Duncan, Daniel (2024) Does ChatGPT have sociolinguistic competence? Journal of Computer-Assisted Linguistic Research 8: 51-75. ISSN: 2530-9455.

Does ChatGPT have sociolinguistic competence?

Daniel Duncan*
Newcastle University, United Kingdom
*Corresponding author: daniel.duncan@ncl.ac.uk

Received: 21 June 2024 / Accepted: 24 September 2024 / Published: 15 November 2024

Abstract
Large language models are now able to generate content- and genre-appropriate prose with grammatical sentences. However, these targets do not fully encapsulate human-like language use. For example, set aside is the fact that human language use involves sociolinguistic variation that is regularly constrained by internal and external factors. This article tests whether one widely used LLM application, ChatGPT, is capable of generating such variation. I construct an English corpus of "sociolinguistic interviews" using the application and analyze the generation of seven morphosyntactic features. I show that the application largely fails to generate any variation at all when one variant is prescriptively incorrect, but that it is able to generate variable deletion of the complementizer *that* that is internally constrained, with variants occurring at human-like rates. ChatGPT fails, however, to properly generate externally constrained complementizer *that* deletion. I argue that these outcomes reflect bias both in the training data and Reinforcement Learning from Human Feedback. I suggest that testing whether an LLM can properly generate sociolinguistic variation is a useful metric for evaluating if it generates human-like language.

Keywords: large language models, ChatGPT, variation, morphosyntactic variation, sociolinguistics

1. Introduction

Large language models (LLMs) have recently progressed to the point that, given a prose input, a model is able to probabilistically generate grammatical sentences which are content- and genre-appropriate. This improved text parsing and generation has been accompanied by increased availability to the general public through applications such as ChatGPT, which is based on GPT-3.5 (OpenAI 2022). These models are typically trained in two phases: first on large volumes of text (Brown et al. 2020), and secondly on human input. For example, ChatGPT is trained by human input first through supervised fine-tuning on human-generated

DOI: HTTPS://DOI.ORG/10.4995/JCLR.2024.21958
PUBLISHER: UNIVERSITAT POLITÈCNICA DE VALÈNCIA.

conversations, and then on human judgements of potential outputs through Reinforcement Learning from Human Feedback (RLHF) (OpenAI 2022).

LLMs have found success in tasks designed to assess their capabilities with syntactic phenomena like long-distance dependencies (Chowdhury and Zamparelli 2018; Gulordava et al. 2018). These successes suggest that language models learn syntactic generalizations at least to a degree (Linzen and Baroni 2021). It is unclear, however, to what extent these models have a full linguistic competence. For example, recent results suggest that LLMs do not successfully make broader generalizations based on semantic patterns (Asher et al. 2023; Lam et al. 2023). In this sense, an important area of current research concerns which aspects of linguistic knowledge these models do or do not have (Zhuang et al. 2023).

It has been widely known for some time that grammatical competence is necessary but not sufficient for generating language in a human-like manner (e.g., Hymes 1966). For example, human speakers have sociolinguistic competence (cf. Bayley and Regan 2004), in which they know what features are variable in their community, how this variation is structured, and the social meaning of linguistic features. That models such as ChatGPT are able to generate text in a variety of registers may suggest that they have human-like sociolinguistic competence. However, recent work has noted that both social factors and language variation more generally have usually been neglected in NLP research (Yang and Eisenstein 2017; Hovy and Yang 2021; Kulkarni and Raheja 2023). Indeed, Zhuang et al. (2023) omit sociolinguistic competence from their discussion of core linguistic competencies that LLMs should be evaluated against. It is therefore unclear overall whether LLMs can perform a range of tasks indicative of sociolinguistic competence.

Linguistic variation is often seen as a problem, or error source, in NLP research, and research which does engage with it emphasizes that social meaning may be adduced from it (Nguyen, Rosseel, and Grieve 2021). More typically, when variation has been included in NLP research, it is typically related to efforts to identify and eliminate social bias from models (Mendelsohn, Sap, and Le Bras 2022; Rauh et al. 2022). With respect to LLM capabilities, ChatGPT performs rather poorly at classifying dialect features (Ziems et al. 2023). That said, given the size of the training data and probabilistic outputs of LLMs, we might expect LLMs to generate variation that approximates human sociolinguistic competence as an emergent product (Bowman 2023) of these “stochastic parrots” (Bender et al. 2021). However, to the best of my knowledge, whether LLMs have achieved sociolinguistic competency with respect to language variation in generated text has not been tested.

In this paper, I briefly outline a set of criteria that LLMs should be able to successfully satisfy in order to be considered sociolinguistically competent. I focus on sociolinguistic competence from a variationist perspective, with the primary consideration of whether an LLM can generate human-like patterns of linguistic variation. I then demonstrate how to test an LLM for sociolinguistic competence through a case study. I use ChatGPT to generate a corpus of English-language “sociolinguistic interviews” and examine the model’s generation of seven morphosyntactic variables. As I show, the model generally falls far short of human-like variation patterns. However, it is able to generate variable deletion of the complementizer *that* at human-like rates, constrained by grammatical context in a human-like manner. I discuss what the overall failure, yet success with respect to complementizer deletion, tells us about biases in ChatGPT's model training, and advocate for further work assessing LLMs' capabilities to achieve sociolinguistic competence.

2. Assessing sociolinguistic competence

In the sociolinguistics of second language acquisition, one question that may be asked of an L2 speaker is whether their language use is native speaker-like with respect to variation in production and perception (cf. Bayley and Regan 2004). In the same way, we can ask whether the language use of an LLM is human-like. From a variationist perspective, this question refers first and foremost to human speakers' ability to vary in their linguistic production and evaluate the meaning of linguistic variation (cf. Weinreich, Labov, and Herzog 1968). At the same time, this question involves the ability of human speakers to agentively and/or performatively use third-order indexicals (Johnstone, Andrus, and Danielson 2006)—in other words, to know the social meaning of a sociolinguistic variant and use that variant to convey that meaning—as well as to innovate novel variants and command the discourse-pragmatic functions of their language. In studies of human communities, these abilities often have a temporal component (Labov 1972; Johnstone, Andrus, and Danielson 2006): an innovative sociolinguistic feature is used variably by speakers, who subsequently learn to attribute social meaning to the variable as its use becomes consciously associated with groups of speakers. Only then can speakers learn to use the variable agentively.

These competencies in the production and evaluation of linguistic variation form the basis of tasks that an LLM with sociolinguistic competence would be able to complete. In this sense, a fully sociolinguistically competent LLM should be able to output human-like patterns of variation as though it were a member of a prompted speech community, identify input text as associated with a particular group of speakers in a human-like manner, and output text that utilizes sociolinguistic variation to evoke a prompted persona. Following findings from community-based studies that the latter two tasks are downstream from the former (Labov 1972; Johnstone, Andrus, and Danielson 2006), I suggest that of these tasks the most basic one associated with sociolinguistic competency from a variationist perspective is to generate human-like variation. That is, if the presence of robust variation precedes the assignment of social meaning, to replicate such patterns would be the basic associated task for an LLM.

This task can be broken into four subtasks: a) generating dialect-specific features (for example, regional variation), b) generating variable outputs, c) generating variation constrained by language-internal factors (i.e., factors associated with the grammatical context of the utterance), and d) generating variation constrained by language-external factors (i.e., factors associated with social and/or cognitive processes separate from grammatical context). Success or failure at each of these tasks can be assessed quantitatively. For this reason, I suggest that quantitative analysis of variation in LLM-generated text, compared to a baseline of past sociolinguistic research into a given linguistic feature, can serve as a test for determining whether an LLM meets a basic threshold for sociolinguistic competence. This does not mean that satisfying this test means an LLM is fully sociolinguistically competent. Indeed, an LLM may prove to be able to generate variation without being able to use variants to index a persona, or vice versa. However, an LLM failing this test certainly is not fully sociolinguistically competent. Because this test is concerned with factors constraining variation in generated text, I suggest that any appropriate testing of an LLM's sociolinguistic competence will concern morphosyntactic variation rather than lexical or phonological variation. Morphosyntactic variation is particularly appropriate for such a test because it displays language-internal conditioning, whereas lexical variation is less likely to depend on the grammatical context of the utterance. At the same time, variants are more consistent orthographically than folk spellings of phonological variation.

Additionally, LLMs can be trained at least in part on internet language use (for example, the GPT-3 LLM was trained in part on the CommonCrawl internet corpus, see Brown et al. 2020). As Squires (2016, 2) notes, computer-mediated language use is "just like language used outside of it…in the way that it participates in linguistic and social processes". This point extends both to variation appearing and being constrained by grammatical context much like as in spoken language (Eisenstein 2015) and to linguistic variants being used agentively to convey social meaning (Ilbury 2019). Given this, it is reasonable to expect that an LLM trained on written language alone would nevertheless be capable of generating variation if it were sociolinguistically competent. That internet language use may comprise a subset of an LLM's training data means that morphosyntactic variation in particular is quite likely to be present in the training data. Indeed, such variation appears in informal online text (Bleaman 2020), including local dialect variants on sites associated with specific local communities (Duncan 2019; Pearce 2021). Furthermore, regional dialectal variation is present even in formal writing (Grieve 2016), which indicates that morphosyntactic variation would appear in training data even if an LLM were solely trained on formal writing.

3. Sociolinguistic competence as a controllable text generation task

The baseline task outlined above, in which the sociolinguistic competence of an LLM is tested by quantitatively examining generated text for the presence and patterning of linguistic variation, casts sociolinguistic competence in part as a "controllable text generation" task. In contrast to uncontrolled text generation, in which output text is generated as a probabilistic function over previously generated text, in controlled text generation the output text is generated as a probabilistic function over previously generated text *given a particular constraint* (Yu, Yu, and Sagae 2021). Chen et al. (2024) divide these into hard and soft constraints. Hard constraints are related to the form of the output itself, specifying matters such as lexical items or syntactic features found in the output text, or the length of the output text itself. Meanwhile, soft constraints relate to the content of the output, specifying matters such as the topic or sentiment of the output. Another soft constraint on the output text may be the intended demographic profile of the "speaker" (Prabhumoye, Black, and Salakhutdinov 2020); such demographic profiles have typically focused on generating specific personas or personality types (Li et al. 2016; Zhang et al. 2018). Given such constraints, the goal of controllable text generation is effectively to consistently generate context-acceptable outputs.

As with other tasks set to language models, current approaches to controllable text generation work to achieve this through the use of large pre-trained models built with the Transformer architecture (Vaswani et al. 2017; see Qiu et al. 2020; Han et al. 2021; and Li et al. 2024 for comprehensive reviews). These models are often initially trained through unsupervised learning over massive text corpora. However, because they are simply trained on text at this point, such models do not initially have the capacity for controllable text generation as they lack the evidence of contexts or attributes that a human user may be seeking to generate (although at the largest scales these models often have emergent abilities such as generating text based on zero- or few-shot natural language prompts, see Wei et al. 2022). Outside of building the model through supervised learning and tagging training data for desired attributes (Yu, Yu, and Sagae 2021), there are a few ways to advance a pre-trained model to better generate controlled outputs. One is to utilize a second model which includes the desired attributes, so that the initial pre-trained language model is not changed (Dathathri et al. 2020; Yu, Yu, and Sagae 2021). However, this approach requires the desired attributes to be discrete and pre-defined, and as such is less viable with natural language prompts (Chen et al. 2024). For controllable text generation using natural language prompts, current approaches involve either fine-tuning the model or prompt engineering. Fine-tuning is updating the initial model parameters, often through supervised training on task-specific annotated data. For example, ChatGPT is fine-tuned from GPT-3 using conversational data to better generate conversational text (OpenAI 2022). Prompt engineering, on the other hand, involves testing a variety of prompts to find which template and phrasing best yields the desired output. This can be difficult, as Gao, Fisch, and Chen (2021, 3816) note that it requires "both domain expertise and an understanding of the language model's inner workings". Prompt engineering may go hand in hand with fine-tuning, as a pre-trained model may be fine-tuned to better respond to a particular style of prompt. More recently, approaches to prompt engineering have included the automatic generation of prompts (Jiang et al. 2020; Gao, Fisch, and Chen 2021).

In the context of this case study, ChatGPT has already been fine-tuned to generate text based on natural language prompts. The research question, then, is fundamentally whether the model as fine-tuned is able to generate sociolinguistic variation in text.[1] As noted, to properly generate variation involves not only generating regional dialect features, but constraining variability for language-internal and -external factors. Generating language-internally constrained variation is an uncontrolled text generation task: Does the rate at which a model outputs a particular linguistic variant depend on the previously generated text (i.e., grammatical context) in a human-like manner? Generating language-externally constrained variation, whether as a matter of regional dialectal variation or some other social constraint, is a controllable text generation task: Given the demographic context and previously generated text, does the model generate a linguistic variant at human-like rates? The evaluation of success at this task differs from other proposed benchmarks for controlled task generation. For example, Chen et al.'s (2024) CoDI-Eval benchmark evaluates a model's ability to generate text based on sentiment, topic, keyword, length, and toxicity avoidance. While rigorously quantifiable, evaluating controllable text

[1] Because we lack exact details of how the model was fine-tuned, I will not speculate as to whether this question is about a task with zero-shot, few-shot, or other training.

generation for sociolinguistic competence in the sense of properly generating variation is a novel criterion.

Controllable text generation has a wide range of applications, from machine translation to dialogue systems, to text summarization, to story generation (Li et al. 2024). Properly generating sociolinguistic variation is a task that is crucial to most, if not all, of these applications. To a large extent, the evaluation of a model's performance on a controllable text generation task is a question of whether the output is acceptable to humans (Belz and Reiter 2006). In this sense, human-like text, such as that which displays variation at human-like rates and patterns, is a necessary feature of output text, as it will read as more acceptable to a human evaluator. After all, as Li et al. (2024, 19) note, "To optimize [pre-trained language models] for real-world deployment, the most important consideration is to align the behaviors of [pre-trained language models] with human expectations". This is to say that sociolinguistic competence is a core element of controlled text generation; a sociolinguistically competent language model will generate controlled outputs more acceptably than a non-sociolinguistically competent model. At the same time, because sociolinguistic competence is part of linguistic competence, (Hymes 1966), the close connection between sociolinguistic competence and controlled text generation means that controllable text generation is a core part of modeling language more generally. That is, testing a model's capability for controllable text generation is as important as other tasks for demonstrating the linguistic competency of a language model.

4. Methods

Here I illustrate a prompt-based approach to obtaining text suitable for addressing the overall task of generating sociolinguistic variation and four accompanying subtasks outlined in section 2 of generating dialect-specific features, generating variable outputs, generating variation constrained by language-internal factors, and generating variation constrained by language-external factors. Because success or failure at the subtasks is assessed from a variationist perspective, the approach here is designed to mimic best practices in variationist data collection. Such data collection typically involves the sociolinguistic interview. This is a "controlled speech event" (Becker 2013, 92) in which one component includes informal conversation. In this conversation, the researcher utilizes prompts such as childhood memories and family traditions to elicit naturalistic speech. Sociolinguistic interviews are analyzed in aggregate, comparing production across a range of speakers sampled from a larger community in order to determine how linguistic variation is patterned within that community (Becker 2013). The aim is for the naturalistic speech in a sociolinguistic interview to approach the "vernacular," defined as how one speaks when they are in a normal setting. Crucially, as Becker (2013) notes, targeting the vernacular should not be confused with targeting non-standard speech. In this sense, an appropriate prompt will elicit speech-like text without overtly cueing the model to target non-standard speech (i.e., asking for a "strong accent" or the like), or indeed any specific linguistic features.[2]

[2] Although many sociolinguistic interviews are collected for phonetic rather than morphosyntactic analysis, the methodology of the sociolinguistic interview is the same regardless of the linguistic feature(s) ultimately under

In fact, to include such cues would conflate the different tasks associated with sociolinguistic competence and thus cause significant difficulties for interpreting the model output. For example, it is possible that prompting a "strong, working-class" accent (or the like) would simply instruct an LLM to output text as though it were uttered by a member of the particular community described in the prompt. More likely, however, such instructions are prompting the described person as a character to be performed. In other words, the outputted text will be ambiguous as to whether it models the variation in the vernacular or speech which utilizes features agentively to convey social meaning. The difficulty with such a prompt, then, is that an LLM's success or failure in response to it cannot be evaluated as we do not know what task it succeeded or failed at.

The same issue applies if we instruct the LLM to output text that includes specific linguistic features. For example, prompting the model to "use regional pronunciations and grammatical forms"(or something similar) may be instructing the model to output utterances from a member of a particular community, but may also be asking it to output features that it has associated with a particular group of speakers. There is a subtle, but crucial distinction between these tasks; the former output would constitute variation in vernacular speech, while the latter output privileges features which index a specific social meaning. As a general point, then, the more information that we provide to the model in the prompt, the less able we are to interpret what the output represents. For this reason, in this case study I aim to provide as little information in the prompt as possible in order to better establish a baseline understanding of LLMs' ability to generate variation. While I acknowledge that providing more detail in a prompt may result in what at face value appears to be a better output, I contend that we need the baseline provided by a more minimal prompt to successfully understand "better" outputs.

To this end, the May 24, 2023, release of ChatGPT[3] was used to generate the informal conversation component of 80 "sociolinguistic interviews" using the following prompt:

> Transcript of a sociolinguistic interview[4] with a *65 year old, working-class woman* from Newcastle-upon-Tyne. The interview should elicit narratives about the following topics: interviewee demographic information, the neighborhood the interviewee grew up in, events that occurred in the interviewee's community, childhood games, school days, adventures with friends, the interviewee's family, holiday traditions, the interviewee's opinion of community changes, and local identity.

study given the goal of eliciting naturalistic speech. Likewise, interview transcripts are typically in standard orthography regardless of the feature under study. While some morphosyntactic variables are too rare to be reliably elicited through this methodology, this is not an issue for the variables considered in this case study.

[3] The online interface at https://chat.openai.com/ was used. Release notes to distinguish this release version from previous/future releases are available at https://help.openai.com/en/articles/6825453-chatgpt-release-notes.

[4] One might question the use of the phrase "sociolinguistic interview," as we do not know whether the model's generation of such a genre approximates sociolinguists' understanding of the genre. I do not find this issue troubling; variationist sociolinguists quite often use multiple styles of interview recording under the guise of "sociolinguistic interview" rather than the formally structured sociolinguistic interview (see Becker 2013 for discussion). In other words, sociolinguists themselves are not clear on what this genre entails. Given this, model outputs approximating semi-structured interviews (taken broadly) may therefore be used in comparison to community-based studies regardless of whether they are "true" sociolinguistic interviews.

The italicized material was edited across the text generation to create a corpus of text balanced for three binary language-external factors (10 "interviewees" per unique combination). These included INTERVIEWEE AGE (25 vs. 65 years old), SOCIAL CLASS (working-class vs. middle-class), and INTERVIEWEE SEX (male vs. female). The framing of this prompt was designed to place the interviewee as member of a community that fit a particular demographic profile without overtly instructing the model to output text evocative of that kind of speaker. Had that type of instruction been present, it would have introduced the task confounds discussed above. Effectively, this prompt amounts to giving the model speaker demographics and no other guidance regarding linguistic form.

Because the output was relatively short, additional text was generated by subsequently prompting "Can you make the interview twice as long?". This approach yielded a corpus of approximately 135,000 words of interviewee text. However, because the model created extended interviews by generating new text in some cases but by repeating the initial text with additional content in others, the full available text for analysis is likely closer to 100,000 words.[5]

Before proceeding further, we must confirm whether this process yielded interview-like text, and whether the prompt was sufficient to generate text localized to social groups within Newcastle. That is, setting the question of linguistic variation aside, did ChatGPT appropriately generate texts controlled for style, topic, and interviewee demographics? Manual inspection of the generated texts shows that each of the 80 texts is effectively the same material, as the same stories and similar phrasing occur across all of them. This limited range of content echoes that found in, for example, Jentzsch and Kersting's (2023) exploration of ChatGPT's ability to generate humorous material. As such, it may be indicative of a low model temperature leading the model to favor a small output range (Chung, Kamar, and Amershi 2023).[6] That said, this material is structured like a sociolinguistic interview, and the text describes memories and attitudes that would likely appear in a sociolinguistic interview. In this sense, the prompt successfully led the model to output interview-like text. Note that this question is only a matter of content; because the sociolinguistic interview does not target any particular linguistic features, their presence/absence is not indicative of interview-like text.

Contentwise, the texts include details which clearly place speakers within Newcastle. The texts differ by age group with respect to the inclusion of specific historical events. Interviewee sex is less distinguished; the texts tend to describe the same childhood activities regardless of sex or age. The texts accurately maintain interviewee sex throughout any given interview. Social class is perhaps the most clearly distinguished factor across texts. Interviewees are described as growing up in specific neighborhoods within Newcastle (Table 1). Working-class interviewees are primarily placed in Byker, Walker, and Benwell, three neighborhoods both popularly known as working-class (WC) and quantitatively found to include Lower-level Super Output Areas among the 10% most deprived neighborhoods in the UK on the Index of Multiple

[5] Assessing this is not simply a matter of deleting repeated material, because added material often takes the form of clauses added onto a short utterance. In such a case, the researcher is required to subjectively decide whether the initial instance has been repeated, or whether the utterance is entirely new.

[6] I used the free research preview of ChatGPT; this does not have a setting to vary the model temperature.

Deprivation (Department for Levelling Up, Housing and Communities and Ministry of Housing, Communities, & Local Government 2019; Noble et al. 2019). In contrast, the majority of middle-class (MC) interviewees are placed in Gosforth, Jesmond, and Heaton, which include LSOAs among the 10% least deprived neighborhoods in the UK and are popularly known as middle-class neighborhoods.

	Neighborhoods		
	Gosforth/Heaton/Jesmond	Benwell/Byker/Walker	Other
MC	22	11	7
WC	0	36	4

TABLE 1. NEIGHBORHOODS ASSIGNED TO INTERVIEWEES IN GENERATED TEXTS

Social class and age are also reflected in the size of family that interviewees are assigned in the text (Figure 1). Linear regression shows that older and working-class interviewees are assigned significantly larger numbers of children in families than younger and middle-class interviewees (baseline: older middle-class; age: β = -1.125, p = 0.0172; class: β = 2.1118, p << 0.0001; r^2 = 0.5068).[7]

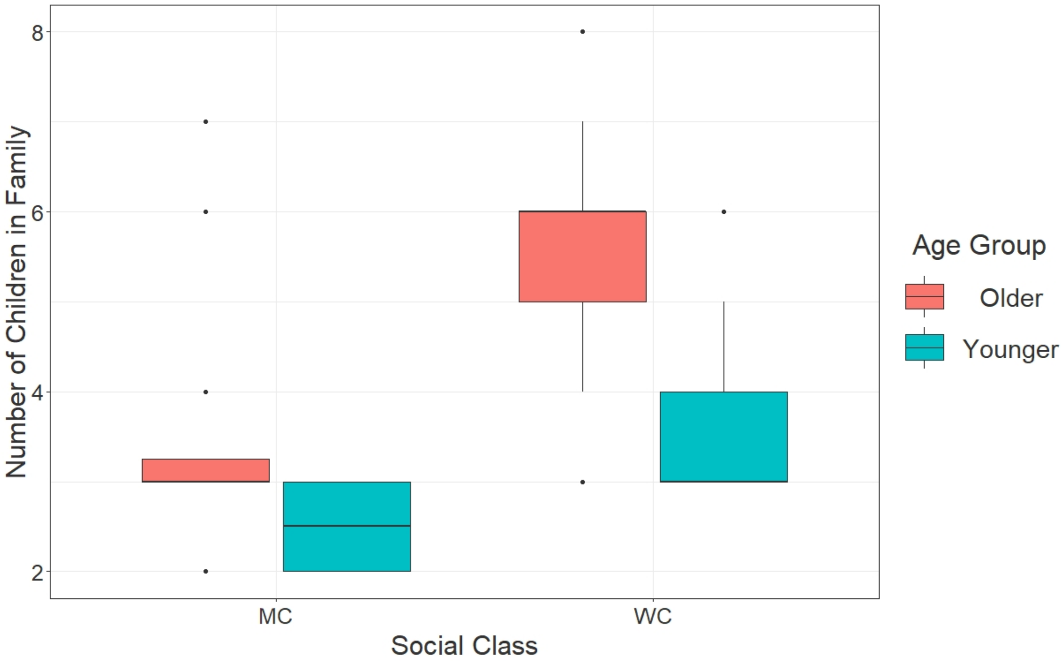

FIGURE 1. NUMBER OF CHILDREN IN INTERVIEWEES' FAMILY IN GENERATED TEXTS

These correlations do not reflect British population trends across the 20th century; in particular, the linking of large family size to midcentury working-class families is a gross overestimate (Laslett 1969; Hunt 2009). This suggests that the model is reproducing assumptions found in the source text or human training. Regardless, the correlations illustrate that the prompt given to the model was sufficient to generate interviewees categorized by language-external factors.

Given that the texts appear to sufficiently resemble a balanced corpus of sociolinguistic interviews, ChatGPT has evidently succeeded at the controllable text generation task with

[7] Some interviewees did not explicitly enumerate the number of children. I have treated these instances as NAs.

respect to style, topic, and interviewee demographics. As such, we can proceed to examining the language use within them. Here I focus primarily on morphosyntactic variation as opposed to lexical variation. Although for some interviewees the model did generate some lexical items associated with Tyneside English such as *lad*, *lass*, and *bairn* (Beal, Burbano-Elizondo, and Llamas 2012), as noted above lexical variation does not lend itself to examining language-internal constraints in the same way that morphosyntactic variation does. Because generating variation constrained by language-internal factors is a key task an LLM must succeed at to be sociolinguistically competent, lexical variation is not viable for analysis despite the presence of localized lexical items in some interviews. In order to test whether the model could generate morphosyntactic variation, I searched for seven variables which are well-described in the sociolinguistic literature (including documented usage on Tyneside) and display robust intra-speaker variation. By robust variation, I mean that each variant of each variable discussed here occurs in over 10% of tokens in studies of human speech. Given this, the presence of variability in LLM-generated text is expected of these variables if the LLM is sociolinguistically competent. Three of these variables are regionally restricted to Tyneside English in form or constraints (Beal 2004):

(1) *Us* as first-person singular object pronoun
He's always been like a mentor to *me/us*. ['Interview' BV]

(2) *We* as first-person plural object pronoun
It's always been a special time for *us/we*. ['Interview' BL]

(3) *Wor* as first-person plural possessive pronoun
Holidays were always special in *our/wor* family. ['Interview' AA]

The other four variables are widespread cross-dialectally in English:

(4) Complementizer *that* deletion (Tagliamonte and Smith 2005; Kearns 2007b)
I believe (*that*) the sense of belonging here is strong. ['Interview' AY]

(5) Nonstandard preterites *come*, *done*, *seen* (Tagliamonte 2001; Anderwald 2009)
…we *saw/seen* each other every day. ['Interview' BF]

(6) Participle-to-preterite leveling (Eisikovits 1987; Chatten et al. 2022)
I've *seen/saw* quite a few changes over the years. ['Interview' BA]

(7) *-body/-one* variation (D'Arcy et al. 2013)
I miss the days when every*one/body* knew each other. ['Interview' F]

In addition to the reasons for selection outlined above, each variable was selected because it is relatively frequent and thus would appear in sufficient numbers to be able to quantitatively assess whether the model can generate variation, and if so, whether the variation is constrained by language-internal and -external factors.

5. Results

This section assesses ChatGPT's ability to generate morphosyntactic variation in a human-like manner. I begin by considering whether the model can generate regional variation, and then expand this to a consideration of whether variation can be generated at all. In general, the model fails to appropriately generate variation. However, one linguistic feature under examination here, complementizer *that* deletion, is produced variably in the model output. I test whether this variable is constrained in a human-like manner.

5.1. Regional variants

The three Tyneside English-specific variables of interest are not generated beyond the token usage of a well-stereotyped feature. Singular *us*, plural *we*, and possessive *wor* are all well-attested in Tyneside English (Beal 2004), and human speakers within the North East of England associate the latter two of these with Newcastle specifically (Childs, Llamas, and Watt 2021). However, both plural *we* and possessive *wor* are categorically absent from the ChatGPT-generated texts despite the variables occurring frequently enough (n = 421 *us*; n = 1068 *our*)[8] that we would reasonably expect to see tokens of the local variant among human speakers.

In contrast to the plural pronouns, the first-person singular pronoun does display limited variability (n_{us} = 2-6/159-163, 1.3-3.7%). The total number of singular *us* tokens is given as a range due to ambiguity. Two tokens, which occur as the final utterance in the text, are unambiguous:

(8) Anytime you want to chat, just give *us* a shout.

The remaining four potential instances are less clear; while *us* co-occurs with a subject *I* in a nearby clause, it is unclear from context whether the speaker is referencing themselves or a generic plural group:

(9) *I* had some great teachers who really cared about *us*.

Regardless, singular *us* would be expected to occur far more frequently than the model generates. Thus, the model does associate Newcastle with a single regional linguistic feature to a highly limited degree, but this association does not extend to consistently producing this feature at human-like rates.

5.2. Cross-dialectal variation

The model's inability to properly generate regional morphosyntactic variation reflects a general difficulty with generating morphosyntactic variation. For example, noncanonical verb morphology is attested in Tyneside English both with respect to the participle variably surfacing as the preterite (Chatten et al. 2022) and preterites of verbs like *come, do*, and *see* variably surfacing as the participle (Serbicki, Lan, and Duncan 2023). Variation in the participle occurs much more widely across Englishes in informal speech (Eisikovits 1987; Bloomer 1998; Kemp et al. 2016; Chatten et al. 2022), large corpora, and the English-speaking internet at large (Geeraert and Newman 2011). Meanwhile, preterite *come* is so widespread cross-dialectally (Tagliamonte 2001; Levey, Fox, and Kastronic 2017; Jankowski and Tagliamonte 2022) that Chambers (1995, 240) considers it to be a "vernacular universal" of the language. Thus, even if the model was generating a generalized English variety, we would expect to see variable noncanonical verb morphology. However, both participles (n = 314) and preterites (n = 296)

[8] N's omit any duplicates introduced when prompting a longer interview.

appear categorically in the canonical form.

Like verbal morphology, whether groups quantified with *no*, *any*, etc., appear with *-body* or *-one* is variable cross-dialectally (D'Arcy et al. 2013), and we would expect to see variation within the ChatGPT-generated text. This variable nearly categorically surfaces as *-one* (n_{one} = 182/187, 97.3%), the variant which D'Arcy et al. (2013) suggest is effectively the prescriptive norm as it is the variant most widely used in formal writing. However, it should be noted that the vast majority of tokens (n = 151) have the quantifier *every*. Because this variable is strongly conditioned by the quantifier (D'Arcy et al. 2013), it is possible that were the variable generated with other quantifiers, the rate of *-body* would increase. Complementizer *that* deletion displays comparatively robust variation ($n_{deleted}$ = 110/150, 73.3%). This high deletion rate is in line with previous studies of the variable (Tagliamonte and Smith 2005; Bleaman et al. 2014). Overall then, the model largely fails to generate variation, but can do so for a restricted number of linguistic features.

5.3. Constrained variation

For an LLM, sociolinguistic competence entails not simply generating linguistic variation, but generating variation constrained by language-internal and -external factors. In this data, the question of whether the model can properly generate constrained variation is best tested on complementizer *that* deletion, as ChatGPT produced both the overt *that* variant and null variant of this sociolinguistic variable. In human speech, this variable is robustly constrained by grammatical and cognitive factors related to the complexity of the utterance (Tagliamonte and Smith 2005; Bleaman et al. 2014). The matrix verb also influences production: shorter verbs with Germanic etymologies tend to favor deletion (Kroch and Small 1978), while collocations such as *I think*, *I mean*, etc., strongly favor deletion (Tagliamonte and Smith 2005; Kearns 2007a). Given this, tokens in the generated text were coded for a set of similar factors: the VERB TYPE (*think* vs. other Germanic vs. Romance vs. copular construction), NUMBER OF SYLLABLES in the verb (monosyllabic vs. polysyllabic), and matrix/embedded subject COREFERENTIALITY (same vs. different). In addition, the language-external factors of AGE GROUP, SEX, and CLASS were considered.

The data was analyzed using logistic regression in R (R Core Team, 2020). I began with a mixed-effects regression, using the lme4 package (Bates et al., 2015), which included each of the above factors as fixed effects and interviewee as a random intercept. The model was then stepped down by removing factors to obtain the lowest possible AIC value in the final model. This process removed the random intercept and number of syllables from the final model. The lack of a random interviewee intercept indicates that there are no individual differences between interviewees. Table 2 shows the overall model results; the overt complementizer was set as the application value. As seen, there are significant effects of subject coreferentiality, interviewee age, and interviewee sex. The null complementizer is favored with coreferential subjects, while the overt complementizer is favored by younger or female interviewees.

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept (Older WC Male, Germanic matrix verb, non-coreferential subjects)	-1.931	0.61	-3.18	**0.0015**
Age Group—Younger	1.421	0.50	2.84	**0.0045**
Sex—Female	1.019	0.48	2.14	**0.0322**
Class—MC	0.726	0.49	1.48	0.1396
Verb Type—copula	0.961	0.94	1.02	0.3078
Verb Type—think	-19.611	1668.50	-0.01	0.9906
Verb Type—Romance	-0.427	0.61	-0.70	0.4812
Coreferential Subjects—Yes	-3.351	1.10	-3.05	**0.0023**

TABLE 2. LOGISTIC REGRESSION OF VARIABLE COMPLEMENTIZER DELETION IN CHATGPT-GENERATED TEXT

The complementizer was categorically deleted when the matrix verb was *think*, yielding the extreme estimate and p-value in Table 2. However, it appears to be a real effect as there are 31 tokens with *think* as the matrix verb. Post hoc comparisons of the factor levels using Tukey contrasts in the *multcomp* R package (Hothorn, Bretz, and Westfall. 2008; R Core Team 2020) show no significant differences in production among other Germanic verbs, Romance verbs, and copular constructions. The coreferentiality effect shows that complementizer deletion is strongly favored when the matrix and embedded clause subjects are the same (Figure 2). Both the effects of *think* and subject coreferentiality replicate previous studies of human speech (Tagliamonte and Smith 2005; Bleaman et al. 2014).

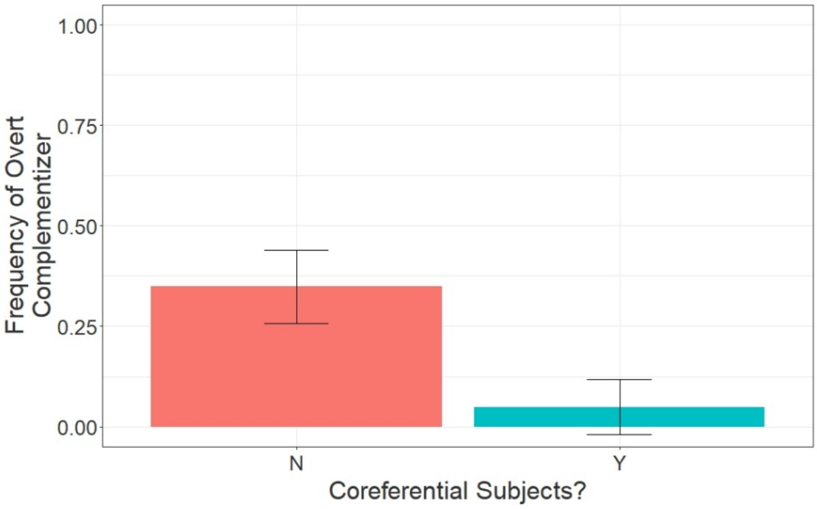

FIGURE 2. COMPLEMENTIZER DELETION IN CHATGPT-GENERATED TEXT BY SUBJECT COREFERENTIALITY

Age group and sex significantly influence rates of complementizer deletion, with model runs prompting a younger or female interviewee favoring the overt complementizer.

6. Discussion

As seen, ChatGPT-generated text produces common Tyneside and cross-dialectal sociolinguistic variables mainly by categorically generating the canonical variant. This categorical and near-categorical usage of the prescriptive norm is likely indicative that the model is insufficiently generating variation. This interpretation of the data is particularly clear with respect to the Tyneside English-specific features of first-person singular object *us*, first-person plural object *we*, and first-person plural possessive *wor*, as well as the cross-dialectally widespread features of nonstandard preterites and participle-to-preterite leveling. The cross-dialectal *-body/-one* variable, which was generated nearly categorically as the prescriptive *-one* variant, can likely be viewed as insufficiently generated variation as well. However, because nearly 81% of *-body/-one* tokens had the quantifier *every*, it should be noted that additional data with the quantifiers *any*, *no*, and *some* may reveal variability not present in the current data.

For each of the variables generated categorically or near-categorically, such production is indicative of a failure to generate constrained variation as well. After all, if variability among human speakers is constrained by a factor in any given context, categorical generation of only one variant simultaneously overgenerates the favored variant and undergenerates the disfavored variant. In contrast, the cross-dialectal complementizer *that* deletion variable does display human-like rates of variability. There is limited evidence of ChatGPT properly generating language-internal constraints on this variable, as it properly produces the null variant comparatively more often when matrix verb is *think* and when the matrix and embedded subjects are identical. However, the lack of further lexical effects does not replicate the patterns found in previous studies of human speech. Overall, with respect to language-internal factors, ChatGPT is able to properly generate constrained variation, but in a limited capacity and less robustly than human speakers. Similarly, with respect to language-external constraints, ChatGPT generates statistically significant differences in variant rates based on the demographic profile of the interviewee; model runs with younger and female interviewees comparatively favored the overt *that* complementizer. At first glance, this would suggest that the model successfully generates language-external constraints on variation. However, these constraints are not human-like because collocation effects in human speech are driving adoption of the *null* complementizer (Tagliamonte and Smith 2005), whereas this data would be indicative of change toward the *overt* complementizer in a study of human speech. These effects thus represent an overall failure of the model to properly generate language-externally constrained variation. As such, although ChatGPT can successfully generate texts controlled for style, topic, and interviewee demographics, it fails at controllable text generation with respect to linguistic variation within the text. This is particularly noteworthy in comparison to ChatGPT's relative success at generating variation constrained by language-internal factors, which amounts to an uncontrolled text generation task.

In sum, ChatGPT-generated text largely fails to replicate human-like patterns of linguistic variation when narrowly prompted to generate speech by a member of a particular community. Of the seven sociolinguistic variables tested here, four appear categorically as a single variant, while two more appear nearly categorically as a single variant. Each of these categorical or near-categorical productions is of the prescriptive norm. This failure to generate variation occurs both in cases of robust cross-dialectical variation and local, dialect-specific variation. Despite this, ChatGPT was able to generate variation relatively successfully with complementizer *that* deletion. The variants appeared in the generated text at rates in line with human speech, and the variation was constrained by language-internal factors. While the constraints were not as robust as in human speech, this variable represents a clear success in comparison to the other six variables. It is worth considering why this may be.

One key difference between this variable and most of the others is that variables such as the surface form of the participle or preterite have variants which are prescriptively incorrect and potentially stigmatized. In contrast, the null complementizer does not appear to be negatively evaluated upon even though the overt complementizer is evaluated as formal (Tagliamonte and Smith 2005, 290). At the same time, the variable shows relatively little language-external conditioning in comparison to a variable such as *-body/-one* variation, which is undergoing change in many communities (D'Arcy et al. 2013). It is quite possible that, unlike other variables, both variants of the complementizer occur frequently enough in training data that an LLM may generate variation in its role as a "stochastic parrot" (Bender et al. 2021).

At the same time, because there are fewer prescriptive norms surrounding this variable, it is likely that human input would provide evidence of variability both through the creation of conversations for fine-tuning with variable complementizer deletion and through the rating of examples with either variant as correct during RLHF training. In contrast, prescriptively incorrect variants have three opportunities to be biased out of the model: in the training data itself, in the creation of human-generated conversations, and in rating data of output appropriateness. That the training data would bias out some variability is a known issue in some respects. The GPT-3 LLM, for example, was trained on the CommonCrawl internet corpus, English Wikipedia, and two online book corpora (Brown et al. 2020). As OpenAI (2020) note, this means that "GPT-3 will by default perform worse on inputs that are different from the data distribution it is trained on, including…specific dialects of English that are not as well-represented in training data". This may explain why Tyneside-specific variables are not generated well by ChatGPT; however, it does not explain why variables that widely occur cross-dialectally also are not generated well, as these would be expected to appear in at least the training data from the CommonCrawl corpus. Their absence is suggestive of some sort of bias in the training. This could take a few forms: training data may be processed to standardize the input, human-generated fine-tuning data may come from speakers who strongly favor canonical variants, and/or RLHF responses may assign prescriptively incorrect variants low ratings. In this sense, ChatGPT's (in)ability to generate variation reflects the model being highly sensitive to frequency. Variables for which a variant has been biased out of the data provide little-to-no evidence of variability for the model to generate, while highly frequent variables such as complementizer deletion provide enough information for the model to generalize both the basic pattern of variation and language-internal constraints that structure the variation.

Although ChatGPT is able to replicate language-internal constraints for complementizer deletion, it still fails to replicate language-external constraints. Importantly, this failure is in generating human community-like patterns, rather than in generating statistically significant patterns. Indeed, outputs with a female and younger prompt generate the overt complementizer more than male and older prompts, and class is kept within the regression model even though it is not a significant predictor. I suggest this is because the overt complementizer is viewed by human speakers as formal (Tagliamonte and Smith 2005, 290), despite there being fewer norms surrounding its usage in comparison to other variables. The few noncanonical tokens in other variables support this hypothesis, as the prompt generating them always included at least two categories of the male/older/working-class combination. For example, each of the singular *us* tokens came from a prompt of an older working-class interviewee. In general, it appears that to the extent that the LLM underlying ChatGPT has associations between linguistic form and social categories, it associates prescriptive correctness with younger/female/middle-class and prescriptive incorrectness with older/male/working-class. While this association is sometimes accurate for a particular variable, it is not a correct generalization overall.

This problem regarding language-external constraints is a difficult problem to solve because it involves overgeneralization both in text generation (most variables near-categorically favor one variant, and the few instances of the other variant are categorically from a subset of social categories) and in application to variables (not every sociolinguistic variable is externally conditioned in the same way). A larger volume of language data alone is unlikely to help; while it may lead ChatGPT to generate more human-like variation with respect to usage rates and language-internal constraints, it will not address the overgeneralizations with language-external factors. Effectively, the issue is that both broad phases of how ChatGPT was trained—the language data itself and comparative rating of potential outputs—appear to be biased against non-canonical linguistic variants. Given this, human-like outputs are not truly possible without substantial changes to the input. With respect to RLHF training, the solution would be to explicitly train ChatGPT on sociolinguistic evaluations. However, it should be noted that if there is a correlation in the model already between social categories and prescriptive correctness, *human raters may have already trained it on sociolinguistic evaluations that yield these correlations*. Thus, this solution may be less to add a new input to the model and more to fix a potential flaw of it. With respect to the training language data itself, the solution would be to restructure training corpora away from formal, written sources such as Wikipedia and books and towards sources more representative of natural speech (cf. Warstadt and Bowman 2022). This may additionally require tagging the training data for speakers/writers' demographic information, which would raise serious ethical issues (Hovy and Yang 2021). Both including sociolinguistic evaluations in RLHF training and restructuring training corpora would require retraining the model from the ground up.

7. Conclusion

A great deal of recent work (Gulordava et al. 2018; Linzen and Baroni 2021; Asher et al. 2023; Zhuang et al. 2023; *inter alia*) has explored whether LLMs have linguistic competence. Such competence is typically framed within a grammatical perspective: Can LLMs make generalizations about and generate text adhering to syntactic and semantic properties of human language? This article suggests that investigating the linguistic competence of LLMs should extend to sociolinguistic competence, or whether they can generate and evaluate sociolinguistic variation in a human-like manner. I outline criteria that an LLM must satisfy in order to be considered sociolinguistically competent, and illustrate how to test these criteria through a case study of controlled text generation using ChatGPT. Overall, we find that, when prompted with information effectively limited to speaker demographics, ChatGPT fails the basic test set here of properly generating human-like sociolinguistic variation.

Note, however, that this is a single case study which tests a single LLM's ability to generate human-like text representing a single dialect of English. This means that the data and interpretation of it naturally have some key limitations. These limitations, however, point toward directions for further research. If, as I suggest, ChatGPT's difficulty with the task of properly generating sociolinguistic variation lies in how the model was trained, the results here do not necessarily indicate that all LLMs, including those with a quite similar architecture to ChatGPT, will have the same difficulties. Indeed, testing a wide range of LLMs has the potential to shed light on not only whether the general architecture(s) of such models can generate variation, but what aspects of model pre-training and fine-tuning, if any, promote or impede the generation of variation and which settings (for example, higher vs. lower model temperature) influence the generation of variation. In this sense, testing whether a given LLM can properly generate sociolinguistic variation can serve both as a test of that model's capabilities and as a window into how and to what extent such models learn human-like language. Likewise, if the dialect studied here is particularly underrepresented in the training data, the failure to properly generate specific regional variants may not be replicated with other dialects better represented in the training data. Exploring ChatGPT and other LLMs' capabilities to generate a wider range of dialectal variation will be necessary to fully assess their competence in generating regional variation. Regardless, ChatGPT's failure to properly generate sociolinguistic variation that is widely found across English dialects is notable.

At the same time, the information provided in the model prompt is worthy of further exploration, particularly because prompt engineering is one current approach to achieving controllable text generation. This case study intentionally provided relatively minimal information in the prompt used for text generation. While this is another limitation to the study, I maintain that such minimal information represents an important baseline condition for assessing the sociolinguistic competence of ChatGPT and other LLMs, particularly when it comes to generating variation. As the results show, if it is to yield outputs indicative of the model having sociolinguistic competence, ChatGPT may need more information than I provided. Indeed, it is quite possible that prompts which include explicit requests for local dialect usage, particular kinds of accents, etc., will find more success in generating outputs suggestive of sociolinguistic competence. As noted above, such prompts are not viable for testing LLMs' sociolinguistic competence because they conflate the task of generating variation with the tasks of performatively using third-order indexicals and directly linking speakers' social categories and linguistic usage. However, given the baseline established here that demographic information alone is not sufficient for ChatGPT to properly generate human-like variation, such prompts suggest two lines of inquiry for further research: how much background information is necessary for an LLM to display signs of sociolinguistic competence, and given this, how does the sociolinguistic competence of that LLM compare to that of humans? Given the current performance of ChatGPT, the basic test illustrated in this case study will be useful both as a first step in evaluating additional/future models' sociolinguistic competence and for setting a baseline of necessary information when prompting any given model to produce sociolinguistic variation.

Availability of Deposited Data

The ChatGPT-generated transcripts, Python code used in processing transcripts, raw data containing sociolinguistic variable tokens, and R code used to analyze "family" size and complementizer *that* deletion may be found in an OSF repository with a CC-By Attribution 4.0 International license (Duncan 2024).

Acknowledgements

Thank you to Mary Robinson, Chrissy Cuskley, and the *JCLR* editors and reviewers for helpful discussion and feedback.

References

Anderwald, Lieselotte. 2009. *The Morphology of English Dialects: Verb Formation in Non-standard English*. Cambridge: Cambridge University Press.

Asher, Nicholas, Swarnadeep Bhar, Akshay Chaturvedi, Julie Hunter, and Soumya Paul. 2023. "Limits for Learning with Language Models." In *Proceedings of the The 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, edited by Alexis Palmer and Jose Camacho-Collados, 236–48. Stroudsburg, PA: Association for Computational Linguistics. doi:10.18653/v1/2023.starsem-1.22.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-effects Models Using *lme4*." *Journal of Statistical Software* 67(1): 1–48. doi:10.18637/jss.v067.i01.

Bayley, Robert, and Vera Regan. 2004. "Introduction: The Acquisition of Sociolinguistic Competence." *Journal of Sociolinguistics* 8(3): 323–38. doi:10.1111/j.1467-9841.2004.00263.x.

Beal, Joan. 2004. "English Dialects in the North of England: Morphology and Syntax." In *A Handbook of Varieties of English*, edited by Bernd Kortmann and Edgar Schneider, 114–41. Berlin: Mouton de Gruyter.

Beal, Joan, Lourdes Burbano-Elizondo, and Carmen Llamas. 2012. *Urban North-eastern English: Tyneside to Teesside*. Edinburgh: Edinburgh University Press.

Becker, Kara. 2013. "The Sociolinguistic Interview." In *Data Collection in Sociolinguistics*, edited by Christine Mallinson, Becky Childs, and Gerard van Herk, 91–100. New York: Routledge.

Belz, Anja, and Ehud Reiter. 2006. "Comparing Automatic and Human Evaluation of NLG Systems." In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 313–20. Stroudsburg, PA: Association for Computational Linguistics. doi: 10.1.1.60.8276.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, 610–23. New York: Association for Computing Machinery. doi:10.1145/3442188.3445922.

Bleaman, Isaac L. 2020. "Implicit Standardization in a Minority Language Community: Real-time Syntactic Change Among Hasidic Yiddish Writers." *Frontiers in Artificial Intelligence* 3: Article 35. doi:10.3389/frai.2020.00035.

Bleaman, Isaac L., Daniel Duncan, Shelley Feuer, Gregory Guy, Zachary Jaggers, and Matthew Stuck. 2014. "'She Said {That/ø} She Couldn't Take a Complement': Complementizer *That* Omission in American English." Paper presented at New Ways of Analyzing Variation 43, Chicago, Illinois, October 23–26.

Bloomer, Robert K. 1998. "*You Shoulda Saw Me*: On the Syntactic Contexts of Nonstandard Past Participles in Spoken American English." *American Speech* 73(2): 221–24. doi:10.2307/455743.

Bowman, Samuel. 2023. "Eight Things to Know about Large Language Models." Ms. Accessed July 7, 2023. https://arxiv.org/abs/2304.00612.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. "Language Models are Few-Shot Learners." In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, Virtual*, edited by Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin.

Chambers, Jack K. 1995. *Sociolinguistic Theory: Linguistic Variation and Its Social Significance*. Oxford: Blackwell.

Chatten, Alicia, Kimberly Baxter, Erwanne Mas, Jailyn Pena, Guy Tabachnick, Daniel Duncan, and Laurel MacKenzie. 2022. "'I've Always Spoke Like This, You See': Preterite-for-participle Leveling in American and British Englishes." *American Speech* 99(1): 3–46. doi:10.1215/00031283-9940654.

Chen, Yihan, Benfeng Xu, Quan Wang, Yi Liu, and Zhendong Mao. 2024. “Benchmarking Large Language Models on Controllable Generation under Diversified Instructions.” *Proceedings of the AAAI Conference on Artificial Intelligence 38*(16): 17808–16. doi:10.1609/aaai.v38i16.29734.

Childs, Claire, Carmen Llamas, and Dominic Watt. 2021. “Pronoun Exchange in North-Eastern English: Perception Versus Use.” Paper presented at United Kingdom Language Variation and Change 13, Glasgow, United Kingdom, September 8–10.

Chowdhury, Shammur Absar, and Roberto Zamparelli. 2018. “RNN Simulations of Grammaticality Judgments on Long-distance Dependencies.” In *Proceedings of the 27th International Conference on Computational Linguistics*, edited by Emily M. Bender, Leon Derczynski, and Pierre Isabelle, 133–44. Stroudsburg, PA: Association for Computational Linguistics.

Chung, John Joon Young, Ece Kamar, and Saleema Amershi. 2023. “Increasing Diversity While Maintaining Accuracy: Text Data Generation with Large Language Models and Human Interventions.” In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers*, edited by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, 575–93. Stroudsburg, PA: Association for Computational Linguistics.

D’Arcy, Alexandra, Bill Haddican, Hazel Richards, Sali Tagliamonte, and Ann Taylor. 2013. “Asymmetrical Trajectories: The Past and Present of *-body/-one*.” *Language Variation and Change* 25(3): 287–310. doi:10.1017/S0954394513000148.

Dathathri, Sumanth, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. “Plug and Play Language Models: A Simple Approach to Controlled Text Generation.” In *Proceedings of International Conference on Learning Representations 2020*. https://openreview.net/pdf?id=H1edEyBKDS.

Department for Levelling Up, Housing and Communities, and Ministry of Housing, Communities and Local Government. 2019. “Indices of Deprivation: 2019 and 2015.” Accessed July 5, 2023. http://dclgapps.communities.gov.uk/imd/iod_index.html.

Duncan, Daniel. 2024. “Do LLMs Have Sociolinguistic Competence?” OSF [Data set]. doi:10.17605/OSF.IO/KHF6M.

Duncan, Daniel. 2019. “Grammars Compete Late: Evidence from Embedded Passives.” *University of Pennsylvania Working Papers in Linguistics* 25(1): 89–98.

Eisenstein, Jacob. 2015. “Systematic Patterning in Phonologically-motivated Orthographic Variation.” *Journal of Sociolinguistics* 19(2): 161–88. doi: 10.1111/josl.12119.

Eisikovits, Edina. 1987. “Variation in the Lexical Verb in Inner-Sydney English.” *Australian Journal of Linguistics* 7(1): 1–24. doi:10.1080/07268608708599371.

Gao, Tianyu, Adam Fisch, and Danqi Chen. 2021. “Making Pre-trained Language Models Better Few-shot Learners.” In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, edited by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, 3816–30. Stroudsburg, PA: Association for Computational Linguistics. doi:10.18653/v1/2021.acl-long.295.

Geeraert, Kristina, and John Newman. 2011. “I Haven’t Drank in Weeks: The Use of Past Tense Forms as Past Participles in English Corpora.” In *Corpus-based Studies in Language Use, Language Learning, and Language Documentation*, edited by John Newman, Harald Baayen, and Sally Rice, 11–33. Leiden: Brill.

Grieve, Jack. 2016. *Regional Variation in Written American English*. Cambridge: Cambridge University Press.

Gulordava, Kristina, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. “Colorless Green Recurrent Networks Dream Hierarchically.” In *Proceedings of NAACL-HLT 2018*, Vol. 1, edited by Marilyn Walker, Heng Ji, and Amanda Stent, 1195–1205. Stroudsburg, PA: Association for Computational Linguistics. doi:10.18653/v1/N18-1108.

Han, Xu, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. 2021. “Pre-trained Models: Past, Present and Future.” *AI Open* 2: 225–50. doi:10.1016/j.aiopen.2021.08.002.

Hothorn, Torsten, Frank Bretz, and Peter Westfall. 2008. “Simultaneous Inference in General Parametric Models.” *Biometrical Journal* 50(3): 346–63. doi:10.1002/bimj.200810425.

Hovy, Dirk, and Diyi Yang. 2021. “The Importance of Modelling Social Factors of Language: Theory and Practice.” In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, edited by Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, 588–602. Stroudsburg, PA: Association for Computational Linguistics. doi:10.18653/v1/2021.naacl-main.49.

Hunt, Stephen A., ed. 2009. *Family Trends: British Families Since the 1950s.* London: Family & Parenting Institute.

Hymes, Dell. 1966. “Two Types of Linguistic Relativity.” In *Sociolinguistics: Proceedings of the UCLC Sociolinguistics Conference, 1964*, edited by William Bright, 114–58. The Hague: Mouton.

Ilbury, Christian. 2019. “’Sassy Queens’: Stylistic Orthographic Variation in Twitter and the Enregisterment of AAVE.” *Journal of Sociolinguistics* 24(2): 245–64. doi:10.1111/josl.12366.

Jankowski, Bridget L., and Sali A. Tagliamonte. 2022. "'He Come Out and Give Me a Beer but He Never Seen the Bear': Vernacular Preterites in Ontario Dialects." *English World-Wide* 43(3): 267–96. doi:10.1075/eww.20014.jan.

Jentzsch, Sophie, and Kristian Kersting. 2023. "ChatGPT is Fun, But it is not Funny! Humor is Still Challenging Large Language Models." In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, edited by Jeremy Barnes, Orphée De Clercq, and Roman Klinger, 325–40. Stroudsburg, PA: Association for Computational Linguistics.

Jiang, Zhengbao, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. "How Can We Know What Language Models Know?" *Transactions of the Association for Computational Linguistics* 8: 423–38. doi:10.1162/tacl_a_00324.

Johnstone, Barbara, Jennifer Andrus, and Andrew E. Danielson. 2006. "Mobility, Indexicality, and the Enregisterment of 'Pittsburghese.'" *Journal of English Linguistics* 34(2): 77–104. doi:10.1177/0075424206290692.

Kearns, Kate. 2007a. "Epistemic Verbs and Zero Complementizer." *English Language and Linguistics* 11(3): 475–505. doi:10.1017/S1360674307002353.

Kearns, Kate. 2007b. "Regional Variation in the Syntactic Distribution of Null Finite Complementizer." *Language Variation and Change* 19(3): 295–336. doi:10.1017/S0954394507000117.

Kemp, Renee, Emily Moline, Chelsea Escalante, Alexander Mendes, and Robert Bayley. 2016. "Where Have All the Participles Went? Using Twitter Data to Teach about Language." *American Speech* 91(2): 226–35. doi:10.1215/00031283-3633129.

Kroch, Anthony, and Cathy Small. 1978. "Grammatical Ideology and its Effect on Speech." In *Linguistic Variation: Models and Methods*, edited by David Sankoff, 44–55. New York: Academic Press.

Kulkarni, Vivek, and Vipul Raheja. 2023. "Writing Assistants Should Model Social Factors of Language." Paper presented at The Second In2Writing Workshop @ CHI '23, Hamburg, Germany. https://arxiv.org/abs/2303.16275v1 (accessed 7 July 2023).

Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Lam, Suet-Ying, Qingcheng Zeng, Kexun Zhang, Chenyu You, and Rob Voight. 2023. "Large Language Models are Partially Primed in Pronoun Interpretation. In *Findings of the Association for Computational Linguistics: ACL 2023*, edited by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, 9493–9506. Stroudsburg, PA: Association for Computational Linguistics.

Laslett, Peter. 1969. “Size and Structure of the Household in England over Three Centuries.” *Population Studies* 3(2): 199–223. doi:10.2307/2172902.

Levey, Stephen, Susan Fox, and Laura Kastronic. 2017. “A Big City Perspective on *Come/came* Variation: Evidence from London, U.K.” *English World-Wide* 38(2): 181–210. doi:10.1075/eww.38.2.03lev.

Li, Jiwei, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. “A Persona-based Neural Conversation Model.” In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, edited by Katrin Erk and Noah A. Smith, 994–1003. Stroudsburg, PA: Association for Computational Linguistics. doi:10.18653/v1/P16-1094.

Li, Junyi, Tianyi Wang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. “Pre-trained Language Models for Text Generation: A Survey.” *ACM Computing Surveys* 56(9): Article 230. doi:10.1145/3649449.

Linzen, Tal, and Marco Baroni. 2021. “Syntactic Structure from Deep Learning.” *Annual Review of Linguistics* 7: 195–212. doi:10.1146/annurev-linguistics-032020-051035.

Mendelsohn, Julia, Maarten Sap, and Ronan Le Bras. 2022. “Sounding the Bullhorn: Surfacing and Analyzing Dogwhistles with Language Models.” Paper presented at New Ways of Analyzing Variation 50, San Jose, California, October 13–15.

Nguyen, Dong, Laura Rosseel, and Jack Grieve. 2021. “On Learning and Representing Social Meaning in NLP: A Sociolinguistic Perspective.” In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, edited by Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, 603–12. Stroudsburg, PA: Association for Computational Linguistics. doi:10.18653/v1/2021.naacl-main.50.

Noble, Stefan, David McLennan, Michael Noble, Emma Plunkett, Nils Gutacker, Mary Silk, and Gemma Wright. 2019. *The English Indices of Deprivation 2019: Research Report*. London: Ministry of Housing, Communities & Local Government.

OpenAI. 2020. “GPT-3 Model Card.” Accessed July 7, 2023. https://github.com/openai/gpt-3/blob/master/model-card.md.

OpenAI. 2022. “Introducing ChatGPT.” Accessed July 7, 2023. https://openai.com/blog/chatgpt.

Pearce, Michael. 2021. “The Participatory Vernacular Web and Regional Dialect Grammar.” *English Today* 147, 37(4): 196–205. doi:10.1017/S0266078420000243.

Prabhumoye, Shrimai, Alan W. Black, and Ruslan Salakhutdinov. 2020. "Exploring Controllable Text Generation Techniques." In *Proceedings of the 28th International Conference on Computational Linguistics*, edited by Donia Scott, Nuria Bel, and Chengqing Zong. International Committee on Computational Linguistics. doi:10.18653/v1/2020.coling-main.1.

Qiu, XiPeng, TianXiang Sun, YiGe Xu, YunFan Shao, Ning Dai, and XuanJing Huang. 2020. Pre-trained Models for Natural Language Processing: A Survey. *Science China Technological Sciences* 63(10): 1872–97. doi:10.1007/s11431-020-1647-3.

R Core Team. 2020. "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing, Vienna, Austria. Accessed July 7, 2023. http://www.R-project.org/.

Rauh, Maribeth, John Mellor, Jonathan Uesato, Po-Sen Huang, Johannes Welbl, Laura Weidinger, Sumanth Dalthathri, Amelia Glaese, Geoffrey Irving, Iason Gabriel, William Isaac, and Lisa Anne Hendricks. 2022. "Characteristics of Harmful Text: Towards Rigorous Benchmarking of Language Models." In *36th Conference on Neural Information Processing Systems (NeurIPS 2022): Track on Datasets and Benchmarks*, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. San Diego: Neural Information Processing Systems.

Serbicki, Sofia, Ruijin Lan, and Daniel Duncan. 2023. "Participle-for-preterite Variation in Tyneside English." *English World-Wide* 45(1): 30–60. doi:10.1075/eww.00081.ser.

Squires, Lauren (Ed.). 2016. *English in Computer-mediated Communication: Variation, Representation, and Change* (Topics in English Linguistics volume 93). Berlin: De Gruyter Mouton.

Tagliamonte, Sali. 2001. "*Come/came* Variation in English Dialects." *American Speech* 76(1): 42–61. doi:10.1215/00031283-76-1-42.

Tagliamonte, Sali, and Jennifer Smith. 2005. "No Momentary Fancy! The *Zero* 'Complementizer' in English Dialects." *English Language and Linguistics* 9(2): 289–309. doi:10.1017/S1360674305001644.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Warstadt, Alex, and Samuel R. Bowman. 2022. "What Artificial Neural Networks Can Tell Us about Human Language Acquisition." In *Algebraic Structures in Natural Language*, edited by Shalom Lappin and Jean-Philippe Bernardy, 17–60. Boca Raton, FL: CRC Press.

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. "Emergent Abilities of Large Language Models." *Transactions on Machine Learning Research*. https://openreview.net/pdf?id=yzkSU5zdwD.

Weinreich, Uriel, William Labov, and Marvin I. Herzog. 1968. "Empirical Foundations for a Theory of Language Change." In *Directions for Historical Linguistics: A Symposium*, edited by W. P. Lehmann and Yakov Malkiel, 95–195. Austin, TX: University of Texas Press.

Yang, Yi, and Jacob Eisenstein. 2017. "Overcoming Language Variation in Sentiment Analysis with Social Attention. *Transactions of the Association for Computational Linguistics* 5: 295–307. doi:10.1162/tacl_a_00062.

Yu, Dian, Zhou Yu, and Kenji Sagae. 2021. "Attribute Alignment: Controlling Text Generation from Pre-trained Language Models." In *Findings of the Association for Computational Linguistics: EMNLP 2021*, edited by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, 2251–68. Stroudsburg, PA: Association for Computational Linguistics. doi:10.18653/v1/2021.findings-emnlp.194.

Zhang, Saizheng, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. "Personalizing Dialogue Agents: I Have a Dog, Do You Have Pets Too?" In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by Iryna Gurevych and Yusuke Miyao, 2204–13. Stroudsburg, PA: Association for Computational Linguistics. doi:10.18653/v1/P18-1205.

Zhuang, Ziyu, Qiguang Chen, Longxuan Ma, Mingda Li, Yi Han, Yushan Qian, Haopeng Bai, Weinan Zhang, and Ting Liu. 2023. "Through the Lens of Core Competency: Survey on Evaluation of Large Language Models." In *Proceedings of the 22nd China National Conference on Computational Linguistics*, edited by Jiajun Zhang, 88–109. Chinese Information Processing Society of China.

Ziems, Caleb, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. "Can Large Language Models Transform Computational Social Science?" *Computational Linguistics* 50. doi:10.1162/coli_a_00502.