Annalea Luther, Jeremy Moser, & Andrin Müller

Annalea.luther@students.unibe.ch | Jeremy.moser@students.unibe.ch | Andrin.müller@students.unibe.ch

Bachelor Minor Digitalization and Applied Data Science

University of Bern

Text as Data and Automated Content Analysis

Dr. Ernesto de León

The Formality Of Cinematic Discourse

An Exploration of Formality Trends in Movie Scripts Using Computational Analysis

Table of Contents

[Table of Figures 1](#_Toc160554578)

[Introduction 2](#_Toc160554579)

[Theoretical Framework 2](#_Toc160554580)

[Impact of Cinema On Society 2](#_Toc160554581)

[Language and Society 3](#_Toc160554582)

[Formal Vs. Informal Language 3](#_Toc160554583)

[Methodology 4](#_Toc160554584)

[Data 4](#_Toc160554585)

[Analysis 4](#_Toc160554586)

[Distribution of the formality score 5](#_Toc160554587)

[Results 6](#_Toc160554588)

[Conclusion 7](#_Toc160554589)

[Bibliography 8](#_Toc160554590)

[Literary Sources 8](#_Toc160554591)

[Data Sets 9](#_Toc160554592)

# Table of Figures

[Figure 1: Formality Score Distribution 5](#_Toc160553965)

[Figure 2: Score of all Movies Over the Years 6](#_Toc160553966)

[Figure 3: Score of all Movies Decade-wise 6](#_Toc160553967)

[Figure 4: Score per Genre over the Years 7](#_Toc160553968)

# Introduction

Movies have long been a cornerstone of modern media, shaping societal perceptions and behaviors through their influence. In recent years, concerns have been raised regarding the evolving tone of language within cinema and its potential impact on cultural norms (Cressman et al., 2009). Thus, this paper seeks to address the research question:

Has the tone of language (formal/informal) changed in cinema over the last century?

Drawing on theories of media cultivation and the relationship between language and society, we aim to understand how formal language has evolved in cinema. We used a machine learning model trained to differentiate between formal and informal language with a diverse dataset of movie dialogues, political speeches, and subreddits. Our analysis focused on examining trends in language formality across time and genres.

Our research reveals a trend towards informality in movie language over the last century. The observed decline in language formality raises concerns about language learning and proficiency, particularly in academic settings. With this analysis, we offer insights into the changing formality of language within cinema and its broader societal implications.

# Theoretical Framework

## Impact of Cinema On Society

Movies play a central role in our daily lives, with research showing that children and adolescents typically spend approximately 2 to 3 hours watching TV each day (Nielsen Media Research, 2000). This exposure to movies seems to have a considerable impact on an individual's behavior and cultural norms.

One theoretical framework that delves into the impact of media and television on cultural perceptions is the cultivation theory, formulated by Gerbner and colleagues (1986). They put forward the concept that media sources shape an individual’s views of social reality (Gerbner et al., 1986). Central to this theory is the notion that individuals learn from their interactions within a social context, with observed behaviors influencing their behavior (Cressman, 2009; Nabavi, 2012).

Movies are believed to play a profound role in this cultivation process. While direct studies on the influence of movies on language are limited, many studies have highlighted their efficacy in language learning contexts (Pegrum et al., 2005; Shcherbakova & Filippova, 2022; Omar & Razi, 2022). This suggests that variables within language in cinema may indeed shape individuals' speech and language patterns.

## Language and Society

The way we talk matters. Language serves as a fundamental medium through which individuals interact, express themselves, and even shape societal attitudes and norms (Wardhaugh & Fuller, 2021). This intricate relationship between language and society has been a subject of research in linguistics for decades (Mgbemena & Ewurum, 2017).

Various empirical studies have demonstrated the impact of language on societal attitudes and behaviors. One study has shown that language influences perceptions of color (Kay et al., 2009), while another study showed that language in job advertisements can reinforce gender biases, impacting who applies and perpetuating discrimination (Bem & Bem, 1973). Moreover, the use of gender-neutral language has been found to mitigate gender bias and promote equality in mass opinion (Tavits & Pérez, 2019).

## Formal Vs. Informal Language

While the analysis of language is very complex, the distinction between formal and informal language can be made. Formality is influenced by various factors, including situational context, social norms, and individual preferences (Heylighen & Dewaele, 1999). For our definition of formal and informal language, we relied heavily on the definitions by Hopkins (2022). The following list shows characteristics that indicate informal language.

**Colloquialisms**: Examples include "gonna" instead of "going to”, “wanna” and “want to”

**Slang**: Expressions that are specific to particular social groups or subcultures.

**Contractions**: Omitting specific words, mostly the subject, such as "I'm" for "I am," or "won't" for "will not."

**Interjections and Exclamations**: Words to express emotions or emphasis, such as "Wow!", and "Oops!.

# Methodology

## Data

For model training, we used different datasets. Such as the "movie-corpus" from the Python toolkit Convokit (Danescu-Niculescu-Mizil & Lee, 2011), a collection of conversations from over 600 movie scripts. We scraped informal language samples from the subreddit r/ask (https://www.reddit.com/r/ask/), known for its informal discourse and without a specific direction of topics. For formal language data, we turned to the Hansard dataset from R, which provides formal speeches from the UK Parliament (the SAMUELS project, 2011).

Subsequently, a manual labeling process was undertaken, classifying 700 dialogues as informal and another 700 as formal. As well as labeling the Subreddit posts and speeches with the respective formality. Utilizing the Quanteda library, the 1,400 labeled dialogues together with the Subreddit and Hansard data sets were partitioned into an 80% training dataset and a 20% testing dataset. A supervised machine-learning model was then developed to determine the formality of a dialogue. The performance metrics of our model are shown in the following figure:

|  |  |
| --- | --- |
| Accuracy | 90% |
| Precision | 81.15% |
| Recall | 94.29% |
| F1 | 94.29% |

## Analysis

We used the model on the unlabelled movie data set. Every individual dialog was categorized into “informal” or “formal”. Given the binary nature of this classification, we derived a formality score for each movie based on the following equation:

This ratio yields the percentage of formal dialogues in a movie, thereby providing a representative formality score.

For our analysis we wanted to study specifically (a) the median formality score of all movies over the years, (b) the median formality score decade, and (c) the median formality score for each genre over the years. The following manipulations were made to reduce biases in each analysis:

1. Removing all years with less than seven movies (in that year)
2. Removing all decades with less than 20 movies (in that decade)
3. Removing all categories in a year with less than 5 movies (per category) and removing the categories “film-noir”, “short”, “adult”, “sport”, “western” and “comedy” due to underrepresentation and then we combined individual genres into broader categories, combining:
   1. Action and Adventure
   2. Drama and Romance
   3. Sci-Fi and Fantasy
   4. Crime, Mystery, and Thriller
   5. Musical and Music
   6. Biography and Documentary
   7. Animation and Family
   8. History and War

## Distribution of the formality score

We examined the distribution of formality scores across all movies. The mean value is approximately 0.044, while the median value is approximately 0.027, indicating the presence of outliers skewing the mean value upwards—a characteristic of a right-skewed distribution. This observation, illustrated in Figure 1, prompted our decision to utilize the median value for further analysis, ensuring a more robust measure of central tendency.

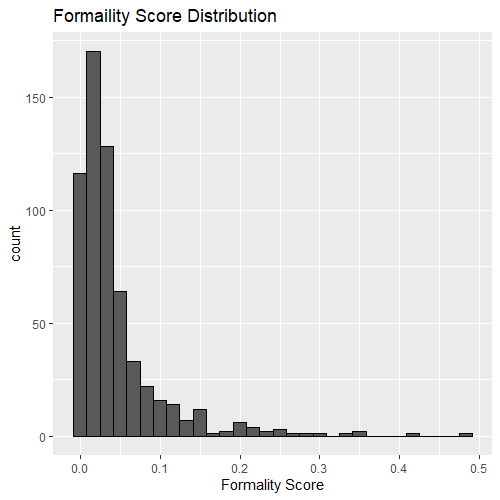
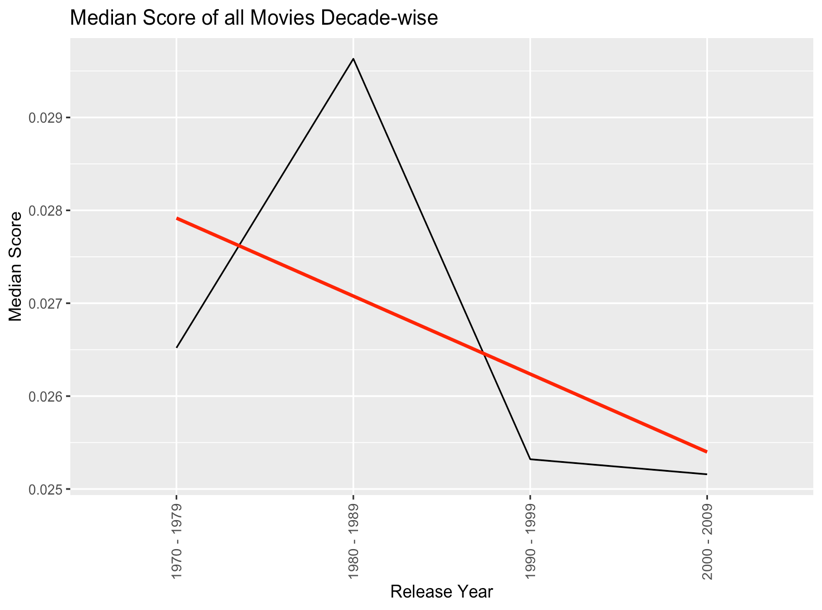


Figure 1: Formality Score Distribution

# Results

Our research sheds light on the evolving trends in language formality within movies. Using a machine learning model with an accuracy rate of 90%, we found an informality trend across movie dialogues in the last century. The relationship between language formality and time does not appear to be strictly linear, with fluctuations observed over different periods.

A graph showing the release of movies

Description automatically generatedThere are some notable deviations in the model such as in the 1980s (see Figure 3) as well as in the year 1990 (see Figure 2). These exceptions could be explained by various factors, like fluctuations in movie releases or shifts in cultural norms and cinematic trends at that time. Some interesting insights can also be gained by splitting the movies into genres, with the highest average formality score in the genre action/adventure and the lowest in drama/romance (see Figure 4).

Figure 2: Score of all Movies Over the Years

Several limitations must be considered. Firstly, our dataset is merely a sample of movies and does not necessarily represent all the movies produced accurately. Secondly, the data used to train our model might not fully meet the criteria of formal and informal language, as not every data entry was manually labeled.

Figure 3: Score of all Movies Decade-wise

Lastly, the study looks at language formality based on a binary classification of formal and informal, but language complexity consists of various factors beyond this binary distinction, such as vocabulary diversity, syntactic structures, and discourse patterns. Despite these limitations, our research provides valuable insight into how language in movies has changed over the years.

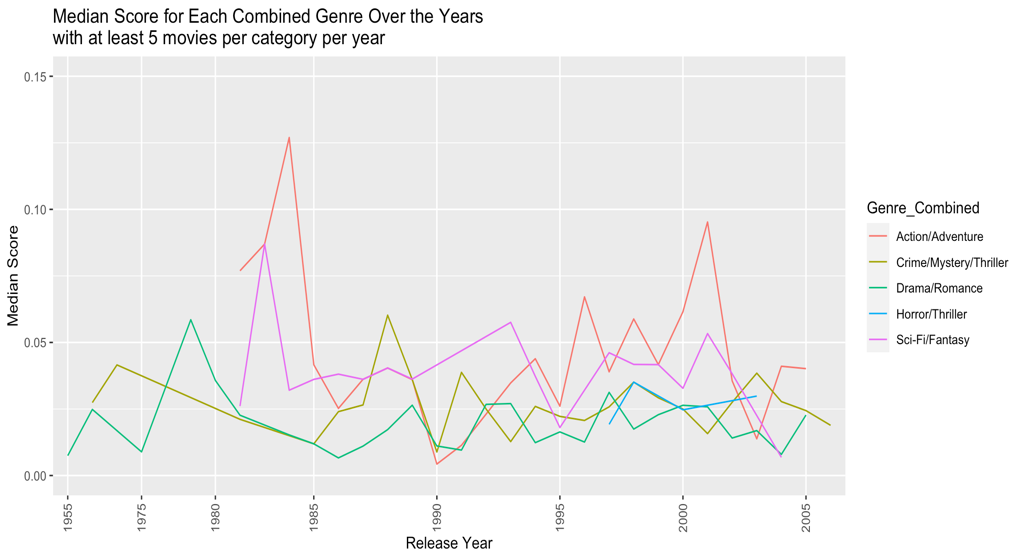


Figure 4: Score per Genre over the Years

# Conclusion

We addressed our research problem by employing a machine learning model to examine language formality trends in movies. Through our analysis, we discovered indications of a decline in language formality over time, which we hypothesize may be attributed to the commercialization of cinemas. As movies increasingly cater to mass audiences and commercial interests, there may be a tendency to prioritize informal language to appeal to broader demographics. These findings hold significant implications for various fields, particularly in education and language learning (Al-Salman & Saeed, 2017; Johannesson, 2022). With a decline in formality language proficiency, especially in academic contexts could heavily decline.

# Bibliography

## Literary Sources

Al-Salman, S. M., & Saeed, A. T. (2017). Effects of text-messaging on the academic writing of Arab EFL students. *Research in Language*, *15*(3), 237-252.

Bem, S. L., & Bem, D. J. (1973). Does sex‐biased job advertising “aid and abet” sex discrimination? 1. *Journal of Applied Social Psychology*, *3*(1), 6-18.

Cressman, D. L., Callister, M., Robinson, T., & Near, C. (2009). Swearing in the cinema: An analysis of profanity in US teen‐oriented movies, 1980–2006. *Journal of Children and Media*, *3*(2), 117-135.

Gerbner, G., Gross, L., Morgan, M., & Signorielli, N. (1986). Living with television: The dynamics of the cultivation process. *Perspectives on media effects*, *1986*, 17-40.

Heylighen, F., & Dewaele, J. M. (1999). Formality of language: definition, measurement and behavioral determinants. *Interner Bericht, Center “Leo Apostel”, Vrije Universiteit Brüssel*, *4*(1).

Hopkins, R. (2022). Formal and Informal Language. In *Educating for Action* (pp. 120-136). Brill.

Johannesson, E. (2022). Informal Language Online-A Danger to Students' Formal Language Awareness?: A Quantitative Study on Informal Language Online and Upper Secondary Students’ Awareness of Informal and Formal Language Features in Relation to Extramural English.

Kay, P., Berlin, B., Maffi, L., Merrifield, W. R., & Cook, R. (2009). *The world color survey*. Stanford, CA: CSLI Publications.

Mgbemena, J. A., & Ewurum, R. (2017). A Sociolinguistic Study of the Language of the Announcements of Obituary and Obsequies. *Ahyu: A Journal of Language and Literature*, *1*, 9-22.

Nabavi, R. T. (2012). Bandura’s social learning theory & social cognitive learning theory. *Theory of Developmental Psychology*, *1*(1), 1-24.

Nielsen Media Research. (2000). *Report on television*. Nielsen Media Research.

Omar, F. R., & Razı, Ö. (2022). Impact of instruction based on movie and TV series clips on EFL learners’ pragmatic competence: Speech acts in focus. *Frontiers in Psychology*, *13*, 974757.

Pegrum, M., Hartley, L., & Wechtler, V. (2005). Contemporary cinema in language learning: From linguistic input to intercultural insight. *Language learning journal*, *32*(1), 55-62.

Shcherbakova, O., & Filippova, N. (2022). Neurolinguistic and Semiotic Film Structure Modelling as the Linguistic Student Engagement Technique.

Tavits, M., & Pérez, E. O. (2019). Language influences mass opinion toward gender and LGBT equality. *Proceedings of the National Academy of Sciences*, *116*(34), 16781-16786.

Wardhaugh, R., & Fuller, J. M. (2021). *An introduction to sociolinguistics*. John Wiley & Sons.

## Data Sets

Danescu-Niculescu-Mizil, C, & Lee, L. (2011). Cornell Movie-Dialogs Corpus. *Cognitive Modeling and Computational Linguistics Workshop* [Data Set].https://www.cs.cornell.edu/~cristian/Chameleons\_in\_imagined\_ conversations.html

The SAMUELS project. (2016). The Hansard Corpus. [Data Set].https://www.hansard-corpus.org/.