Classist Tools: Social Class Correlates with Performance in NLP

Amanda Cercas Curry
MilaNLP
Bocconi University
amanda.cercas@unibocconi.it

Giuseppe Attanasio
Instituto de Telecomunicações
Lisbon, Portugal
giuseppe.attanasio@lx.it.pt

Zeerak Talat
Mohamed Bin Zayed
University of Artificial Intelligence
z@zeerak.org

Dirk Hovy
MilaNLP
Bocconi University
dirk.hovy@unibocconi.it

Abstract

The field of sociolinguistics has studied factors affecting language use for the last century. Labov (1964) and Bernstein (1960) showed that socioeconomic class strongly influences our accents, syntax and lexicon. However, despite growing concerns surrounding fairness and bias in Natural Language Processing (NLP), there is a dearth of studies delving into the effects it may have on NLP systems. We show empirically that NLP systems' performance is affected by speakers' SES, potentially disadvantaging less-privileged socioeconomic groups. We annotate a corpus of 95K utterances from movies with social class, ethnicity and geographical language variety and measure the performance of NLP systems on three tasks: language modelling, automatic speech recognition, and grammar error correction. We find significant performance disparities that can be attributed to socioeconomic status as well as ethnicity and geographical differences.[1] With NLP technologies becoming ever more ubiquitous and quotidian, they must accommodate all language varieties to avoid disadvantaging already marginalised groups. We argue for the inclusion of socioeconomic class in future language technologies.

1 Introduction

The relationship between social class and language was first examined by Labov (1964). He discovered that New Yorkers with greater socioeconomic status pronounce the /R/ sound after vowels, whereas individuals with lower socioeconomic position drop it. He quantified this observation by asking employees in various department shops (a proxy for socioeconomic rank) for things found on

---

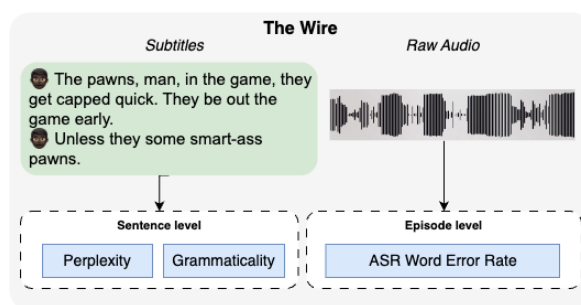[1]We make our artifacts available on our Github repository.

Figure 1: Overview of the empirical framework. We use OpenSubtitles as truth-value data to calculate perplexity, grammaticality, and word-error-rate in Automatic Speech Recognition (ASR) for each episode.

the *fourth floor*. Then he kept track of how many Rs were dropped in those two words. He discovered a clear anti-correlation between the store's (and presumably the speakers') socioeconomic standing and the quantity of dropped Rs: the higher the status, the fewer dropped Rs.

Since Labov's (1964) analysis of social stratification and language, linguistics has made concerted efforts to understand how different socio-demographic factors influence language production and perception, and how speakers use them to create identity (Rickford, 1986; Bucholtz and Hall, 2005; Eckert, 2012); using both naturally produced language and language from media like TV shows and films (Stamou, 2014) to study sociolects. Despite a considerable body of evidence demonstrating links between language and demographic characteristics (the "first wave" of socio-linguistic variation studies), comparatively few socio-demographic factors have been studied in the context of Natural Language Processing (NLP) technology (Cercas Curry et al., 2024). But NLP is a product and a tool for understanding other phe-

12643

nomena and must represent all language varieties.

Existing research on socio-demographic characteristics has primarily concentrated on the signaling effect of certain linguistic variables like age, ethnicity, regional origin, and gender (Johannsen et al., 2015). Most of these publications focus on any bias toward the particular variable they study. However, NLP hardly interacts with the second and third waves of sociolinguistics, i.e., how variation 1) shapes local identity and 2) drives language development. To narrow this gap, we use NLP to focus on socioeconomic position.

Using a dataset of 95K utterances from English-language television episodes and movies, we empirically investigate performance as a function of socioeconomic class. We study its effect on vocabulary, Automated Speech Recognition (ASR), language modelling, and grammatical error correction. In all applications, we see strong correlations. Movie characters are usually typecast to present as a certain class, providing a representative sample without the privacy issues associated with regular subjects. The authenticity of these representations is backed by work in sociolinguistics such as McHoul (1987) and Quaglio (2008).

We discover that the performance of NLP tools is associated with socioeconomic class, geographic variety (US vs. UK), and race across all tasks. Our findings highlight an important lack of flexibility of NLP tools, and a more fundamental issue: What does it mean if NLP technologies only reliably work for a limited segment of society?

Contributions

- We construct a corpus of 95K utterances from television shows and movie scripts, coded for socioeconomic status, geography, and race.

- We empirically show that socioeconomic status measurably impacts the performance of NLP systems across four measures.

2 Social Class and Language

Social stratification is the grouping of people based on their socioeconomic status (SES), which is determined by factors such as income, education, wealth, and other characteristics. Different groups are distinguished in terms of power and prestige. There are various social stratification systems, such as the Indian caste system, indigenous American clans or tribes, and the Western hierarchical class system.

The exact number of social strata is unknown and varies by country and culture. However, at least three strata are commonly used to refer to different groups in Western societies: upper, middle, and lower-class people. Other systems distinguish between blue collar and white collar jobs. Recently, the Great British Class Survey (GBCS, Savage et al., 2013) has taken an empirical approach to understanding the different social strata, and they propose a seven-level system for the United Kingdom. Their stratification is based on economic, social, and cultural capital: elite, established middle-class, technical middle-class, new affluent workers, traditional working class, emergent service workers, and precariat. They derive these classes from a survey conducted over British citizens that received more than 160K responses.

Social class influences people's daily lives by granting or limiting access to resources. Beyond power and prestige, social stratification has a significant impact on people. For example, lower socioeconomic status has been linked to poorer health outcomes and higher mortality. (Saydah et al., 2013).

SES also affects language use from the very early stages of development. Bernstein (1960) theorised that language takes on a different role in families of different socioeconomic status, where middle-class parents encourage language learning to describe more abstract thinking leading to broader vocabularies. In working-class families, parents are *limited* to more concrete and descriptive concepts. Parents from lower SES tend to interact less with their children, with fewer open-ended questions than parents from higher SES, which shapes language development (Clark and Casillas, 2015). While Usategui Basozábal et al. (1992) show that education reduces this language gap, education has traditionally been one of the most important factors in determining social class and potential for upward social mobility.

Given the well-documented effects of socioeconomic status on language development and use, it stands to reason that social class should be carefully considered as a variable in NLP.

3 Dataset

Here, we want to assess the impact a user's social class has on the performance of NLP systems. Although there have been concerted efforts to document wide varieties of English, existing datasets do not currently capture all social strata or other sociodemographic aspects sufficiently. For example, the National Speech Corpus (Clopper and Pisoni, 2006) provides education information

for the talker and their parents but consists of almost exclusively white and educated speakers;[2] the CORAAL corpus (Kendall and Farrington, 2018) on the other hand is focused on African American English (AAE), not capturing other ethnicities. To the best of our knowledge, no datasets are available that would suit our purposes: a large enough sample that is annotated with the speakers' socioeconomic background. We thus collect our own.

We identify an initial sample of English language television series by selecting the 250 highest-rated series on the International Movie Database (IMDB).[3] We annotate each show according to attributes,[4] i.e., class, race, gender, dialect, and geography of the main characters, as well as the genre and time period of the show. Three annotators familiar with the shows agree on the labels unanimously. We only consider shows that are scripted to ensure that the language production of each character is controlled by the script-writers and actors to produce a particular vernacular and to avoid collecting highly sensitive information about private individuals. We therefore exclude cartoons, documentaries. We further exclude science fiction and historical fiction, as these are likely to operate without strict adherence to existing social structures and language variation. After our filtering and selection, 63 shows remain in consideration. Of those, 34 are predominantly about white men, and 29 feature a mixture of white women and men. Only three predominantly feature female characters, and none black or lower-class female characters. Based on this initial sample, we identified additional shows to augment gaps (e.g., the lack of Black stories) in class and dialect. Because we want to measure the effect of class separately from gender, race and regional accent, we select the TV shows from the list representing each category.

In our final selection of movies and television shows, we include a total of 19 shows that are based in the United States of America and the United Kingdom. To ensure racial diversity, we emphasise movies and shows that predominantly represent Black characters in addition to movies and shows that represent white characters.[5] We include the U.S.-based shows *The Wire* and *When They See Us* (predominantly lower-class Black characters), *Fargo* (middle-class Black characters), *The Fresh Prince of Bel-Air* (upper-class Black characters). For the shows' centre white characters, we include *Trailer Park Boys* (lower-class white), *The Sopranos* (lower-class white), *Breaking Bad* (middle-class white), and *Arrested Development* (upper-class white). The list of shows above, however, predominantly feature male characters. For this reason, we also include shows whose characters represent a wider array of genders. The shows that we include are *Pose* (lower-class Black and Latinx), *Big Little Lies* (middle-class white), and *Sex and the City* (upper middle-class white). Finally, to compare with other regional varieties of English, we include movies and shows from different regions in the United Kingdom. The movies and shows that we include are *Train Spotting* and *T2* (lower-class white Scottish, predominantly male characters), *Derry Girls* (lower-class white Irish, primarily female), *Downton Abbey* (English), *The IT Crowd* (white, middle-class) and *The Crown* (upper-class white English). We report full details in Table 8, Appendix A.

We collect the first season for each of the shows on our list, except for Fargo and The Wire. For Fargo, we collect the fourth season, which includes Black characters. For The Wire, we collect seasons 1 and 3, which feature Black characters but have a minority representation of white characters.

We deliberately chose this approach to data selection from "artificial sources" since collecting class information is highly sensitive and can introduce privacy risks. Fictional characters, however, are often scripted into particular roles with explicit SES. A drawback of this methodology is the question of authenticity - the extent to which actors are representative of particular sociolects. Actors reportedly undergo significant training to learn different lects.[6] By selecting highly rated shows, we hope to capture good, realistic performances. Moreover, TV show and film data is increasingly used in sociolinguistics (Stamou, 2014) and Quaglio (2008) found important similarities between real and TV show dialogue. In NLP, media data is broadly used

---

[2]Education is one aspect of social class. For a further discussion see Saunders (1990) and Savage et al. (2013).

[3]IMDB top 250 TV shows.

[4]Note that we are using *ascribed* characteristics rather than *self-reported* since we are dealing with fictional characters, however these two are found to be highly correlated based on linguistic features (Weirich and Simpson, 2018)

[5]At least one of the authors is familiar with every show and movie. We annotated them for race and class based on the setup and main characters.

[6]See for example, https://www.newyorker.com/magazine/2009/11/09/talk-this-way

Geography	Class	Episodes	Utterances
USA	Low	13	20119
	Low, middle	13	22947
	Middle	7	4140
	Middle, Upper	25	13221
	Upper	15	8561
EN	Middle	6	2271
	Upper	3	1515
	Upper, Low	7	5214
NE	Low	7	3929
	Middle	2	1238

Table 1: Summary of statistics by geographic variety and class. EN: England, NE: non-English U.K.

in tasks like emotion analysis (Plaza-del Arco et al., 2024) and dialogue (Lison and Tiedemann, 2016). We therefore have reason to believe our dataset provides a reasonable proxy for real speakers.

In total, our dataset contains 95K utterances, text and speech, from 19 TV shows and movies (see Table 1 for dataset statistics).

3.1 Lexical Analysis

Bernstein (1960) showed that children from working-class families had significantly smaller vocabularies even when general IQ was controlled for, and Flekova et al. (2016) showed that there is significant lexical and stylistic variation between different social strata, with lexical features being stronger predictors of income than even age.

We first assess whether our dataset has sufficient linguistic cues to capture lexical differences. We generally follow the methodology set out in Flekova et al. (2016). However, we model class as a categorical value. We calculate:

Surface: Turn length, mean word length per turn, ratio of words longer than five letters, type-token ratio. Flekova et al. (2016) also measure the use of emojis, but this is not relevant to our analysis.

Readability Metrics: Readability metrics aim to measure the complexity of a text generally based on the number of syllables per word and the number of words per sentence. Using Textstat,[7] we calculate the Automatic Readability Index (ARI) (Senter and Smith, 1967), the Flesch-Kincaid Grade Level (FKG) (Kincaid et al., 1975), the Coleman-Liau Index (CLI) (Coleman and Liau, 1975), the Flesch Reading Ease (FRE) (Flesch, 1948), the LIX Index (LIX) (Anderson, 1983), and the Gunning-Fog Index (FOG) (Gunning, 1968). We normalise the text before calculating these metrics by lower-casing

---
[7]https://github.com/textstat/textstat

	Class	Gender	Race	Geography
Length	-0.033	-0.034	-0.025	0.052
Chars	**-0.437**	**0.243**	**0.203**	**0.155**
>5	0.073	-0.024	0.017	-0.022
TTR	-0.019	-0.047	-0.046	-0.085
FRE	-0.096	0.033	0.032	0.074
ARI	-0.214	0.068	0.059	0.110
CLI	**0.467**	**-0.213**	**-0.201**	**-0.137**
FKG	0.007	-0.035	-0.029	-0.005
FOG	0.137	-0.031	-0.040	-0.025
LIX	0.106	0.013	-0.007	-0.065
NOUN	**-0.181**	0.076	-0.035	0.163
VERB	-0.059	0.039	-0.002	**0.082**
PROPN	-0.017	-0.064	-0.099	-0.067
ADV	-0.066	-0.068	-0.103	-0.037
ADJ	-0.051	0.044	0.021	0.076
DET	0.037	0.059	0.024	0.026
INTJ	0.121	**-0.079**	**-0.093**	-0.044
PRON	-0.004	0.014	0.008	0.020
NER	0.078	-0.001	0.035	-0.010
abst	**0.052**	-0.034	-0.026	-0.034
HL	-0.034	**0.109**	**0.142**	**0.066**

Table 2: Coefficients for the logistic regression model. For each class and type of feature, the strongest coefficient is in bold face.

	Val	Test
Lexical		0.290
TF-IDF	**0.528**	**0.524**
TF-IDF + Lexical	0.297	0.290
Sentence Embed	0.457	0.454

Table 3: F1 Macro for different representations on social class prediction. Best results in bold.

and removing punctuating except for periods since these are used by CLI. Note that these were designed to evaluate long-form text.

Syntax: Texts with more nouns and articles as opposed to pronouns and adverbs are considered more formal. We measure the ratio of each POS per turn using Stanza (Qi et al., 2020).

Style: Finally, to capture some notion of speaker style, we measure the number of abstract words[8], the ratio of hapax legomena (HL), and the number of named entities (NER).

We run a multi-class logistic regression to understand whether the features are correlated to class. After normalisation, we run a logistic regression with social class as the predicted class and the

---
[8]https://onlymyenglish.com/list-of-abstract-nouns/

features above as the predictive features. In contrast with Flekova et al. (2016), we find our set of features only mildly predictive in this dataset. Although CLI and the number of characters have strong coefficients, the overall performance is very low. Our best model achieves a macro F1 of 0.16, only slightly better than a most-frequent label baseline (macro-F1 of 0.07). In addition, we run the same regression for the other sociodemographic aspects: gender, race, and geographical origin. Overall, we find these less predictive with F1 scores similar to the baseline, suggesting the features better capture differences caused by SES than other sociodemographic factors. Table 2 shows the coefficients for each sociodemographic factor.

Lexical features do not capture strong differences in SES indicating one of two things: the features are better suited for written text (for example, many of them capture formality), and/or our dataset does not accurately capture linguistic differences. To assess whether linguistic cues in our dataset might help us differentiate between speakers of SES which the previous study did not capture, we experiment with alternative text encodings such as TF-IDF, and sentence embeddings (Reimers and Gurevych, 2019).[9] The results are shown in Table 3. Logistic regression based on TF-IDF achieves an F1 of 0.52 (x1.8 lexical features). We hypothesise two reasons for the improvement over the lexical features: (1) there are linguistic differences that the metrics do not capture, or (2) the logistic regression model can better differentiate between TV shows. Given the existing literature in support of linguistic differences and the (albeit small) predictive power of the metrics, we ascribe the improvement to a mixture of both. With this in mind, we continue our experiments with user-facing NLP systems.

4   Social Class and Speech

Beyond written language, our dataset affords testing whether speech models understand and process accent unevenly. Therefore, we test disparities in ASR word error rates across social classes. We calculate the WER for all shows and movies to measure how well ASR technologies capture different variants of English in our sample.

Models: We use Wav2Vec2 (Baevski et al., 2020), XLS-R (Babu et al., 2021), and Whisper-medium (Radford et al., 2023) as implemented in

transformers (Wolf et al., 2020).

Processing Pipeline: We use movies and TV shows which are available online. We first collect the audio and subtitles from OpenSubtitles[10] and use them as gold reference.[11] Then, we used one ASR model to transcribe the audio.

The timestamps from the model-generated transcription and the subtitles often do not match, and neither source provides information about which character is speaking. We therefore group all utterances from a single episode or movie together into one long string and calculate the WER using Jiwer (Morris et al., 2004). Ideally, we would like to separate by character to get a more fine-grained analysis. However, this approach would require us to keep mappings from utterance to character consistent across episodes. Despite several approaches, we were not able to achieve this goal with the computational power we had. Our analysis is therefore group-based, rather than character-based.

Results: Figure 2 shows the differences in word-error-rate by race and class in US TV shows. We find clear trends across models, with effects from both race and class: **lower ASR error rates are associated with higher SES and with whiteness.** These trends are stronger for the Wav2Vec2 model.

5   Language Modelling

Language modelling plays an essential role in downstream applications in NLP, so it is imperative that language models can accurately and equally represent different speakers' lects. Language models are evaluated through perplexity, a measure of how expected a given sentence is: the higher the perplexity, the more unexpected. Perplexity can give us an indication of how well a model might perform in downstream tasks: Gonen et al. (2023) show that the lower the perplexity of a prompt, the better the prompt can perform the task. Here, we calculate perplexity as a measure of linguistic acceptability, that is, how 'expectable' a particular language variant might be to a given language model. Should models display differing perplexities for different groups, they would put such groups at a disadvantage.

Models: We experiment with two state-of-the-art base models, Llama 2 (Touvron et al., 2023)

---

[9]We use scikit-learn's TfidfVectorizer collecting uni- and bi-grams for TF-IDF, and https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

[10]https://www.opensubtitles.org/

[11]Here, we manually checked a random sample to verify the quality. We henceforth assume that the subtitles are an accurate transcription of what is being said.
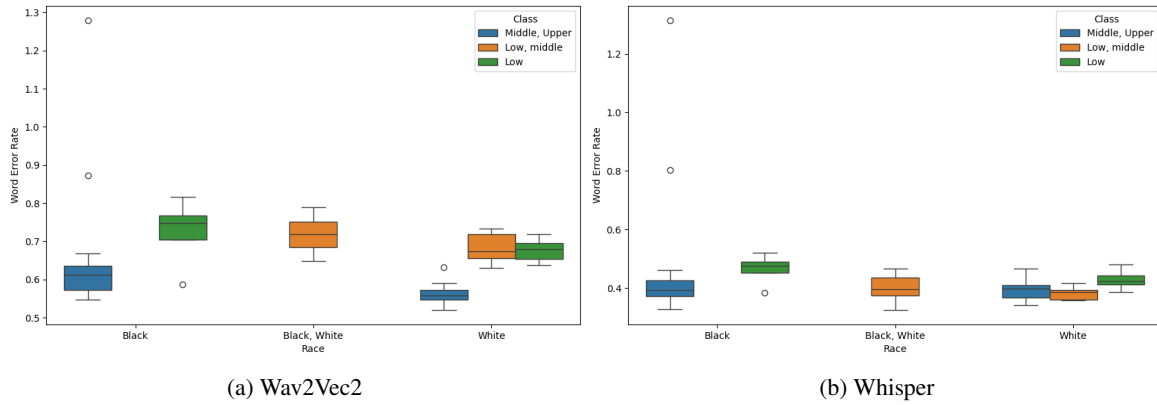
(a) Wav2Vec2	(b) Whisper

Figure 2: Word error rate for the different models, grouped by the speakers' SES and race, for US-based TV shows. The race in the chart refers to that of the majority of characters in the show.
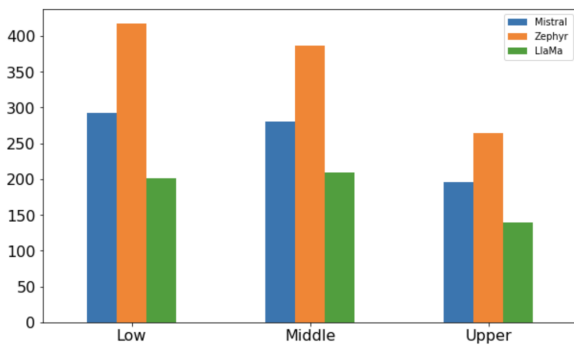

Figure 3: Mean perplexity among U.K.-based shows by model.

and Mistral-7B (Jiang et al., 2023). Moreover, we include Zephyr-7B (Tunstall et al., 2023), a Mistral-7B-based chat model to test the effect of alignment.[12]

We calculate perplexity for each sentence in our dataset after lower-casing them and removing numbers and punctuation. We exclude turns shorter than five tokens as they generally do not differentiate between classes.

Results: Table 4 shows the mean perplexity and standard deviation for each model and each class for the entire dataset. We do not find significant differences based on geographical location (U.K. vs U.S.A) across the models. We see small differences across groups based on class alone, however, because language is affected by other socio-demographic besides class, we also consider the effect of race and geographic variety.

We calculate mean perplexity by SES in the U.K. and the U.S.A. Figure 3 shows that class and per-

plexity are correlated and we find similar patterns across geographical dialects: higher SES leads to lower perplexity also in U.S.A.-based shows.

Next, we consider U.S.-based shows. The breakdown of by class and race are shown in Table 5. Once both factors are taken into account, a clearer picture emerges: lower SES leads to higher perplexity. To confirm these differences are significant, we run a one-tailed student's t-test and find significant differences between classes, particularly for white speakers for both Mistral-7B and Zephyr-7B. In the case of Mistral, lower-class and lower-middle class speakers have significantly higher perplexities than Mid-Upper (*) and Middle class speakers (*). In the case of Zephyr, the model shows higher perplexities for lower classes than it does for Middle (*), Mid-Upper (**) and Upper class speakers (**).[13] In sum, **models show significantly higher perplexities for Lower class speakers than for higher prestige groups.**

6 Grammar Correction

Finally, we consider the grammaticality of each sentence. We suspect that grammar features correlate strongly with class, as "correct" grammar is typically a hallmark of signalling higher SES. We calculate the edit distance between the original sentence and the corrected one. Lower edit distance means higher grammaticality.

Models: Since we are concerned with potential end-user experience, we choose to evaluate the four most downloaded models on Huggingface since they will reach the largest audience. We use the following models for grammar correction:

---

[12]HuggingFace: meta-llama/Llama-2-7b, mistralai/Mistral-7B-v0.1, and HuggingFaceH4/zephyr-7b-beta

[13]Significance: *: $p < 0.05$, **: $p < 0.01$.

Class	Mistral-7B		Zephyr-7B		Llama 2	
	Mean	Std	Mean	Std	Mean	Std
Low	294.606	690.361	415.641	989.066	189.804	361.815
Low, middle	442.756	911.300	649.462	1441.759	265.114	477.377
Middle	252.729	536.903	353.667	822.553	177.900	398.560
Middle, Upper	**241.923**	683.345	**332.224**	999.137	**164.807**	450.446
Upper	288.324	693.994	399.854	1063.942	190.536	370.958
Upper, Low	282.815	575.833	385.273	876.126	190.696	358.952

Table 4: Mean perplexity and standard deviation per class and model. Best mean result per model in bold.

		Perplexity
Black	Low	328.848
	Middle, Upper	**259.673**
White	Low	275.337
	Low, middle	342.865
	Middle	229.278
	Middle, Upper	**205.585**
	Upper	302.057

Table 5: Mean perplexity and standard deviation per class and race. Best results per race in bold.

	Happy	CoEdit	Flan-T5	Gramf.
Low	2.455	1.963	6.424	2.192
Mid-Low	2.590	2.138	12.050	1.553
Middle	2.068	1.785	6.117	1.653
Mid-Upper	2.866	1.944	7.952	2.091
Upper	1.122	1.212	5.518	0.808
% corrected	19.76	35.94	66.42	19.11

Table 6: Mean edit distance between the subtitle and the grammar-corrected utterances generated by each model, and percentage of sentences with at least one correction.

- T5 Grammar Correction[14]: A model based on the HappyTransformer[15] trained on the JF-LEG dataset (Napoles et al., 2017).

- Gramformer:[16] A seq2seq model fine-tuned on a dataset of WikiEdits.

- CoEdit-large[17]: A Flan-T5 (Large) based model, finetuned on the CoEdit dataset (**?**).

- Grammar-synthesis-large[18]: A fine-tuned version of Google's flan-t5-large for grammar correction, trained on an expanded version of the JFLEG dataset.
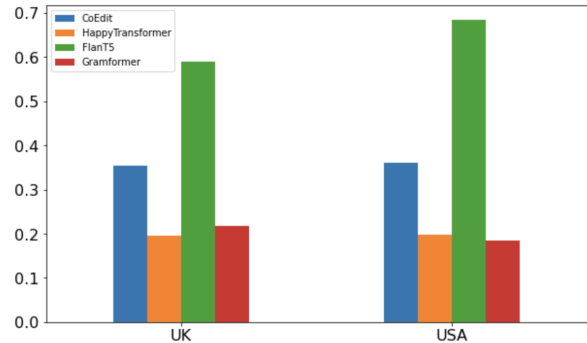

Figure 4: Proportion of utterances that are grammar-corrected per model and geographical language variety.

Results: Figure 4 shows the percentage of utterances that each model corrected, grouped by geographical variation. The Flan-T5-based model generates significantly more corrections than the other models, followed by CoEdit-large. Overall, U.K. utterances are corrected slightly more often by the models. Table 6 shows the mean edit distances between the original sentence extracted from the subtitles and the models' proposed corrections. We find little difference in grammar error correction for most models, even when controlling for geographical and racial differences. However, on further analysis, we find that models generally produce corrections for relatively few utterances.

Instead, we consider whether models are more likely to produce edits for different classes. In Figure 5 we plot the count of utterances with corrections per model and per class. From this, **we can see a clear pattern: models produce corrections more frequently for those of lower SES.**

In addition, we notice that often some of these corrections are performed on in-group slang or regional linguistic phenomena, see Table 7.

---

[14]T5-based Grammar Correction
[15]HappyTransformer
[16]Gramformer
[17]CoEdit
[18]Flan-T5 Correction

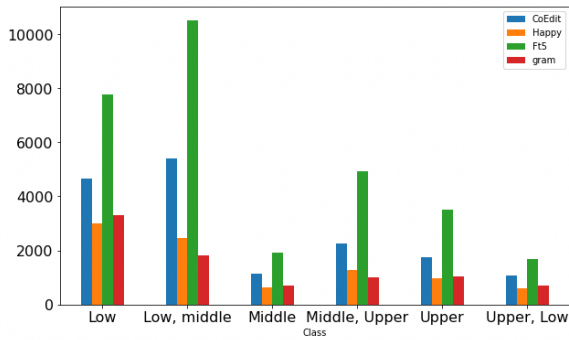[19]*Skag* refers to heroin in Scottish slang.

Figure 5: Distribution by class of utterances with at least one correction.

7 Discussion

Following established work in sociolinguistics, our results support our hypothesis that NLP systems display biases towards different sociolects, beyond geographical and ethnic differences.

On the difference between life and art: The use of media (if not synthetic) data as a stand-in for real-world examples is not uncommon in NLP (e.g. Plaza-del Arco et al., 2024; Lison and Tiedemann, 2016). Our work presents a starting point to study an aspect of linguistic variation that has thus far received little attention in the NLP literature. To substantiate our call to action, the characters' language need not be an exact replica but a good approximation of a given sociolect. However, we do not claim that all media portrayals are accurate representations: TV shows and films are written by certain people, for specific audiences, and the quality of acting varies. Through our selection process, we selected shows that would not caricature certain groups, including TV shows like *The Wire* and *Pose*, which feature a cast representative of the characters. Nevertheless, these results should be validated with real speakers.

NLP as an enforcer of "standard" language: Native speakers of any language have different registers available to them, including when interacting with NLP systems. While this may help mitigate some of the potential harms shown in this paper, our position is that humans should not have to adapt their language to interact with technology, but rather technology should adapt to humans. By dint of their use, NLP systems are setting a standard of language: NLP systems are becoming more common, not only in social contexts calling for formal (i.e. standardised) language, or in fields like Social Science (Ziems et al., 2024), but in everyday sce-

The Crown

And what a bunch of ice-veined monsters my family are.
And what a bunch of ice-veined monsters my family is.

Pose

The tittiness of it all.
The quality of it all.

Trainspotting

So, I'm off the skag .[19]
So, I'm off the grid.

Trailer park Boys

We be dope when the new NW A come out, know what I'm sayin'
We will be happy when the new NWA comes out, know what I'm saying?

Table 7: Examples of grammar correction in different shows. The corrected examples are from the T5 Gramar correction model.

narios. For example one of the proposed use cases for the Gramformer is correcting text messages as a built-in feature for messaging apps, language models inform systems like automatic correction. In Section 6, we show grammar-correction models erase particularities of certain language varieties like regional slang or norms. Likewise, LLMs like chatGPT now boast hundreds millions of users. Keeping in mind the findings from Gonen et al. (2023) and ours in Section 5, we may find users of lower SES have worse user experience with such systems. This widespread interaction with language technologies reinforces the lect of middle class, white English speakers as the standard even in quotidian activities.

Going forward While reporting on the socioeconomic background of study participants and dataset contributors is more complicated than other factors such as age or ethnicity (as it is not a one-point measurement and it is culture-dependent), we encourage researchers to report on this to draw a more accurate picture of the language varieties NLP will capture and to better benchmark systems across sociolects. Given the limitations of our study (most notably, restriction to English and group rather than individual scores, see the Limitations section below), we envision several possible avenues for

future research in this area: (1) the findings of this study should be validated for other languages where such differences have been observed; (2) the collection of a (possibly larger) corpus of unaggregated speakers; and (3) while we find support from previous research such as Flekova et al. (2016) for social media data, the findings should also be validated for other data forms such as long-form written texts.

8 Related Work

Works in NLP considering social class in the context of language use are few and far between (Cercas Curry et al., 2024), mostly focusing on predicting social class from texts (e.g. Basile et al., 2019; Lampos et al., 2014). Nguyen et al. (2016) ascribe the dearth of work in bias and social class to the lack of self-reported explicit labels in online user profiles. The difficulty in obtaining such data hinders efforts in addressing this issue. More recently, some work has focused on representational harms with regards to social class: Curto et al. (2022) investigate bias against the poor in NLP in terms of the language that surrounds poor people, considering the associations of 'poor' and 'rich' in embeddings and language models; Kiritchenko et al. (2023) instead focus on *aporophobia*, classist attitude stigmatising the poor, in the context of toxic language. To the best of our knowledge, we are the first to investigate how NLP systems perform with speakers from different social strata.

9 Conclusion

Social class plays an important role in people's identity construction, and is consequently strongly reflected in their language use. Despite its central status as a variable in sociolinguistics, there is so far little work in NLP engaging with social class and its impact. Our contributions are twofold: following methods from sociolinguistics, we use a data set of 95K movie transcripts that we annotate for portrayed class, race, and geographical variety (UK vs. US). We use this data to show that linguistic class markers greatly affect the performance of various NLP tools. Our findings suggest that speakers of low-prestige sociolects experience lower application performance on a range of tasks. These results suggest that we should actively incorporate social class as a variable in NLP systems if we want to make them more equitable and fair. Our data set also provides a starting point for more explorations of social class in NLP.

Limitations

Our study presents several limitations: First, we are limited to the study of class, region, and race in English. TV shows and films are biased in their representations of different identities—not all groups are (equally) represented and those that are may be stereotyped into certain roles. For example, in the initial top-200 TV shows list, we did not find any TV shows about lower-class women, Latinxs, upper-class Non-English Britons, or non-white Britons. For this reason, we excluded a gender-based analysis. Moreover, in this list, the majority of TV shows about Black Americans revolve around the drug trade.

Second, characters are not necessarily written by, nor perfectly depict the group they represent and this is potentially reflected in their language. To ensure high quality portrayals, we sourced highly rated and realistic shows, ensuring characters were not caricatures of the groups they portrayed. Having said that, a sociolinguistics review found that TV show and film data is increasingly common as data for the study of sociolects, particularly with regards to race and class (Stamou, 2014). More specifically, work in sociolinguistics has also defended the use of film data in conversation analysis (McHoul, 1987; Quaglio, 2008). Furthermore, TV shows like *The Wire* have been praised for their linguistic authenticity (Lopez and Bucholtz, 2017), reportedly primarily hiring Baltimore natives. For these reasons, we believe our study sufficiently captures the relevant linguistic phenomena.

Despite our best efforts, we were unable to reliably separate the utterances by character across episodes. Instead of introducing more sources of potential mix-ups, we therefore estimate the metrics over all characters in a show. Although this is a limitation in terms of the accuracy of the results, we expect that character-level annotations would strengthen our findings rather than negate them. Generally speaking, each TV show focuses on a single social stratum. A single character belonging to a different stratum would add noise and make the classes less clearly defined, so we suspect our results would be more marked with character-level annotations, not less. We leave the character-based separation for future work.

Finally, though our dataset is large in terms of utterances, it is limited in the number of characters. However, we present a case study hoping to motivate further research that may validate our findings.

Ethical Considerations

Our paper is arguing for fairness in the services provided by NLP systems for people of all socioeconomic backgrounds. However, we note that by showing measurable differences in language use across groups, we run the risk of profiling speakers. In our work, we used fictional characters to avoid this, but in order to ensure fairness of service all peoples must be represented. Future work should ensure that this is done respectfully and with consent from any participants.

Furthermore, measuring socioeconomic status is not a straightforward process. The class system used in this paper is fuzzy and tied to Western social structures that are not valid across the world. Future work should focus on measuring in the most appropriate way by following established metrics and guidelines from economics and other fields, such as Savage et al. (2013). We urge any future work to follow ethical guidelines when it comes to dataset and system development.

Acknowledgements

References

Jonathan Anderson. 1983. Lix and rix: Variations on a little-known readability index. *Journal of Reading*, 26(6):490–496.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Angelo Basile, Albert Gatt, and Malvina Nissim. 2019. You write like you eat: Stylistic variation as a predictor of social stratification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2583–2593, Florence, Italy. Association for Computational Linguistics.

Basil Bernstein. 1960. Language and social class. *The British journal of sociology*, 11(3):271–276.

Mary Bucholtz and Kira Hall. 2005. Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, 7(4-5):585–614.

Amanda Cercas Curry, Zeerak Talat, and Dirk Hovy. 2024. Impoverished language technology: The lack of (social) class in NLP. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8675–8682, Torino, Italia. ELRA and ICCL.

Eve V Clark and Marisa Casillas. 2015. First language acquisition. In *The Routledge handbook of linguistics*, pages 311–328. Routledge.

Cynthia G Clopper and David B Pisoni. 2006. The nationwide speech project: A new corpus of american english dialects. *Speech communication*, 48(6):633–644.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Georgina Curto, Mario Fernando Jojoa Acosta, Flavio Comim, and Begoña Garcia-Zapirain. 2022. Are AI systems biased against the poor? A machine learning analysis using Word2Vec and GloVe embeddings. *AI & society*, pages 1–16.

Penelope Eckert. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology*, 41(1):87–100.

Lucie Flekova, Daniel Preoţiuc-Pietro, and Lyle Ungar. 2016. Exploring stylistic variation with age and income on Twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 313–319, Berlin, Germany. Association for Computational Linguistics.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Hila Gonen, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. 2023. Demystifying prompts in

language models via perplexity estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10136–10148, Singapore. Association for Computational Linguistics.

Robert Gunning. 1968. *The Technique of Clear Writing*. McGraw-Hill Book Company, New York.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 103–112, Beijing, China. Association for Computational Linguistics.

Tyler Kendall and Charlie Farrington. 2018. The corpus of regional african american language. *Version*, 6:1.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Svetlana Kiritchenko, Georgina Curto Rex, Isar Nejadgholi, and Kathleen C. Fraser. 2023. Aporophobia: An overlooked type of toxic language targeting the poor. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 113–125, Toronto, Canada. Association for Computational Linguistics.

William Labov. 1964. *The social stratification of English in New York city*. Ph.D. thesis, Columbia University.

Vasileios Lampos, Daniel Preoţiuc-Pietro, Sina Samangooei, Douwe Gelling, and Trevor Cohn. 2014. Extracting socioeconomic patterns from the news: Modelling text and outlet importance jointly. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 13–17, Baltimore, MD, USA. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Qiuana Lopez and Mary Bucholtz. 2017. "How my hair look?" Linguistic authenticity and racialized gender and sexuality on The Wire. *Journal of Language and Sexuality*, 6(1):1–29.

Alec W McHoul. 1987. An initial investigation of the usability of fictional conversation for doing conversation analysis. *Semiotica*, 67(1-2):83–104.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. In *Interspeech 2004*, pages 2765–2768. ISCA.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234, Valencia, Spain. Association for Computational Linguistics.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Survey: Computational sociolinguistics: A Survey. *Computational Linguistics*, 42(3):537–593.

Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710, Torino, Italia. ELRA and ICCL.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Paulo Quaglio. 2008. Television dialogue and natural conversation. *Corpora and discourse*, pages 189–210.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

John R Rickford. 1986. The need for new approaches to social class analysis in sociolinguistics. *Language and communication*, 6(3):215–221.

Peter Saunders. 1990. *Social class and stratification*. Routledge.

Mike Savage, Fiona Devine, Niall Cunningham, Mark Taylor, Yaojun Li, Johs Hjellbrekke, Brigitte Le Roux, Sam Friedman, and Andrew Miles. 2013. A new model of social class? findings from the bbc's great british class survey experiment. *Sociology*, 47(2):219–250.

Sharon H. Saydah, Giuseppina Imperatore, and Gloria L. Beckles. 2013. Socioeconomic status and mortality. *Diabetes Care*, 36(1):49–55.

RJ Senter and Edgar A Smith. 1967. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories*, pages 1–14.

Anastasia G Stamou. 2014. A literature review on the mediation of sociolinguistic style in television and cinematic fiction: Sustaining the ideology of authenticity. *Language and Literature*, 23(2):118–140.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Elisa Usategui Basozábal et al. 1992. La sociolingüística de basil bernstein y sus implicaciones en el ámbito escolar. *Revista de educación*.

Melanie Weirich and Adrian P Simpson. 2018. Gender identity is indexed and perceived in speech. *PLoS One*, 13(12):e0209226.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Annotations per show

Table 8 shows the annotated demographics for each season/film.

	Class	Gender	Race	Geography
Arrested Development	Upper	M, F	White	USA
Sex and the City	Middle, Upper	F	White	USA
The Fresh Prince of BelAir	Middle, Upper	M, F	Black	USA
Big Little Lies	Middle	F	White	USA
Breaking Bad	Low, middle	M	White	USA
The Wire (S01)	Low, middle	M	Black, white	USA
The Wire (S03)	Low, middle	M	Black, white	USA
Fargo (S04)	Low, middle	M, F	Black, white	USA
Trailer Park Boys	Low	M	White	USA
When They See Us	Low	M, F	Black	USA
The Sopranos	Low	M	White	USA
Pose	Low	F, Trans	Black,Latino	USA
The Crown	Upper	M, F	White	UK
Downton Abbey	Upper, Low	M, F	White	UK
The IT Crowd	Middle	M	White	UK
Shetland	Middle	M, F	White	Scotland (UK)
T2.trainspotting	Low	M	White	Scotland (UK)
Trainspotting	Low	M	White	Scotland (UK)
Smother	Middle	M, F	White	Ireland (UK)
Derry Girls	Low	F	White	Ireland (NE)

Table 8: Annotated demographics for each movie or TV show. M denotes Men and F, women.