Examining Language Modeling Assumptions Using an Annotated Literary Dialect Corpus

Craig Messner
Center for Digital Humanities
Johns Hopkins University
cmessne4@jhu.edu

Tom Lippincott
Center for Digital Humanities
Johns Hopkins University
tom.lippincott@jhu.edu

Abstract

We present a dataset of 19th century American literary orthovariant tokens with a novel layer of human-annotated dialect group tags designed to serve as the basis for computational experiments exploring literarily meaningful orthographic variation. We perform an initial broad set of experiments over this dataset using both token (BERT) and character (CANINE)-level contextual language models. We find indications that the "dialect effect" produced by intentional orthographic variation employs multiple linguistic channels, and that these channels are able to be surfaced to varied degrees given particular language modelling assumptions. Specifically, we find evidence showing that choice of tokenization scheme meaningfully impact the type of orthographic information a model is able to surface.

1 Introduction

Orthographic variation, the deviation from one system of spelling in favor of another, occurs due to a range of intentional and unintentional motivations. Unintentional variation may occur when a writer misspells a word relative to their intended system, or when an optical character recognition system misidentifies a particular character. Intentional deviations are instead used to create a desired political or literary effect (Sebba, 2007). For example, adhering to a system of simplified spelling may signal one's dedication to egalitarian politics, while embedding a literary character's speech in a particular orthographic form may signal an authorial desire to present that character as belong to a particular race, class, region or gender (Ives, 1971) (Jones, 1999).

This latter class of intentional variations proves especially diverse. Supported by the availability of surrounding context and reader-familiar stereotypes of speech, literary orthographic edits are frequently unsystematic ("eye dialect") or not fully beholden to phonetics or morphology (Krapp, 1925). Instead, the means by which they convey a "dialect effect" is likely multidimensional.

We present a dataset that includes a novel human-annotated layer of dialect family tags designed to support investigations into these varied signalling pathways. We perform an initial set of experiments and discover indications that literary orthographic variation communicates its dialect effect by modifying information along multiple axes: word-level semantics, context-level semantics, and character edits. In the spirit of previous work investigating the phonetic (Agirrezabal et al., 2023), semantic (Rahman et al., 2023) and contextual (Ethayarajh, 2019) information token and character level models capture, we also provide analysis of the literary orthographic understanding of these model types. We additionally offer evidence that character-level models distinguish between intentional literary orthovariants and constructed unintentional variants.

2 Experiments

2.1 Setup

Data. The data for the following experiments consists of 4032 orthovariant tokens paired with their standard forms and sentence-level context, drawn from a 19th century American literature subset of the Project Gutenberg corpus. This corpus is further described in (Messner and Lippincott, 2024). Messner extended the tag set by providing an additional "Dtag" drawn from a set of 31 possibilities, indicating the dialect form ascribed to each observed token.

Messner used the authorially intended subject-position of speaking characters to assign Dtags to tokens. As a result, the Dtag set mostly represents perceived race, nationality, and region. The most populous category (1726 tokens) is the backwoods (BW) tag which combines samples from white-identified northeastern, western and central plains characters. These subcategories are of BW

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 325–330
November 16, 2024.

are often only subtly disjoint; distinguishing them is likely to cause confusion. Other frequent tags include AA (African American: 653), AR (intentionally archaic: 549), GA (Gaelic: 336) and DE (German: 220).

Models. We employ six models for the following experiments. One, fastText-pretrained (Mikolov et al., 2018) is a subword-aware type level embedding model provided by Facebook and trained on CommonCrawl. We use four pretrained token-level contextual models. Two, BERT-large-uncased and BERT-base-uncased (Devlin et al., 2019) use WordPiece tokenization, while CANINE-c and CANINE-s (Clark et al., 2022) are character-level, with the latter utilizing an additional subword loss function during training. Finally, BERT-forced is BERT-base-uncased configured to encode input strings using only single character WordPiece tokens.

2.2 Procedure

Embeddings: the absolute set. We truncate the dataset, keeping only samples that fit the BERT-forced limit of 512 characters, resulting in 3871 observed-standard pairs. For each pair we generate four additional synthetic tokens:

1. **rev:** The standard word in reversed character order. Ex: circus -> sucric
2. **ocr:** A mutated version of the standard word produced using the nlpaug (Ma, 2019) library's OCR error engine. Ex: circus -> cikcos
3. **swp:** The standard word with a single character swap Ex: circus -> icrcus
4. **rnd:** The standard word with a randomly mutated single character Ex: circus -> circun

We collect embeddings for this full token set. For the type-level model, we embed each individual word. For the contextual models, we insert each variant into the context sentence in turn, embed the full sentence, and extract the set of embeddings that represent the target word. For the BERT family of models, we use the last four hidden layers of the model as the embedding values, while for CANINE we use the final hidden layer. If the target word is embedded as more than one subword or character we mean pool the sub-embeddings to generate a final word embedding.

Data augmentation: the relative set. We use these embeddings to produce additional datapoints consisting of the difference between the embedding of a token's standard form and the embedding of each of the variant forms. Similar to the analogy test of (Mikolov et al., 2013), we use these relative datapoints to investigate a given model's ability to preserve the intuition that similar types of orthographic transformation should produce similar differences in n-dimensional space.

For each model, we cluster the relative and absolute sets using k-means clustering for each $k \in \{1, .., 20\}$. [1]

2.3 Evaluation

We use the following measures to evaluate the efficacy of a given k clustering.

Purity. We calculate purity (Manning, 2008) over the clustering of the full relative token set to gain insight into each model's ability to distinguish between synthetic and observed variants. We also calculate it over the absolute and relative sets of only the observed token to track how well the models cluster embeddings or embedding differences that bear the same Dtag.

Overall accuracy and SO accuracy evaluate a given k clustering's ability to group token variants from the same datapoint into the same cluster. Overall accuracy is the average percentage of correct groupings of all elements of a datapoint into a single cluster. SO accuracy is the average percentage of correct groupings of only the standard and observed tokens into a single cluster.

Cluster semantic coherency measures the overall semantic similarity of the tokens gathered into a cluster k. We calculate this using the average pointwise cosine similarity of the Word2Vec (Mikolov, 2013) embeddings of each token in a cluster. To support this we train a Word2Vec model on the full corpus using the Gensim (Řehůřek and Sojka, 2010) library.

Cluster Mphone similarity measures the phonetic similarity of the tokens gathered into a cluster k. We calculate this using the average pointwise Levenshtein Distance (LD) of the Metaphone (Philips, 1990) encoded version of each token in the cluster. A lower score indicates that the members of the cluster are more phonetically similar.

[1] Code and data for these experiments can be found at https://github.com/comp-int-hum/orthography-embedding-clustering

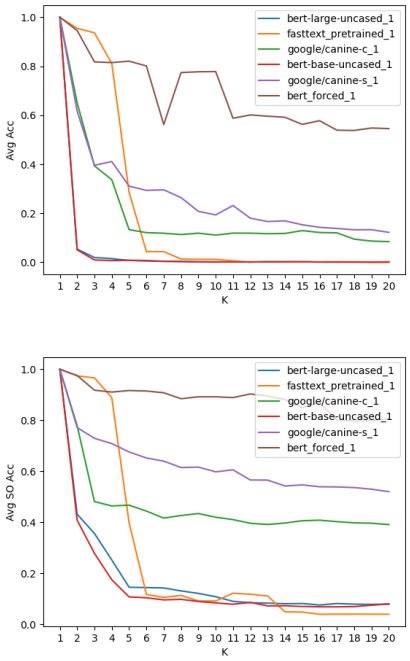

Figure 1: Full absolute set (T), SO absolute set (B) accuracy by k.

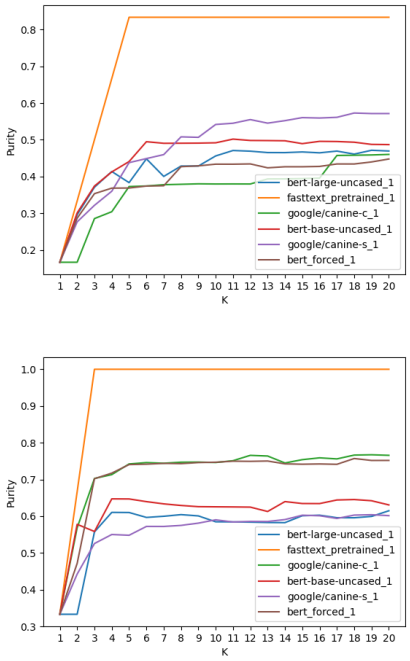

Figure 2: Purity across the full relative set (T) and across non order-swapped tokens (B)

3 Results and Discussion

3.1 Evaluating absolute

Only BERT-forced consistently embeds all variants into a similar region. Figure 1 demonstrates that all of the models except for BERT-forced perform uniformly poorly on both overall accuracy across all k, barring the uninformative case where $k < 6$.

The models that perform best on SO accuracy are character-level.

Again barring the uninformative $K < 6$ cases, BERT-forced, CANINE-s and CANINE-c best separate observed-standard pairings from other tokens in their datapoints (Figure 1). Analysis of their shared error reveals that both models perform poorly on a set of high Levenshtein Distance (LD) edit pairs (average LD 2.67). Correspondingly, their shared correct token transformation set has a lower average LD of 1.66. BERT-forced performs better on higher LD transformations, with average correct and error set average LD of 2.2 and 1.9 respectively, implying that BERT-forced preserves difference information beyond character edits.

3.2 Evaluating relative

Of the character and token level models, the CANINE series most distinctly separates constructed and non-constructed variants into clusters. The type-level fastText-pretrained model most accurately separates the variants (Figure 2). However closer examination reveals that it does not separate individual constructed forms, instead grouping them into a single cluster. Notably, character-level models treat order-swapped tokens as functionally similar, while token-level models do not. Removing the rev and swp tokens benefits all models, but overall benefits character-level models the most. Ultimately, this indicates that character-level models preserve information about the distinctions between standard/constructed token differences and standard/observed token differences. It also implies that they rely to a greater degree on the character-edit information stream of the dialect effect to make this determination.

3.3 Evaluation in the light of Dtag and semantic information

High performance on Dtag clustering relies on a mixture of word-semantic, context-semantic and character edit information. As K increases, BERT-base performs best on absolute and CANINE-s on relative (Figure 3). However, both models ultimately only reach purity scores of $\sim .5$, in part at least due to the dominance of the BW tag. Investigating the proportion of individual dtags on

a per-cluster basis at the jointly performant k=17 reveals how both models capture a partial mix of these signals.

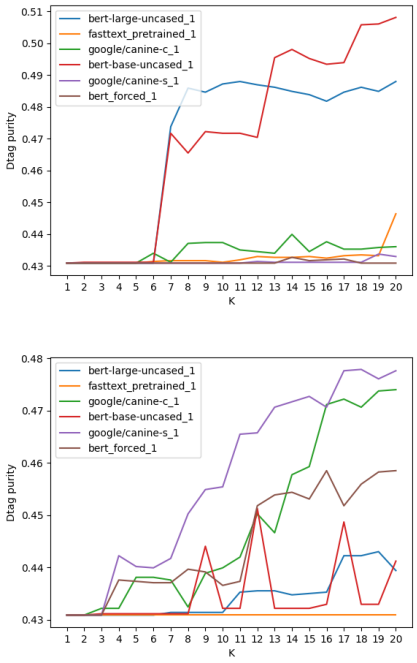

Figure 3: Dtag purity over the obv token embedding (T) Dtag purity over the std-obv relative set (B)

Certain cluster compositions have potential literary significance. Clusters 3 and 7 of the CANINE-s relative set contain high proportions of both AA (African-American) and WS (White Southern) labeled tokens (Table 1).

K	Count	aa	bw	ws	Mphone
3	160	0.36	0.19	0.24	3.1
7	53	0.59	0.08	0.19	2.5

Table 1: Excerpted Dtag proportions and Mphone similarities of CANINE-s relative set clusters at $K = 17$

Both clusters have low word-level semantic coherence scores (.25 and .27 respectively), consistent with the bulk of of the other clusters at $k = 17$, indicating that this grouping likely does not emerge from word-semantics. This Dtag clustering is particularly striking, as it suggests that period authors took a position on the debate surrounding the origins of southern speech (Bonfiglio, 2010).

The edits shared by WS and AA in these clusters (Table 2) largely impact "r"-related graphemes, demonstrating that this clustering likely occurs due to character edit information. Notably, cluster 3 ranks as less sel-similar than cluster 7 by average

Edit	Standard	Observed
er -> ah	after	aftah
er -> a	rather	ratha
r -> ’	quarters	qua’ters

Table 2: Characteristic edits and examples shared by AA and WS in CANINE-s relative clusters 3 and 7

Mphone LD. Upon inspection, cluster 3 contains a wider variety of "r"-related edits, including r -> y and r -> w. In combination with the somewhat more broad distribution of Dtags found in cluster 3, this implies that these "r" edits are somehow nearer to the sorts of "r" edits characteristic of other Dtag groupings found in the cluster, potentially for contextual reasons.

A similar type of distribution also occurs over GA (Gaelic) tokens. Clusters 8 and 12 contain uniquely high proportions of GA tokens (.27 each) while retaining typically low word-level semantic coherence (.24 and .25). Inspection of the tokens reveals that these clusters collect a variety of edits to the "i" and "e" graphemes. However, unlike the WS and AA clusters examined above, both share similar Mphone LD averages of 3.1 and 3.3 respectively. This may signal that these "i" and "e" transformations are more broadly indicative of a variety of dialect contexts.

Context semantics in part determines accurate literary variant clustering. Notably, the BERT-base absolute set at k=17 centralizes clusters around different tags while diffusing WS and GA tokens. For example, cluster 14 has a significantly higher proportion (.33) of DE (German) tagged tokens than any DE-containing cluster found in the CANINE-s relative set.

Edits	Standard	Observed
b -> p	poem	boem
-g	blooming	bloomin
u -> oo	hunters	hoonters
f -> v	falls	valls

Table 3: Characteristic edits and examples of DE tagged tokens in BERT-base absolute cluster 14

The DE tokens in this cluster (Table 3) represent a diverse set of edits, including one (-g in the terminal position) associated with numerous Dtags, including BW (Backwoods) and AA. Given this cluster’s low semantic coherence (.34), a likely conclusion is that this cluster emerges due to the similarity of orthographic contexts in which these tokens appear – say an utterance laden with other

characteristic DE edits.

Low performance on Dtag clustering correlates with high word-semantic cluster coherence in the relative set. For example, BERT-large relative contains multiple clusters with semantic coherency $> .5$, while CANINE-s relative has only one cluster with a score $> .4$. This implies these models favor preserving word-semantic analogical relationships over character edit and context semantics relationships, destabilizing the blend of information needed to successfully cluster over Dtags.

4 Conclusions and Further Work

These experiments offer indications that the dialect effect presented by literary orthographic variation utilizes multiple channels of information: contextual semantics, word semantics and character edits. They also offer evidence that while both contextual token and character level language models can capture all of these aspects, they do so unevenly, justifying further work on the best combination of their information streams.

5 Limitations

The primary limitation of this study emerges from the data. Beyond the inherent limitation of self-restriction to works by 19th century American authors, the coherence of a given observed token and its assigned Dtag is also limited by the inventory of tags chosen. Authors of this period grant their characters multidimensional subject-positions that are reasonably described by but not fully reducible to the granularity of tags like WS and AA. Analysis done in a Dtag-to-cluster direction where the assigned tags are taken as full ground truth limits access to these subtleties.

References

Manex Agirrezabal, Sidsel Boldsen, and Nora Hollenstein. 2023. The hidden folk: Linguistic properties encoded in multilingual contextual character representations. In *Proceedings of the Workshop on Computation and Written Language (CAWL 2023)*, pages 6–13, Toronto, Canada. Association for Computational Linguistics.

Thomas Paul Bonfiglio. 2010. *Race and the rise of standard American*, volume 7. Walter de Gruyter.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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 55–65, Hong Kong, China. Association for Computational Linguistics.

Sumner Ives. 1971. A theory of literary dialect. *A various language: Perspectives on American dialects*, pages 145–177.

Gavin Jones. 1999. *Strange talk: The politics of dialect literature in Gilded Age America*. Univ of California Press.

George Philip Krapp. 1925. *The English Language in America*, volume 1. Century Company, for the Modern language association of America.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Christopher D Manning. 2008. Introduction to information retrieval.

Craig Messner and Thomas Lippincott. 2024. Pairing orthographically variant literary words to standard equivalents using neural edit distance models. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 264–269, St. Julians, Malta. Association for Computational Linguistics.

Tomas Mikolov. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Lawrence Philips. 1990. Hanging on the metaphone.

Md Mushfiqur Rahman, Fardin Ahsan Sakib, Fahim Faisal, and Antonios Anastasopoulos. 2023. To token or not to token: A comparative study of text representations for cross-lingual transfer. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 67–84, Singapore. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Mark Sebba. 2007. *Spelling and society: The culture and politics of orthography around the world*. Cambridge University Press.