Spelling convention sensitivity in neural language models

Elizabeth Nielsen[†] **Christo Kirov**[°] **Brian Roark**[°]
[†]School of Informatics, University of Edinburgh, UK [°]Google
e.nielsen@ed.ac.uk {ckirov,roark}@google.com

Abstract

We examine whether large neural language models, trained on very large collections of varied English text, learn the potentially long-distance dependency of British versus American spelling conventions, i.e., whether spelling is consistently one or the other within model-generated strings. In contrast to long-distance dependencies in non-surface underlying structure (e.g., syntax), spelling consistency is easier to measure both in LMs and the text corpora used to train them, which can provide additional insight into certain observed model behaviors. Using a set of probe words unique to either British or American English, we first establish that training corpora exhibit substantial (though not total) consistency. A large T5 language model does appear to internalize this consistency, though only with respect to observed lexical items (not nonce words with British/American spelling patterns). We further experiment with correcting for biases in the training data by fine-tuning T5 on synthetic data that has been debiased, and find that finetuned T5 remains only somewhat sensitive to spelling consistency. Further experiments show GPT2 to be similarly limited.

1 Introduction

The probabilities that neural language models (LMs) assign to strings can be used to assess how effectively they capture linguistic dependencies found in their training data. Much as in psycholinguistic experiments on human language speakers, we can present LMs with strings both with and without agreement in key dependencies and measure the assigned probabilities to determine whether the model has learned these linguistic generalizations or not (see e.g., Futrell et al. 2018). For example, sentences both with and without subject/verb number agreement (but otherwise identical) can be used to assess whether the model accounts for that particular dependency, even over long distances. Various long-distance dependencies have been investigated in this manner, from purely linguistic phenomena such as syntactic dependencies (e.g., Gulordava et al. 2018) to extra-linguistic phenomena such as socio-cultural biases (e.g., Rudinger et al. 2018).

In this paper, we examine dependencies based on orthographic cues to language variety. Many LMs are trained on large corpora scraped from the web, and data from different language varieties are often combined. For example, LMs trained on web-scraped English (e.g., the WebText Corpus of Radford et al. 2019) encounter British English, North American English, and multiple World Englishes. Likewise, Spanish web corpora may include several distinct varieties of Latin American Spanish, as well as Iberian Spanish (e.g., Kilgarriff and Renau 2013). Here we use differences between British and American English spelling conventions to ask whether LMs trained on large and diverse collections of English learn to apply these conventions consistently within the same span of text. For example, if the British spelling of the word *labour* appears in a sentence prefix, will the LM assign higher probabilities to continuations that maintain British spelling conventions (e.g., *organisation*) over those that have American-spelled forms (*organization*)? To the extent that such models are used within response generation systems or for next word prediction in virtual keyboards, maintaining such consistency would be strongly desirable so users receive results appropriate for their locale.

Of course, as with any such dependencies, models can only learn generalizations that are present in the data, so we also look at the degree to which corpora used to train the large LMs (LLMs) that we investigate (as well as a few others) demonstrate spelling convention consistency. Assessing whether syntactic or semantic generalizations are learned by models trained on noisy, errorful and inconsistent data is complicated by the difficulty in

Findings of the Association for Computational Linguistics: EACL 2023, pages 1334–1346

quantifying the actual degree of consistency of the dependency in the data itself. In contrast to structural linguistic generalizations or other implicit information, the explicitness of spelling conventions permits straightforward corpus analysis in addition to model probing, providing another avenue for explaining model performance.

The results of our data analysis are presented in §4. We find that relevant web-scraped English text used to train LLMs unsurprisingly does not provide perfect consistency — and further that it is heavily skewed towards American spelling conventions — but that it provides as much or more consistency than some curated corpora such as the British National Corpus (BNC Consortium, 2007). We then present methods, in §5, to measure the degree to which two neural LLMs – T5 (Raffel et al., 2020) (both with and without additional finetuning) and GPT2 – exhibit spelling variation consistency. We find that T5 without finetuning demonstrates a general preference for consistency, but that this preference is weaker for British than American English and does not extend robustly to nonce words. Finetuning T5 on a synthetically modified portion of the British National Corpus reduces the preference for American English. We then modify our conditional probability calculations to allow demonstration of similar patterns of model behavior for GPT2, a very differently architectured and trained LLM (Radford et al., 2019). Lastly, in §6, we take a slightly deeper dive into the kinds of (and reasons for) spelling convention inconsistencies in some corpora analyzed in §4.

Overall, we demonstrate that, while T5 and GPT2 display some sensitivity to spelling convention differences, this cannot be relied on to produce consistent generated output. If reliable spelling consistency is an application requirement, additional post-processing may need to be applied to LLM output.

This paper makes several key contributions. First, we provide methods for straightforwardly assessing the ability of LLMs to capture certain well-attested long-distance dependencies in English, and demonstrate the strengths and shortcomings of two well-known models in doing so. This opens up the possibility of exploratory studies in languages where such conventions are less well documented. In contrast to the most heavily investigated types of long-distance dependencies (e.g., syntactic), the (previously unexplored) dependency of spelling convention consistency is directly observable in the surface string and hence is relatively easy to assess in both models and data. As a result, it can be seen as a useful task for assessing LM learning in general. We also document the degree to which web-scraped corpora exhibit spelling consistency, making clear that the models have plenty of room for improvement. However, American English is shown to be far more heavily represented in the training corpora than British English, to the point that performance for British English is demonstrably far worse than for American English, something that language generation or word prediction systems must address for equitable performance.

2 Background

2.1 Dependencies and LMs

Much of the work investigating whether large language models capture long-distance linguistic generalizations has focused on non-surface dependencies, such as co-reference. In order to correctly identify that two expressions refer to the same entity, models often need to identify complex syntactic relationships (e.g., c-command), or build a model of entities over an entire discourse (e.g., Clark and Manning 2016). Despite this complexity, LLMs have shown some promise as general-purpose co-reference resolvers (Joshi et al., 2019). This suggests that they can learn to model complex long-distance dependencies.

Other research has shown more directly that LLMs model syntactic dependencies. A common methodology is to compare an LM's surprisal directly to psycholinguistic data (Futrell et al., 2018). If the LM still performs like a human on examples that require modeling hierarchical relationships between tokens, this suggests that the LM has learned some part of the more complex syntactic structure of the language. Work such as Futrell et al. (2018) has shown that a recurrent neural network language model achieves surprisal rates that mimic human processing, including in these syntactically complex situations. This suggests that an RNN LM can be sensitive to complex syntactic relationships as well. Similar methods have been used to show LMs learning syntactic dependencies in Linzen et al. (2016), Frank et al. (2016), and Brennan et al. (2020).

Another class of methods for assessing whether LMs learn complex syntactic dependencies involves probing the models themselves to evaluate

whether syntax-like relationships between tokens can be discovered. Details of their methods vary widely, but Clark et al. (2019), Hewitt and Manning (2019), and Lin et al. (2019) all suggest that many LMs learn complex syntactic dependencies.

In contrast, the topic of the current paper – spelling convention dependencies – is a relatively surface-level dependency. A model does not need to capture the syntactic or semantic relationship between two words in order to evaluate spelling consistency, rather simply their co-occurrence. Given prior results showing that LMs can and do learn complex semantic and syntactic relationships between words, one might expect that a relatively simple dependency like spelling convention should be easy for an LM to learn.

2.2 Spelling variation

As discussed by Berg and Aronoff (2017), the orthography of English has never been regulated by an official body, but has rather emerged dynamically over time. Dictionaries played a key role in settling spelling conventions, with Samuel Johnson's (1755) dictionary being the key source of contemporary British spelling conventions and Webster's (1828) dictionary the key source of contemporary American spelling. The latter included spelling reforms such as using the suffix *-or* instead of *-our* for certain words, e.g., *labor* instead of *labour*. These reforms were adopted in American spelling but not in British spelling conventions.

This history makes English an interesting case study for spelling variation in particular. Languages that have historically had centralized regulatory institutions, such as the French or Royal Spanish Academies, have much less purely orthographic variation. For example, despite many lexical differences, there are few spelling differences between Iberian and Latin American Spanish. On the other hand, there are many language situations that have considerably more spelling variation. For example, speakers of South Asian languages that are traditionally written with Brahmic or Arabic scripts often write using the Latin alphabet in contexts like SMS messages and social media (Roark et al., 2020). This kind of informally romanized text presents many spelling variations due to these languages' lack of orthography in the Latin script. The well-documented nature of English spelling variation and its close ties to standardized regional varieties make it a good initial case study for whether LLMs learn systematic variation in the data. If so, such models may be useful in more exploratory studies, such as the above-mentioned scenario where no official orthography exists.

As far as we are aware, the issue of spelling convention consistency in language models has not been investigated. Nguyen and Grieve (2020) looked at whether word embeddings are *robust* to spelling variation, not whether generative language models capture spelling consistency. That paper focused mainly on the kinds of variation that arise in informal social media text, but they also examined British versus American spelling. Unsurprisingly, they found that cosine similarity between British and American spelled variants are high relative to other patterns of informal spelling variability.

2.3 Prompting LMs

In the present work, we construct prompts to measure the probability assigned to various tokens by LLMs. In constructing these prompts, we take into account the findings of recent work on prompting LMs. Our work is different from the sort of prompting described by these papers, which generally includes features such as task-specific prefixes containing instructions (e.g., Raffel et al. 2020), verbalized class labels (e.g., Schick and Schütze 2021), or in-context learning (e.g., Brown et al. 2020), none of which are present in our approach. However, work such as Webson and Pavlick (2022) has shown large effects due to small variations in the wording of prompts, even if the reasons for these effects are not apparent. Therefore, we choose to present the model with several different prompts and average the probabilities over all prompts, in order to account for possible variation.

3 Data and models

To assess the spelling convention consistency of data and models, we use a list of British and American English spelling differences that is part of the open-source American British English Translator.[1] We used the 1706 word pairs in the `data/american_spellings.json` file at that site. This list includes American and British spelling variants for words with common differences such as *-or*/*-our* (e.g., vapor/vapour), *-ize*/*-ise* (realize/realise), consonant doubling (modeling/modelling), *-er*/*-re* (liter/litre), along with

[1] https://github.com/hyperreality/American-British-English-Translator

some number of term-specific spelling differences (aluminum/aluminium). We use this list to create prompts for probing the language models and to establish the consistency of usage within corpora, i.e., whether strings found in this list consistently follow one convention or the other when they co-occur.

For model probing, we examine T5 (Raffel et al., 2020), a general purpose encoder-decoder model. We use the t5-large architecture variant on the T5X codebase,[2] which has approximately 770M parameters. For English, T5 is (pre-)trained using a span corruption objective on the Colossal Clean Crawled Corpus (C4), an English language collection derived from Common Crawl (Raffel et al., 2020).[3]

We also examine GPT2, for which we use the open-source HuggingFace implementation (Radford et al., 2019). Unlike T5, GPT2 is a purely autoregressive language model rather than an encoder-decoder sequence-to-sequence model. It is trained to perform next-word prediction rather than fill in corrupted spans of text. GPT2 is built on OpenAI's WebText corpus (Radford et al., 2019), of which there is an open-source variant available.[4]

We examine C4 and OpenAI's WebText corpus for spelling convention consistency, along with several other corpora: English Wikipedia (downloaded 06-21-2020); the Billion Word Benchmark (Chelba et al., 2013), which is a collection of newswire text; and the British National Corpus (BNC Consortium, 2007),[5] which is a balanced corpus of both written and spoken material.[6]

4 Training corpora consistency

To examine spelling consistency in training data, we made use of the list of spelling variants and the five corpora mentioned in Section 3: C4, the OpenWebText Corpus (OWT), English Wikipedia (EngWiki), the Billion Word Benchmark (BWB), and the British National Corpus (BNC). We convert all strings in each corpus to lowercase, and treat all characters outside of the a–z range as whitespace for tokenization. We look for exact matches of list items in the resulting whitespace-delimited tokens.

Corpus	total # of word pairs	X-matched % US	UK	mis
C4	542,755,756	74.6	14.7	10.8
OWT	42,255,261	79.7	11.5	8.8
EngWiki	1,527,529	58.0	26.5	15.4
BWB	442,733	67.5	23.6	8.9
BNC	74,072	14.5	64.8	20.8

Table 1: Study of word pairs found in the same string from either UK or US spelling list over corpora of different sizes and characteristics, with percent of *US*-matched, *UK*-matched and *mis*matched US/UK pairs.

Let V_{US} be the US spelling variants[7] of the words in the list and V_{UK} the UK spelling variants. For each corpus C, let $s^k = w_1 \dots w_{|s^k|}$ represent the kth string in the corpus, consisting of $|s^k|$ words. We extract all pairs of words (w_i, w_j) from s^k such that $i < j$ and $w_i, w_j \in V_{\text{US}} \bigcup V_{\text{UK}}$. Each extracted pair (w_i, w_j) is placed into one of three classes: the pair is (1) *US-matched* if $w_i, w_j \in V_{\text{US}}$; (2) *UK-matched* if $w_i, w_j \in V_{\text{UK}}$; and (3) *mismatched* otherwise. We then aggregate the counts for pairs in these three bins across all strings in the corpus.

Table 1 presents the number of pairs extracted from each corpus and the percentage of those within each class. Several things jump out from these results. First, all of the corpora, other than the British National Corpus, have significantly more US-matched pairs than UK-matched pairs, with OWT and C4 being the most skewed towards US-matched pairs. This likely indicates a heavy overall skew towards US spelling variants, leading to a high prior probability of US spelling variants in LLMs. Second, the percentage of extracted pairs that are mismatched are non-negligible, however there is a lot of consistency. For example, in the C4 corpus, if a word from V_{UK} is the first word of a pair, the probability that the next word will also be from V_{UK} is nearly three times the probability that it is from V_{US}.[8] Finally, both English Wikipedia and the British National Corpus have somewhat elevated levels of mismatch compared to the other corpora, something we look at more closely in Section 6.

Having established that the level of mismatch in the C4 corpus used to train T5 is at the lower end

[2] https://github.com/google-research/t5x/blob/main/docs/models.md#t5-checkpoints

[3] http://commoncrawl.org/

[4] https://skylion007.github.io/OpenWebTextCorpus/

[5] http://www.natcorp.ox.ac.uk/

[6] Code for querying corpora and generating prompts, as well as other relevant data and code, can be found at https://github.com/google-research/google-research/tree/master/spelling_convention_nlm.

[7] For convenience, we use US as shorthand for American and UK as shorthand for British.

[8] Mismatched pairs in all corpora are roughly equally split between having V_{US} or V_{UK} words first. Hence, for C4, 5.4% of pairs are V_{UK} followed by V_{US} words (half of the mismatched probability), while 14.7% are V_{UK} followed by V_{UK}.

observed in the data we examined,[9] we now move on to examine whether the trained models pick up on these dependencies.

5 Language model consistency

From the dictionary presented in Section 3, we kept only the words that can be described by a small number of rules, e.g., the variation between *-ize* and *-ise*, etc, leaving us with 1266 options. For efficiency, we sample ≈16k prompt-target pairs (16028) from all possible 1266^2 combinations.

To eliminate all sources of variation besides the pair of words being tested, we created several template sentences into which we can insert pairs of words. The full set of templates is presented in Table 9 in Appendix A. Several considerations informed how we formulated the templates so that they work for all the tokens we wanted to test.

First and most obviously, we need to ensure that all tokens in a template are variety-neutral. This ensures that the probability of any of our test words being British or American will not be swayed by any regional bias in the template. While neutrality is difficult to enforce perfectly within a single frame, we hope that by using multiple different templates, we can mitigate unknown sources of bias via averaging.

Second, we need templates that will be syntactically and semantically acceptable, regardless of the inserted tokens. LLMs may assign low probability to tokens that result in grammatically unacceptable or semantically unlikely sentences, and we want to avoid introducing this source of variation. This is challenging, since the tokens we are testing include different parts of speech and come from very different semantic domains, hence there are few contexts where all tokens would be acceptable.

Fortunately, this problem has an analogue in linguistics: linguists interested in detailed phonetic description often elicit tokens in set contexts to eliminate extraneous sources of acoustic variation (Bowern, 2015). The approach these linguists often take is to use a template that *mentions* the tokens in question rather than *using* them. We follow this approach, and use templates similar to (1), which contain a list of word mentions.

(1) *My preferred words are ..., ..., and tree.*

[9] We note again the benefit of these explicit surface-level dependencies – we can easily assess the prevalence/consistency of the training data, in contrast to structural dependencies.

We then substitute pairs of words from our dictionary into the spaces marked with ellipses, both with consistent and inconsistent spelling conventions. In other words, given the pair of dictionary entries *realize*/*realise* and *center*/*centre*, we use the template above to generate the four test sentences:

(2)
a. US/US: *My preferred words are **realize**, **center**, and tree.*
b. US/UK: *My preferred words are **realize**, **centre**, and tree.*
c. UK/US: *My preferred words are **realise**, **center**, and tree.*
d. UK/UK: *My preferred words are **realise**, **centre**, and tree.*

We use T5 to score the probability of generating the second bolded word, as shown in Example (2), given the first.

In the above template, the two words are adjacent in the string. We also include a non-adjacent condition, which augments the templates by adding ten variety-neutral tokens between the bold-face words. For the above sample, the non-adjacent variant would be:

(3) *My preferred words are ..., flower, interesting, jump, ponderous, sky, skipping, desk, small, ladder, lovely, ..., and tree.*

Since T5 is a seq2seq model trained on a span-corruption objective, we present a prompt that includes a priming word and a blank span token representing the second word:

(4) My preferred words are **flavour**, <BLANK-SPAN-1>, and tree

The decoder then scores an output string that replaces the blank, but represents the known inputs with span markers instead:

(5) <INPUT-SPAN-1> **harbour** <INPUT=SPAN-2>

Thus we are effectively computing the probability that the blank span will be filled with a particular word (with a US or UK spelling), given the visible input sentence (which contains a US or UK primer) — P("harbour" | "My preferred words are flavour, ..., and tree.").

We report a few different measures to give a picture of how strongly each model prefers spelling consistency: mean conditional probabilities, prediction accuracy and mutual information. We then

Condition	Word 1	T5 Word 2		T5+FT Word 2		C4 Word 2	
		US	UK	US	UK	US	UK
Adjacent	US	0.86	0.14	0.66	0.34	0.91	0.09
	UK	0.39	0.61	0.44	0.56	0.38	0.62
Non-adjacent	US	0.83	0.17	0.69	0.31	0.93	0.07
	UK	0.48	0.52	0.43	0.57	0.27	0.73

Table 2: Conditional probability of Word 2, given a template with Word 1, given by T5 (no finetuning) and T5+FT (finetuned on synthetic balanced BNC data). For each instance, the probability has been normalized over each condition (corresponding to each row for the model). We also present the conditional probabilities from pairs found in the training corpus C4.

examine behavior with nonce words.

5.1 Measure 1: conditional probability tables

The first measure we use to show the preferences of each model is a 2x2 table of the conditional probability of the second probe word, given the first. For ease of interpretation, we normalize the conditional probabilities for each conditioning word as though the two alternative second words (US and UK) are the only possibilities, i.e., the two conditional probabilities are made to sum to 1. That is, $P(UK|US) + P(US|US) = 1$ and $P(US|UK) + P(UK|UK) = 1$ for each example. These conditional probabilities are then averaged over the whole test corpus (16028 word pairs replicated across 29 template sentences[10] for a total of 464812 samples) for both the adjacent and non-adjacent conditions. Table 2 presents these mean conditional probabilities for base T5 and T5 finetuned (TF+FT) on a synthetic balanced corpus derived from the BNC (see §5.2), alongside conditional probabilities calculated from the pairs extracted for the analysis in Table 1 from their training corpus (C4), under the same adjacent and non-adjacent conditions.[11]

As can be seen from these results, T5 shows a preference for spelling consistency in both the adjacent and non-adjacent conditions — probabilities for both the consistent US and consistent UK conditions are higher than the probabilities for the respective inconsistent conditions. The differences are notably larger in the adjacent conditions than the non-adjacent conditions, indicating that the preference for spelling consistency attenuates somewhat over longer strings. The model also shows a preference for US forms overall, assigning a higher probability to a US form after a UK form than to a UK form after a US form. This is likely due to US forms being over-represented in the training data, leading to high prior probability.

Comparing the model and corpus columns in Table 2, the degree of consistency preference displayed by T5 in the adjacent condition is actually very similar to the consistency levels in the C4 training corpus (similarly replicating the bias for US forms). However, C4 is much more consistent in the non-adjacent condition than T5, indicating that the model is failing to capture some long-distance dependencies.

5.2 Finetuning on synthetic data

Finding naturally occurring English text using perfectly consistent spelling conventions of sufficient size to help improve a model's consistency may be difficult, given the results presented in Table 1. It would be useful, however, to determine if T5 could be finetuned with some resource to exhibit better spelling consistency. To that end, we created a synthetic version of the BNC, which was changed to exhibit perfect consistency of British and American spelling conventions for the words in our lexicon.

This synthetic BNC corpus was produced as follows. Using our list of spelling variants, we identified strings in the corpus that contained an instance of either the American or British spelling. We then produced a synthetic consistent American spelling version of these strings by using the American spelling of all of the words, along with a synthetic consistent British spelling version of these strings by using the British spelling of all of the words. The resulting corpus is thus balanced between American and British spelling for these 1706 words, and every sentence is consistent in using one convention or the other. In total, the syn-

[10] For information on the variance across prompts, see Appendix A.

[11] The conditional probabilities from C4 are simply the probability that Word 2 is from the UK or US class given the class of Word 1, with extracted pairs split by whether the words were adjacent or not in the string. Adjacent pairs account for roughly 1% of all pairs in the corpus.

	Word 1 = US		Word 1 = UK	
Condition	T5	T5+FT	T5	T5+FT
Adjacent	92.2	71.1	65.1	63.4
Non-adjacent	88.7	77.7	54.3	63.2

Table 3: Percent of test set examples for which each model prefers consistent over inconsistent spelling.

thetic corpus contains 954238 sentences,[12] equally split between US and UK spelling conventions. A small random subset of 2560 sentences was reserved for validation, and T5 was finetuned on the rest. Finetuning used the same span-filling masked LM task used for pretraining, with dropout set to 0.1, and the loss normalizing factor set to 233472 as suggested in the T5 documentation. Fine-tuning started at the default T5-large checkpoint, which represents 1000700 steps, and proceeded another 99300 steps at a batch size of 128.

As seen in Table 2, finetuning on this synthetic corpus does not appear to improve overall spelling consistency – quite the opposite. However, it does have at least two interesting effects. First, as might be expected, the overwhelming preference for US English shown by base T5 is reduced. Furthermore, the finetuned model is better able to retain long-distance information — there is no dropoff in consistency between the adjacent and non-adjacent conditions as seen for T5 without finetuning.

5.3 Measure 2: prediction accuracy

While the conditional probabilities in Table 2 show the overall preferences of the models over the test set, we also want a measure that captures how often the LLMs assign consistent pairs a higher probability than inconsistent pairs. In Table 3 we show the percentage of the test set examples for which each model predicted consistency over inconsistency. The results show a similar pattern as the conditional probability measures in Table 2. Again, finetuning lowers overall consistency, but results in less drop-off in non-adjacent vs. adjacent conditions.

5.4 Measure 3: mutual information

We also calculated the average mutual information (MI) across all prompt/target pairs in order to measure the strength of association between spelling conventions in both words. For each pair, we calculate four joint probabilities — P(US,US), P(US,UK), P(UK,US), p(UK,UK). We assume these four probabilities make up the entire universe with respect to a particular prompt/target pair, and normalize them so they sum to 1. This also allows us to easily calculate marginal probabilities simply by adding the appropriate joint probabilities – e.g., P(US prompt) = P(US,US) + P(US,UK). To calculate MI, we use a formula based on the log-likelihood ratio calculation in Moore (2004), but equivalent to the standard formulation for mutual information, where x, y are the two probe words:

$$\sum_{x \in \{\text{UK,US}\}, y \in \{\text{UK,US}\}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Since T5 is trained on masked token prediction, to measure the joint probability $p(x,y)$ of each pair of probe words x, y we can simply mask both probing tokens and measure the probability of generating both of them. That is, we present T5 with (6-a) and measure the probability of (6-b):

(6) a. My preferred words are <BLANK-SPAN-1>, <BLANK-SPAN-2>, and tree.
b. <INPUT-SPAN-1> flavour <INPUT-SPAN-2> harbour <INPUT-SPAN-3>

Condition	T5	T5+FT
Adjacent	0.0048	0.0017
Non-adjacent	0.0044	0.0015

Table 4: Average mutual information in the adjacent and non-adjacent conditions.

Table 4 presents these mutual information values. There doesn't seem to be a significant difference between adjacent and non-adjacent conditions for either T5 variant, though finetuning does seem to cause an overall drop in MI, in line with the overall drop in consistency seen in the measures above.

5.5 Nonce forms

We want to determine if T5 assigns the probabilities reported above on the basis of dependencies between specific lexical items, or if it is learning sub-word generalizations. In other words, does the model learn that specific words like *flavour* and *realise* are more likely to co-occur than *flavour* and *realize*? Or does it learn that words containing *-our* are more likely to co-occur with words containing *-ise*? Since the model is trained using Sentence-

[12] In our testing, this was not enough data to reliably train a T5-large LLM from scratch.

British	American	British	American
glavour	glavor	reptalise	reptalize
mentre	menter	amolirise	amolirize
unulise	unulize	sphectre	sphecter
malvour	malvor	imminise	imminize
larbour	larbor	voitre	voiter

Table 5: Nonce forms created by making one to three changes to words in the American-British dictionary.

		Word 2	
		US	UK
Word 1	US	0.68	0.32
	UK	0.56	0.44

Table 6: Conditional probability table for nonce forms given by T5. The table shows the conditional probability of Word 2 (which is a nonce form), given Word 1. For each instance, the probability has been normalized over each condition (i.e., each row in the table).

Piece tokenization (Kudo and Richardson, 2018), it is possible that it exploits sub-word features.

One way of testing if a model can use sub-word features is to create nonce words that contain British- or American-specific sub-words. If the model treats these as being British or American, this is an indication that the model is able to pick up on sub-word features.

We created a list of ten nonce forms by changing, adding, or removing one to three letters in existing words in our dictionary of American and British forms. These forms are shown in Table 5.

We use the same probing template and method as described above. For each probe, we use a real American or British word for the first probe word, and one of the nonce forms shown in Table 5 for the second. For this experiment we queried the base T5 model in the adjacent condition. The resulting conditional probability table is shown in Table 6.

Table 6 shows that the patterns shown in Section 5.1 above do not generalize very strongly to nonce forms. The probabilities assigned to US forms following UK forms are on average higher than UK forms following UK forms. However, the difference between these alternatives is attenuated compared to when Word 1 is a US form, indicating that (a) there is a heavy skew towards US spelling conditions in the training data; but (b) some sensitivity to the UK context, if not enough to counteract the high US form priors. This suggests that the results in Table 2 are to a large extent driven by lexical dependencies rather than any lower-level spelling patterns encoded by wordpieces.

5.6 Autoregressive LLMs

Many commonly-used LLMs (including T5) are trained to predict words in the input that have been masked out. Another common class of LLMs, however, are trained to perform next-word prediction instead. To examine how such autoregressive architectures handle spelling consistency, we experiment with OpenAI's GPT2 (Radford et al., 2019), which has a readily available open-source implementation through HuggingFace.[13]

As GPT2 is purely autoregressive, we cannot compute the probability that a particular probe word will fill a masked sentence span as easily as we could with T5. We can only efficiently compute the probability of a suffix given a prefix. Given this caveat, we have at least two options for assigning conditional probability scores, neither of which should be treated as exactly comparable to the T5 scores above. First, we can count only the logits corresponding to the target word:

P("harbour" | "My preferred words are flavour,"). This local score ignores any words in the template occurring after the target word. Second, we can compute from the start of the target to the end of the sentence: P("harbour, and tree" | "My preferred words are flavour,"), which accounts for the post-target suffix of the sentence.

Tables 7 and 8 show results for both of these methods for calculating the conditional probability, compiled in the same way as the T5 results in Tables 2 and 3. Table 7 also includes the conditional probabilities from GPT2's training corpus, OWT. We see that GPT2 shows a similar preference for consistency as T5, but only very locally. There is a large drop-off in preference for consistency when moving from adjacent to non-adjacent conditions, or when including the completion of the sentence in the calculation. For UK English in particular, any preference for consistency completely disappears beyond the immediate vicinity of the priming word, and the model returns to chance performance on the task.

6 Further analysis of corpora

We now return to a slightly more detailed examination of two of the corpora presented in Table 1,

[13] https://huggingface.co/gpt2

Condition	Word 1	GPT2 (tgt only) Word 2		GPT2 (to EOS) Word 2		OWT Word 2	
		US	UK	US	UK	US	UK
Adjacent	US	0.87	0.13	0.69	0.31	0.95	0.05
	UK	0.36	0.64	0.51	0.49	0.34	0.66
Non-adjacent	US	0.83	0.17	0.66	0.33	0.95	0.05
	UK	0.49	0.51	0.54	0.46	0.28	0.72

Table 7: Conditional probability of Word 2, given a template with Word 1, given by GPT2 scored until the end of the target word only (tgt only) and scored until the end of the sentence (to EOS). We also present the conditional probabilities from pairs found in the training corpus, OWT.

	Word 1 = US		Word 1 = UK	
Condition	GPT2 target	GPT2 EOS	GPT2 target	GPT2 EOS
Adjacent	94.2	70.1	70.8	49.4
Non-adjacent	92.5	67.5	54.6	45.1

Table 8: Percent of test set examples for which each GTP2 scoring variant prefers consistent over inconsistent spelling.

English Wikipedia and the British National Corpus, both of which had relatively high levels of mismatch compared to the other corpora.

Wikipedia is an interesting case, since the documents are collectively edited by potentially a large number of contributors, which may lead to higher expected mismatch than in other corpora. For example, one version of the article on *air lock* used both US-spelling of the word *vapor* and the UK-spelling (*vapour*). This is explained via three versions of the introductory sentence to the page, shown in Table 11 in Appendix B, where the two spellings are added to the sentence at different times, years apart.

The amount of mismatch in the British National Corpus is perhaps more surprising, given the provenance of the materials and intent of the collection. However the diversity of sources, which include things such as journal articles and edited volumes, likely leads to similar issues to those found in Wikipedia, along with simple human error and/or inconsistency. Table 12 in Appendix B presents a few examples of sentences with words from both spelling conventions, with American *-ize* spellings mixed with British *-ise* or *-our* versions.

7 Conclusion and Future Work

We have presented results showing that T5 does tend towards consistency in spelling, but not to the degree that could be relied upon should such consistency be desired in generated text. We show that this general preference for consistency reflects the data that the model is trained on, which also is mostly consistent, but with a significant proportion of exceptions. The model's behavior is also shown to be affected by the relative frequency of language varieties in the training data. We took advantage of the explicit and surface-accessible nature of these dependencies to attribute some model performance to the training data, while also demonstrating that modeling improvements should be possible, since the training data itself is substantially more consistent than the models.

These results suggest several possible avenues for future work. First, methods for addressing bias in training data should yield improvements for British spelling consistency in these models. We also intend to extend these results to languages other than English and investigate how spelling variation in other language situations is learned by LLMs. Some of the methods we used here rely on the fact that English spelling variation is quite thoroughly catalogued. Extending this work to less-documented cases of language variation will require us to either (1) collect data about spelling variation from language informants or data, or (2) develop methods that require less prior knowledge. In the interest of finding methods that are extensible to the greatest number of cases, we intend to pursue path (2), working on methods to mine information about language variation from large corpora and LLMs that have been trained on them.

Acknowledgements

Thanks to Alexander Gutkin, Shankar Kumar, Arya McCarthy and Richard Sproat for useful discussion and comments, and to the anonymous reviewers for helpful suggestions.

Limitations

Our work is focused on just a single case study of spelling variation. As detailed in Section 2, English is a good candidate for a case study for several reasons, but it would be beneficial to extend this work to other language situations.

Another limitation was our choice to focus on already existing pre-trained models, rather than directly controlling the training data that is input to each model. This means some of the conclusions about the connection between training data and outcome are tentative, pending experimental confirmation.

Ethics Statement

This work does not propose a new model or dataset, but rather probes the behavior of existing models. Thus novel ethical considerations about model behavior and dataset contents are not directly raised by this work. While not explicitly focused on ethical considerations, this paper's methods hopefully contribute to better understanding model behavior, and could be used to understand the ways in which large language models treat underrepresented and marginalized language varieties.

References

Kristian Berg and Mark Aronoff. 2017. Self-organization in the spelling of English suffixes: The emergence of culture out of anarchy. *Language*, 93(1):37–64.

BNC Consortium. 2007. The British National Corpus, XML edition. Oxford Text Archive, http://www.natcorp.ox.ac.uk.

Claire Bowern. 2015. *Linguistic fieldwork : a practical guide*, 2nd edition. edition. Palgave Macmillian, Basingstoke, Hampshire.

Jonathan R. Brennan, Chris Dyer, Adhiguna Kuncoro, and John T. Hale. 2020. Localizing syntactic predictions using recurrent neural network grammars. *Neuropsychologia*, 146:107479.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? An analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Kevin Clark and Christopher D. Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Empirical Methods on Natural Language Processing*.

Stefan L. Frank, Thijs Trompenaars, and Shravan Vasishth. 2016. Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? *Cognitive Science*, 40(3):554–578.

Richard Futrell, Ethan Wilcox, Takashi Morita, and Roger Levy. 2018. RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency. *arXiv preprint arXiv:1809.01329*.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Samuel Johnson. 1755. *A Dictionary of the English Language*. J. & P. Knapton, London.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. 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 5803–5808, Hong Kong, China. Association for Computational Linguistics.

Adam Kilgarriff and Irene Renau. 2013. esTenTen, a vast web corpus of Peninsular and American Spanish. *Procedia - Social and Behavioral Sciences*, 95:12–19. Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC2013).

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Robert C. Moore. 2004. On log-likelihood-ratios and the significance of rare events. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 333–340, Barcelona, Spain. Association for Computational Linguistics.

Dong Nguyen and Jack Grieve. 2020. Do word embeddings capture spelling variation? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 870–881, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Brian Roark, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J. Mielke, Cibu Johny, Isin Demirsahin, and Keith Hall. 2020. Processing South Asian languages written in the Latin script: the Dakshina dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2413–2423, Marseille, France. European Language Resources Association.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Noah Webster. 1828. *An American Dictionary of the English Language*. S. Converse, New York.

A Prompts

Table 9 presents the 29 prompt templates that were used in this study. Table 10 gives macro-averaged conditional probabilities for T5 runs across the different prompts, along with standard deviations to indicate how much performance varies due to the choice of prompt.

B Examples of corpus mismatch

Tables 11 and 12 present examples illustrating the mixture of American and British spelling in Wikipedia and the British National Corpus, respectively, as discussed in Section 6.

My preferred words are <CUE> and <FILLER>.
My preferred words are <CUE>, <FILLER>, and tree.
She wrote the words <CUE> and <FILLER>.
She wrote the words <CUE> and <FILLER> in her notebook.
She wrote the words <CUE>, <FILLER>, and cabbage.
I wrote the words <CUE> and <FILLER>.
I wrote the words <CUE> and <FILLER> in my notebook.
I wrote the words <CUE>, <FILLER>, and cabbage.
He wrote the words <CUE> and <FILLER>.
He wrote the words <CUE> and <FILLER> in his notebook.
He wrote the words <CUE>, <FILLER>, and cabbage.
We wrote the words <CUE> and <FILLER>.
We wrote the words <CUE> and <FILLER> in our notebook.
We wrote the words <CUE>, <FILLER>, and cabbage.
Mary wrote the words <CUE> and <FILLER>.
Mary wrote the words <CUE> and <FILLER> in her notebook.
Mary wrote the words <CUE>, <FILLER>, and cabbage.
Please spell <CUE> and <FILLER>.
Please spell <CUE>, <FILLER>, and panther.
Please spell <CUE> and <FILLER> correctly.
Say <CUE> and <FILLER>.
Say <CUE>, <FILLER>, and tapestry.
Say <CUE> and <FILLER> again.
The first words on the list were <CUE> and <FILLER>.
The first words on the list were <CUE>, <FILLER>, and oligarchy.
The easiest words on the list were <CUE> and <FILLER>.
The easiest words on the list were <CUE>, <FILLER>, and oligarchy.
The hardest words on the list were <CUE> and <FILLER>.
The hardest words on the list were <CUE>, <FILLER>, and oligarchy.

Table 9: Prompts used for model evaluation. Non-adjacent versions of each prompt were created by inserting the sequence ", flower, interesting, jump, ponderous, sky, skipping, desk, small, ladder, lovely," between the <CUE> and <FILLER> word slots.

Condition	Word 1	T5 Word 2		T5+FT Word 2	
		US	UK	US	UK
Adjacent	US	0.86 (0.01)	0.14 (0.01)	0.66 (0.03)	0.34 (0.03)
	UK	0.39 (0.06)	0.61 (0.06)	0.44 (0.03)	0.56 (0.03)
Non-adjacent	US	0.83 (0.02)	0.17 (0.02)	0.69 (0.02)	0.31 (0.02)
	UK	0.48 (0.05)	0.52 (0.05)	0.43 (0.04)	0.57 (0.04)

Table 10: Conditional probability of Word 2, given a template with Word 1, given by base T5 and T5 with additional finetuning. Each cell includes a macro-average and standard deviation across 29 prompts.

version date	sentence version
4 Aug. 2017 (neither vapor nor vapour)	An **air lock** is a restriction of, or complete stoppage of liquid flow caused by gas trapped in a high point of a liquid-filled pipe system.
6 Sept. 2017 (vapour replaces gas)	An **air lock** is a restriction of, or complete stoppage of liquid flow caused by vapour trapped in a high point of a liquid-filled pipe system.
15 Feb. 2020 (vapor added)	An **air lock** (or **vapor lock**) is a restriction of, or complete stoppage of liquid flow caused by vapour trapped in a high point of a liquid-filled pipe system.

Table 11: Three versions of a Wikipedia page: (1) no use of *vapor* or *vapour* in the sentence; (2) the term *vapour* replaces *gas*; and (3) the alternative name for the phenomenon "*vapor lock*" is introduced.

Doc ID	sentence
CPD	'What this guy will do is get a **demoralized** sales **organisation revitalised**...' said John Jones, analyst at Salomon Brothers.
CLW	They **conceptualize** these differences in terms of 'separate local **labour** market cultures' (ibid., p. 104).
CBH	It is a metaphor which attempts to create a reality of **organization** whereby cooperation is **mobilised** for fight with the outside world.

Table 12: Examples of spelling convention mismatches in the British National Corpus, sampled from varied books and periodicals.