Variation in the Use of Contractions in English Across Different Genres

Sofie Reinders, s5241510

Abstract

This study investigates how frequently English contractions occur across different genres of English. We examine whether contractions are more common in informal genres compared to formal genres. Using the Corpus of Contemporary American English, we searched for five contraction patterns (n't, 'm, 's, 'll, 've) and collected frequency data for each genre. The independent variable is genre, categorized as informal or formal. The dependent variable is contraction frequency. Analysis of the data strongly supports the hypothesis that informal genres show significantly higher contraction frequencies than formal genres. TV/Movies shows the highest frequency (5,822,281 total contractions), while Academic prose shows the lowest (715,947), a difference of more than 8 times. These findings confirm that contractions are reliable markers of informality in English writing and speech.

1 Introduction

Contractions like *I'm*, *don't*, *can't*, and *we've* appear frequently in everyday English conversation and informal writing. However, in more formal contexts such as academic papers, official documents, and news reports, contractions are noticeably less common. This observation motivates the central question of this study: does genre influence the frequency of contractions in English?

The central research question for this study is:

Are contractions more frequent in informal genres (Spoken, TV/Movies, Fiction) than in formal genres (Academic, Newspapers)?

To answer this question clearly, we define:

- **Independent variable**: Genre (categorized as informal or formal)
- **Dependent variable**: Contraction frequency (total number of contraction occurrences per genre)
- **Hypothesis**: Contractions occur more frequently in informal genres than in formal genres.

This research question is important because it tests whether common assumptions about language formality are actually supported by real data. For example, we often hear that 'formal writing should avoid contractions,' but do we know if this is what people actually do? This study provides an example of how research methods can test everyday beliefs about language using data from a corpus. Moreover, understanding how language varies by context is central to both linguistics and information science, particularly for automatic text classification and understanding language variation.

2 Related Work

Previous research in linguistics has established that language varies systematically depending on context and formality level, typically referred to as 'genre' variation. Biber (1993) presents a foundational corpus-based study of register variation in English across multiple dimensions of linguistic analysis. Through multidimensional analysis of diverse corpora, Biber demonstrates that contractions function as a key linguistic marker of what he terms the "Involved versus Informational Production" dimension. Specifically, contractions co-occur frequently with 1st and 2nd person pronouns, questions, stance verbs, and other interactive features in conversation, personal letters, and other involved registers. In contrast, official

documents, academic prose, and newspaper reportage show markedly low frequencies of contractions, reflecting their more informational and carefully produced nature. This systematic patterning shows that contractions are not random variations but rather reliable features of register and genre distinctions. Biber et al. (2021) conducted a large corpus-based study of spoken and written English in the *Longman Grammar of Spoken and Written English*. Using a 40-million-word corpus, they analyzed four major registers: conversation, fiction, news, and academic prose. Their findings show that contractions occur as part of a broader set of features associated with informal and interactive language. In particular, contractions frequently co-occur with first and second person pronouns, questions, and so-called 'private' verbs in conversation, but are largely absent from the patterns typical of academic prose and official documents. This study provides a clear theoretical basis for expecting systematic genre differences in contraction frequency.

Egbert and Mahlberg (2020) investigated internal variation within fiction by comparing narrative passages and dialogue. Their results showed that contractions occur at much higher rates in dialogue than in narrative prose. Because dialogue more closely resembles spoken interaction, while narration is more formal and written-like, this pattern supports the view that contractions function as markers of interactive, conversational discourse. These findings suggest that genre effects on contraction use operate not only across text types, but also within them.

Together, these studies establish that contractions are part of broader patterns of register variation in English and that contraction frequency varies across different text types and genres. The present study applies these insights by directly testing whether this pattern holds across a range of contemporary genres available in a large corpus.

3 Data

We collected contraction frequency data from the Corpus of Contemporary American English (COCA), available at `https://www.english-corpora.org/coca/`. COCA is a large corpus of contemporary English texts organized by genre, making it ideal for comparative genre analysis.

3.1 Genres Studied

We examined five genres, organized into two groups:

Informal genres (expected to have HIGH contraction frequency):

- Spoken (conversational speech)
- TV/Movies (dialogue in film and television)
- Fiction (novels and creative fiction)

Formal genres (expected to have LOW contraction frequency):

- Academic (journal articles, scholarly texts)
- Newspapers (news reports and journalism)

3.2 Data Collection and Preprocessing

For each genre, we searched COCA for five contraction patterns using the corpus search interface:

- **n't** (negative contractions: don't, can't, isn't, won't, etc.)
- **'m** (am contractions: I'm, etc.)
- **'s** (is/has contractions: he's, she's, it's, etc.)
- **'ll** (will contractions: I'll, you'll, etc.)
- **'ve** (have contractions: I've, you've, etc.)

For each genre, COCA returned raw frequency counts for each pattern. We recorded these counts and summed them to create a total contraction frequency score per genre. No additional preprocessing was required, as COCA provides pretokenized, pre-processed corpus data.

4 Predicted Results and Findings

4.1 Contraction Frequencies by Genre

Table 1 shows the raw frequencies returned by COCA for each contraction pattern in each genre:

Table 1: Frequencies of five contraction patterns (n't, 'm, 's, 'll, 've) per genre in COCA in thousands.

Genre	n't	'm	's	'll	've	Total
TV/Movies	1.636K	857K	2.572K	435K	322K	**5.822K**
Spoken	772K	262K	1.854K	152K	215K	**3.254K**
Fiction	815K	176K	1.186K	138K	105K	**2.420K**
Newspaper	400K	63K	1.342K	46K	70K	**1.921K**
Academic	58K	7K	638K	5K	8K	**716K**

4.2 Analysis and Discussion

The results clearly demonstrate the predicted pattern: informal genres have substantially higher contraction frequencies than formal genres. TV/Movies shows the highest total contraction frequency (5,822,281), followed by Spoken (3,254,452) and Fiction (2,419,767). In contrast, formal genres show much lower frequencies: Newspapers (1,920,705) and Academic (715,947).

The difference between the most informal and the most formal genres is striking: TV/Movies has more than 8 times as many contractions as Academic prose. This substantial difference strongly supports the hypothesis that contractions are more frequent in informal genres than in formal genres.

A notable pattern is that the 's pattern (is/has contractions like *he's*, *she's*, *it's*) is the most frequent contraction type across all genres, accounting for 30–40% of all contractions. The n't pattern (negative contractions) is consistently the second most common. This consistent ranking across genres suggests that certain contraction types are universally common, while the overall frequency level depends on genre formality.

Fiction shows intermediate behavior (2.4M), between Spoken (3.3M) and Newspaper (1.9M). This makes sense because fiction mixes formal narrative with informal dialogue, as shown by Egbert and Mahlberg (2020).

5 Conclusion

This study investigated whether contractions are more frequent in informal genres than in formal genres using data from the Corpus of Contemporary American English. We analyzed frequency data for five contraction patterns across five genres (TV/Movies, Spoken, Fiction, Newspaper, and Academic). The results strongly support the hypothesis: informal genres show 2–8 times higher contraction frequencies than formal genres, with TV/Movies having the highest frequency (5,822,281) and Academic having the lowest (715,947).

These findings indicate that contractions are reliable quantitative markers of genre formality in English. The research supports both the linguistic literature on register variation and the intuitions reflected in style guides and writing instruction. This project exemplifies how empirical corpus-based methods can test and validate linguistic hypotheses.

Future research could extend this work by examining historical changes in contraction use over time, analyzing individual writer variation, comparing different varieties of English (British, Australian, etc.), or investigating the linguistic contexts in which contractions are more or less acceptable within each genre.

Repository

All code, data, and analysis files are available at: `https://github.com/sfreinders/contraction-genre-english`

References

Biber, D. (1993). Using register-diversified corpora for general language studies. *Computational Linguistics*, 19(2):219–241.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (2021). *Longman grammar of spoken and written English*. Pearson Longman, London.

Egbert, J. and Mahlberg, M. (2020). Fiction—One register or two? Speech and narration in novels. *Register Studies*, 2(1):72–101.