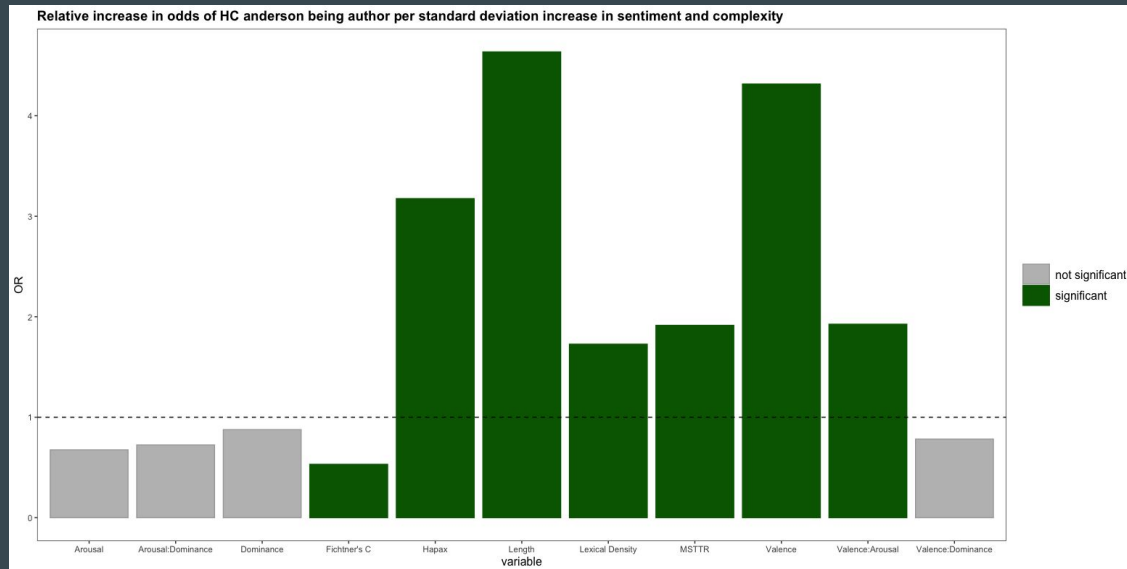
Modeling

- Logistic regression to predict authorship
- Based on complexity measures and sentiment
- Adding interaction terms between sentiment variables
- Removing Flesch Kincaid due to multicollinearity



Odds Ratios

- 1 SD increase in valence/length \rightarrow 4 times increase in odds of the text being written by Andersen



Conclusion

- Andersen wrote longer and more complex stories
→ supports the idea of him being a more independent writer
- Significant and relatively large difference in Valence
→ supports hypothesis of love characterizing the stories of H.C. Andersen
- Grimm stories score higher in Fichtner's C
→ possibly an effect of German being the source language for translations
- Grimm score lower in dominance and higher in arousal
→ Grimm stories could be more “action-filled” rather than moral preaching

Thank you for your attention!

References cited in the presentation:

Mohammad, Saif. (2018) "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words."
Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers).

O. Althobaiti M.A. (2023). The Evolution of European Fairy Tales: A Comparative Analysis of the Grimm Brothers and Hans Christian Andersen. ESI Preprints.

HCA texts: <http://hca.gilead.org.il/#list>

Grimm texts: <https://www.cs.cmu.edu/~spok/grimmtmp/>

A complete list of references can be found in our project report

GitHub repository:



Interactive 3D Plot:

