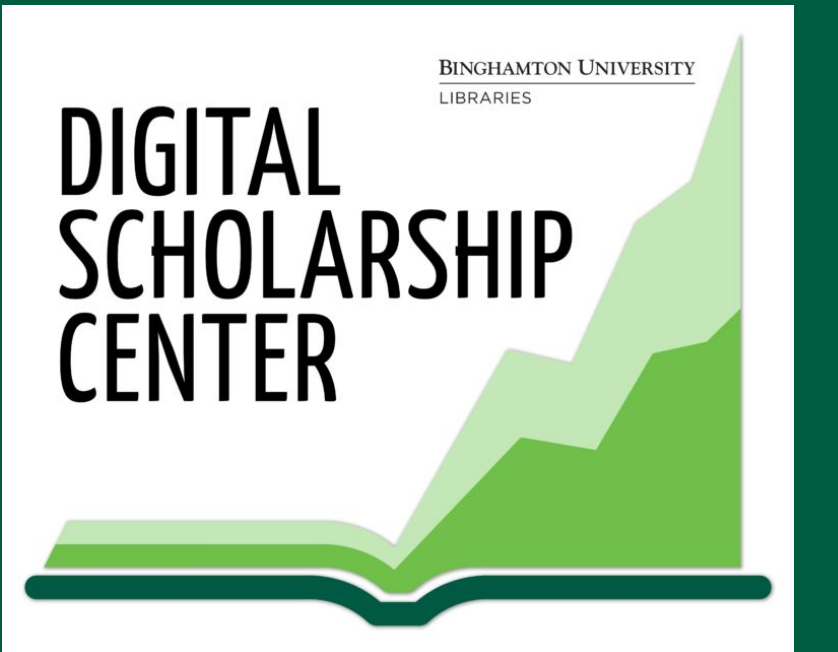


# Chronicling language, gender, and sexuality research through digital text analysis

**BINGHAMTON**  
UNIVERSITY  
STATE UNIVERSITY OF NEW YORK

**Aidan L. Niebauer, Ruth Carpenter & Halie Kerns**  
Libraries Research Scholars Program



## Background

This project is a historical analysis of language, gender, and sexuality (LGS) research using the journal *Language in Society*, a prolific publication in the fields of sociolinguistics and linguistic anthropology. Using a corpus of 4,190 documents from the journal from 1970 to the present day, the project aimed to answer the following research questions:

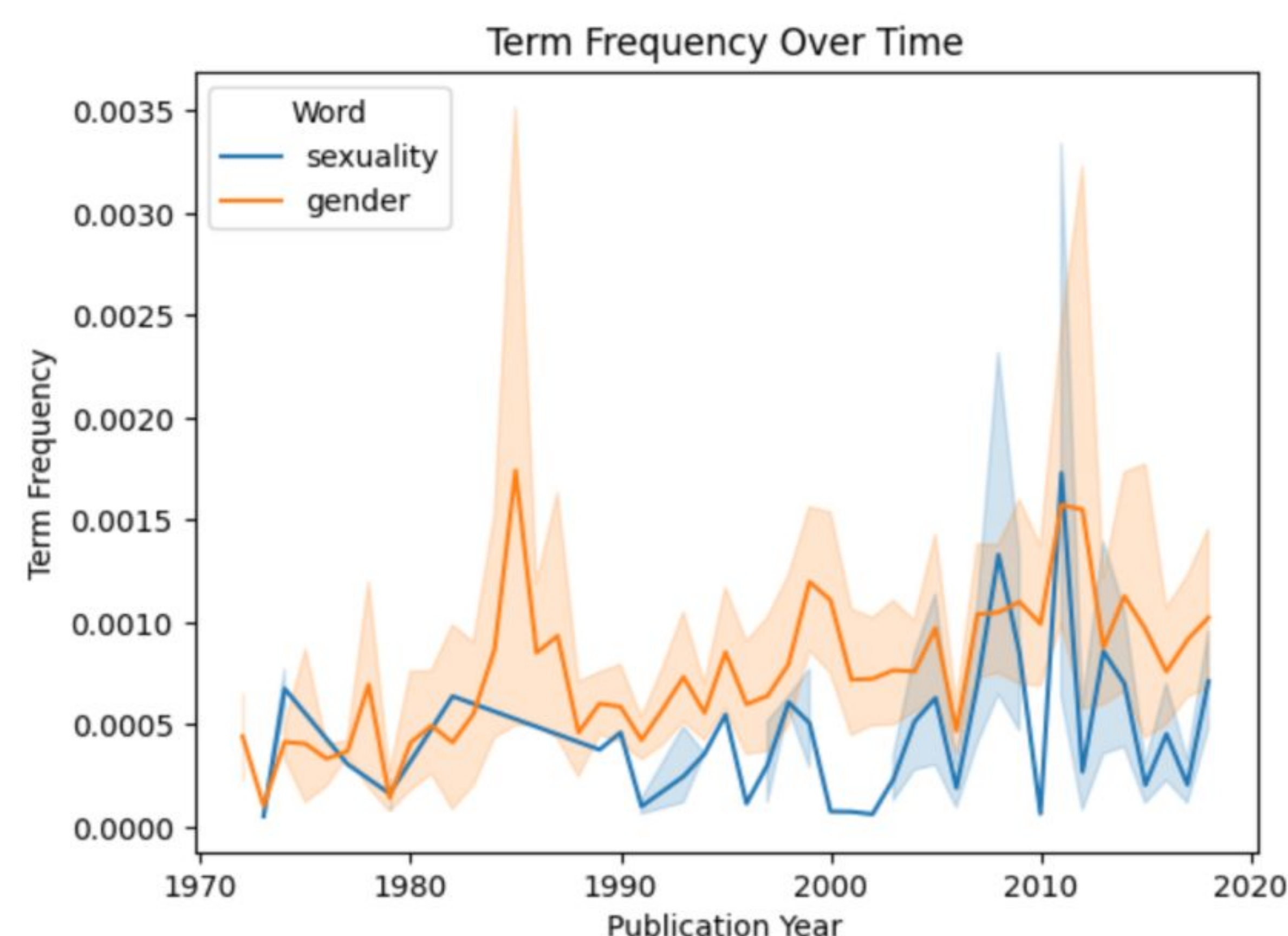
1. How has the field of language, gender, and sexuality been historically represented over the life of the journal?
2. How has language regarding sexual and gender identity evolved over the life of the journal?

## Methods

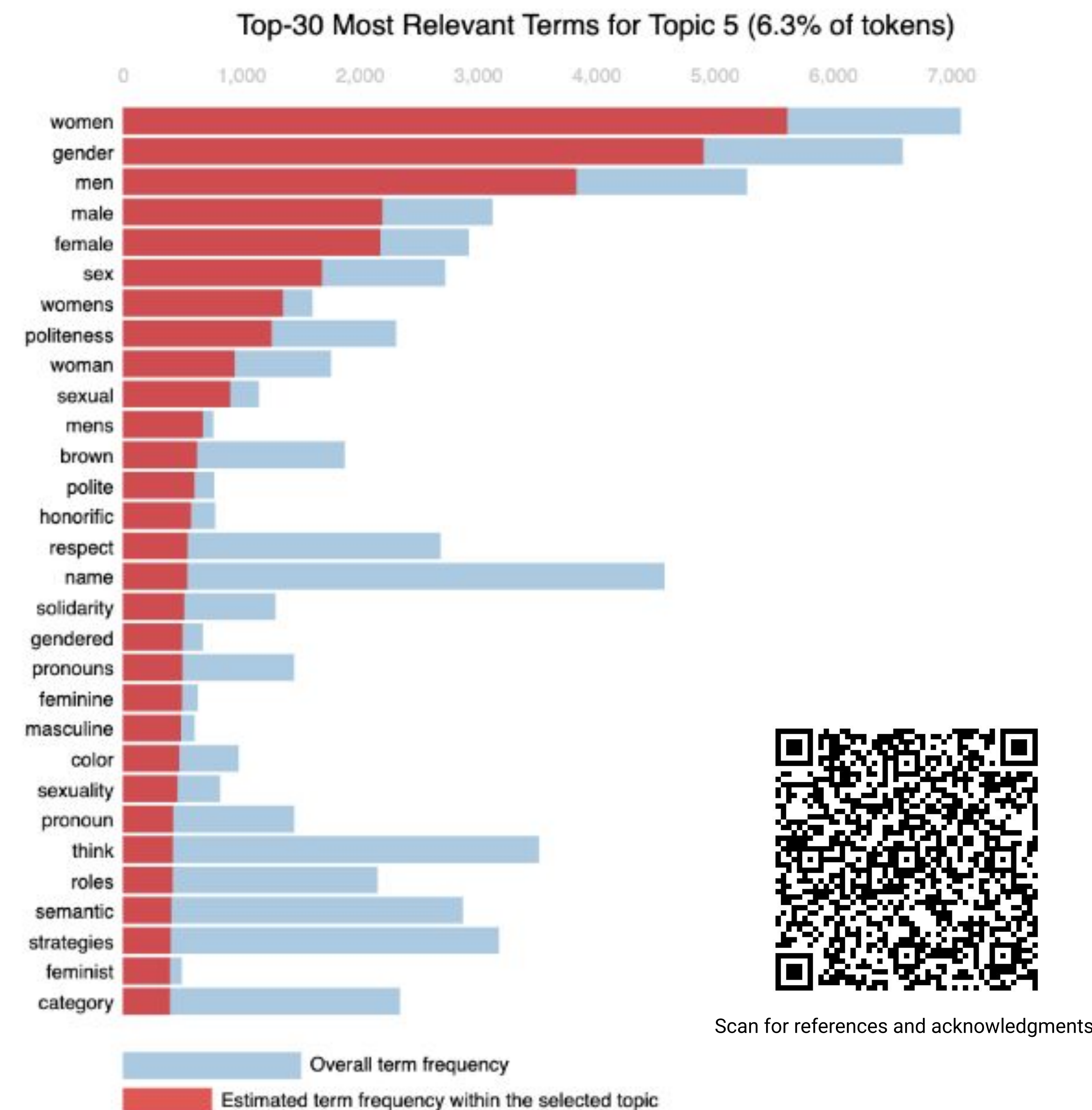
The data analysis was completed using Python. Several Python libraries were used to analyze and visualize the data, such as *NumPy*, *Matplotlib*, *pyLDAvis* and *scikit-learn*. The research involved extensive organization and cleaning of the text, such as utilizing “stop words”, cleaning undefined or missing values in the dataset, deleting whitespace and trailing characters.

The main technique used was **Latent Dirichlet allocation (LDA)**, a topic modeling technique used to uncover trends and topics in a corpus. Table 1 utilizes LDA to create a list of topics featured and their attributed saliency, and Table 3 utilizes a temporal LDA model to track general topics of the journal over its history.

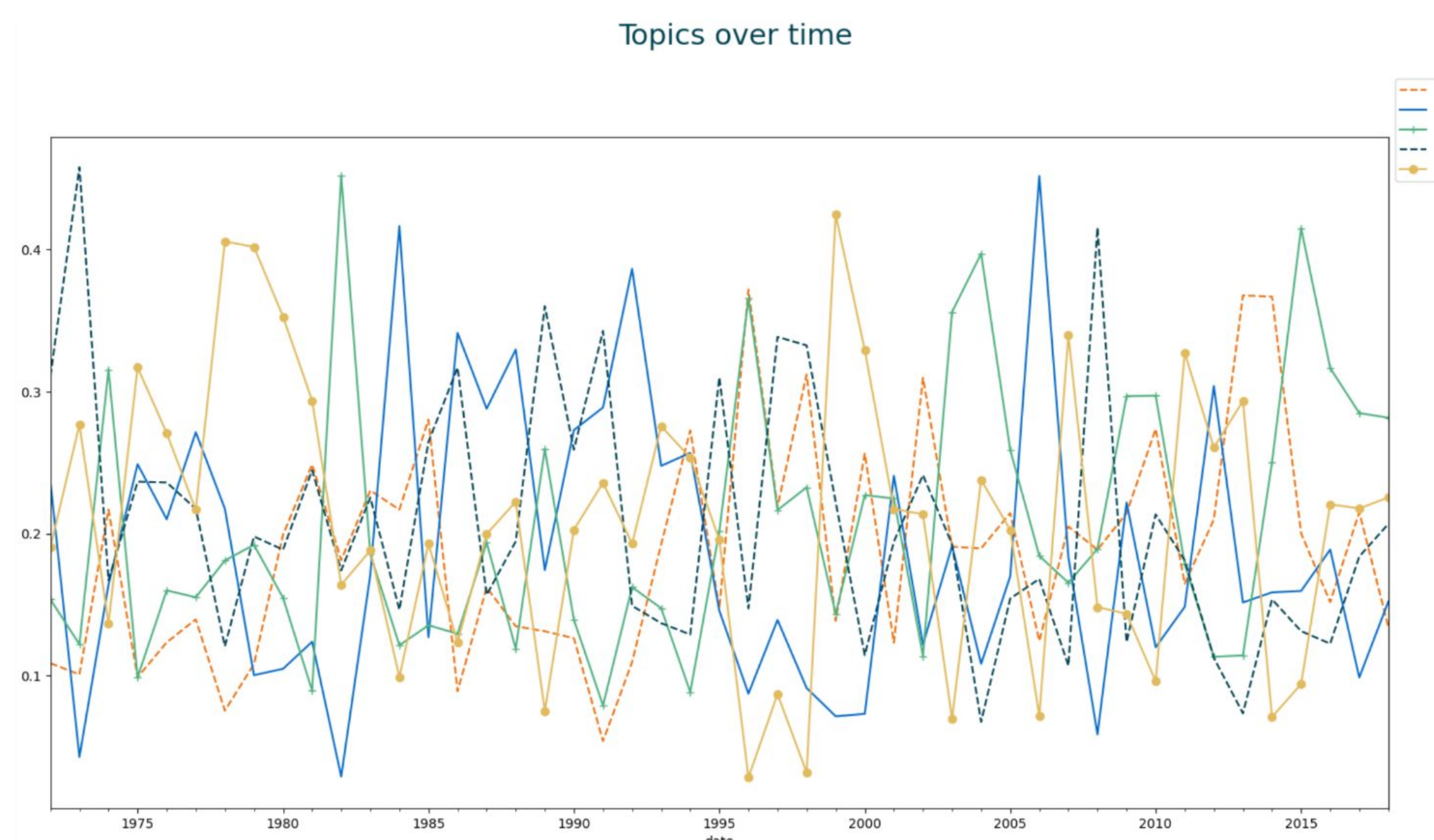
The second technique used was **TF-IDF (Term Frequency - Inverse Document Frequency)**, which tracks the distinct frequencies of terms in a corpus. This allowed us to trace word frequencies over time. (Table 2)



**Table 2.** Frequency of the terms “gender” and “sexuality” in the journal



**Table 1.** LDA analysis; blue represents overall frequency of the term and red represents saliency to the topic



**Table 3.** Historical analysis of topics based on article titles; 1 (orange) = language and gender; 2 (blue) = language and identity; 3 (green) = research methods; 4 (dark green) = language ideologies; 5 (yellow) = variationist sociolinguistics

## Results

Table 1 represents the salience (or relevance) of the terms in a set of articles focusing on LGS. The data reveals a prominence of the terms “women” and “gender” over “men” and “male”. This reflects the historical foci of the field; in the 1970s, the field was born (Lakoff 1973)<sup>2</sup>, and research focused on the dominance approach to LGS, which stated that men and women speak differently because of men’s dominant role in society.<sup>3</sup> The analysis reveals that women’s speech and the dichotomy of gender were primary areas of analysis as opposed to men’s speech and masculinity.

Table 2 also highlights the prevalence of the topic of gender as opposed to “sexuality” in the history of the field. While gender has remained a prevalent topic of discussion, sexuality was not as prevalent in the data until the 2010s, with the emergence of subfields such as queer linguistics and trans linguistics (thanks to the efforts of Lal Zimman).

Table 3 represents the prevalence of specific topics in article titles. As shown, LGS has remained prevalent in the journal, with certain influxes over key periods of scholarship (e.g. Scott Kiesling’s work on masculinity in the 1990s.)

## Discussion

The analysis shows the salience of women and gender roles to the formation of the field and its ongoing prevalence, yet the data does suggest shortcomings in scholarship regarding studies of masculinity and sexuality. Sexuality was not relevant in the field until the 1990s. Before this, studies focused only on gender differences or considered sexuality to be a category under gender. Studies that focused solely on masculinities and men as a topic were not as prolific until the 1990s and 2000s, with scholars like Scott Kiesling (1997<sup>4</sup>, 2001<sup>5</sup>, 2004<sup>6</sup>) and Deborah Cameron (1997<sup>7</sup>) advancing studies in masculinities.

## Conclusion and next steps

Overall, LGS has established itself as a distinct subfield in sociolinguistics and linguistic anthropology, but its history focused predominantly on women’s speech and women’s gender roles. Further studies can focus on prevalence of masculinities, sexuality, and the emergence of queer and trans linguistics within the field.

**Limitations:** This dataset consists of texts solely from the journal *Language in Society*, which features a broad range of topics in sociolinguistics and linguistic anthropology. These methods also generalize the data by focusing solely on word frequency rather than contextualized data.