BERT's Insights Into the English Dative and Genitive Alternations

Qing Yao
College of Creative Studies
& Department of Linguistics,
University of California Santa Barbara
qyao@ucsb.edu

Simon Todd
Department of Linguistics,
University of California Santa Barbara
& NZILBB, University of Canterbury
sjtodd@ucsb.edu

Abstract

We construct two models that encode varying degrees of context to predict noun phrase order in English dative constructions from their BERT embeddings. The models can successfully predict dative alternations, even without access to context. They are sensitive to features such as animacy, definiteness, and pronominality, suggesting that BERT embeddings encode such information. The best-performing model also shows reasonable success in zero-shot transfer to predicting genitive alternations, indicating some understanding of the shared factors that shape the two alternations. However, the effects of features on the transfer results are not always consistent with known influences on genitive alternations, suggesting that the model may also be drawing from other information encoded in BERT's embeddings. These findings provide insights into the extent to which BERT exhibits human-like word order preferences and demonstrate the potential application of large language models in replacing hand-annotated features for corpus-based studies of syntactic knowledge.

1 Introduction

In the literature on language and cognition, much attention has been paid to syntactic alternations: situations where language users have an apparent choice between two ways of putting together the same words without radically altering meaning. Two such situations that have gained prominence are the English dative (Bresnan et al., 2007; Bresnan and Ford, 2010; Gropen et al., 1989; Theijssen et al., 2013) and genitive (Rosenbach, 2014; Szmrecsanyi et al., 2017; Szmrecsanyi and Hinrichs, 2008) alternations, exemplified in (1) and (2).

(1) **Dative alternation**

a. **NP-dative:** Bob gives [Alice]$_{\text{recipient}}$ [the money]$_{\text{theme}}$

b. **PP-dative**: Bob gives [the money]$_{\text{theme}}$ to [Alice]$_{\text{recipient}}$

(2) **Genitive alternation**

a. **s-genitive:** [a car]$_{\text{possessor}}$'s [tires]$_{\text{possessum}}$ are very durable

b. **of-genitive:** [the tires]$_{\text{possessum}}$ of [a car]$_{\text{possessor}}$ are very durable

In this paper, we study the processing of the dative alternation in BERT (Devlin et al., 2019), in two ways. First, we ask whether pre-trained BERT embeddings can be used to predict alternant choice in dative constructions in a corpus of New Zealand English. We compare models based on BERT embeddings with different degrees of context to a model based on the array of features identified as relevant in the linguistic literature, and find that all models are similarly successful, showing that BERT embeddings encode information that is relevant to the dative alternation. Second, we use the BERT embeddings to assess how the underpinnings of the dative alternation may relate to that of the genitive alternation, by asking how well a model trained to predict the dative alternation can be zero-shot transferred to predict the genitive alternation. The degree to which transfer is possible reflects the degree to which the two alternations are shaped by shared factors, including both general-purpose considerations such as accessibility and construction-specific considerations that are paralleled between them (Diessel, 2020).

Studying the dative and genitive alternations through the lens of BERT has both theoretical and practical implications. On the theoretical side, it can help us to model the cognitive basis of probabilistic sentence production and processing preferences, including the extent to which such preferences are construction-specific and how they can be learned in a highly general way. On the practical side, it can allow us to assess the potential of using large language models to replace time-consuming hand-annotation of features for corpus-based studies of syntactic knowledge.

Proceedings of the Society for Computation in Linguistics (SCiL) 2024, pages 52-62.
Irvine, California, June 27-29, 2024

2 Background

2.1 The dative and genitive alternations

The dative and genitive alternations have figured into many proposals about the nature of the cognitive representations and processes that underpin syntactic knowledge, production, and processing. For example, rule- (Gropen et al., 1989) and construction-based approaches (Gries and Stefanowitsch, 2004) to the dative alternation have appealed to subtle differences in meanings represented by the verb in each alternant, giving cognitive representations of lexical semantics a central role. At the other extreme, accessibility-based approaches (Bock, 1982; MacDonald, 2013) have appealed to the cognitive bottleneck of serial lexical retrieval and highlighted a tendency to prefer alternants that order easily-retrieved arguments first, thus downplaying the role of the precise nature of representations in comparison to general information-processing constraints. In recent years, corpus, experimental, and modeling investigations (Bresnan, 2007; Bresnan and Ford, 2010; Theijssen et al., 2013) have generally supported a middle ground, in which syntactic production and processing are seen as probabilistic, influenced by an array of features including both lexical semantics and determinants of accessibility.

Extensive work has been done in understanding what factors drive these alternations (Bresnan et al., 2007; Rosenbach, 2014; Szmrecsanyi et al., 2017; Szmrecsanyi and Hinrichs, 2008) and how humans learn these alternations (Bresnan, 2007; Bresnan and Ford, 2010; Campbell and Tomasello, 2001; De Marneffe et al., 2012). For both datives and genitives, the alternation can be predicted with high accuracy through a logistic regression model on hand-labeled features including the animacy, definiteness, givenness, pronominality, and length of noun phrase arguments (Bresnan et al., 2007; Szmrecsanyi and Hinrichs, 2008). While these features are universally important in determining these alternations in English, they are sensitive to the variety of English and the era that it is spoken in (Szmrecsanyi et al., 2017).

The similarity between datives and genitives is evident in terms of both semantics and predictive modeling. In terms of semantics – at least for the instances that are typically included in alternation analyses – both can attribute one nominal argument to another in a possession-type relation: prototypical genitives state such a relation, while datives often express a change in such a relation (Wolk et al., 2013). This semantic overlap is further evidenced by the fact that the dative and the genitive cases have merged into one in some Indo-European languages such as Greek or Bulgarian (Catasso, 2011; Stolk, 2015). In terms of predictive modeling, both alternations are sensitive to a common set of features, in similar ways, which is reflected in qualitatively similar coefficients for such features in logistic regression models (Szmrecsanyi et al., 2017; Wolk et al., 2013). In both datives and genitives, there are probabilistic tendencies to order short, animate, and/or definite noun phrase arguments before long, inanimate, and/or indefinite ones.

2.2 BERT and syntactic knowledge

BERT (Bidirectional Encoder Representations from Transformers; Devlin et al., 2019) utilizes a bidirectional attention-based architecture to capture dependencies between words. Its design is particularly well suited for capturing relations between words that are linearly distant in a stream of text, which can present issues for traditional sequence-to-sequence (RNN and LSTM) models. Such long-distance relations are invoked in probabilistic accounts of the dative and genitive alternations through the comparison of features across phrasal arguments (e.g., animacy of the recipient and theme), since those features are typically primarily cued by just one word in a phrase that may be arbitrarily long. We expect BERT to possess an understanding of English word order preferences because previous work has shown that they can be learned by structurally-simpler RNN models (Futrell and Levy, 2019).

Past studies have established that BERT's embeddings encode information about syntactic structure and semantic roles (Jawahar et al., 2019; Manning et al., 2020; Rogers et al., 2021), including at the construction level (Tayyar Madabushi et al., 2020). They also encode information about higher-order organization of the grammatical system that cannot be inferred from any single sentence (Papadimitriou et al., 2021). This information is represented in a multifaceted and gradient manner, much like is posited for human syntactic knowledge, suggesting that insights from human syntactic knowledge may help us understand BERT embeddings and that modeling based on BERT embeddings may help us test hypotheses about human syntactic knowledge.

3 Methods

3.1 Data

Our experiments make use of dative constructions (for training and testing) and genitive constructions (for transfer). To control for effects of variety and era, we restrict our focus to constructions taken from contemporary New Zealand English, as represented by the Canterbury Corpus component of the Origins of New Zealand English corpus (ONZE; Gordon et al., 2007). These constructions occurred in sociolinguistic interviews with New Zealand English speakers born between 1926 and 1987, which were conducted between 1994 and 2007.

Our data consists of 790 datives (680 NP-datives and 110 PP-datives) and 1842 genitives (664 s-genitives and 1178 of-genitives). These are largely the same constructions contained in the data shared by Szmrecsanyi et al. (2017), with minor differences in numbers due to slightly different inclusion criteria. There are two main differences between our data and Szmrecsanyi et al.'s: (1) for the datives, our data is focused on contemporary constructions across a wide range of dative verbs, whereas Szmrecsanyi et al.'s data includes historical constructions and is restricted to datives involving the verb *give*; and (2) for both the datives and genitives, our data contains a brief context for each construction, consisting of the entire line in the corpus from which the construction was extracted, whereas Szmrecsanyi et al.'s data has no context for New Zealand English constructions.

We preprocessed the data by removing transcription annotations that marked pauses, hesitations, and disfluencies. We kept filler words such as 'um' and 'uh', which are argued to be planned components of an utterance (Clark and Fox Tree, 2002).

3.2 Models

We use two models to predict the relative order of two arguments in a dative construction. Both models consist of a binary classifier that uses pretrained BERT embeddings as input. The embeddings used by each model represent different syntactic entities and have access to different amounts of context. The *contextless* model uses embeddings that represent the phrasal arguments, each taken in isolation without consideration of the construction or any broader context. The *preference* model uses embeddings that represent different alternants of the entire construction, considered within a broader context. The corresponding formulations of the prediction task undertaken by each model are as follows:

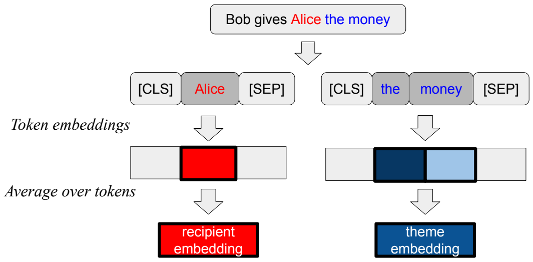

Figure 1: Extraction of embeddings for the contextless model (a)

(a) **CONTEXTLESS: predict phrase order from out-of-context phrasal embeddings.** Given the BERT embeddings of the recipient and theme extracted in isolation, i.e. the embeddings of BERT("[CLS] **[recipient]** [SEP]") or BERT("[CLS] **[theme]** [SEP]"), determine the order in which the noun phrases appear in a dative construction. See Figure 1 for an illustration of the recipient and theme embeddings.

(b) **PREFERENCE: predict attested alternant from contextual construction embeddings.** Given the BERT embeddings of both alternants of a dative construction extracted in context, i.e. the average of embeddings over the bolded tokens in BERT("[CLS] [context] **[verb] [recipient] [theme]** [SEP]") and BERT("[CLS] [context] **[verb] [theme] to [recipient]** [SEP]"), determine which alternant is attested. See Figure 2 for an illustration of the attested and unattested construction embeddings.

The classifier in each model is implemented as a multilayer perceptron with a single hidden layer of size 64 and a sigmoid output layer. For the contextless model, the input is the embedding of the theme concatenated to the embedding of the recipient, and the expected output is 0 if the input is from an NP-dative and 1 if the input is from a PP-dative. For the preference model, the input is the embedding corresponding to the PP-dative concatenated to the embedding corresponding to the NP-dative, and the expected output is 0 if the NP-dative is attested and 1 if the PP-dative is attested.

Each classifier is trained with a binary cross-entropy loss function, via stochastic gradient descent with learning rate 0.01 over 25 epochs. The

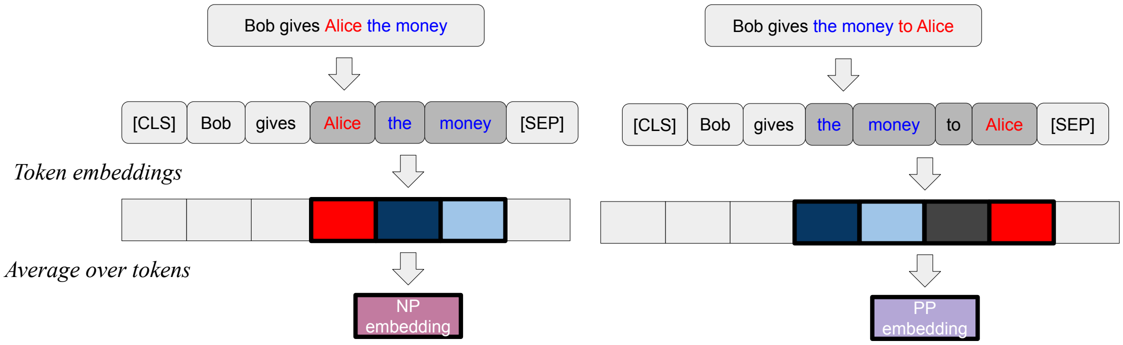

Figure 2: Extraction of embeddings for the preference model (b)

training set for each classifier consists of the same fixed sample of 50 NP-datives and 50 PP-datives, and the held-out test set consists of the same fixed sample of 60 NP-datives and 60 PP-datives. These sizes were chosen to maintain a balance between NP- and PP-datives in training and testing; they are forced to be small by the fact that the data contains only 110 PP-datives. Despite the small size of the training set, we show that the dative alternation can still be reliably predicted without overfitting.

3.3 Embeddings

The embeddings used as input to the models are obtained from the pretrained `BERT-base-uncased` model. To obtain a single embedding for a phrase or construction, we average the embeddings of all tokens it contains.

For both models, the embeddings are obtained from text sequences that are not single, complete sentences. Since BERT is trained on complete sentences, the embeddings therefore represent unaccounted-for situations and may not be entirely robust. Nevertheless, this situation is unavoidable for various reasons. In the contextless model, embeddings are obtained from phrases, paralleling the use of decontextualized phrases in analyses using hand-labeled features; using complete sentences would introduce context, breaking this parallelism, and would allow the model to 'cheat' by referring to information about the relative position of the phrases in position embeddings. In the preference model, embeddings are obtained from lines in the transcripts of a spoken conversational corpus, which may correspond to a fragment of a sentence or several sentences; using complete sentences is not feasible as the transcripts do not indicate sentence boundaries, since utterances in spontaneous speech are not consistently structured into sentences (e.g., Miller and Weinert, 1998).

The BERT model has a lexical layer (layer 0) and 12 Transformer layers (layers 1–12), meaning that it can produce 13 embeddings for each token, each integrating context to different extents. Our analysis compares the results of using these different embeddings in each model. Thus, we train 26 distinct classifiers in total, corresponding to each of the two prediction tasks (a) and (b), and each BERT layer $l = 0, 1, \cdots, 12$.

4 Experiment I: Predicting the dative alternation

In our first experiment, we examine how well the two BERT models are able to predict the dative alternation in the test set. In this examination, we consider the BERT models relative to a logistic regression model based on hand-labeled features, which is the predominant model used to analyze and interpret the alternation in past literature (e.g. Bresnan et al., 2007; Szmrecsanyi et al., 2017). This baseline both establishes how to interpret the performance of the BERT models and highlights the features that are particularly predictive in our training data.

4.1 Baseline logistic model

The baseline logistic regression model is trained on the same balanced training set of 100 dative constructions as the BERT models. Like the contextless BERT model, it receives representations of the recipient and theme as input and must predict the order in which they occur, where the expected output is 0 if the recipient comes first (NP-dative) and 1 if the theme comes first (PP-dative). However, unlike the contextless model, the input representations it uses are not machine-learned embeddings but rather vectors of hand-labeled features, derived from variables that have been established as relevant in past work. These variables include definiteness (*indefinite* or definite), pronominality (*nonpronoun* or pronoun), animacy (*inanimate* or animate), and number (*plural* or singular) of both the recipient and the theme, person (*nonlocal* or local) of the recipient, concreteness (*nonconcrete* or concrete) of the theme, and the length difference in orthographic words between the recipient and the theme (log recipient.length − log theme.length).[1]

For each categorical variable listed above, the italicized level serves as the reference level; that is, the italicized level has a feature value of 0, while the non-italicized level has a feature value of 1. In each case, the non-reference level is the one that has been argued to be 'easier' for lexical retrieval in production planning. Consequently, according to accessibility-based approaches such as Easy First (Bock, 1982; MacDonald, 2013), in which 'easy' elements are ordered before 'hard' ones, we expect recipient-oriented coefficients to be negative when significant and theme-oriented coefficients to be negative when significant. Similarly, given that shorter phrases are 'easier' than longer ones, we expect the length difference coefficient to be positive when significant.

The coefficients learned by the logistic regression model are shown in Table 1. They are qualitatively consistent with results from Bresnan et al. (2007) in terms of both directionality[2] and significance. There is only one difference, in that recipient definiteness is significant in Bresnan et al.'s results but not in ours; this is likely due to the differences in training data size. This difference notwithstanding, the coefficients are consistent with expectations from Easy First, indicating that Easy First preferences are learnable from our training set.

Table 1: Logistic regression coefficients learned from the training set; bolded coefficients are significant at $p < .05$

	Coeff	z
constant	1.71	
rec.def	-0.20	-0.65
rec.pron	**-2.31**	-4.75
rec.person	0.31	0.67
rec.anim	**-0.67**	-2.23
rec.number	0.62	1.55
thm.def	**1.00**	1.99
thm.pron	**0.95**	2.32
thm.anim	0.00	0.03
thm.number	-0.15	-0.37
thm.conc	-0.25	-0.50
length diff (log)	1.42	1.69

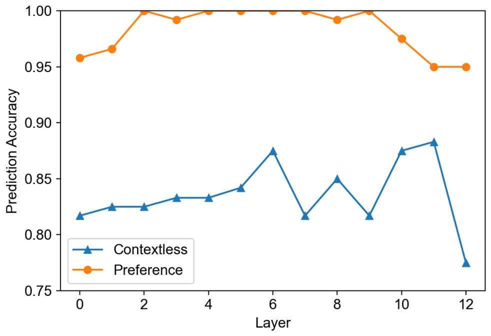

Figure 3: Dative alternation prediction accuracy on the test set by layer

4.2 Results: model comparison

The baseline logistic regression model achieves an accuracy of 0.86 on the test set. The contextless BERT model achieves a similar accuracy and the preference BERT model far exceeds it, in both cases regardless of the BERT layer that is used to provide input embeddings (Figure 3). In all cases, the accuracies are far above those expected from random chance (0.5), indicating that any overfitting due to the small size of the training set is limited.

At the best BERT layers, the contextless model's prediction accuracy is 0.88, which exceeds that of the baseline logistic regression model. As the confusion matrices in Table 2 show, the pattern of responses from the contextless model is very similar to the pattern from the baseline model. Thus, the use of contextless BERT embeddings yields classifications that are equivalent to, or better than, the use of hand-labeled features, at a fraction of the annotation cost.

The predictions made by the contextless model are also highly consistent with those made by the

[1] Note that our list of variables differ from that of Bresnan et al. (2007), since we have only included variables pertaining to the recipient and theme and have omitted variables pertaining to the dative verb.

[2] Note that our coefficients are designed to have the opposite signs to those reported by Bresnan et al. (2007), because we have chosen opposite reference levels.

Table 2: Confusion matrices for the logistic model and contextless model on the dative test set

		True Labels		
		NP	PP	Total
Logistic Predictions	NP	53	10	63
	PP	7	50	57
Contextless Predictions	NP	56	10	66
	PP	4	50	54
	Total	60	60	120

logistic model. The models agree on all but 7 constructions in the test set, consisting of 5 NP-datives that are correctly predicted by the contextless model but not by the logistic model and 2 PP-datives that are correctly predicted by the logistic model but not by the contextless model. Thus, the similarity in overall accuracy reflects a similarity in predicting individual alternations, which may imply that the contextless model is self-discovering sensitivities to a similar set of features as the logistic model (i.e., those listed in Table 1).

The preference model does even better than the contextless model, with near-perfect[3] accuracy on the test set over several BERT layers. We suspect that this increase in performance of the preference model over the contextless model is due to its incorporation of information about the dative verb and the broader context. Because the accuracy is so high, we do not decompose it further.

5 Experiment II: Zero-shot transfer to genitives

Section 4 showed that the BERT models could successfully predict the dative alternation. In particular, the preference model showed near-perfect classification performance on the test set. Here, we ask whether this best-performing model seems to have learned preferences that are specific to the dative alternation or more general preferences that also apply to the genitive alternation.

5.1 The transfer setup

Table 3: Confusion matrices for the adjusted outputs of the preference model on the genitive dataset

		True Labels		
		S	Of	Total
Preference Predictions	S	489	312	801
	Of	175	866	1041
	Total	664	1178	1842

To enact transfer, we created input embeddings for the genitive data in the same way as for the dative data Section 3.3, under the alignment of s-genitives with NP-datives and of-genitives with PP-datives. That is, for each attested genitive in our dataset, we manually created its unattested alternant and obtained embeddings for both the attested and unattested alternants in context. We then formed the input to the preference model by concatenating the embedding corresponding to the of-genitive to the embedding corresponding to the s-genitive.

We measure the success of the transfer by how well the classifier separates the s- and of-genitive constructions. To do so, we manually move the decision threshold by applying an additional linear translation before the final sigmoid layer. We pick the threshold value that yields equal accuracy for s- and of-genitives and treat the overall accuracy obtained under this threshold as our measure of success.

5.2 Results: transfer accuracy

The preference model trained on layer 2 of BERT achieves the best adjusted transfer accuracy of 0.74, which is significantly better than the baseline accuracy of 0.64 achieved by only predicting of-genitives ($p < 0.001$ by exact binomial test). The confusion matrix of the transfer is shown in Table 3, and a graph of its prediction outputs over the entire genitive dataset is shown in Figure 4. While the model is able to separate s- and of-genitives fairly well, suggesting that it has learned general ordering constraints from datives that are applicable to genitives, its output probabilities are compressed, suggesting that these general constraints may yield only weak preferences that could be further adapted for specific constructions.

5.3 Association between labels and features

To dig into the general constraints underpinning the transfer performance, we now consider how the preference model is influenced by the features that have been recognized as (potentially) relevant for predicting both dative and genitive alternations. These target features are animacy and definiteness of the possessor (recipient), animacy of the possessum (theme), and difference in argument lengths

[3] We do not interpret accuracies of 1 as 'perfect' due to the limited sample size of the test set. In a larger and more diverse test set, we expect the preference model's accuracy to be high but not quite this extreme.

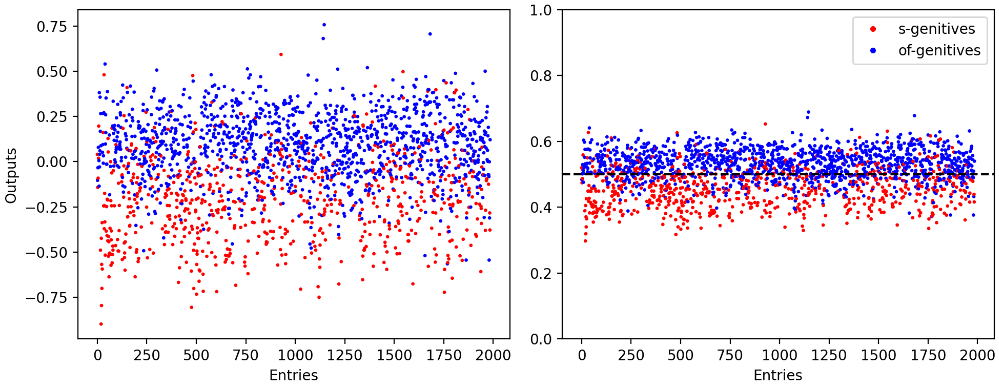

Figure 4: Genitive alternation predictions by the preference model. Left: pre-sigmoid outputs; Right: sigmoid adjusted outputs

(possessor − possessum)[4].

For each target feature, we restrict attention to a subset of constructions differing only in that feature, to minimize confounds. The other features are fixed at levels that maximize the size of this subset and ensure that each level of the target feature is (maximally) attested. Then, for each level of the target feature, we calculate its pointwise mutual information (PMI) with both the BERT labels predicted by the preference model and true labels of the chosen alternant in each construction in the subset. If the associations are consistent with the principle of Easy First, we expect animate and definite possessors, inanimate possessums, and small/negative length differences to yield positive PMI with s-genitives and negative PMI with of-genitives, and vice versa for the opposite levels of each feature. If the preference model has learned associations that are present in genitives, despite being trained on datives, then we expect the PMIs with the BERT labels to pattern similarly to the PMIs with the true labels.

The results are shown in Tables 4 to 7. The associations between features and genitive alternant choice do not consistently align with expectations from Easy First, either for the alternant labels predicted by the preference model or for the true labels. It is hard to know whether this unexpected behavior indicates a real quirk of New Zealand English or is just an artifact of the sparse data and/or the specific levels at which non-target features were fixed.

Regardless, the PMIs with the model's labels almost always agree in sign and relative magnitude with the PMIs with the true labels, which suggests that the model has learned general associations that are transferable between the dative and genitive alternations. The associations seem to be weaker for the model than for the true labels, consistent with the idea that general constraints on order preferences are weaker than construction-specific constraints. However, the associations with possessor definiteness (Table 6) appear to be stronger for the model than for the true labels, which is especially surprising given that recipient definiteness was not strongly correlated with alternant choice in the datives training data (Table 1).

[4] In order to permit alternation, genitive constructions must have a definite possessum (Rosenbach, 2014). This definiteness can be marked by determiner in of-genitives, but not in s-genitives (e.g., [*the tires*] *of the car* vs. *the car's* [*tires*]). To account for this difference when calculating length, we followed past work (e.g., Szmrecsanyi and Hinrichs, 2008) in not counting *the* at the beginning of the possessum in of-genitives.

Table 4: Transfer accuracy and PMIs on genitive constructions differing only in possessor animacy. These constructions all have definite possessors, inanimate possessums, and a length difference of 1.

	Possessor animacy		
s-genitive	Inanim	Anim	Total
# Correct	8	78	86
# Total	15	100	115
Accuracy	0.53	0.78	0.75

of-genitive	Inanim	Anim	Total
# Correct	379	23	402
# Total	457	28	485
Accuracy	0.83	0.82	0.83

BERT-labels PMI	Inanim	Anim
s-genitive	-0.63	1.20
of-genitive	0.19	-1.03

True-labels PMI	Inanim	Anim
s-genitive	-2.59	2.03
of-genitive	0.26	-1.89

Table 5: Transfer accuracy and PMIs on genitive constructions differing only in possessum animacy. These constructions all have animate and definite possessors and a length difference of 1.

	Possessum animacy		
s-genitive	Inanim	Anim	Total
# Correct	78	35	113
# Total	100	46	146
Accuracy	0.78	0.76	0.77

of-genitive	Inanim	Anim	Total
# Correct	23	1	24
# Total	28	2	30
Accuracy	0.82	0.50	0.80

BERT-labels PMI	Inanim	Anim
s-genitive	-0.06	0.15
of-genitive	0.12	-0.37

True-labels PMI	Inanim	Anim
s-genitive	-0.09	0.21
of-genitive	0.36	-2.03

Table 6: Transfer accuracy and PMIs on genitive constructions differing only in possessor definiteness. These constructions all have animate possessors, inanimate possessums, and a length difference of 1.

	Possessor definiteness			
s-genitive	Indef	Def	Def-pn	Total
# Correct	30	78	5	113
# Total	35	100	10	145
Accuracy	0.86	0.78	0.50	0.78

of-genitive	Indef	Def	Def-pn	Total
# Correct	4	23	2	29
# Total	8	28	3	39
Accuracy	0.50	0.82	0.67	0.74

BERT-labels PMI	Indef	Def	Def-pn
s-genitive	0.24	-0.04	-0.53
of-genitive	-0.66	0.08	0.70

True-labels PMI	Indef	Def	Def-pn
s-genitive	0.05	-0.01	-0.03
of-genitive	-0.19	0.05	0.12

Table 7: Transfer accuracy and PMIs on genitive constructions differing only in length difference (possessor − possessum). These constructions all have inanimate and definite possessors and inanimate possessums.

	Length difference			
s-genitive	≤ 0	$= 1$	≥ 2	Total
# Correct	3	8	3	14
# Total	11	15	5	31
Accuracy	0.27	0.53	0.60	0.45

of-genitive	≤ 0	$= 1$	≥ 2	Total
# Correct	88	379	90	557
# Total	130	457	127	714
Accuracy	0.68	0.83	0.71	0.78

BERT-labels PMI	≤ 0	$= 1$	≥ 2
s-genitive	0.48	-0.33	0.40
of-genitive	-0.18	0.09	-0.14

True-labels PMI	≤ 0	$= 1$	≥ 2
s-genitive	0.91	-0.39	-0.14
of-genitive	-0.06	0.01	0.01

6 Discussion & Conclusion

In this paper, we have presented two models designed to predict dative alternations from BERT embeddings. In Section 4, we found that the dative alternation can be predicted with high accuracy from BERT embeddings, and in a manner mostly consistent with traditional logistic regression models based on hand-annotated features. In Section 5, we explored the zero-shot transferability of our context-aware dative alternation model to genitive alternations. The transfer was relatively successful, and we explored both its success and limitations by analyzing the pointwise mutual information between assigned labels and features. Our findings suggest that BERT-based alternation models perform comparably to traditional approaches utilizing hand-annotated features, and that they are capable of recognizing general principles that yield similarities between the dative and genitive alternations.

Our experiments showcase potential approaches for understanding how word-order preferences are encoded in BERT's embedding space and the extent to which they are construction-specific. The success of our preference model in the zero-shot transfer from datives to genitives suggests that it is not solely relying on (dative) construction-specific constraints to derive word-order preferences, but rather appealing to more general constraints. One possible such general constraint is Easy First (Bock, 1982; MacDonald, 2013), which showed reasonable explanation of patterns of alternant choice in our datives training set. However, the fact that the transferred model captures the apparent patterns in genitive alternant choices even when they do not seem to be consistent with Easy First suggests that the general constraints it learned from the datives cannot be boiled down just to Easy First. Given that the preference model utilizes pre-trained embeddings of entire alternants, which plausibly reflect in some way the extent to which lexical subsequences within that alternant are evidenced in BERT's training data, it is possible that the model's choices may be influenced by local surprisal statistics based on the different lexical subsequences that are formed when the noun phrase arguments are placed in different orders. That is, the general constraints being invoked may involve some degree of 'episodic memory-matching' based on BERT's pre-training data, as well as consideration of more abstract features.

One interesting future study could consider a direct comparison between the alternation preferences of the preference model with that of humans. In the present work, we focused on analyzing the extent to which our BERT-based models can determine the order in which humans produce two noun phrases in dative and genitive constructions. To what extent does learning to match these categorical production preferences enable the prediction of gradient human perceptual preferences? Humans have preferences about reading the arguments in one order relative to the other, which varies between individuals and across contexts (Bresnan and Ford, 2010). By evaluating the similarities and differences between these preferences and the probabilities output by the preference model, we may be able to further understand both BERT embeddings and human syntactic knowledge.

7 Limitations

Although a small training set of 100 dative constructions appears to be sufficient for predicting dative alternations and for zero-shot transfer to genitive alternations, we ideally want a larger training set to improve the robustness of our models. Also, due to the strong correlation between animacy and alternation type in both the dative and genitive datasets, obtaining a sufficient number of constructions that differ minimally in features for the PMI analysis is challenging. Some of the feature labeling in our dataset may also be too coarse to capture the gradient nature of the features. For instance, rather than treating animacy to be binary, Szmrecsanyi et al. (2017) considers human and animals, collective, temporal, locative, and inanimate as distinct categories. All of these data-related issues can add variability to our analysis.

On the model side, our interpretation of results has generally made the assumption that our models are actually making predictions from self-discovered versions of the features that the literature has shown to be relevant to the dative and genitive alternations, rather than from something else entirely. Although our models' predictions are consistent with known associations between features and alternations, it does not necessarily imply that they are learning to be sensitive to those features, since the training labels are themselves correlated with the features. In addition, we have interpreted our results very generally, but the restriction to contemporary New Zealand English may limit the generalizability of our findings.

Acknowledgements

Cloud computing credits for this study were provided through a UCSB Undergraduate Research & Creative Activities grant to Qing Yao. The data used in this study was collected by Simon Todd with support from the National Science Foundation under Grant No. BCS-1025602, with the help of Kirsty Canillo and Daniel Bürkle and the guidance of Joan Bresnan, Anette Rosenbach, and Jen Hay. We gratefully acknowledge the students and researchers from the Linguistics Department of the University of Canterbury who contributed to the creation of the contemporary Canterbury Corpus portion of the ONZE Corpus, as well as Robert Fromont for programming and maintaining the LaBB-CAT interface to the corpus that was used to extract our data. We thank Marthe Midtgaard and Shialan Yu for their contributions to early work from which this study developed, Mary Bucholtz and members of the UCSB Linguistics undergraduate thesis seminar for their support, and members of the UCSB CEILing research group for their feedback on an early draft.

References

J. Kathryn Bock. 1982. Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review*, 89(1):1–47.

Joan Bresnan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston and Wolfgang Sternefeld, editors, *Roots: Linguistics in Search of its Evidential Base*, pages 77–96. Mouton de Gruyter, Berlin.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Bouma, Irene Krämer, and Joost Zwarts, editors, *Cognitive Foundations of Interpretation*, pages 69–94. KNAW, Amsterdam.

Joan Bresnan and Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language*, 86(1):168–213.

Aimee L. Campbell and Michael Tomasello. 2001. The acquisition of English dative constructions. *Applied Psycholinguistics*, 22(2):253–267.

Nicholas Catasso. 2011. Genitive-dative syncretism in the Balkan sprachbund: An invitation to discussion. *SKASE Journal of Theoretical Linguistics*, 8(2):70–93.

Herbert H. Clark and Jean E. Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Marie-Catherine De Marneffe, Scott Grimm, Inbal Arnon, Susannah Kirby, and Joan Bresnan. 2012. A statistical model of the grammatical choices in child production of dative sentences. *Language and Cognitive Processes*, 27(1):25–61.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171—-4186. Association for Computational Linguistics.

Holger Diessel. 2020. A dynamic network approach to the study of syntax. *Frontiers in Psychology*, 11:604853.

Richard Futrell and Roger P. Levy. 2019. Do RNNs learn human-like abstract word order preferences? *Proceedings of the Society for Computation in Linguistics*, 2:50–59.

Elizabeth Gordon, Margaret Maclagan, and Jennifer Hay. 2007. The onze corpus. In Joan C. Beal, Karen P. Corrigan, and Hermann L. Moisl, editors, *Creating and Digitizing Language Corpora: Volume 2: Diachronic Databases*, pages 82–104. Palgrave Macmillan, Basingstoke.

Stefan Th. Gries and Anatol Stefanowitsch. 2004. Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics*, 9(1):97–129.

Jess Gropen, Steven Pinker, Michelle Hollander, Richard Goldberg, and Ronald Wilson. 1989. The learnability and acquisition of the dative alternation in English. *Language*, 65(2):203–257.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651—-3657. Association for Computational Linguistics.

Maryellen C. MacDonald. 2013. How language production shapes language form and comprehension. *Frontiers in Psychology*, 4:226.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Jim Miller and Regina Weinert. 1998. *Spontaneous Spoken Language: Syntax and Discourse*. Oxford University Press, Oxford.

Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. 2021. Deep subjecthood: Higher-order grammatical features in multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2522–2532. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Anette Rosenbach. 2014. English genitive variation -– the state of the art. *English Language and Linguistics*, 18(2):215–262.

Joanne Vera Stolk. 2015. Dative by genitive replacement in the Greek language of the papyri: A diachronic account of case semantics. *Journal of Greek Linguistics*, 15(1):91–121.

Benedikt Szmrecsanyi, Jason Grafmiller, Joan Bresnan, Anette Rosenbach, Sali Tagliamonte, and Simon Todd. 2017. Spoken syntax in a comparative perspective: The dative and genitive alternation in varieties of English. *Glossa: a journal of general linguistics*, 2(1).

Benedikt Szmrecsanyi and Lars Hinrichs. 2008. Probabilistic determinants of genitive variation in spoken and written English. In Terttu Nevalainen, Irma Taavitsainen, Päivi Pahta, and Minna Korhonen, editors, *The Dynamics of Linguistic Variation: Corpus Evidence on English Past and Present*, pages 291–309. John Benjamins, Amsterdam.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets Construction Grammar. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4020–4032. International Committee on Computational Linguistics.

Daphne Theijssen, Louis Ten Bosch, Lou Boves, Bert Cranen, and Hans Van Halteren. 2013. Choosing alternatives: Using Bayesian networks and memory-based learning to study the dative alternation. *Corpus Linguistics and Linguistic Theory*, 9(2):227–262.

Christoph Wolk, Joan Bresnan, Anette Rosenbach, and Benedikt Szmrecsanyi. 2013. Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica*, 30(3):382–419.