

# Character Comparison of Bart, Lisa, and Marge in The Simpsons

Ryan Hull<sup>1</sup>, Nicholas Imperioli<sup>1</sup>, Youmin Son<sup>1</sup>

<sup>1</sup>McGill University, COMP 370 Introduction to Data Science

ryan.hull@mail.mcgill.ca, nicholas.imperioli@mail.mcgill.ca, youmin.son@mail.mcgill.ca

## Introduction

Our team has been hired by a production company to analyze the dialogue patterns of side characters in The Simpsons to understand what topics they discuss and how much attention they give to each. Scripts from 61 episodes of The Simpsons were analyzed to conduct a topic analysis of spoken words from the iconic show's most prevalent side characters. Through extensive data cleaning, structured manual annotation, and TF-IDF/LLM analysis, these characters' speech patterns were compared and contrasted based on a comprehensive 8-topic categorization system. Certain topics (events, objects, locations) were given equal attention by Bart, Lisa, and Marge. Others (different character types, emotions, and opinions) served as the most useful topics in differentiating these side characters' speech, helping lay out the following telling traits: (1) Bart's speech is the most self-centered, while he is the most diversely social, (2) Marge is most comfortable conversing about emotions and sharing opinions, and (3) Lisa occupies a middle ground in all these aspects. These findings aid the production of The Simpsons as they provide a base from which to ensure a realistic and interesting character progression of these important Simpsons family members.

## Data

### Data Collection and Sources

We used 61 raw Simpsons episode scripts in .txt format, all pulled from a nohomers.net post. This post was made by a super fan in order to share scripts of the show we all love. This collection represents one of the most extensive publicly available archives of the show's dialogue, spanning multiple seasons and providing a substantial amount of dialogue that can be used for character analysis.

### Script Selection and Preprocessing

Among available scripts in the nohomers.net collection, not all were in the same draft format. Scripts existed in various stages of production, including final shooting scripts, revised drafts, and earlier draft versions. To maximize data

quality and consistency, final versions were preferentially chosen whenever available as these represent the dialogue that actually aired and thus most accurately reflects the characters as the audience knows them. When final versions were unavailable, draft and revised draft scripts were randomly selected to obtain sufficient data for robust statistical analysis. This random selection approach for non-final scripts helped mitigate potential systematic bias in episode representation.

### Potential Limitations and Biases

The limited number of available scripts and our preferential selection of final versions could lead to biased data in several ways.

First, certain characters and topics may be over or underrepresented within the 61 episodes selected as our sample represents only a fraction of the show's 700+ episodes produced across multiple decades. Episodes featuring particular characters prominently or focusing on specific themes like special episodes like Treehouse of Horrors may be disproportionately represented in the available scripts. Special episodes like Treehouse of Horrors were deemed acceptable in our project since the core values and ideals the cast embodies are still present in these non-canon episodes.

Second, our random selection of non-final scripts could be biased as in choosing those scripts we had to read the titles of each episode which could have subconsciously influenced us. This makes our random selection not fully random. This situation could have been mitigated by using a script to pull random scripts from the post, but the format of the site didn't allow us to easily scrape the info wanted and as such our method was deemed easier and non-biased enough for the purposes of this project.

Finally, the mixing of final and draft scripts could introduce inconsistencies as draft versions may contain dialogue that was ultimately revised or removed before broadcast as well as characters with minor roles or name changes could be

counted as different entities depending on how varied the names were.

### Data Cleaning and Standardization

Once inconsistent script structure and formatting was homogenized through our formatting script `format.py`, 13,815 clean lines of dialogue were produced. From these lines, characters with less than 10 lines were removed and characters with 300 or more lines were considered candidates for side character analysis. The 300-line minimum also helped filter out guest characters or minor recurring roles whose limited appearances would not provide reliable insight into consistent speech patterns.

## Methods

### Data processing

We applied a formatting program `format.py` to standardize character names, dialogue, and stage directions into a consistent format and to immensely reduce the amount of unwanted narration. Nonetheless, some lines still contained non-dialogue data. These lines were kept as there was no feasible method to fully remove them beyond manual annotation.

These cleaned scripts were processed using `extract.py` to produce two datasets:

- A **dialogue CSV** which contains the character name, dialogue text, and episode source for every extracted line.
- A **summary CSV** containing the total number of lines spoken by each character.

We excluded characters with fewer than 10 total spoken lines in the scripts to debloat our data and remove any anomalies from the extract script identifying duplicates or dialogue from non-characters. For instance, the script would sometimes mark stage direction lines as distinct characters. Dialogue was considered non-trivial if 10 or more words were used in each line. Certain trivial lines passed this filter due to appended narrative sections - these were manually omitted during annotation.

To choose side characters to investigate, we ran a pie chart script on the summary CSV to compare characters line counts. Though the entire Simpsons family is arguably the center focus of the show, Homer has substantially more lines compared to all other family members, who appear more sparsely across episodes. We thus chose Bart, Lisa, and Marge Simpson as our side characters of focus. Due to their relatedness, we were able to develop a more specific topic categorization system than would have been possible with other characters exceeding the 300-line threshold.

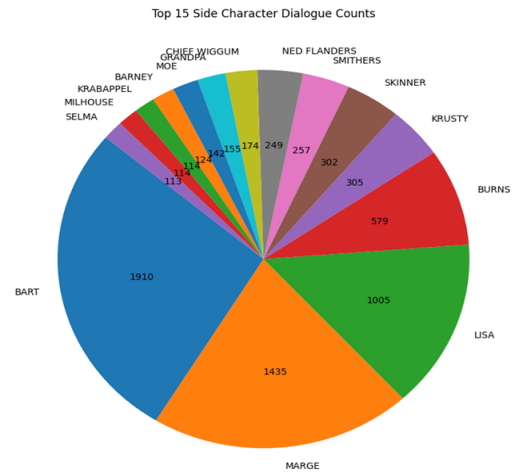


Figure 1: Distribution of lines spoken by various side characters across 61-episode scripts from The Simpsons, showing the proportion of dialogue contributed by the top 15 characters with the most lines except Homer Simpson.

### Topic design

We performed an open coding on 100 lines of meaningful dialogue for each character, noting the topics discussed in each. The most frequent topics of conversation were then generalized to create an 8-category classification system. Since more general topics are less insightful, we chose the most specific topics that could still lead to a comprehensive categorization system across hundreds of lines. Inevitably, some topics were more general than others, and thus less insightful (for example, “objects” category); this was necessary to ensure every line could be reliably tied to a topic.

### Annotation process

A typology was constructed, providing precise definitions for each topic along with positive, negative, and edge cases with explanations regarding inclusion or exclusion. The main annotation difficulty was in selecting one topic based on importance when many were mentioned in a given line. This inherently subjective process was made more objective through consulting typology examples which used sentence structure to decide on importance. Annotation bias is also possible as due to the volume of data and time available, every line was only annotated once.

### Topic Characterization & Analysis

We characterized topics using a two-step process. First, we computed TF-IDF scores for aggregated dialogue lines to identify the top 10 representative keywords per topic. Second, we used ChatGPT to generate descriptive summaries based on these keywords, combining statistical analysis with AI-driven interpretation. The number of lines attributed to

each topic were analyzed with simple plotting features. Specific character analysis was performed within the Simpsons nuclear family category by totaling the occurrences of various forms of each name within each character’s lines attributed to this topic.

## Results

Our analysis yielded two key sets of findings: a data-driven characterization of the topics discussed in The Simpsons, and the distinct engagement patterns of Bart, Lisa, and Marge with these topics.

### Topic Characterization

Using the TF-IDF and LLM methodology, we defined the 8 topics as follows. These definitions reveal the specific vocabulary characters use when discussing these subjects.

- **Themselves:** Defined by keywords like ill (I'll), simpson, im (I'm), birthday. This topic captures dialogue where characters refer to themselves, assert their identity or discuss personal milestones.
- **Core Family:** Characterized by bart, homer, dad, lisa, mom. This represents the domestic sphere. The keywords are almost exclusively names and familial titles, indicating that when this topic arises, it is usually in the context of direct address or discussing the immediate family unit.
- **Non-Core Family:** Includes mr, krusty, burns, ned, flanders. These keywords map the social geography of Springfield. The presence of "Krusty" and "Burns" highlights the external forces (entertainment, employment) that impact the family.
- **Location:** Defined by school, town, library, house, kitchen. This topic delineates the physical settings of the show. "Kitchen" and "House" represent the domestic stage, while "School" and "Library" are key settings for Bart and Lisa respectively.
- **Object:** Keywords include sauce, cranberry, big, box, money. This focuses on tangible items, props, and their physical descriptors.
- **Event:** Characterized by tonight, tomorrow, party, parade, contest. This involves time-sensitive planning or anticipation of specific non-trivial occurrences.
- **Emotion:** Defined by sorry, love, hate, worry, happy. This captures the show's emotional core. The balance of positive ("love", "happy") and negative ("hate", "worry") terms reflects the show's blend of cynicism and sentimentality. "Sorry" being a top word highlights the frequency of conflict and reconciliation.
- **Opinion:** Keywords like think, like, know, stupid, best. This represents cognitive and evaluative language expressing personal stances or judgments.

### Topic Engagement

We analyzed the distribution of these topics across the three characters to understand their distinct narrative roles.

#### • Bart Simpson

Bart's most frequent topic is "Non-Core Family" (83 lines), accounting for approximately 28% of his annotated dialogue. This is the highest count for this category among all three characters. His second most frequent topic is "Opinion/Judgement" (53 lines). Notably, Bart has the highest engagement with the "Themselves" topic (33 lines), which is more than three times higher than Lisa (9) and six times higher than Marge (5).

#### • Marge Simpson

Marge shows a distinct pattern where "Core Family" is her dominant topic (97 lines), representing 32.3% of her dialogue. She also records the highest counts in "Opinion/Judgement" (74 lines) and "Emotion" (35 lines) compared to Bart and Lisa. Her engagement with "Themselves" is minimal, with only 5 lines (1.7%) attributed to this category, making it her lowest topic overall.

#### • Lisa Simpson

Lisa's topic distribution often falls between Bart and Marge. Her most frequent topic is "Core Family" (87 lines), closely followed by "Opinion/Judgement" (68 lines). Her "Non-Core Family" count (58 lines) sits between Bart's (83) and Marge's (38). Like Marge, her lowest engagement is with "Themselves" (9 lines), though she discusses it slightly more than her mother.

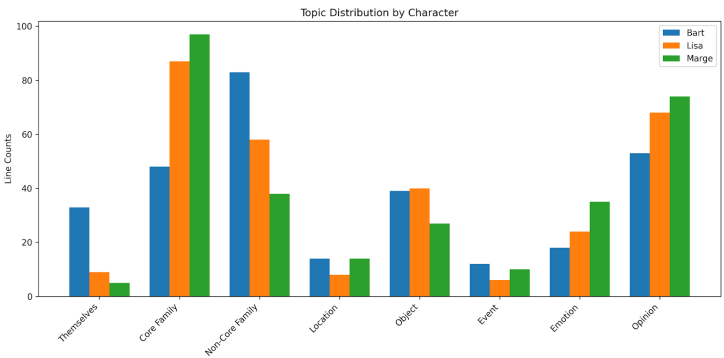


Figure 2: Line counts by character for each topic after manually attributing 300 nontrivial lines of dialogue to one topic, for each of Bart, Lisa, and Marge Simpson.

### Family engagement

Taken together, dialogue in the direct family category accounted for 232 lines, or nearly 24% of all lines. We further investigated this important topic of conversation, wishing to characterize what characters exactly each of

Lisa, Bart, and Marge tend to reference. The heatmap below (figure. 3) shows that Homer is the most often discussed, with a particularly strong presence in Marge’s speech. Maggie, Marge, and Lisa are the least discussed, while Bart figures prominently in both his sister’s and mother’s dialogue.

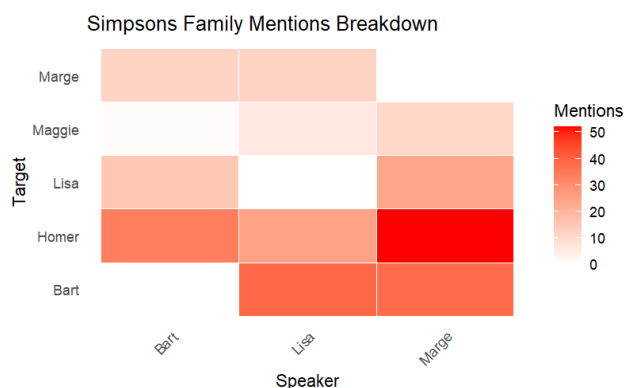


Figure 3: Heatmap analysis of each family member’s importance in family-oriented dialogue. Mentions were calculated by counting the number of occurrences of various forms of each name within each character’s lines attributed to the immediate-family category.

## Discussion

Our topic analysis of 900 lines of script from Bart, Lisa, and Marge provides insight into their interests and means of expression, helping writers remain consistent and design realistic character developments in future seasons of The Simpsons.

Bart led the way in the Themselves section, with 2.4 times the amount of self-centered lines than both Marge and Lisa combined. This underscores Bart’s tendency to focus on his own identity and actions. He also leads the way in the third category, with 28% of his lines focused on non-family characters (e.g., Principal Skinner, Krusty). This showcases Bart’s social proclivity, and perhaps his disruption-seeking behavior, within the broader Springfield community. When he does speak about his family, he most often mentions his father, highlighting their unique relationship in this show.

Nearly a third of Marge’s lines revolved around Bart, Lisa, Homer, and Maggie, making her the character who is most involved with the family. She also leads in the Emotion and Opinion categories. This confirms her role not just as a caretaker, but as the primary source of emotional expression and moral guidance within the household. A large part of her emotional lines also involve Homer, showing that Marge likely carries out lots of emotional labor caused by her husband’s mishaps. In fact, lines where Marge mentions Homer were more numerous than any other speaker-target combination analyzed.

In all character categories, Lisa placed between Bart and Marge. Lisa almost reaches Marge’s totals in the Immediate family and Opinion categories, which could indicate a strong maternal influence and shared tendency to voice judgments. Lisa is much more outgoing than her mother with non-family members (58 lines vs. 38), suggesting that she bridges the gap between the domestic world and the rest of the town more actively than her mother. Lisa most often mentioned Bart in her family dialogue, showcasing the formative role these two characters share.

Beyond revealing each character’s prominent characteristics, this analysis identified which topics serve to distinguish between the characters. Relatively few lines were attributed to the location and events categories, with marginal differences between characters. Similarly, all characters gave similar attention to objects. Characters seem to be more strongly defined by what other characters they speak about and spend time with. Along with the character topics, the emotion and opinion sections serve as useful ways for the producers of The Simpsons to maintain interesting differences between their characters.

The principal takeaway from this analysis was in identifying where each character allocates their attention, namely (1) Marge appears to serve as the emotional and moral anchor of the family, (2) Bart is the most self-centered, but also interacts with the community more than others, and (3) Lisa tends to find a middle ground, showing a more balanced character.

To conclude, our analysis of 900 lines of script from the three most important side characters of The Simpsons shows distinctive speech patterns in each character. To ensure a realistic and interesting continuation of this family’s story, producers should concentrate on preserving each of Lisa, Bart, and Marge’s speech tendencies, especially regarding character groups, emotion, and judgement.

## Group Member Contributions

The work was split evenly, and we held regular meetings to confer and talk about any issues that occurred after every step of the project. One person worked on each of the following: Data Collection phase, Manual Annotation Phase, and AI/TF-IDF Phase. Finally, we all worked on the report together, each writing up the parts we worked on.

### Data Collection Phase

This person worked on extracting the data from the site and created the `format.py`, `extract.py`, and `piechart.py` scripts.

When an issue of bias or inconsistency arose, we would confer as a group on what would be best in terms of the scope of our project. After collecting the raw data and running it through the extraction script, the csv was handed to the next person for manual annotation.

### **Manual Annotation Phase**

This person performed the open coding, topic design, and typology building. He conferred with teammates and TAs about the topic design process. Based on this typology, he performed the manual annotation of the lines. He wrote `family_mentions_analysis.R` to generate fig. 3. Again, if issues or doubts arose, we conferred as a group.

### **AI/TF-IDF Phase**

This person worked on the automated analysis pipeline and created the `analyze_topics.py` script to compute TF-IDF scores and identify top keywords. This work also involved integrating ChatGPT to generate objective topic summaries, creating the `generate_chart.py` script to visualize character engagement, and drafting the Results section.

## **References**

No Homers Club. 2024, Simpsons Script Collection: < <https://no-homers.net/forums/index.php?threads/simpsons-script-collection.57493>> (accessed December 5, 2025).