An LLM Investigation into Inherent and Structural Case Representation: a German Case Study

Iona Carslaw[1, 2], András Bárány[2], Itamar Kastner[2], Mark Steedman[1]

[1]School of Informatics, University of Edinburgh
[2]School of Philosophy, Psychology & Language Sciences, University of Edinburgh
i.c.a.carslaw@sms.ed.ac.uk {andras.barany, itamar.kastner, m.steedman}@ed.ac.uk

Abstract

A question for computational linguistics has been to what degree do language models encode case information. However, the majority of the work has focused on structural cases (cases which change when the syntactic configuration changes). On the other hand, inherent cases (which are assigned by specific lexical items and do not change if the syntactic configuration changes) have been overlooked. This paper sets out to investigate if German language models distinctly encode inherent dative from structural accusative and nominative. We conducted a linguistic probing investigation where probes are trained on contextual word embeddings of active nominative, accusative, and dative arguments to predict if passivised datives are analysed as a structural nominative. We provide a cased and caseless version of the experiment. Our results suggest that when case information is removed language models can distinguish between inherent dative and structural accusative, regardless of argument position, due to verb information. However, language models cannot distinguish between structural nominative and inherent dative when the dative appears in a position where there is an expected nominative, due to over-relying on surface patterns.

1 Introduction

A large amount of literature has been concerned with testing the syntactic abilities of language models as a means to understand neural net architectures and what (if anything) this can say about languages and how we learn them (Wilcox et al., 2018; Marvin and Linzen, 2018; Futrell et al., 2019; Papadimitriou et al., 2021; Kamerath and Santo, 2025; Kennedy, 2025, a.o.). A question amongst this work is to what degree do language models encode case information. Case here is defined as how DPs are morphologically marked to indicate their role in a sentence (Baker, 2015). Consider the active German sentence in 1a with the subject receiving nominative, and the object receiving accusative. If we passivise 1a, the case assignment changes with the new subject receiving nominative (1b[1]). Thus, a change in the syntactic configuration and the argument's syntactic position leads to a change in case assignment. These case types (accusative and nominative for German, but in other languages ergative and absolutive) are called *structural cases*.

(1)

a. *Der* *Metallurge* *hat* *den* *Cricketschläger* *geklaut*
the.NOM metallurgist.NOM has the.ACC cricket-bat.ACC stolen

'The metallurgist stole the cricket bat'

b. *Der* *Cricketschläger* *wurde* *geklaut*
the.NOM cricket-bat.NOM PASS.AUX stolen

'The cricket bat was stolen' (McFadden, 2024, p.177)

Given that a particular question is how language models encode arguments and what information is used to distinguish them (Papadimitriou et al., 2022; Mahowald et al., 2023; Lee et al., 2024; Wilson et al., 2023; Conia and Navigli, 2022), understanding how language models represent structural cases has become a research area (Papadimitriou et al., 2021; Mahowald et al., 2023; Ozaki et al., 2025). However, a language's case system can consist of more than just structural cases. German has dative which can be assigned by specific predicates to their objects (2a). However, unlike the structural cases, when the dative object gets passivised, the dative remains despite the change in syntactic configuration (2b). These types of cases are called *inherent cases*. Their representation within language models has yet to be investigated.

(2)

a. *Der* *Metallurge* *hat* *dem* *Dekan* *nicht*
the.NOM metallurgist.NOM had the.DAT dean.DAT not
gehorcht
obeyed

'The metallurgst didn't obey the dean'

[1]Glosses have been adapted.

b. *Dem* *Dekan* *ist* *nicht* *gehorcht* *worden*
the.DAT dean.DAT is not obeyed PASS.AUX

'The dean was not obeyed' (McFadden, 2024, p.177)

Past research has shown that language models favour utilising word-order strategies over morphological strategies (Czarnowska, 2022), and struggle generalising morphological information (Temesgen et al., 2025; Ismayilzada et al., 2025). Given the different behaviours of inherent and structural case, they provide a fruitful testing ground for how much morphological information (i.e. case) is being used by language models versus just word-order/surface patterns when encoding arguments. For instance, a German language model when generalising a structural case system can generalise that subjects receive nominative and objects receive accusative, and therefore just rely on syntactic positions. However, to predict the inherent dative, a model cannot just rely on syntax. Thus, this paper sets out to investigate how language models distinguish German structural and inherent cases from one another, by answering the following two questions: (i) are arguments which receive inherent cases versus those that receive structural cases distinguished from one another in the language model's representation; and (ii) do models default to a structural representation of arguments instead of a case representation. Our results suggest that German language models distinguish between structural accusative and inherent dative due to verb information, but when it comes to structural nominative and inherent dative, the models rely on surface patterns[2]

2 Background

Previous work shows that language models encode structural case information. Papadimitriou et al. (2021) conducted a probing investigation to see if language models can learn high-order structural case information. They found that *mBERT* represents arguments differently depending on how a language's structural cases are organised. Mahowald et al. (2023) questioned if language models can predict subject and objects when word order information is not present. They found that language models trained on cased languages performed better than those that do not. Although, they do state that the margin of improvement is minimal. Ozaki et al. (2025) found that a Hindi-Urdu LSTM had a linear encoding of structural cases which the model used in computing agreement tokens. However, doubts have been raised about the ability of language models to generalise morphological information. Czarnowska (2022) found that neural parsers have a tendency to over-rely on word order and lexical semantics, even if the language's primary way of indicating syntactic relations is through morphology. Weissweiler et al. (2023) conducted the famous Wug test (Berko, 1958) on *ChatGPT* and found that it fell short from human achievement and models that are trained for morphological analysis. Haley (2020) showed similar results with *BERT*. Many other studies have also raised doubts that language models have an abstract representation of morphological rules - inflectional or derivational (Ismayilzada et al., 2025; Temesgen et al., 2025; Weller-Di Marco and Fraser, 2024). Given there is evidence that language models encode structural case, but do not generalise morphological rules, we should investigate instances where case information diverges from structural positioning (i.e. inherent case).

[2]Part of the code is available at: https://github.com/ionacarslaw/German_Inherent_Structural_Probe

Train		**Test**
Label: 0	**Label: 1**	
1: Nom + Acc	Dat	Dat
2: Nom	Dat	Dat
3: Acc	Dat	Dat
4: Nom	Dat + Acc	Dat
5: Acc	Dat + Nom	Dat

Table 1: Different data configurations for the binary training classes.

3 Experimental Design

3.1 Probing Design

To investigate how inherent and structural cases are represented in German language models, we undertook a probing experiment. We followed closely to the methodology in Papadimitriou et al. (2021), where we take contextual word embeddings of arguments to train classifiers. We are interested in answering the following: (i) if structural and inherent cases are meaningfully distinguished from one another in the language model's representation, and (ii) if, instead of a case representation, models default to a structural/word-order representation of arguments. To test this, we used a training set of active structural nominatives (1a), structural accusatives (1a), and inherent datives (2a). Our test set consisted of passivised inherent datives (2b).

The main task in answering (ii) is, given class 0 of active structural nominatives and accusatives, and class 1 of active inherent datives, what would a probe classify passivised inherent datives as? If a model has just structural case representation and relies purely on structural argument encoding, then the passivised inherent argument will be classified alongside the nominatives/structural cases (class 0), if the models do not just use structural sequences and correctly model inherent case behaviour, then it will be classified alongside the other inherent datives (class 1). This set-up also helps answer (i) with the classes being along an inherent-structural case split. However, to fully answer if an inherent-structural dimension is a meaningful dimension that language models encode case, we include other data configurations, as it may be that a language model does not specifically distinguish between inherent and structural cases, but nonetheless the probe could extrapolate that information through other means. Therefore, we include a data configuration where nominative is class 0 and dative and accusative are class 1, and another where accusative is class 0 and nominative and dative are class 1 (see configuration 4 and 5 in Table 1). If the models meaningfully distinguish inherent and structural cases, the inherent and structural split (configuration 1) will be the most accurate. We have also included configurations where there are only two cases to see how much the classifier is training on subjecthood and objecthood (configuration 2 and 3).

3.2 Argument Selection

In German, structural nominatives and accusatives are the canonical case assignment, therefore specific criteria is not needed for them to arise. However, inherent cases are distinct in that they need certain lexical items to assign them. There are three distinct verbs groups in German which assign an inherent dative[3]; malefactive/benefactive verbs, where the malefactive/benefactive object is dative (2a), ditransitives, where the recipient object is dative (3a), and experiencer-stimulus verbs, where the experiencer object is dative (3b). These verbs differ in their argument structure (two versus three arguments), their theta roles (benefactive/malefactive, recipient, experiencer), and what syntactic operations they can undertake (e.g. experiencer-stimulus verbs cannot undergo passivisation).

(3)

a. *Er* *gibt* *seinem* *Meerschwein* *Futter*
he.NOM gives his.DAT guinea.pig.DAT food.ACC

'He gives food to the guinea pig'

b. *Der* *Tänzerin* *gefällt* *die* *Musik*
the.DAT dancer.DAT pleases the.NOM music.NOM

'The music pleases the dancer' (Haspelmath and Baumann, 2013)

A case system requires generalising high-order information (Papadimitriou et al., 2021). Therefore, to actually test the models case-generalising ability, we have made sure that the inherent datives in the training set are distinct from the test set. Therefore, our training set datives consisted of active datives from ditransitives and experiencer-stimulus verbs. Our test set consisted of passivised datives from malefactive/benefactive verbs (2b). Setting up the test set as so means we can determine if the probe has merely learnt that datives are a class due to them occurring with ditransitives and experiencer-stimulus verbs; if so the error rating on the test set will be around chance as how the dative are classified will not generalise over to malefactive/benefactive verbs.

4 Experiment 1: Cased Representation

4.1 Datasets

To build the datasets, we used the German UD corpora (McDonald et al., 2013) and native speakers. From UD, we removed any sentences that had sentence or argument coordination by looking for `ccomp`, and any with subordinate clauses by looking for `xcomp`. To gather the nominatives, we searched for `nsubj` and checked if `Case=Nom`. If the relationship was `nsubj:pass`, or the argument was a proper noun (`PROPN`) - which in German does not exhibit overt case - we skipped it. All other instances were added to the nominative dataset. For the accusatives, we searched for `obj` and checked if `Case=Acc`. If the argument was a proper noun we skipped it, otherwise all others were added to the accusative set. To gather the datives, we searched for sentences which contained verbs which assign dative. If an argument had `obj`

[3]German has inherent genitive and (arguably) inherent accusative. For the inherent genitive, we decided to not include them in our experiment, specifically because German genitive can naturally occur in a subject due to non-inherent case assignment means. For the inherent accusative, it is argued to be in the following: *mich.ACC friert* 'I'm freezing' (McFadden, 2008, p.64). However, *mich* cannot undergo A-movement (of which passivisation is an instance of); because we diagnose inherent cases as those that survive under A-movement, we do not count these as inherent. Both of these instances were removed from the data set.

or obl:arg and Case=Dat, they were set as potential training instances. If an argument had nsubj or nsubj:pass and Case=Dat they were set aside for potential training instances. Two native speakers went through the instances of datives and checked for the training set if: (i) these were instances of inherent datives, (ii) they were objects, and (iii) these datives were assigned by ditransitives or the experiencer-stimulus verbs[4]. For the test data, the native speakers checked if: (i) these were instances of datives, (ii) they were passivised, and (iii) these datives were assigned from malefactive/benefactive verbs. If they found an instance of a dative object assigned by a malefactive/benefactive verb, they passivised the sentence and added the entry to the test set. This process meant there were 39 instances of passivised inherent datives and 198 instances of active inherent datives. To increase the training and test size of the dative, the native speakers created new sentences. They had access to the nominative and accusative instances to emulate the subject areas and sentence structures the UD corpora has. For all the tasks involving the annotators, both checked the others work and if they disagreed on a judgement then they removed those data instances from the datasets. The total amount of training datives (including the UD instances) was 2021, and the total amount of test datives (including the UD instances) was 144. To make sure there were no class imbalances, a random selection of the nominatives and accusative occurred so that the training set had 2021 instances of accusatives and 2021 instances of nominatives. For data configurations where a single class was made up of two cases, 50% of each case was randomly selected so there was not a class imbalance. Approx 10% of the training set was withheld as a validation set so after the classifiers were trained, the layers that were investigated were chosen on accuracy of the validation set rather than the test set.

Control Datasets Given that the data has come from two different sources, we created control datasets to see if dataset characteristics played a role in the classifier accuracy. Alongside the randomly selected conditions, we built a *noun only* dataset (1234 training instances per case of which ∼10% for validation, 192 dative test instances), a *pronoun only* dataset (214 training instances per case of which ∼10% for validation, 52 dative test instances), and a dataset where the overall sentence *token amount* per case had the same distribution (1336 training instances per case of which ∼10% for validation, 244 dative test instances). For the *same token amount* control dataset we tokenised sentences with the *mBERT* tokeniser, counted token per data entry for token amounts, and made sure the overall distribution of token amounts were the same across cases. We conducted a one-way ANOVA to confirm that there was not a difference between case and token amount for this control dataset, and it confirms that there is not a statistically significant difference ($F(2, 4005) = 0.31$, $p = 0.736$).

[4]For a full list of the datives predicates see appendix E.

4.2 Models and embeddings

BERT We obtained contextual word embeddings from *bert-base-german-cased* (Chan et al., 2019); a case-sensitive German only language model with a bidirectional transformer encoder architecture (Vaswani et al., 2017), with 12 hidden layers and 110 Million parametres. Hyperparametres were set at default. If an argument had multiple tokens, the token embeddings were averaged.

mBERT We obtained contextual word embeddings from *bert-base-multilingual-cased* (Devlin et al., 2018); a case-sensitive multilingual language model with a bidirectional transformer encoder architecture (Vaswani et al., 2017), with 12 hidden layers and 110 Million parametres. Hyperparametres were set at default. If an argument had multiple tokens, the token embeddings were averaged.

LLäMmlein 1B We obtained contextual word embeddings from *LSX-UniWue/LLaMmlein_1B* (Pfister et al., 2025); a German only language model based on the decoder-only transformer architecture (Vaswani et al., 2017), with 22 hidden layers and 1 Billion parametres. Hyperparametres were set at default. If an argument had multiple tokens, the token embeddings were averaged.

Word2Vec Baseline We used static word embeddings as a baseline because the embeddings do not represent the surrounding sentence context of the arguments. If the Word2Vec embeddings represents the different cases, it is then due to the distribution of the lexical items. We used a case-sensitive German only static model (Müller, 2015) which was trained with gensim's word2vec library. A skip-algorithm was implemented with a sliding window of 5. Words with a frequency lower than 50 were ignored. The total vocabulary size of the

model is 608,130. For noun phrases, the noun embedding and determiner embedding were averaged.

Given this model does not include a tokeniser, and relies on a word appearing in its training data, some arguments could not be embedded. To mitigate this, if a word did not have an embedding, we undertook three processes to see if one was available. The first was capitalising pronouns (ignoring pronouns which have a distinction with uppercase, e.g. sie/Sie). The second was removing accents in case a data-cleaned version of the word was in the vocabulary. The third concerned German's productive ability to create new words through compounding. Here we used the compound splitter *Charsplit* (Tuggener, 2016) on the nouns in our dataset; if the first suggested compound split had a probability above 0, we checked if the two suggested words had embedding representation. If so they were averaged together. All other instances were removed from the static training, this meant 267 datives (of which 35 were from the test set), 313 accusatives, and 299 nominatives were removed from the Word2Vec datasets.

4.3 Classifier

Our classifier was a neural network with a single hidden layer of size 64, trained for 20 epochs on our training data. The classifier was trained to predict 0 or 1, with 1 always representing the dative in every data configuration. A probe was trained for each model (*BERT*, *mBERT*, *LLäMmlein 1B*, *Word2Vec*), and for each model's layers, for every different data configuration (*Acc* $\sim$ *Dat*, *Nom* $\sim$ *Dat*, *Nom/Dat* $\sim$ *Acc*, *Nom/Acc* $\sim$ *Dat*, *Nom* $\sim$ *Acc/Dat*), and for every control dataset (*Random*, *Same Token Amount*, *Noun Only*, *Pronoun Only*). This meant that for *BERT* we trained 13 layers $\times$ 5 data configurations $\times$ 4 datasets = 260 classifiers; for *mBERT* we trained 13 layers $\times$ 5 data configurations $\times$ 4 datasets = 260 classifiers; for *LLäMmlein 1B* we trained 23 layers $\times$ 5 data configurations $\times$ 4 datasets = 460 classifiers; and for *Word2Vec* we trained 1 layer $\times$ 5 data configurations $\times$ 4 datasets = 20 classifiers. This totals 1000 classifiers.

4.4 Results

The error rates on the dative test sets are given in Figures 1 and 2. For each model, only the layer which achieved the most accurate classifications on the validation set in the *Nom/Acc* $\sim$ *Dat* configuration is shown; for *BERT* this is layer 11, for *mBERT* this is layer 10, and for *LLäMmlein 1B* this is layer 21[5]. Across the board, all the models performed well. Although German exhibits some case syncretism (where the same case morpheme is used for different cases, e.g. *euch* in German is a second person plural pronoun for accusative and dative), the majority of the case morphemes are distinct from one another (e.g. *ich/mich/mir* is first person singular for nominative, accusative, and dative respectively). It is likely that the classifier has trained on the existence of these specific case tokens rather than actual case representation[6]. Due

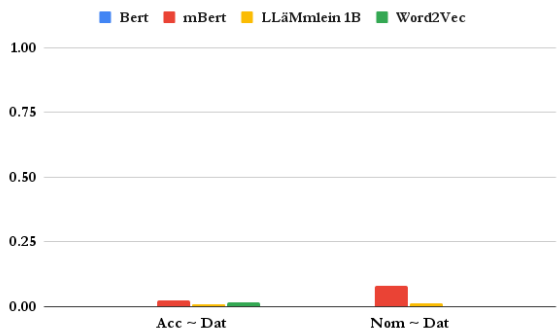

Figure 1: Error rate for the test set. Results show when the classifier trained on two cases.

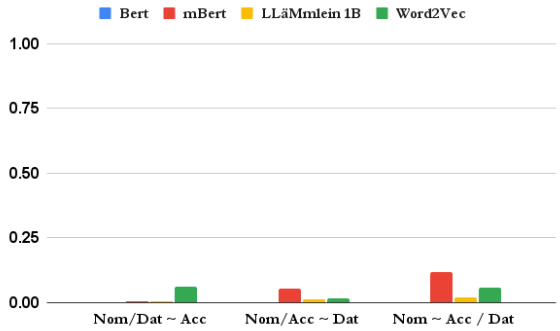

Figure 2: Error rate for the test set. Results show when the classifier trained on three cases.

[5]To see the full results, including the error ratings on the validation set and the control datasets, consult the appendix.

[6]An anonymous reviewer suggested that due to the embeddings coming from deep layers within all the models, specific token information would not be available for the classifiers to train on, and therefore an alternative explanation is needed for why they perform so well here. To play hypothetical, it could be that indeed a language model has encoded inherent cases, and the classifiers are picking up on such. Given the second experiment's results, that when you remove explicit case information, datives get analysed as a structural nominative, what could be argued is that the German language models have overgeneralised structural case assignment, in other words without a specific case signal the models default to a structural case representation as a strategy instead of what we claim to be overfitting to surface patterns. This is a testable claim; in child learners of German it has been noted that children

to this, we put these results to the side, and undertook a caseless version of this experiment, where we removed the case morphology.

5 Experiment 2: Caseless Representation

5.1 Caseless Datasets

We assume that the language models can handle caseless arguments given that proper nouns in German do not exhibit overt case marking. To create a caseless representation of the arguments, we removed case information and replaced it with a bundle of features. For pronouns, we had a bracket notation which included information such as gender (masculine, feminine, neuter), person (first, second, third), number (singular, plural), and pronoun type (personal, relative, interrogative, demonstrative). For instance, *wir/uns* (we/us - nominative and accusative/dative respectively) would become: `[Person=1Plural|Personalpronomen]`

For noun phrases, we lemmatised the noun to remove case declensions, and replaced the article with a bundle of features. The article features included article type (indefinite, definite, demonstrative, possessive), if it was a possessive article it included gender, person, number, and in general if the articles were plural that information was shown. However, although articles do agree in gender with the noun, because the noun was still represented, we did not include this information as it is still recoverable. For an example, *Seine Visionen* 'his visions' would become: `[Numerus=Plural|Geschlecht=Maskulin|Person=3Singular|Possessivartikel] Vision`

We achieved this by using SpaCy's morphological parser (Honnibal et al., 2020). If the predicted case from the parser did not match the case given, then it was done by hand.

5.2 Controls, Classifiers, and Models.

The control datasets *same token amount*, *noun only*, and *pronoun only* were kept the same, bar now having a caseless argument. The classifier architecture, the classifier training, and models used were all kept the same from the first experiment. However, with *Word2Vec*, given the lemmatising and the insertion of new words with the feature bundles, some of the words were not represented within its dictionary. Therefore, 492 datives (46 from the test set), 520 accusatives, and 511 nominatives were dropped from the *Word2Vec* datasets.

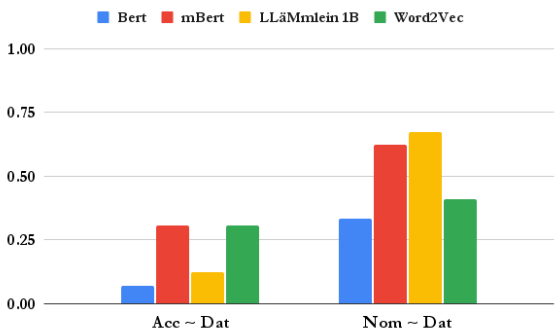

Figure 3: Error rates for the caseless test set. Results show when the classifier trained on two cases.

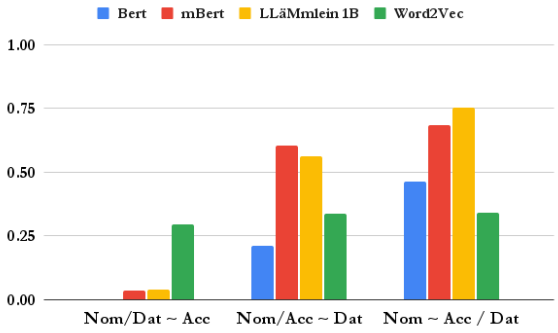

Figure 4: Error rates for the caseless test set. Results show when the classifier trained on three cases.

5.3 Results

The error rates on the caseless dative test sets are given in Figures 3 and 4. The layers shown are 11 for *BERT*, 10 for *mBERT*, and 21 for *LLäMmlein 1B*. For the data configurations that used only two cases (Figure 3), each model had a lower error rate with the *Acc* $\sim$ *Dat* probes, than the *Nom* $\sim$ *Dat* probes. Only *BERT* did better than the *Word2Vec* baseline in both configurations, with *mBERT* and *LLäMmlein 1B* performing worse than *Word2Vec* with the *Nom* $\sim$ *Dat* configuration. Concerning the control datasets, the same general trend of the *Acc* $\sim$ *Dat* probes performing better than the *Nom* $\sim$ *Dat* probes was seen, bar for *Word2Vec* in the *pronoun only* condition where both data configurations had the same error rate.

will overgeneralise the structural case system (nominative and accusative) rather than the dative, however in contrast Icelandic children overgeneralise the dative (see section 1.2 of Nowenstein (2023) for an overview). Given this, if it's true that the structural case system has been overgeneralised for the German language models, then we would expect the inherent case systems to overgeneralise for the Icelandic language models. If instead it's about surface patterns then we should expect the same results for both languages - although some controls would be needed for Icelandic given dative subjects can naturally occur.

For the data configurations that used three cases (Figure 4), all the models showed the same trends, where the data configuration with the lowest error rate was *Nom/Dat* ∼ *Acc*, and the data configuration with the highest error rate was *Nom* ∼ *Acc/Dat*. Concerning the control datasets, the same general trends were seen excluding the following exceptions: *mBERT* in the *pronoun only* condition had the same error rates for *Nom/Acc* ∼ *Dat* and *Nom* ∼ *Acc/Dat*; *LLäMmlein 1B* in the *pronoun only* condition had the highest error rate be *Nom/Acc* ∼ *Dat*; *Word2Vec* in the *noun only* and *pronoun only* condition had the highest error rate be *Nom/Acc* ∼ *Dat*.

6 Discussion

The NLP Take: *BERT* better distinguishes German inherent and structural cases when compared to *mBERT* and *LLäMmlein 1B*. This indicates that smaller unilingual models encode a more consistent inherent case representation. For why a unilingual model represents case better than a multilingual model, it could be that a consistent inherent case generalisation across languages is harder to learn. Consider the dative; a cross-linguistic definition can be described as indicating (i) low-transitivity and (ii) the affectedness/volition of an argument (Næss, 2008); this provides a good description of the German instances, which are used on experiencers, benefactives, and recipients. However, many languages which have dative utilise other cases for the supposed dative function; e.g. in Kalkatungu (an Australian indigenous language [Blake 1983]) dative represents benefactives, but allative represents recipients (Blake, 2001). The syntactic position of inherent datives can also differ between languages; e.g. German inherent datives occur on objects, whilst Icelandic can assign dative to subjects (Jónsson, 2003). Thus, dative case is not homogenous crosslinguistically - making it harder for multilingual models to generalise. Of course, as shown in Papadimitriou et al. (2021) multilingual models can generalise case information crosslingually, but crucially these were structural cases being probed, not inherent. Inherent cases require more information than just structural sequencing; they require language-specific lexical knowledge, by knowing which specific lexical items take an inherent case. This specific knowledge is truly to do with lexical items rather than more general semantic roles, given that there are German verbs which have the same semantic roles, but do not take inherent cases, e.g. *gratulieren* and *beglückwünschen* both mean 'to congratulate', but the former takes a dative object, and the latter accusative (McFadden, 2008); meaning a model cannot generalise inherent cases to semantic roles. Therefore, it may be that unilingual models better represent language-specific lexical requirements, compared to multilingual models.

For why a smaller language model represents case better than a larger model, this could be because the larger model has overfit more to surface syntactic patterns. A strategy to distinguish between structural accusative and inherent dative (regardless of if the dative is passivised) in the caseless condition is to train on which verbs are present in the larger context, as transitive verbs that assign an inherent dative and those that default to structural case (and therefore get accusative objects) are in a complementary distribution. Of course, given that the test datives used verbs which were distinct from the training dative verbs, for the classifier to achieve a low error rate on test set, the benefactive/malefactive verbs have to be encoded as being the same as the stimulus-experiencer verbs and ditransitives case-wise. In Figure 3 we can see that both *BERT* and *LLäMmlein 1B* are using such information given that they are below the static baseline in *Acc* ∼ *Dat*, where such context information is not available.

When we come to the *Nom* ∼ *Dat* classifiers however, verb information cannot be utilised because nominatives occur with predicates which select an inherent dative. For both *BERT* and *LläMmlein 1B* we now see a higher error rate than the *Acc* ∼ *Dat* condition, with *LläMmlein 1B* performing worse than the static baseline. This is in contrast with low error rates on the *Nom* ∼ *Dat* validation set (0 and 0.008 for *BERT* and *LläMmlein 1B* respectively). However, this can be explained if the language models are utilising a structural strategy combined with lexical knowledge to distinguish arguments. Consider that a way to split structural nominative, structural accusative, and inherent dative in German, is by encoding nominatives as what occurs on subject, datives as what occurs on objects when specific group of verbs X occurs, and accusatives as what occurs on objects

with all other verbs[7][8]. This strategy explains why for both *BERT* and *LläMmlein 1B* have in all conditions (including the three case configurations) low error rates on the validation sets. However, when the dative argument gets passivised, it gets realised as a subject and therefore patterns with the nominative. This is interesting given that there are distinct patterns between passivised inherent datives and passivised arguments which receive nominative; for instance German does not allow subject-verb agreement with passivised datives and instead opts for default agreement. Thus, the difference in performance from *BERT* and *LläMmlein 1B* is that *LläMmlein 1B* has overfit to surface patterns concerning passivised arguments, in that passivised inherent datives are passivised subjects, when *BERT* has less done so. However, this is not to give *BERT* too much credit given that in the *Nom* $\sim$ *Acc/Dat* condition it is worse than the static baseline and is basically at random chance, and therefore has also overfit, just not to the same degree.

This structural and lexical strategy of distinguishing cases can also explain the pattern for all the models in the three case conditions where they performed best in the *Nom/Dat* $\sim$ *Acc* condition and worst in *Nom* $\sim$ *Acc/Dat*, because in *Nom/Dat* $\sim$ *Acc* you can use the verb information and structural positioning to distinguish the arguments, in *Nom/Acc* $\sim$ *Dat* only the verb information, and in *Nom* $\sim$ *Acc/Dat* no such information. However, there is an interplay with the different distributions of pronouns and nouns per case as seen by the static baseline being better than random chance and due to the control datasets of *pronoun only* and *noun only* changing the error rates for *Nom/Acc* $\sim$ *Dat* and *Nom* $\sim$ *Acc/Dat* with *LläMmlein 1B* and *mBERT*. However, crucially, with the control datasets, the *Nom/Dat* $\sim$ *Acc* condition was always the best performing configuration. So although there is some argument type distribution per case, this is not the main information the models use to distinguish the case. Of course, to confirm this structural and lexical encoding of inherent and structural case information, a follow-up experiment should conduct a causal intervention experiment (Geiger et al., 2021).

The Linguistics Take: A recent minimalist theory in how to account for the distinct behaviours of inherent and structural case, has been to synthesise two competing case accounts (McFadden, 2024). We refer the reader to the paper for a full lay out of the synthesis, but in simple terms inherent cases have their own syntactic head (i.e. in the syntactic domain), whilst structural cases do not have their own syntactic head but instead are the morphemes that are inserted on bare DPs post-spellout (i.e. in the phonological domain). A prediction of such is that when encoding case information, inherent and structural cases should be distinctly represented because they are different phenomena.

Although the field is debating where language models should fall in theory creation (e.g. Chomsky et al. (2023) versus Piantadosi (2023)), we can agree that language models produce grammatical text and *seemingly* do not over-generate. Given this, if inherent and structural cases have such distinct behaviours/distributions, we may expect a language model to utilise this information when representing cased argument, and thus to distinctly represent inherent and structural cases. Our results show that this is not true, as the three-way case configuration with a distinction between inherent and structural cases did not have the most accurate probe. Although language models cannot prove any linguistic theory, they can at least show us if distributions are prevalent enough to learn from and if certain information is needed for a grammatical producing system. Here, the distribution between structural accusative and inherent dative is distinct and thus beneficial to encode, but the same cannot be said for the inherent dative and structural nominative. What we may want to take from this is that although structural and inherent cases have distinct behaviours they are not distinct enough to warrant being encoded by language models, and therefore we may not want to theoretically include such a strong distinction between the two case types.

To answer our questions: are arguments which receive inherent cases versus those that receive structural cases distinguished from one another? No, although the three cases can be distinguished from one another, there does not seem to be a specific encoding to distinguish inherent status from structural status. Do models default to a structural

[7]The hard part is figuring out what is the subject, especially given that German can have variable word order. However, language models have been shown to accurately predict subject and objects when word order information is removed and instead rely on lexical cues (Mahowald et al., 2023).

[8]To account for the ditransitive, you would need to add an extra stipulation that accusative occurs with all other verbs/objects which are left over with no case. This does raise an interesting question of how models deal with indirective or secundative marking on ditransitives - see Haspelmath (2005).

representation of arguments instead of a case representation? Yes, as passivised inherent datives are interpreted as a passive subject despite not having full subjecthood status. Overall, these results fall in line with Czarnowska (2022) in that models overrely on word patterns and lexical semantics despite a language's morphology, and raises doubts of how well language models generalise case information.

7 Conclusion

Previous investigations into how language models represent case has mainly focused on structural cases, with inherent cases being overlooked. However, inherent cases have distinct behaviours and require different generalisations to accurately represent them. By utilising linguistic probing, we found that when case tokens are removed, language models can distinguish between inherent dative and structural accusative, regardless of argument position, due to verb information. However, language models struggle to distinguish between structural nominative and inherent dative when the dative argument is passivised due to over-relying on surface patterns. This adds to a larger body of work which shows that language models do not utilise morphology information, but instead use syntactic sequences and word items to disambiguate language.

8 Limitations and Ethical Considerations

There are three limitations to this project: (i) the difficulty of expanding into other languages, (ii) the datasets coming from two different sources, and (iii) a lack of models with differing tokeniser strategies. In response to the first limitation, there are other languages which also have inherent cases (e.g. Hungarian, Polish, Icelandic). However, inherent cases can be particularly language-specific, with other cases than dative being used, or occurring on subjects instead of objects. It is also not guaranteed that the inherent case can be passivised (consider the German stimulus-experiencer verbs), and instead a different test construction would have to be used (e.g. some other A-movement construction). Given this, considerable work would need to be undertaken to expand to a different language. Not only this, but considering the frequency problem with these cases, it would be likely that anyone attempting a different language would face the same difficulty of UD not having a significant size of the specific cases/constructions needed. As a means to address how other languages may have different case representations, we included a multilingual model *mBERT* in our investigation which does include other languages with inherent cases. There is also the other side of this limitation which cannot be solved with multilingual models, in that there are many languages with case systems which do not have language models. To best mitigate this problem, we have made sure that our paper specifically mentions we work on German, and have also not overstated our claims to apply to all languages. Concerning the second limitation of how the majority of our dative training, validation, and test set comes from a different source to the accusative and nominative dataset; we dealt with this by including control datasets which tested for if different dataset characteristics played a role in the representation. For the third limitation concerning lack of differing tokeniser strategies, we do test sub-word tokenisation versus word tokenising (e.g. *Word2Vec*), however, given that *Word2Vec* creates a static embedding it is unclear how much of an impact the tokenising has had versus the static nature of the embedding. Simply, the experiment can be repeated, however, instead of changing model size and architecture, you would test along different tokenising strategies.

On the ethical considerations, it is always good to consider the broader ethical implications of using language models in the first place. Concerning this, we used a minimal set of models which still represented different conditions as a means to reduce computing power and resources, and thus carbon emissions. We also used models where the original authors were transparent in their architecture and training, and can be used without having to give user details to the companies/authors of said models. There is, as stated in the limitations, the ethical problem of only working on high-resource languages (e.g. German), and how the work is not reproducible on low-resource languages, which are also commonly minoritised languages. Unfortunately, this is a broader problem in computational linguistics, and how we have attempted mitigate this problem is to clearly state we are working on German and to not overstate our claims.

Acknowledgements

This work was supported by the UKRI Centre for Doctoral Training in Natural Language Processing, and Iona Carslaw was funded by the UKRI (refer-

ence: EP/S022481/1). We would like to thank Pia Pratscher and Theresia Michel for their work on the dataset and German judgements. We would also like to thank the three anonymous reviewers, the participants at the LELPGC25 at the University of Edinburgh (where the first experiment of this paper was presented), Ella Markham for reading an initial draft, and Frank Keller for insightful comments.

References

Matt Baker. 2015. *Case: Its Principles and its Parameters*. Cambridge University Press.

Jean Berko. 1958. The child's learning of english morphology. *Word*, 14(2-3):150–177.

Barry Blake. 1983. Structure and word order in kalkatungu: The anatomy of a flat language. *Australian Journal of Linguistics*, 3(2):143–175.

Barry Blake. 2001. *Case*, 2nd edition edition. Cambridge University Press.

Branden Chan, Timo Möller, Malte Pietsch, and Tanay Soni. 2019. German bert. Accessed on 12th January 2026.

Noam Chomsky, Ian Roberts, and Jeffrey Watumull. 2023. The false promise of chatgpt. *The New York Times*.

Simone Conia and Roberto Navigli. 2022. Probing for predicate argument structures in pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4622–4632.

Paula Czarnowska. 2022. *Morphological competence in neural natural langauge processing*. Ph.D. thesis, The University of Camrbidge.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Robert Dixon. 1994. *Ergativity*. Cambridge University Press.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. 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 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. *Advances in neural information processing systems*, 34:9574–9586.

Coleman Haley. 2020. This is a BERT. now there are several of them. can they generalize to novel words? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 333–341, Online. Association for Computational Linguistics.

Martin Haspelmath. 2005. Argument marking in ditransitive alignment types. *Linguistic discovery*, 3(1):1–21.

Martin Haspelmath and Luisa Baumann. 2013. German (standard). In Iren Hartmann, Martin Haspelmath, and Bradley Taylor, editors, *Valency Patterns Leipzig*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Mete Ismayilzada, Defne Circi, Jonne Sälevä, Hale Sirin, Abdullatif Köksal, Bhuwan Dhingra, Antoine Bosselut, Duygu Ataman, and Lonneke Van Der Plas. 2025. Evaluating morphological compositional generalization in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1270–1305, Albuquerque, New Mexico. Association for Computational Linguistics.

Jóhannes Gísli Jónsson. 2003. Not so quirky: On subject case in icelandic. *New perspectives on case theory*, 156:127–163.

Michael Kamerath and Aniello De Santo. 2025. Do llms disambiguate italian relative clause attatchment. In *Society for Computation in Linguistics*, volume 8, page 33.

Mary Kennedy. 2025. Evidence of hierarchically-complex syntactic structure within bert's word representation. In *Society for Computation in Linguistics*, volume 8, page 18.

Eun-Kyoung Rosa Lee, Sathvik Nair, and Naomi Feldman. 2024. A psycholinguistic evaluation of language models' sensitivity to argument roles. *arXiv preprint arXiv:2410.16139*.

Kyle Mahowald, Evgeniia Diachek, Edward Gibson, Evelina Fedorenko, and Richard Futrell. 2023. Grammatical cues to subjecthood are redundant in a majority of simple clauses across languages. *Cognition*, 241:105543.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. *Preprint*, arXiv:1808.09031.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Clausia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

Thomas McFadden. 2008. German inherent datives and argument structure. In *Datives and other cases: Between argument structure and event structure*, pages 49–77. John Benjamins Publishing Company.

Thomas McFadden. 2024. A synthesis for the structural/inherent case distinction and its comparative and diarchronic consequences. In Christina Sevdali, Dionysios Mertyris, and Elena Anagnostopoulou, editors, *The Place of Case in Grammar*. Oxford Academic.

Andreas Müller. 2015. Analyse von Wort-Vektoren deutscher Textkorpora.

Iris Edda Nowenstein. 2023. *Building yourself a variable case system: The acquisition of Icelandic datives*. Ph.D. thesis.

Åshild Næss. 2008. Varieties of datives. In *The Oxford Handbook of Case*. Oxford Academic.

Satoru Ozaki, Rajesh Bhatt, and Brian W Dillon. 2025. A lstm language model learns hindi-urdu case-agreement interactions, and has a linear encoding of case. In *Proceedings of the Society for Computation in Linguistics 2025*, pages 64–73.

Isabel Papadimitriou, Ethan A Chi, Richard Futrell, and Kyle Mahowald. 2021. Deep subjecthood: Higher-order grammatical features in multilingual bert. *arXiv preprint arXiv:2101.11043*.

Isabel Papadimitriou, Richard Futrell, and Kyle Mahowald. 2022. When classifying grammatical role, bert doesn't care about word order... except when it matters. *arXiv preprint arXiv:2203.06204*.

Jan Pfister, Julia Wunderle, and Andreas Hotho. 2025. LLäMmlein: Transparent, compact and competitive German-only language models from scratch. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2227–2246, Vienna, Austria. Association for Computational Linguistics.

Steven T Piantadosi. 2023. Modern language models refute chomsky's approach to language. *From fieldwork to linguistic theory: A tribute to Dan Everett*, 15:353–414.

Roland Schäfer. 2015. Processing and querying large web corpora with the COW14 architecture. In *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, Lancaster. UCREL, IDS.

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 486–493, Istanbul, Turkey. European Language Resources Association (ELRA).

Tsedeniya Kinfe Temesgen, Marion Di Marco, and Alexander Fraser. 2025. Extracting linguistic information from large language models: Syntactic relations and derivational knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27210–27226, Suzhou, China. Association for Computational Linguistics.

Don Tuggener. 2016. Incremental coreference resolution for german. Master's thesis, The University of Zurich.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schuetze, Kemal Oflazer, and David Mortensen. 2023. Counting the bugs in ChatGPT's wugs: A multilingual investigation into the morphological capabilities of a large language model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6508–6524, Singapore. Association for Computational Linguistics.

Marion Weller-Di Marco and Alexander Fraser. 2024. Analyzing the understanding of morphologically complex words in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1009–1020, Torino, Italia. ELRA and ICCL.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do rnn language models learn about filler-gap dependencies? *arXiv preprint arXiv:1809.00042*.

Michael Wilson, Jackson Petty, and Robert Frank. 2023. How abstract is linguistic generalization in large language models? experiments with argument structure. *Transactions of the Association for Computational Linguistics*, 11:1377–1395.

A Full Bert Results

				layers												
				0	1	2	3	4	5	6	7	8	9	10	11	12
Random	Cased	Acc0_Dat1	Val	0.015	0.0175	0.0125	0.01	0.0125	0.01	0.01	0.005	0.0025	0	0	0	0
			Test	0.0204918	0.0204918	0.01639344	0.01639344	0.00409836	0.00409836	0.01229508	0	0	0	0	0	0
		Nom0_Dat1	Val	0.0175	0.0075	0.01	0.0075	0.0025	0.0025	0	0	0	0	0	0	0
			Test	0.09016393	0.07786885	0.08196721	0.06967213	0.04098361	0.04098361	0.0204918	0.0204918	0	0.00409836	0	0	0
		Nom1_Acc0_Dat1	Val	0.12166667	0.11	0.07333333	0.075	0.06333333	0.06166667	0.05	0.02833333	0.01	0.00833333	0.01166667	0.01166667	0.015
			Test	0.02459016	0.0204918	0.0204918	0.01639344	0	0	0	0	0	0	0	0	0
		Nom0_Acc0_Dat1	Val	0.02166667	0.01166667	0.01333333	0.01333333	0.005	0.00666667	0.00666667	0.00333333	0.00166667	0	0.00166667	0	0.00166667
			Test	0.05737705	0.05327869	0.03688525	0.03688525	0.03278689	0.0204918	0.02459016	0.00409836	0	0.00409836	0.00409836	0	0
		Nom0_Acc1_Dat1	Val	0.12166667	0.11333333	0.08166667	0.06333333	0.05166667	0.04333333	0.035	0.02333333	0.00666667	0.00833333	0.005	0.00333333	0.005
			Test	0.1352459	0.0942623	0.12704918	0.09016393	0.07377049	0.04508197	0.04098361	0.03278689	0.00409836	0.00409836	0	0	0.00819672
		Acc0_Dat1	Val	0.1875	0.1325	0.1325	0.095	0.1025	0.1025	0.09	0.0775	0.06	0.0675	0.0625	0.035	0.0375
			Test	0.14344262	0.22131148	0.20491803	0.26639344	0.2295082	0.22540984	0.19262295	0.13934426	0.1557377	0.24590164	0.19672131	0.06967213	0.04918033
		Nom0_Dat1	Val	0.2275	0.19	0.1825	0.155	0.135	0.1275	0.11	0.0825	0.0575	0.06	0.0625	0.0375	0.0275
			Test	0.62295082	0.46721311	0.46721311	0.4795082	0.51229508	0.52459016	0.5204918	0.51229508	0.43032787	0.59836066	0.63114754	0.33196721	0.19672131
	Caseless	Nom1_Acc0_Dat1	Val	0.20166667	0.2	0.19333333	0.16166667	0.14	0.13666667	0.14	0.115	0.09833333	0.10833333	0.11166667	0.05333333	0.045
			Test	0.07377049	0.06967213	0.07377049	0.06967213	0.05737705	0.06147541	0.07377049	0.01639344	0.0204918	0.02868852	0.0204918	0	0
		Nom0_Acc0_Dat1	Val	0.235	0.185	0.16666667	0.15166667	0.14333333	0.11833333	0.125	0.10333333	0.06166667	0.08666667	0.07166667	0.03833333	0.04
			Test	0.46311475	0.43852459	0.4057377	0.44262295	0.44672131	0.4057377	0.43852459	0.40163934	0.4057377	0.65163934	0.60655738	0.21311475	0.19262295
		Nom0_Acc1_Dat1	Val	0.21166667	0.19666667	0.19333333	0.17	0.16833333	0.16333333	0.165	0.13	0.09833333	0.09	0.06666667	0.07	0.05333333
			Test	0.72540984	0.58606557	0.5942623	0.6352459	0.68852459	0.65163934	0.68442623	0.66393443	0.68852459	0.65163934	0.67622951	0.46311475	0.35245902
		Acc0_Dat1	Val	0.01712329	0.02054795	0.01369863	0.01027397	0.01712329	0.01369863	0.01369863	0.00684932	0.00684932	0.00342466	0.00342466	0.00342466	0.00342466
			Test	0.03688525	0.01639344	0.01229508	0.00819672	0.01229508	0.00409836	0.01229508	0	0	0	0	0	0
		Nom0_Dat1	Val	0.02054795	0.01027397	0.01027397	0.00684932	0.01027397	0.01027397	0.00342466	0.00342466	0	0	0	0	0.00342466
			Test	0.08606557	0.05737705	0.05327869	0.03688525	0.02868852	0.02868852	0.03278689	0.0204918	0.00409836	0.00409836	0	0	0
	Cased	Nom1_Acc0_Dat1	Val	0.10958904	0.11415525	0.08675799	0.07534247	0.06392694	0.05707763	0.05479452	0.02283105	0.00913242	0.00684932	0.00684932	0.00456621	0.00228311
			Test	0.01229508	0.01229508	0.00409836	0	0	0	0	0	0	0	0	0	0
		Nom0_Acc0_Dat1	Val	0.02283105	0.02054795	0.01598174	0.01826484	0.00684932	0.01141553	0.01369863	0.00684932	0.00456621	0.00684932	0.00456621	0.00684932	
			Test	0.03688525	0.06557377	0.05327869	0.04918033	0.0204918	0.0204918	0.03278689	0.00409836	0.00409836	0.00409836	0	0.00409836	0
		Nom0_Acc1_Dat1	Val	0.11187215	0.10958904	0.08219178	0.06849315	0.05936073	0.06164384	0.05022831	0.02283105	0.00228311	0.00456621	0	0	0.00228311
			Test	0.10655738	0.12704918	0.15983607	0.1147541	0.09016393	0.10245902	0.09836066	0.04918033	0.00409836	0	0	0	0.00409836
Same Token Amount		Acc0_Dat1	Val	0.20547945	0.16780822	0.17808219	0.13356164	0.12671233	0.15753425	0.11643836	0.09589041	0.08219178	0.07191781	0.04452055	0.03082192	0.03767123
			Test	0.19262295	0.20901639	0.21721311	0.20491803	0.18852459	0.18032787	0.18852459	0.13934426	0.15983607	0.26639344	0.20491803	0.03278689	0.05327869
		Nom0_Dat1	Val	0.20547945	0.19178082	0.17123288	0.15068493	0.14726027	0.11986301	0.12328767	0.06849315	0.04109589	0.05136986	0.05479452	0.0239726	0.04452055
			Test	0.63114754	0.55327869	0.52868852	0.58606557	0.59016393	0.54508197	0.5204918	0.52868852	0.54098361	0.57377049	0.62295082	0.32786885	0.30737705
	Caseless	Nom1_Acc0_Dat1	Val	0.21004566	0.18949772	0.19863014	0.16438356	0.15296804	0.14840183	0.14840183	0.12328767	0.11187215	0.10958904	0.07990868	0.05022831	0.0456621
			Test	0.08196721	0.07786885	0.06557377	0.06557377	0.05737705	0.04918033	0.05327869	0.02868852	0.0204918	0.04918033	0.0204918	0.00409836	0
		Nom0_Acc0_Dat1	Val	0.23972603	0.22146119	0.20547945	0.17808219	0.14155251	0.14155251	0.15068493	0.1347032	0.05936073	0.0913242	0.10045662	0.04109589	0.03196347
			Test	0.47540984	0.43442623	0.44262295	0.4795082	0.52868852	0.43852459	0.4795082	0.42213115	0.45491803	0.66393443	0.61065574	0.26229508	0.19262295
		Nom0_Acc1_Dat1	Val	0.21232877	0.20319635	0.21689498	0.1826484	0.16666667	0.15068493	0.1369863	0.12328767	0.07762557	0.08447489	0.10045662	0.0913242	0.06164384
			Test	0.74180328	0.62704918	0.67213115	0.64344262	0.67622951	0.68852459	0.68852459	0.65163934	0.72540984	0.68442623	0.67622951	0.65983607	0.39754098
		Acc0_Dat1	Val	0.02970297	0.01980198	0.0049505	0.01485149	0.0049505	0.0049505	0.0049505	0.0049505	0.0049505	0.0049505	0	0	0
			Test	0.015625	0.03125	0.01041667	0.01041667	0.01041667	0.00520833	0.00520833	0	0	0	0	0	0
		Nom0_Dat1	Val	0.01485149	0.01485149	0.00990099	0.00990099	0	0	0	0	0	0	0	0	0
			Test	0.06770833	0.08854167	0.08333333	0.06770833	0.04166667	0.03645833	0.03645833	0.02083333	0	0	0	0	0.00520833
	Cased	Nom1_Acc0_Dat1	Val	0.16831683	0.15511551	0.1320132	0.11881188	0.08250825	0.0660066	0.05610561	0.0330033	0.01320132	0.01650165	0.01320132	0.00660066	0.00990099
			Test	0.03645833	0.046875	0.03645833	0.02604167	0.01041667	0	0	0	0	0	0	0	0
		Nom0_Acc0_Dat1	Val	0.02970297	0.03630363	0.01650165	0.01320132	0.00990099	0.00990099	0.00660066	0.00660066	0.00660066	0.00660066	0.0033003	0.0033003	0
			Test	0.04166667	0.046875	0.03645833	0.02083333	0.03125	0.015625	0.02604167	0.015625	0	0	0	0	0
		Nom0_Acc1_Dat1	Val	0.13861386	0.11881188	0.0990099	0.07260726	0.04620462	0.0330033	0.02970297	0.01980198	0.00990099	0.00660066	0.00660066	0.00660066	0.00660066
Noun Only			Test	0.1875	0.16666667	0.18229167	0.078125	0.05729167	0.0625	0.05208333	0.015625	0.00520833	0	0	0.00520833	0.015625
		Acc0_Dat1	Val	0.16831683	0.13366337	0.13366337	0.11386139	0.1039604	0.11386139	0.06435644	0.06930693	0.02970297	0.05445545	0.02970297	0.01485149	0.02970297
			Test	0.109375	0.17708333	0.21875	0.24479167	0.21875	0.17708333	0.22395833	0.1875	0.203125	0.21875	0.1875	0.02604167	0.03125
		Nom0_Dat1	Val	0.19306931	0.18316832	0.17326733	0.15841584	0.12871287	0.11881188	0.08910891	0.03960396	0.01485149	0.02475248	0.03960396	0.03465347	0.03465347
			Test	0.69270833	0.578125	0.609375	0.609375	0.67708333	0.57291667	0.609375	0.63020833	0.55208333	0.67708333	0.68229167	0.375	0.33854167
	Caseless	Nom1_Acc0_Dat1	Val	0.22442244	0.21122112	0.17821782	0.15181518	0.13861386	0.13861386	0.11221122	0.09570957	0.08250825	0.07920792	0.06270627	0.0330033	0.04950495
			Test	0.05729167	0.0625	0.05729167	0.04166667	0.046875	0.03125	0.04166667	0.01041667	0	0	0.00520833	0	0
		Nom0_Acc0_Dat1	Val	0.19471947	0.18151815	0.18481848	0.15181518	0.11551155	0.1320132	0.07920792	0.05280528	0.04620462	0.04290429	0.02970297	0.02970297	0.0330033
			Test	0.515625	0.421875	0.41666667	0.48958333	0.67708333	0.44791667	0.5625	0.47916667	0.390625	0.56770833	0.55729167	0.25	0.234375
		Nom0_Acc1_Dat1	Val	0.26732673	0.22772277	0.20792079	0.18151815	0.15841584	0.15181518	0.11881188	0.05940594	0.04620462	0.04950495	0.04290429	0.02310231	0.03630363
			Test	0.71354167	0.72395833	0.71354167	0.77083333	0.8125	0.80729167	0.79166667	0.82291667	0.81770833	0.73958333	0.83854167	0.671875	0.41666667
		Acc0_Dat1	Val	0	0	0	0	0	0	0	0	0	0	0	0	0
			Test	0	0	0	0	0	0	0	0	0	0	0	0	0.01923076923
		Nom0_Dat1	Val	0	0	0	0	0	0	0	0	0	0	0	0	0
			Test	0	0	0	0	0	0.01923076923	0.01923076923	0.03846153846	0.03846153846	0.01923076923	0	0	0.01923076923
	Cased	Nom1_Acc0_Dat1	Val	0.1666666667	0.1166666667	0.08333333	0.05	0.06666667	0.03333333	0.05	0.03333333	0.01666666667	0.01666666667	0	0	0.01666666667
			Test	0	0	0	0	0	0	0	0	0	0	0	0	0
		Nom0_Acc0_Dat1	Val	0	0	0	0	0	0	0	0	0	0	0	0	0
			Test	0	0	0	0	0.01923076923	0.01923076923	0	0.01923076923	0	0.01923076923	0	0	0.01923076923
		Nom0_Acc1_Dat1	Val	0.2	0.2	0.1	0.08333333	0.08333333	0.1	0.06666667	0.01666666667	0	0	0	0	0
Pronoun Only			Test	0.01923076923	0.03846153846	0.05769230769	0.07692308	0.07692308	0.01923076923	0.05769230769	0.01923076923	0.01923076923	0.01923076923	0.01923076923	0	0.01923076923
		Acc0_Dat1	Val	0.075	0.1	0.075	0.125	0.1	0.15	0.075	0.125	0.075	0.125	0.1	0.15	0.1
			Test	0.23076923	0.26923077	0.25	0.23076923	0.23076923	0.25	0.17307692	0.15384615	0.09615385	0.28846154	0.36538462	0.17307692	0.23076923
		Nom0_Dat1	Val	0.2	0.075	0.075	0.05	0.05	0.05	0.05	0.075	0.05	0.1	0.075	0.1	0.05
			Test	0.40384615	0.40384615	0.48076923	0.40384615	0.36538462	0.42307692	0.36538462	0.42307692	0.46153846	0.48076923	0.55769231	0.44230769	0.28846154
	Caseless	Nom1_Acc0_Dat1	Val	0.25	0.3	0.2	0.26666667	0.26666667	0.21666667	0.26666667	0.28333333	0.16666667	0.16666667	0.2	0.16666667	0.23333333
			Test	0.19230769	0.21153846	0.25	0.23076923	0.21153846	0.19230769	0.15384615	0.15384615	0.07692308	0.09615385	0.11538462	0.09615385	0.19230769
		Nom0_Acc0_Dat1	Val	0.15	0.08333333	0.11666667	0.05	0.08333333	0.1	0.06666667	0.06666667	0.03333333	0.06666667	0.08333333	0.05	0.03333333
			Test	0.34615385	0.36538462	0.32692308	0.40384615	0.36538462	0.32692308	0.34615385	0.25	0.34615385	0.42307692	0.59615385	0.19230769	0.34615385
		Nom0_Acc1_Dat1	Val	0.4	0.2833333		0.25	0.23333333	0.25	0.15	0.2	0.25	0.21666667	0.2	0.21666667	0.2	0.16666667
			Test	0.61538462	0.51923077	0.51923077	0.44230769	0.40384615	0.48076923	0.48076923	0.51923077	0.48076923	0.48076923	0.48076923	0.61538462	0.42307692
		Acc_random	Val	-												
			Test	0.47131148	0.6147541	0.44262295	0.57377049	0.50819672	0.41393443	0.4795082	0.38934426	0.31557377	0.56557377	0.38934426	0.60245902	0.46721311
	Cased	Nom random	Val	-												
			Test	0.55737705	0.5204918	0.62704918	0.56147541	0.52459016	0.40983607	0.55327869	0.54918033	0.50819672	0.45081967	0.74590164	0.44672131	0.45491803
		All Random	Val	-												
Control			Test	0.60655738	0.60655738	0.42213115	0.54508197	0.3852459	0.48770492	0.47131148	0.55327869	0.49180328	0.44262295	0.68032787	0.26639344	0.41393443
		Acc_random	Val	-												
			Test	0.38114754	0.61885246	0.67213115	0.63934426	0.39754098	0.38934426	0.42213115	0.4795082	0.40983607	0.17213115	0.17213115	0.26639344	0.35655738
	Caseless	Nom random	Val	-												
			Test	0.44672131	0.50819672	0.45901639	0.45901639	0.52868852	0.52459016	0.52868852	0.53278689	0.55737705	0.45491803	0.52459016	0.35655738	0.5
		All Random	Val	-												
			Test	0.43032787	0.46311475	0.57377049	0.48770492	0.44262295	0.41803279	0.54918033	0.54098361	0.54508197	0.38114754	0.45081967	0.28688525	0.45901639

Table 2: Full results for Bert model. Including the layer by layer test error rate, the layer by layer test error rates for the control datasets (Same Token Amount, Noun Only, and Pronoun Only), and the layer by layer test error rate for the control probe. The table also includes all error rates for the validation set, minus the control probe where there was no validation set. The results reported in the main paper are from the Random dataset, layer 11.

B Full mBert Results

				layers												
				0	1	2	3	4	5	6	7	8	9	10	11	12
		Acc0_Dat1	Val	0.015	0.0125	0.0125	0.01	0.01	0.0125	0.01	0.0075	0.01	0.005	0.0025	0.005	0.01
			Test	0.01229508	0.01229508	0.01639344	0.01639344	0.00819672	0.00409836	0.00409836	0.00409836	0.01639344	0.01229508	0.02459016	0.02459016	0.02868852
		Nom0_Dat1	Val	0.0275	0.02	0.02	0.0125	0.01	0.005	0.0075	0.005	0.005	0.005	0.005	0.0075	0
			Test	0.09836066	0.10245902	0.06147541	0.04918033	0.04098361	0.03278689	0.01229508	0.01229508	0.03688525	0.04098361	0.08196721	0.05327869	0.08196721
	Cased	Nom1_Acc0_Dat1	Val	0.10833333	0.11833333	0.105	0.09166667	0.08833333	0.05166667	0.04333333	0.03166667	0.01166667	0.02	0.02166667	0.02333333	0.03833333
			Test	0.04098361	0.02459016	0.02459016	0.02459016	0.01639344	0.01229508	0	0	0	0	0.00409836	0.00409836	0.00819672
		Nom0_Acc0_Dat1	Val	0.03666667	0.03	0.02333333	0.01833333	0.01	0.01166667	0.01	0.00833333	0.00666667	0.005	0.00333333	0.005	0.00833333
			Test	0.10655738	0.08196721	0.04508197	0.04098361	0.04918033	0.0204918	0.03688525	0.01639344	0.02868852	0.05327869	0.05327869	0.05737705	0.06967213
		Nom0_Acc1_Dat1	Val	0.11666667	0.10333333	0.12	0.10666667	0.09666667	0.06833333	0.03166667	0.02	0.02	0.02666667	0.025	0.03333333	0.03666667
			Test	0.18032787	0.16803279	0.15983607	0.12704918	0.09836066	0.11065574	0.10655738	0.05737705	0.08196721	0.08606557	0.11885246	0.12704918	0.17213115
Random		Acc0_Dat1	Val	0.1775	0.155	0.1425	0.135	0.125	0.1225	0.0975	0.085	0.0875	0.08	0.08	0.0925	0.09
			Test	0.21311475	0.14344262	0.17622951	0.16803279	0.15983607	0.17213115	0.2295082	0.20901639	0.34836066	0.30737705	0.30737705	0.29098361	0.19262295
		Nom0_Dat1	Val	0.27	0.2325	0.23	0.205	0.17	0.1725	0.13	0.12	0.0975	0.085	0.0975	0.1075	0.105
			Test	0.61065574	0.56147541	0.53278689	0.57786885	0.57377049	0.55327869	0.62704918	0.72540984	0.63934426	0.62704918	0.62295082	0.50409836	0.56147541
	Caseless	Nom1_Acc0_Dat1	Val	0.22	0.20666667	0.18833333	0.175	0.17666667	0.17166667	0.15333333	0.12	0.14	0.11833333	0.13333333	0.13166667	0.13333333
			Test	0.12295082	0.06967213	0.06557377	0.06557377	0.07377049	0.06967213	0.06967213	0.01639344	0.03278689	0.02868852	0.03688525	0.03688525	0.04508197
		Nom0_Acc0_Dat1	Val	0.26333333	0.21666667	0.20166667	0.19	0.16666667	0.14833333	0.12833333	0.11833333	0.115	0.11666667	0.09666667	0.11	0.12
			Test	0.48770492	0.37704918	0.42213115	0.35245902	0.45491803	0.39754098	0.39754098	0.68032787	0.64344262	0.56147541	0.60655738	0.53688525	0.53688525
		Nom0_Acc1_Dat1	Val	0.23	0.24	0.23	0.2266667	0.215	0.2	0.16	0.14666667	0.12833333	0.12333333	0.11333333	0.15	0.14166667
			Test	0.68442623	0.68032787	0.71311475	0.70491803	0.68442623	0.65983607	0.68032787	0.76639344	0.78278689	0.62704918	0.68442623	0.69262295	0.68442623
		Acc0_Dat1	Val	0.03767123	0.03767123	0.03424658	0.02739726	0.01712329	0.01369863	0.01712329	0.01369863	0.01027397	0.01369863	0.01027397	0.01027397	0.01712329
			Test	0.03278689	0.01639344	0.02459016	0.0204918	0.00819672	0.00409836	0	0.00409836	0.01639344	0.0204918	0.02868852	0.03688525	0.02459016
		Nom0_Dat1	Val	0.02054795	0.02054795	0.01712329	0.01027397	0.00684932	0.00342466	0.00342466	0.00342466	0	0.00342466	0	0	0.01027397
			Test	0.1147541	0.14344262	0.1352459	0.10655738	0.08606557	0.04508197	0.03688525	0.04098361	0.08196721	0.06557377	0.10245902	0.10245902	0.10245902
	Cased	Nom1_Acc0_Dat1	Val	0.12328767	0.11643836	0.11187215	0.10502283	0.07762557	0.05251142	0.02283105	0.02511416	0.02739726	0.01826484	0.02739726	0.03196347	0.043379
			Test	0.04508197	0.0204918	0.0204918	0.03278689	0.01639344	0.00819672	0	0	0	0	0	0.00409836	0.00409836
		Nom0_Acc0_Dat1	Val	0.04794521	0.0456621	0.03881279	0.03196347	0.01826484	0.02283105	0.00684932	0.00456621	0.00913242	0.00456621	0.00684932	0.00684932	0.01369863
			Test	0.06557377	0.07377049	0.06557377	0.06967213	0.04918033	0.03688525	0.04098361	0.03688525	0.04098361	0.05327869	0.06147541	0.10245902	0.0942623
		Nom0_Acc1_Dat1	Val	0.15068493	0.11415525	0.10502283	0.09360731	0.06849315	0.05251142	0.01369863	0.01598174	0.01141553	0.01826484	0.02054795	0.02511416	0.03196347
			Test	0.24590164	0.28688525	0.26639344	0.17213115	0.14344262	0.1352459	0.11885246	0.09836066	0.15983607	0.08606557	0.1147541	0.14754098	0.22131148
Same Token Amount		Acc0_Dat1	Val	0.21575342	0.17123288	0.15753425	0.17465753	0.15410959	0.16438356	0.14383562	0.14041096	0.1130137	0.09589041	0.11986301	0.1130137	0.11986301
			Test	0.23360656	0.12704918	0.17622951	0.12704918	0.19672131	0.13114754	0.12295082	0.14754098	0.21721311	0.30737705	0.25409836	0.27459016	0.19262295
		Nom0_Dat1	Val	0.23287671	0.22945205	0.19863014	0.16780822	0.1369863	0.1369863	0.12671233	0.10273973	0.0890411	0.06849315	0.10273973	0.11643836	0.10616438
			Test	0.68032787	0.57786885	0.57377049	0.58196721	0.61885246	0.55737705	0.58196721	0.71311475	0.7295082	0.61885246	0.55737705	0.54098361	0.55737705
	Caseless	Nom1_Acc0_Dat1	Val	0.2283105	0.23059361	0.21004566	0.19634703	0.17351598	0.16894977	0.15068493	0.1347032	0.12100457	0.10502283	0.12328767	0.14383562	0.15296804
			Test	0.1147541	0.06147541	0.06147541	0.07377049	0.06967213	0.04918033	0.06147541	0.01229508	0.02868852	0.04918033	0.03278689	0.02459016	0.02868852
		Nom0_Acc0_Dat1	Val	0.2739726	0.23744292	0.22146119	0.20776256	0.19634703	0.18493151	0.14611872	0.12557078	0.13013699	0.1347032	0.1369863	0.1347032	0.14383562
			Test	0.52459016	0.45491803	0.43852459	0.41803279	0.48360656	0.44672131	0.38114754	0.60245902	0.66393443	0.54508197	0.61065574	0.54508197	0.52868852
		Nom0_Acc1_Dat1	Val	0.21689498	0.21917808	0.2260274	0.18721461	0.19178082	0.15981735	0.1369863	0.10273973	0.11187215	0.11187215	0.11415525	0.14383562	0.1347032
			Test	0.78278689	0.67213115	0.70901639	0.74590164	0.7295082	0.72540984	0.78278689	0.78278689	0.82786885	0.75819672	0.70901639	0.68442623	0.67213115
		Acc0_Dat1	Val	0.03465347	0.04950495	0.03465347	0.01980198	0.01980198	0.00990099	0.00990099	0.00990099	0.00990099	0.0049505	0.01485149	0.01485149	0.01980198
			Test	0.03645833	0.02604167	0.03125	0.02604167	0.01041667	0.00520833	0.015625	0.01041667	0.015625	0.02083333	0.03125	0.05208333	0.03645833
		Nom0_Dat1	Val	0.03465347	0.02475248	0.02475248	0.01980198	0.01485149	0.0049505	0	0	0.0049505	0	0	0	0
			Test	0.11979167	0.11458333	0.09895833	0.08333333	0.07291667	0.07291667	0.05729167	0.078125	0.14583333	0.11979167	0.140625	0.20833333	0.171875
	Cased	Nom1_Acc0_Dat1	Val	0.16171617	0.1650165	0.15841584	0.14521452	0.12211221	0.09240924	0.04620462	0.01650165	0.00990099	0.01320132	0.02640264	0.0330033	0.04290429
			Test	0.04166667	0.015625	0.02083333	0.03125	0.015625	0	0.01041667	0	0	0	0	0	0
		Nom0_Acc0_Dat1	Val	0.02970297	0.03960396	0.03630363	0.02970297	0.01650165	0.00990099	0.00660066	0.00660066	0.00660066	0	0.00990099	0.01320132	0.02640264
			Test	0.078125	0.05729167	0.046875	0.05208333	0.046875	0.03125	0.05729167	0.03645833	0.13020833	0.09895833	0.10416667	0.15104167	0.13020833
		Nom0_Acc1_Dat1	Val	0.21452145	0.15511551	0.15181518	0.12541254	0.10231023	0.06270627	0.02640264	0.02970297	0.02310231	0.01980198	0.01980198	0.01980198	0.04290429
			Test	0.22916667	0.27083333	0.27604167	0.203125	0.140625	0.18229167	0.17708333	0.10416667	0.10416667	0.078125	0.13541667	0.17708333	0.23958333
Noun Only		Acc0_Dat1	Val	0.1980198	0.13861386	0.13861386	0.13366337	0.13366337	0.11386139	0.11386139	0.07920792	0.06435644	0.06930693	0.05445545	0.05940594	0.07920792
			Test	0.22916667	0.13541667	0.125	0.16666667	0.11979167	0.14583333	0.15104167	0.14583333	0.19791667	0.24479167	0.23958333	0.24479167	0.15104167
		Nom0_Dat1	Val	0.23762376	0.19306931	0.2029703	0.17326733	0.15346535	0.15841584	0.09405941	0.07425743	0.05940594	0.05445545	0.04950495	0.04455446	0.06930693
			Test	0.63541667	0.61458333	0.6875	0.640625	0.6875	0.64583333	0.67708333	0.80729167	0.703125	0.6875	0.66145833	0.66666667	0.64583333
	Caseless	Nom1_Acc0_Dat1	Val	0.22112211	0.16831683	0.15841584	0.17821782	0.17491749	0.14191419	0.13531353	0.07590759	0.08250825	0.08910891	0.09570957	0.10231023	0.10231023
			Test	0.06770833	0.046875	0.046875	0.046875	0.04166667	0.04166667	0.02604167	0.00520833	0	0.00520833	0.015625	0.015625	0.015625
		Nom0_Acc0_Dat1	Val	0.22442244	0.20132013	0.19471947	0.19141914	0.15841584	0.17161716	0.15181518	0.08250825	0.09570957	0.08580858	0.08910891	0.08250825	0.09240924
			Test	0.53125	0.48958333	0.5	0.4375	0.52604167	0.46354167	0.44270833	0.67708333	0.52604167	0.609375	0.57291667	0.57291667	0.51041667
		Nom0_Acc1_Dat1	Val	0.28052805	0.23762376	0.21782178	0.18811881	0.17491749	0.14191419	0.0990099	0.06930693	0.04950495	0.08580858	0.08580858	0.10891089	0.10891089
			Test	0.765625	0.796875	0.81770833	0.77083333	0.83333333	0.83854167	0.875	0.89583333	0.859375	0.73958333	0.765625	0.76041667	0.73958333
		Acc0_Dat1	Val	0.025	0.025	0.05	0.05	0.075	0.05	0.075	0.05	0.05	0.05	0.025	0.025	0
			Test	0	0	0	0.01923077	0.01923077	0.01923077	0	0	0.01923077	0	0.01923077	0	0
		Nom0_Dat1	Val	0	0	0	0	0	0	0	0	0	0	0	0	0
			Test	0	0	0	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.03846154	0.03846154	0.03846154	0.05769231
	Cased	Nom1_Acc0_Dat1	Val	0.18333333	0.21666667	0.18333333	0.18333333	0.16666667	0.13333333	0.08333333	0.05	0.06666667	0.03333333	0.06666667	0.05	0.08333333
			Test	0	0.03846154	0.01923077	0.01923077	0.01923077	0.01923077	0	0	0	0	0	0	0.01923077
		Nom0_Acc0_Dat1	Val	0	0	0.01666667	0	0	0	0.01666667	0	0	0	0	0	0
			Test	0	0	0	0.01923077	0	0	0	0	0.01923077	0.03846154	0.07692308	0.07692308	0
		Nom0_Acc1_Dat1	Val	0.16666667	0.16666667	0.13333333	0.16666667	0.05	0.05	0.05	0.03333333	0.01666667	0.01666667	0.01666667	0	0.01666667
			Test	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.03846154	0.03846154	0.03846154	0.03846154	0.03846154	0.01923077	0.01923077	0.09615385
Pronoun Only		Acc0_Dat1	Val	0.1	0.125	0.1	0.125	0.175	0.15	0.125	0.15	0.15	0.175	0.125	0.15	0.125
			Test	0.23076923	0.17307692	0.17307692	0.17307692	0.19230769	0.21153846	0.21153846	0.23076923	0.30769231	0.36538462	0.40384615	0.34615385	0.44230769
		Nom0_Dat1	Val	0.2	0.05	0.1	0.075	0.125	0.1	0.125	0.15	0.175	0.125	0.15	0.075	0.075
			Test	0.40384615	0.26923077	0.30769231	0.26923077	0.21153846	0.32692308	0.38461538	0.51923077	0.53846154	0.59615385	0.57692308	0.53846154	0.55769231
	Caseless	Nom1_Acc0_Dat1	Val	0.21666667	0.18333333	0.21666667	0.18333333	0.26666667	0.25	0.21666667	0.2	0.2	0.16666667	0.2	0.21666667	0.2
			Test	0.26923077	0.19230769	0.19230769	0.25	0.13461538	0.25	0.21153846	0.23076923	0.19230769	0.21153846	0.21153846	0.21153846	0.21153846
		Nom0_Acc0_Dat1	Val	0.18333333	0.13333333	0.13333333	0.15	0.16666667	0.18333333	0.13333333	0.11666667	0.16666667	0.13333333	0.16666667	0.11666667	0.11666667
			Test	0.42307692	0.26923077	0.21153846	0.26923077	0.25	0.23076923	0.30769231	0.40384615	0.5	0.42307692	0.57692308	0.5	0.53846154
		Nom0_Acc1_Dat1	Val	0.33333333	0.21666667	0.3	0.3	0.31666667	0.31666667	0.36666667	0.23333333	0.18333333	0.26666667	0.25	0.21666667	0.3
			Test	0.48076923	0.36538462	0.46153846	0.44230769	0.53846154	0.5	0.53846154	0.65384615	0.55769231	0.61538462	0.57692308	0.67307692	0.59615385
		Acc_random	Val	-												
			Test	0.5942623	0.48360656	0.45491803	0.42213115	0.54918033	0.44672131	0.46721311	0.34016393	0.62704918	0.41803279	0.38934426	0.30737705	0.43852459
	Cased	Nom random	Val	-												
			Test	0.40163934	0.42622951	0.52459016	0.47131148	0.44262295	0.45491803	0.61065574	0.5204918	0.35655738	0.37295082	0.43442623	0.37704918	0.33606557
		All Random	Val	-												
Control			Test	0.51229508	0.49180328	0.38114754	0.4057377	0.43032787	0.37704918	0.52459016	0.36885246	0.46311475	0.39344262	0.44262295	0.40983607	0.3647541
		Acc_random	Val	-												
			Test	0.53688525	0.54508197	0.44672131	0.5204918	0.49180328	0.49590164	0.40163934	0.14344262	0.54508197	0.40163934	0.2295082	0.29508197	0.22540984
	Caseless	Nom random	Val	-												
			Test	0.47131148	0.48360656	0.41393443	0.48770492	0.3852459	0.27459016	0.23360656	0.43852459	0.4795082	0.5	0.52459016	0.54098361	0.3442623
		All Random	Val	-												
			Test	0.45081967	0.39754098	0.3647541	0.56147541	0.33606557	0.26229508	0.26229508	0.30737705	0.34836066	0.20491803	0.3647541	0.24590164	0.30327869

Table 3: Full results for mBert model. Including the layer by layer test error rate, the layer by layer test error rates for the control datasets (Same Token Amount, Noun Only, and Pronoun Only), and the layer by layer test error rate for the control probe. The table also includes all error rates for the validation set, minus the control probe where there was no validation set. The results reported in the main paper are from the Random dataset, layer 10.

C Full LläMmlein 1B Results

				layers																							
				0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
		Acc0_Dat1	Val	0.0275	0.0225	0.015	0.03	0.0225	0.0075	0.015	0.01	0.01	0.0075	0.01	0.0025	0.01	0.0075	0.0025	0.0025	0.0025	0.0025	0.0025	0	0.005	0.0025	0
			Test	0.04098361	0.06557377	0.03278689	0.0204918	0.04508197	0.01229508	0.045082197	0	0.00409836	0.00409836	0.00409836	0.00409836	0.01639344	0.00819672	0.00819672	0.00819672	0.00819672	0.00819672	0.00819672	0.00819672	0.00819672	0.00819672	0.00819672
		Non0_Dat1	Val	0.0525	0.035	0.05	0.03	0.0175	0.015	0.015	0.005	0.005	0.005	0.0025	0.0025	0.0025	0.0025	0.005	0	0	0	0	0	0	0.0025	0.0075	0.0025
			Test	0.13114754	0.15983607	0.19262295	0.04508197	0.05737705	0.07377049	0.03278689	0.02459016	0.0204918	0.03278689	0.02459016	0.04508197	0.03278689	0.02868852	0.0204918	0.0204918	0.0204918	0.0204918	0.01639344	0.0204918	0.01229508	0.01639344	0.0204918	0.01229508	0.01639344
	Cased	Non1_Acc0_Dat1	Val	0.15	0.13666667	0.13666667	0.11166667	0.08833333	0.09166667	0.08333333	0.03166667	0.045	0.035	0.03	0.03333333	0.03333333	0.04166667	0.02166667	0.02333333	0.025	0.02333333	0.02333333	0.02	0.03666667	0.02333333	0.04
			Test	0.04508197	0.02868852	0.0204918	0.0204918	0.0204918	0.01639344	0.04098361	0.03278689	0.00819672	0.01639344	0.02868852	0.02459016	0.03688525	0.03688525	0.00409836	0	0	0	0.00409836	0.00409836	0	0.00409836	0.00409836	0.00409836
		Non0_Acc0_Dat1	Val	0.03666667	0.05166667	0.03	0.03	0.03	0.01666667	0.01666667	0.005	0.00666667	0.00666667	0.00666667	0.00333333	0.005	0.005	0.00333333	0.00166667	0.00166667	0.00333333	0.00166667	0.00166667	0.00333333	0.00166667	0.01	
			Test	0.17622951	0.15983607	0.1352459	0.06147541	0.07377049	0.06967213	0.06967213	0.02459016	0.0204918	0.02459016	0.01229508	0.03278689	0.02868852	0.00819672	0.01639344	0.01639344	0.01639344	0.01639344	0.01639344	0.01639344	0.01639344	0.01229508	0.0204918	
		Non1_Acc1_Dat1	Val	0.15666667	0.125	0.095	0.095	0.07166667	0.06666667	0.025	0.02166667	0.02166667	0.01666667	0.01666667	0.02833333	0.025	0.03	0.01833333	0.02166667	0.025	0.01833333	0.02166667	0.02	0.02833333	0.02833333	0.03833333	
	Random		Test	0.19672131	0.19262295	0.18032787	0.18442623	0.16803279	0.16393443	0.10655738	0.06147541	0.02459016	0.02459016	0.03278689	0.06967213	0.06147541	0.04098361	0.01639344	0.01229508	0.0204918	0.02868852	0.0204918	0.01639344	0.0204918	0.01639344	0.0204918	0.0204918
		Acc0_Dat1	Val	0.27	0.34	0.225	0.21	0.2275	0.185	0.175	0.18	0.1525	0.145	0.135	0.14	0.1325	0.1275	0.115	0.11	0.105	0.105	0.1025	0.075	0.08	0.0775	0.095	
			Test	0.07786885	0.4057377	0.08606557	0.09836066	0.08606557	0.07786885	0.08196721	0.06967213	0.05327869	0.05327869	0.06967213	0.07786885	0.07786885	0.07377049	0.27868852	0.2295082	0.18442623	0.29918033	0.28278689	0.1352459	0.1352459	0.12295082	0.08196721	
		Non0_Dat1	Val	0.2875	0.2625	0.2625	0.26	0.2325	0.2475	0.1575	0.27	0.2025	0.205	0.175	0.155	0.15	0.1475	0.1375	0.15	0.1625	0.155	0.1075	0.135	0.1425	0.12	0.115	
			Test	0.6557377	0.75409836	0.82377049	0.74180328	0.7704918	0.82377049	0.72540984	0.8647541	0.68852459	0.71721311	0.70081967	0.69262295	0.68852459	0.70491803	0.70491803	0.72131148	0.7295082	0.72540984	0.76639344	0.68852459	0.66803279	0.67213115	0.71317475	
			Val	0.285	0.235	0.23666667	0.26	0.25	0.22166667	0.22333333	0.21333333	0.21166667	0.22333333	0.2283333	0.165	0.17333333	0.205	0.18	0.16	0.16166667	0.13666667	0.14833333	0.13833333	0.13166667	0.12833333	0.1183333	
		Non1_Acc0_Dat1	Test	0.08196721	0.06557377	0.08196721	0.11065574	0.10245902	0.09016393	0.07786885	0.05737705	0.06967213	0.04508197	0.06147541	0.05737705	0.08196721	0.04508197	0.04918033	0.03688525	0.06557377	0.04508197	0.05737705	0.02868852	0.04098361	0.04098361	0.02459016	
			Val	0.38833333	0.41166667	0.41166667	0.42166667	0.41166667	0.38833333	0.2	0.23166667	0.31166667	0.205	0.19333333	0.18333333	0.195	0.18333333	0.17833333	0.17666667	0.16166667	0.13333333	0.165	0.14666667	0.14166667	0.11	0.12166667	
		Non0_Acc0_Dat1	Test	0.54098361	0.51639344	0.58196721	0.54508197	0.52868852	0.50409836	0.70901639	0.76639344	0.58196721	0.70081967	0.56967213	0.58606557	0.6557377	0.70491803	0.69672131	0.68442623	0.54508197	0.60245902	0.63934426	0.56147541	0.66803279	0.56147541	0.55327869	
			Val	0.25333333	0.24	0.23666667	0.28333333	0.26833333	0.23833333	0.22166667	0.1983333	0.215	0.21166667	0.19166667	0.17	0.16833333	0.15666667	0.14	0.1383333	0.15333333	0.125	0.13666667	0.12	0.13166667	0.13333333	0.12	
		Non1_Acc1_Dat1	Test	0.79098361	0.81557377	0.85655738	0.86065574	0.85245902	0.8442623	0.85655738	0.82786885	0.81147541	0.84016393	0.78688525	0.75819672	0.76229508	0.78688525	0.76639344	0.80737705	0.78278689	0.7704918	0.78278689	0.85655738	0.77868852	0.75409836	0.79918033	
			Val	0.06164384	0.07191781	0.05821918	0.06849315	0.04452055	0.03424658	0.0239726	0.00684932	0.00684932	0.01369863	0.01369863	0.00684932	0.01369863	0.01027397	0.00342466	0.00342466	0.00342466	0.00342466	0.00684932	0.00342466	0.00684932	0.01027397	0.01369863	
		Acc0_Dat1	Test	0.04508197	0.04918033	0.03688525	0.0204918	0.03688525	0.03278689	0.04098361	0.01229508	0.00819672	0.00819672	0.01639344	0.01639344	0.00819672	0.00819672	0.00819672	0.00819672	0.00819672	0.00819672	0.00819672	0.01229508	0.01639344	0.01639344	
			Val	0.06849315	0.07191781	0.04109589	0.04452055	0.02054795	0.00684932	0.01027397	0	0.00684932	0	0	0	0	0	0	0	0.00342466	0.00342466	0.00342466	0.00342466	0.00342466	0	0	0
		Non0_Dat1	Test	0.2295082	0.19672131	0.17213115	0.19672131	0.08196721	0.04918033	0.07377049	0.03688525	0.03278689	0.04098361	0.02459016	0.02868852	0.03278689	0.0204918	0.01229508	0.01639344	0.01639344	0.01229508	0.01229508	0.01229508	0.01229508	0.01639344	0.01639344	
			Val	0.11872146	0.13470032	0.10958904	0.11415525	0.0890411	0.08219178	0.06164384	0.05251142	0.06164384	0.04109589	0.05251142	0.03881279	0.03652968	0.05251142	0.03881279	0.0319635	0.03652968	0.02739726	0.03652968	0.02739726	0.04109589	0.05251142	0.05707763	
		Non1_Acc0_Dat1	Test	0.05327869	0.04098361	0.01639344	0.0204918	0.03278689	0.00819672	0.02868852	0.0204918	0.00409836	0.00409836	0.00409836	0.00819672	0.00819672	0.02459016	0.00819672	0.00819672	0.00819672	0.01229508	0.00819672	0.00819672	0.00819672	0.01229508	0.0204918	
			Val	0.09536073	0.07990868	0.06164384	0.07762557	0.05936073	0.01598174	0.02511416	0.00456621	0.01141553	0.00456621	0.00684932	0.00456621	0.00913242	0.00913242	0.00456621	0.00228311	0.00228311	0.00456621	0.00456621	0.00456621	0.00456621	0.00913242	0.01141553	
		Non0_Acc0_Dat1	Test	0.13934426	0.18032787	0.16803279	0.10245902	0.06967213	0.04918033	0.05327869	0.02868852	0.00819672	0.01639344	0.03688525	0.03278689	0.0204918	0.01639344	0.01639344	0.01639344	0.0204918	0.01229508	0.01229508	0.01639344	0.01639344	0.01639344	0.01639344	
			Val	0.14383562	0.14383562	0.11872146	0.08904411	0.05479452	0.02511142	0.04794521	0.02511416	0.03424658	0.03196347	0.03196347	0.03424658	0.03196347	0.02739726	0.02511416	0.02968037	0.03196347	0.02968037	0.02739726	0.02511416	0.03424658	0.02511416	0.04109589	
		Non1_Acc1_Dat1	Test	0.19262295	0.18442623	0.18442623	0.22540984	0.20901639	0.20491803	0.18032787	0.10655738	0.08606557	0.08606557	0.07377049	0.08606557	0.07786885	0.08606557	0.01639344	0.01639344	0.01639344	0.01229508	0.01229508	0.01229508	0.0204918	0.02459016	0.03688525	
	Same Token Amount		Val	0.30479452	0.35273973	0.32191781	0.30136986	0.29109589	0.27054795	0.26712329	0.2260274	0.23972603	0.23630137	0.21575342	0.20547945	0.21232877	0.21917808	0.18150685	0.17123288	0.14383562	0.18150685	0.14383562	0.14041096	0.13013699	0.12671233	0.13356164	
		Acc0_Dat1	Test	0.07786885	0.45081967	0.09016393	0.07786885	0.07377049	0.07786885	0.0942623	0.07786885	0.07377049	0.07786885	0.07377049	0.08606557	0.09836066	0.08606557	0.31557377	0.08196721	0.34016393	0.10655738	0.33606557	0.1557377	0.23770492	0.20081967	0.07737049	
			Val	0.29452055	0.26712329	0.25	0.24657534	0.23972603	0.23287671	0.2260274	0.25	0.26027397	0.19863014	0.19863014	0.17123288	0.15753425	0.17465753	0.1609589	0.14726027	0.1369863	0.13356164	0.1369863	0.13013699	0.14383562	0.14041096	0.13356164	
		Non0_Dat1	Test	0.70901639	0.79098361	0.81147541	0.84967213	0.76229508	0.82786885	0.82786885	0.86065574	0.75	0.79081967	0.68442623	0.68852459	0.71721311	0.70081967	0.74180328	0.73360656	0.72540984	0.7295082	0.69672131	0.73360656	0.77704092	0.71311475	0.73770492	
			Val	0.27625571	0.29908676	0.30136986	0.27853881	0.29223744	0.26711329	0.25342466	0.26255708	0.26484018	0.25799087	0.23744292	0.23744292	0.23515982	0.23972603	0.20776256	0.20547945	0.19406393	0.19634703	0.20319635	0.13926941	0.1803653	0.16666667	0.1347032	
		Non1_Acc0_Dat1	Test	0.08606557	0.08196721	0.07377049	0.09016393	0.10655738	0.0942623	0.07377049	0.06557377	0.06967213	0.07737705	0.06557377	0.07786885	0.06967213	0.06967213	0.06147541	0.07737705	0.06147541	0.06557377	0.06147541	0.0204918	0.05327869	0.07737705	0.0204918	
			Val	0.42922374	0.42922374	0.37671233	0.41735516	0.42694064	0.42465753	0.41552511	0.2191708	0.31050228	0.2146187	0.32191781	0.20091324	0.22146119	0.18721461	0.23287671	0.18721461	0.1803653	0.13926941	0.17123288	0.16894977	0.16666667	0.17808219	0.13013699	
		Non0_Acc0_Dat1	Test	0.54508197	0.54508197	0.60655738	0.49590164	0.6147541	0.59836066	0.65163934	0.75	0.61065574	0.7704918	0.61065574	0.65557377	0.65983606	0.71721311	0.57377049	0.77459016	0.47131148	0.65163934	0.72131148	0.66393443	0.61885246	0.67213115	0.6352459	
			Val	0.26484018	0.24657534	0.24200913	0.26255708	0.25114155	0.23744292	0.21917808	0.21689498	0.19863014	0.20776256	0.17123288	0.21461187	0.1826484	0.16438356	0.15981735	0.16438356	0.16666667	0.15068493	0.14840183	0.14383562	0.14840183	0.1347032	0.12328767	
		Non1_Acc1_Dat1	Test	0.78278689	0.86065574	0.84016393	0.8442623	0.82786885	0.84016393	0.83606557	0.84016393	0.82786885	0.79918033	0.71721311	0.78278689	0.74180328	0.72540984	0.80737705	0.76639344	0.7663934	0.81557377	0.79098361	0.7704918	0.79508197	0.79508197	0.83196721	
			Val	0.04950495	0.04950495	0.02970297	0.02475248	0.02475248	0.02475248	0.01980198	0.00990099	0.01980198	0.01485149	0.01485149	0.01485149	0.0049505	0.00990099	0.00990099	0.00990099	0.01485149	0.01485149	0.0049505	0.0049505	0.0049505	0.0049505	0.0049505	
		Acc0_Dat1	Test	0.046875	0.03645833	0.04166667	0.03645833	0.05729167	0.06770833	0.046875	0	0.00520833	0.01041667	0	0.015625	0.01041667	0.00520833	0.00520833	0.00520833	0.00520833	0.01041667	0.00520833	0.01041667	0.00520833	0.01041667	
			Val	0.04950495	0.06930693	0.03960396	0.03960396	0.02970297	0.01485149	0.01485149	0	0.00990099	0.0049505	0	0	0.0049505	0.0049505	0	0	0	0	0	0	0	0	0.0049505	
		Non0_Dat1	Test	0.1875	0.25	0.109375	0.13020833	0.09375	0.02604167	0.125	0.02604167	0.02604167	0.03125	0.02083333	0.03645833	0.02604167	0.015625	0.01041667	0.015625	0.015625	0.01041667	0.01041667	0.015625	0.02604167	0.02083333	0.01041667	
			Val	0.16831683	0.18811881	0.17491749	0.1320132	0.11221122	0.08250825	0.08250825	0.0330033	0.07590759	0.02970297	0.02640264	0.0330033	0.02640264	0.02640264	0.01650165	0.02640264	0.02310231	0.01980198	0.01980198	0.01980198	0.02640264	0.02310231	0.02640264	
	Cased	Non1_Acc0_Dat1	Test	0.02083333	0.01125	0.015625	0.01041667	0.00520833	0.02083333	0.02604167	0.046875	0	0.046875	0.046875	0.01041667	0.02083333	0.05729167	0.00520833	0.00520833	0.01041667	0	0	0.00520833	0.00520833	0.01041667	0.00520833	
			Val	0.07260726	0.05280528	0.04290429	0.02970297	0.0330033	0.0330033	0.01980198	0.00330033	0.00990099	0.01980198	0.00660066	0.00660066	0.00990099	0.00660066	0.00660066	0.00990099	0.01320132	0.00660066	0.00330033	0.00990099	0.00990099	0.00660066	0.00990099	
		Non0_Acc0_Dat1	Test	0.28125	0.19270833	0.11458333	0.05729167	0.078125	0.08333333	0.05208333	0.02604167	0.02083333	0.02083333	0.04166667	0.046875	0.02083333	0.03645833	0.015625	0.015625	0.015625	0.01041667	0.01041667	0.01041667	0.02083333	0.015625	0.02083333	
			Val	0.21782178	0.16831683	0.15841584	0.1320132	0.10561056	0.06930693	0.0330033	0.0330033	0.04950495	0.01320132	0.01320132	0.01650165	0.01650165	0.01650165	0.01980198	0.02640264	0.01980198	0.01980198	0.01320132	0.00990099	0.02310231	0.02970297	0.02640264	
		Non1_Acc1_Dat1	Test	0.21354167	0.3229167	0.234375	0.21875	0.20833333	0.16666667	0.20312	0.11458333	0.16145833	0.07529167	0.07529167	0.015625	0.02083333	0.02083333	0.015625	0.01041667	0.02083333	0.015625	0.01041667	0.015625	0.02083333	0.02083333	0.02083333	
	Noun Only		Val	0.26237624	0.21287129	0.21287129	0.19306931	0.26732673	0.21782178	0.18316832	0.16336634	0.14356436	0.16336634	0.16336634	0.16336634	0.16336634	0.13861386	0.14356436	0.12871287	0.13366337	0.12871287	0.11881188	0.07920792	0.08910891	0.09405941	0.08910891	
		Acc0_Dat1	Test	0.08854167	0.08854167	0.06770833	0.078125	0.046875	0.05208333	0.046875	0.046875	0.05208333	0.046875	0.046875	0.046875	0.046875	0.046875	0.19270833	0.140625	0.328125	0.21875	0.3125	0.25	0.328125	0.23958333	0.25	0.16145833
			Val	0.26732673	0.31683168	0.34158416	0.28217822	0.28217822	0.25742574	0.26732673	0.26732673	0.15841584	0.13861386	0.12871287	0.15346535	0.14851485	0.14356436	0.09405941	0.12871287	0.05445545	0.08415842	0.0990099	0.09405941	0.08415842	0.10891089	0.04950495	
		Non0_Dat1	Test	0.703125	0.703125	0.66145833	0.72395833	0.70833333	0.85416667	0.83333333	0.80729167	0.703125	0.74479167	0.68229167	0.76875	0.80729167	0.78645833	0.80729167	0.828125	0.671875	0.796875	0.8125	0.75520833	0.765625	0.828125	0.73958333	
			Val	0.36633663	0.2871287	0.24092409	0.31353135	0.2970297	0.28052805	0.25082508	0.20132013	0.21122112	0.17161716	0.23102231	0.15511551	0.1650165	0.15841584	0.15181518	0.12871287	0.14851485	0.13531353	0.13531353	0.10561056	0.09570957	0.11221122	0.10561056	
		Acc0_Dat1	Test	0.22395833	0.08333333	0.0625	0.09375	0.078125	0.08854167	0.0625	0.03645833	0.05208333	0.04166667	0.05208333	0.05729167	0.04166667	0.05208333	0.05208333	0.01041667	0.015625	0.01041667	0.01041667	0.00520833	0.00520833	0.00520833	0.015625	
			Val	0.33993399	0.3960396	0.38613861	0.40594059	0.33993399	0.23432343	0.28712871	0.24422442	0.21452145	0.26732673	0.19471947	0.18151815	0.19471947	0.14851485	0.1650165	0.16171617	0.16171617	0.1320132	0.14521452	0.12541254	0.14191419	0.10891089	0.11221122	
		Non0_Dat1	Test	0.63541667	0.6093775	0.796875	0.6093775	0.8125	0.76041667	0.85416667	0.71354167	0.81770833	0.64583333	0.69270833	0.66145833	0.7630125	0.71354167	0.77083333	0.765625	0.80208333	0.625	0.54166667	0.578125	0.72916667	0.59895833	0.58333333	
			Val	0.28382838	0.25082508	0.23432343	0.2310231	0.22112211	0.21452145	0.21122112	0.17491749	0.1650165	0.13531353	0.14521452	0.17821782	0.14191419	0.12871287	0.10891089	0.12211221	0.10891089	0.09240924	0.09570957	0.09570957	0.10231023	0.08580858	0.07920792	
		Non1_Acc1_Dat1	Test	0.828125	0.6875	0.86979167	0.828125	0.79166667	0.80208333	0.86458333	0.828125	0.88541667	0.88541667	0.765625	0.85416667	0.91666667	0.92708333	0.93229167	0.91145833	0.91666667	0.90625	0.890625	0.91145833	0.88020833	0.91666667	0.91145833	
			Val	0	0.025	0	0.025	0.025	0	0	0.025	0	0	0	0.025	0	0.025	0	0.025	0.025	0.025	0	0	0	0	0
		Acc0_Dat1	Test	0.03846154	0.03846154	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.05769231	0.03846154	0.07692308	0.03846154	0	0.03846154	0.03846154	0.03846154	
			Val	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.025	0.025	
		Non0_Dat1	Test	0.03846154	0.01923077	0.01923077	0.01923077	0.01923077	0.03846154	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.03846154	0.03846154	0.05769231	0.05769231	0.05769231	
			Val	0.13333333	0.1	0.06666667	0.1	0.05	0.1	0.08333333	0.1	0.03333333	0.06666667	0.13333333	0.06666667	0.1	0.08333333	0.08333333	0.05	0.08333333	0.06666667	0.05	0.06666667	0.05	0.06666667	0.1	
		Non1_Acc0_Dat1	Test	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0	0	0	0.01923077	0.01923077	0.01923077	0.01923077	0	0	0.01923077	0	0.01923077	0.01923077	0.03846154		
			Val	0	0.01666667	0.01666667	0.01666667	0.01666667	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01666667	0	0.01666667	0.01666667	0.01666667	
		Non0_Acc0_Dat1	Test	0.03846154	0.03846154	0.01923077	0.01923077	0.03846154	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.03846154	0.03846154	0.03846154	0.05769231	0.03846154	0.03846154	0.03846154	0.03846154	
			Val	0.03333333	0.16666667	0.1	0.13333333	0.13333333	0.03333333	0.03333333	0.06666667	0.03333333	0.03333333	0.03333333	0.03333333	0.1	0.03333333	0.03333333	0.03333333	0.03333333	0.03333333	0.06666667	0.06666667	0.03333333	0.11666667	
		Non1_Acc1_Dat1	Test	0.03846154	0.01923077	0.01923077	0.01923077	0	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	0.01923077	
	Pronoun Only		Val	0.15	0.15	0.15	0.225	0.225	0.15	0.15	0.125	0.15	0.175	0.2	0.125	0.15	0.125	0.125	0.125	0.1	0.175	0.1	0.075	0.1	0.125	0.075
		Acc0_Dat1	Test	0.25	0.15384615	0.28846154	0.34615385	0.11538462	0.17307692	0.19230769	0.23076923	0.17307692	0.19230769	0.15384615	0.17307692	0.23076923	0.21153846	0.17307692	0.15384615	0.19230769	0.19230769	0.13461538	0.17307692	0.19230769	0.13461538	0.15384615	
			Val	0.25	0.125	0.175	0.25	0.15	0.175	0.15	0.175	0.125	0.25	0.15	0.125	0.225	0.25	0.075	0.075	0.075	0.175	0.075	0.1	0.1	0.075	0.15	
		Non0_Dat1	Test	0.53846154	0.53846154	0.25	0.51923077	0.57692308	0.42307692	0.44230769	0.42307692	0.57692308	0.69230769	0.46153846	0.57692308	0.71153846	0.71153846	0.40384615	0.40384615	0.40384615	0.26923077	0.36538462	0.34615385	0.36538462	0.28846154	0.53846154	
			Val	0.25	0.26666667	0.26666667	0.26666667	0.3	0.25	0.3	0.3	0.23333333	0.23333333	0.25	0.23333333	0.26666667	0.26666667	0.23333333	0.21666667	0.23333333	0.25	0.16666667	0.18333333	0.35	0.16666667	0.21666667	
		Acc0_Dat1	Test	0.09615385	0.09615385	0.15384615	0.11538462	0.23076923	0.21153846	0.25	0.23076923	0.11538462	0.21153846	0.17307692	0.17307692	0.17307692	0.19230769	0.19230769	0.13461538	0.15384615	0.15384615	0.11538462	0.21153846	0.09615385	0.11538462		
			Val	0.2	0.21666667	0.25	0.18333333	0.2	0.25	0.18333333	0.21666667	0.18333333	0.15	0.15	0.15	0.15	0.11666667	0.11666667	0.11666667	0.13333333	0.11666667	0.15	0.13333333	0.11666667	0.2	0.16666667	0.16666667	
		Non0_Acc0_Dat1	Test	0.25	0.34615385	0.34615385	0.28846154	0.25	0.53846154	0.51923077	0.25	0.25	0.44230769	0.42307692	0.32692308	0.40384615	0.38461538	0.38461538	0.40384615	0.21153846	0.36538462	0.38461538	0.30769231	0.13461538	0.36538462	0.19230769	
			Val	0.31666667	0.13333333	0.31666667	0.31666667	0.3	0.3	0.35	0.3	0.33333333	0.26666667	0.35	0.36666667	0.26666667	0.33333333	0.33333333	0.36666667	0.31666667	0.3	0.31666667	0.3	0.31666667	0.31666667	0.3	
		Non1_Acc1_Dat1	Test	0.19230769	0.40384615	0.19230769	0.36538462	0.36538462	0.36538462	0.63461538	0.36538462	0.63461538	0.48076923	0.63846615	0.63461538	0.61538462	0.63461538	0.61538462	0.59615385	0.59615385	0.59615385	0.59615385	0.57692308	0.57692308	0.32692308	0.55769231	
			Val	-																						
		Acc_random	Test	0.479508	0.52868852	0.5	0.3770491	0.52459016	0.40163934	0.35245902	0.46311475	0.43442623	0.45901639	0.46311475	0.59836066	0.52459016	0.34836066	0.6557377	0.5204918	0.55737705	0.6352459	0.53688525	0.58606557	0.52459016	0.63114754	0.5204918	
			Val	-																						
	Cased	Non_random	Test	0.34016393	0.48770492	0.50409836	0.5	0.57377049	0.44262295	0.56147541	0.54591803	0.34016393	0.4795082	0.47131148	0.50409836	0.3647541	0.50819672	0.47131148	0.62295082	0.56147541	0.56557377	0.57377049	0.43032787	0.45081967	0.44262295	
			Val	-																						
		All Random	Test	0.36885246	0.48770492	0.43032787	0.44262295	0.35655738	0.30327869	0.2295082	0.57377049	0.42622951	0.41393443	0.40163934	0.46311475	0.55737705	0.46311475	0.45901639	0.58196721	0.54918033	0.62704918	0.60245902	0.56962213	0.30737705	0.52868852	0.55327869	
	Control		Val	-																						
		Acc_random	Test	0.47131148

D Full Word2Vec Results

Random	**Cased**	**Acc0_Dat1**	**Val**	0.01662049861
			Test	0.01739130435
		Nom0_Dat1	**Val**	0.01333333333
			Test	0
		Nom1_Acc0_Dat1	**Val**	0.1423357664
			Test	0.06086956522
		Nom0_Acc0_Dat1	**Val**	0.02737226277
			Test	0.01739130435
		Nom0_Acc1_Dat1	**Val**	0.147810219
			Test	0.05652173913
	Caseless	**Acc0_Dat1**	**Val**	0.1982758621
			Test	0.3051643192
		Nom0_Dat1	**Val**	0.2841225627
			Test	0.4084507042
		Nom1_Acc0_Dat1	**Val**	0.2129277567
			Test	0.2957746479
		Nom0_Acc0_Dat1	**Val**	0.2585551331
			Test	0.338028169
		Nom0_Acc1_Dat1	**Val**	0.319391635
			Test	0.3427230047
Same Token Amount	**Cased**	**Acc0_Dat1**	**Val**	0.03370786517
			Test	0.02173913043
		Nom0_Dat1	**Val**	0.007272727273
			Test	0
		Nom1_Acc0_Dat1	**Val**	0.1523341523
			Test	0.05217391304
		Nom0_Acc0_Dat1	**Val**	0.03194103194
			Test	0.01304347826
		Nom0_Acc1_Dat1	**Val**	0.1646191646
			Test	0.04782608696
	Caseless	**Acc0_Dat1**	**Val**	0.1984126984
			Test	0.3051643192
		Nom0_Dat1	**Val**	0.2996108949
			Test	0.3849765258
		Nom1_Acc0_Dat1	**Val**	0.2663185379
			Test	0.2723004695
		Nom0_Acc0_Dat1	**Val**	0.2480417755
			Test	0.3568075117
		Nom0_Acc1_Dat1	**Val**	0.3446475196
			Test	0.3990610329
Noun Only	**Cased**	**Acc0_Dat1**	**Val**	0.02312138728
			Test	0.0393258427
		Nom0_Dat1	**Val**	0.01149425287
			Test	0.01685393258
		Nom1_Acc0_Dat1	**Val**	0.2432432432
			Test	0.06179775281
		Nom0_Acc0_Dat1	**Val**	0.03861003861
			Test	0.02808988764
		Nom0_Acc1_Dat1	**Val**	0.2007722008
			Test	0.0393258427
	Caseless	**Acc0_Dat1**	**Val**	0.2075471698
			Test	0.2795031056
		Nom0_Dat1	**Val**	0.28125
			Test	0.3229813665
		Nom1_Acc0_Dat1	**Val**	0.268907563
			Test	0.2670807453
		Nom0_Acc0_Dat1	**Val**	0.256302521
			Test	0.3602484472
		Nom0_Acc1_Dat1	**Val**	0.3655462185
			Test	0.3229813665
Pronoun Only	**Cased**	**Acc0_Dat1**	**Val**	0
			Test	0
		Nom0_Dat1	**Val**	0
			Test	0.01923076923
		Nom1_Acc0_Dat1	**Val**	0.1166666667
			Test	0
		Nom0_Acc0_Dat1	**Val**	0
			Test	0.01923076923
		Nom0_Acc1_Dat1	**Val**	0.15
			Test	0.01923076923
	Caseless	**Acc0_Dat1**	**Val**	0.1052631579
			Test	0.25
		Nom0_Dat1	**Val**	0.175
			Test	0.25
		Nom1_Acc0_Dat1	**Val**	0.2586206897
			Test	0.09615384615
		Nom0_Acc0_Dat1	**Val**	0.2068965517
			Test	0.25
		Nom0_Acc1_Dat1	**Val**	0.3275862069
			Test	0.1923076923
Control	**Cased**	**Acc_random**	**Val**	
			Test	0.4
		Nom random	**Val**	
			Test	0.5347826087
		All Random	**Val**	
			Test	0.5217391304
	Caseless	**Acc_random**	**Val**	
			Test	0.3896713615
		Nom random	**Val**	
			Test	0.4835680751
		All Random	**Val**	
			Test	0.4882629108

Table 5: Full results for Word2Vec model. Including the test error rates for the control datasets (Same Token Amount, Noun Only, and Pronoun Only) and test error rate for the control probe. The table also includes all error rates for the validation set, minus the control probe where there was no validation set. The results reported in the main paper are from the Random dataset.

E Dative Predicates

Benefactive/Malefactive Verbs: *helfen* 'to help', *danken* 'to thank', *gratulieren* 'to congratulate', *folgen* 'to follow', *glauben* 'to believe', *zuhören* 'to listen', *vertrauen* 'to trust', *antworten* 'to answer', *verzeihen* 'to forgive', *drohen* 'to threaten', *widersprechen* 'to contradict', *zustimmen* 'to agree', *raten* 'to recommend', *schaden* 'to harm', *dienen* 'to serve', *begegnen* 'to respond', *zureden* 'to persuade', *nachlaufen* 'to chase', *ausweichen* 'to avoid', *beisteuern* 'to contribute', *ermöglichen* 'to enable'

Ditransitives: *zuordnen* 'to allocate', *zuteilen* 'to allocate, *zuwenden* 'to bestow', *zustecken* 'to slip somebody something', *vermitteln* 'to impart', *verkaufen* 'to sell', *verleihen* 'to loan', *verdanken* 'to owe', *vermachen* 'to leave somebody something, *hinzufügen* 'to add somebody to something', *darbieten* 'to offer somebody something', *geben* 'to give', *schenken* 'to give', *bescheren* 'to give', *(ein)bringen* 'to bring', *unterstellen* 'to leave somebody something', *unterlegen* 'to underlay something with something', *überlassen* 'to transfer', *übergeben* 'to hand somebody something', *überreichen* 'to present somebody with something', *nachempfinden* 'to base something on somebody', *näherbringen* 'to make something more accessible to somebody', *entgegensetzen* 'to counter something with something', *entgegenkommen* 'to make concessions to someone', *erschweren* 'to impede', *weihen* 'to dedicate', *verschreiben* 'to prescribe', *erzählen* 'to tell somebody something', *empfehlen* 'to recommend', *versprechen* 'to promise', *mitteilen* 'to disclose', *anvertrauen* 'to entrust somebody with something', *vorwerfen* 'to accuse somebody of something', *goennen* 'to begrudge something something', *platzmachen* 'to make room', *raten* 'to recommend', *wünschen* 'to want something', *zeigen* 'to explain', *stehlen* 'to steal', *rauben* 'to steal', *entziehen* 'to deprive', *entreißen* 'to steal', *entwenden* 'to steal', *wegnehmen* 'to take something away from someone', *abnehmen* 'to take something away from someone', *vorenthalten* 'to keep something from somebody', *verweigern* 'to refuse', *verschweigen* 'to hide something from someone', *verheimlichen* 'to conceal something from someone'

Stimulus-Experiencer Verbs: *einfallen* 'to think of', *fehlen* 'to miss', *gefallen* 'to enjoy', *leidtun* 'to be sorry', *passen* 'to work/be good for someone', *anpassen* 'to fit', *schmecken* 'to like', *ergehen* 'to indulge', *ist + DP ADJ* 'somebody is something'

F Control Probes

A criticism of linguistic probing is that classifier may be learning the task and not actually showing if the language model has encoded the information. A solution to see if the classifier is just solving the task is to include a control probe (Hewitt and Liang, 2019). These kind of probes are trained on a random organisation of the training data; for us that means class 0 and class 1 being made up of a random selection of all the cases. We included three control probes: *Acc Control Probe* (class 0 = random accusatives and random datives; class 1 = random accusatives and random datives), *Nom Control Probe* (class 0 = random nominatives and random datives; class 1 = random nominatives and random datives), and *All Control Probe* (class 0 = random nominatives, random accusatives, and random datives; class 1 = random nominatives, random accusatives, and random datives). The architecture of these probes and training is the same from the classifiers in the main paper. For each model, we trained a control probe on the random training set (i.e. not on any of the control datasets), and are presenting the control probes that were trained on the layers that were presented in the paper (e.g. for *BERT* layer 11, for *mBERT* layer 10, for *LLäMmlein 1B* layer 21). We then classified the test datives. We have assumed classifying the datives as 1 as the correct answer (of course, given the organisation of the training data, this is an arbitrary assignment). We did this for the cased condition (Figure 5) and the caseless condition (Figure 6).

For the majority of the probes and models in the cased condition (Figure 5), they are at chance or around chance. However, there is a notable exception with *BERT* with the *All Probe*. For the caseless condition (Figure 6), there is once again a noticeable difference with *BERT* as well as *mBERT*. Given that we know *mBERT* was not good at the task in the main paper, we will put those controls aside to focus on *BERT*. When looking closer at the probabilities the control probe has with *BERT* in the cased *All Probe*, the caseless *Acc Probe* and *All Probe*, we can see that for both class 0 and class 1 the probabilities are close to 0.5; this is in contrast with the respective probes in the main experiments. More specifically, in the caseless experiment, *BERT* scored a similar error rate in the

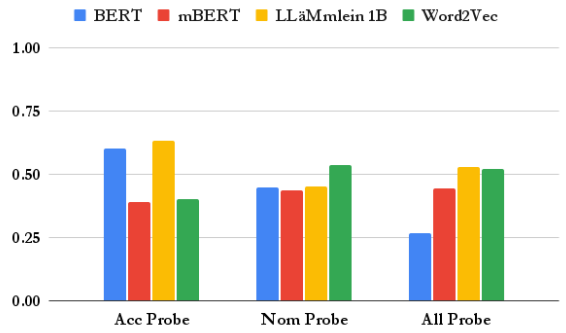

Figure 5: Control probes on the cased representation. Error rates for *BERT* layer 11, *mBERT* layer 10, *LLäMmlein 1B* layer 21, and *Word2Vec* baseline.

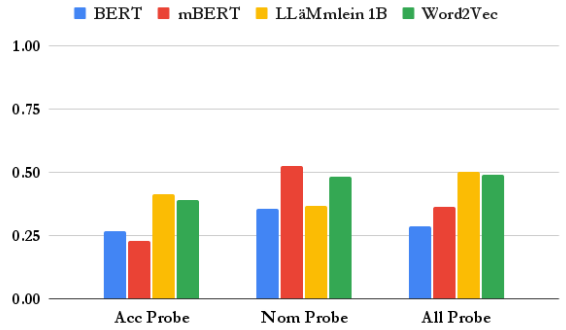

Figure 6: Control probes on the decased representation. Error rates for *BERT* layer 11, *mBERT* layer 10, *LLäMmlein 1B* layer 21, and *Word2Vec* baseline.

Nom/Acc ∼ *Dat* data configuration (Figure 4) to the caseless *All Probe* (Figure 6), however, when we look at the probabilities for class 0 and class 1 on the dative test set, we see a stark difference. The experimental probe (Figure 7) skews to assigning class 1 probabilities as around 0.9, whilst the control probe assigns class 1 around 0.5. What we can take away is that although control probes look like they are solving their task, the probabilities show that they are assigning almost a random chance for either class.

There is a strategy with categorising nominative, accusative, and dative arguments whereby you can utilise argument type information, as seen with the effect of the *pronoun only* and *noun only* datasets and how the *Word2Vec* model was able to do the caseless task relatively well with zero context and with the exact same arguments in all cases/classes (e.g. the caseless pronoun representations). Considering that the *BERT* probes in the *Acc Probe* and *All Probe* in the caseless condition and the *All Probe* in the cased condition are around an error rate of

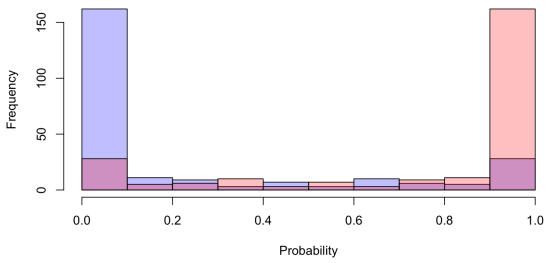

Figure 7: The probabilities for class 0 (blue) and probabilities for class 1 (red) for *BERT* in the caseless *Nom/Acc* ∼ *Dat* condition, on the dative test set.

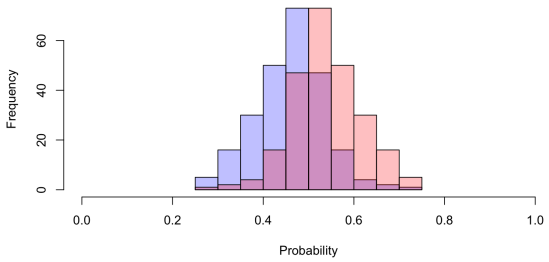

Figure 8: The probabilities for class 0 (blue) and probabilities for class 1 (red) for *BERT* in the caseless *All probe*, on the dative test set.

30% and this is the same of the *Word2Vec* model in the caseless experiment (section 5), it may be that these control probes are training on some sort of argument distribution to achieve the task. Given the probes being close to the *Word2Vec* caseless baseline, the probe's probabilities of class 0 and class 1 both being close to 50%, and the fact that *BERT* performed better in the caseless *Acc* ∼ *Dat* and *Nom/Dat* ∼ *Acc* condition compared to the *Acc Probe* and *All Probe* respectively, and performed worse in the *Nom* ∼ *Acc/Dat* compared to the *All Probe*, we can conclude that the non-control probes are not training on argument type distributions, but that this information can influence the classifiers, something that we state in our discussion.

G Frequency Effects

To test the impact of frequency, we conducted a spearman's Rank Correlation to see if there is a correlation between an argument's frequency and the probability of being assigned class 1. We only looked at *BERT* (as it was the model that performed the best) and the dative test set in the caseless condition (section 5), and we did a test for each data con-

figuration. For DPs, we only used the frequency of the nouns. We obtained the argument frequencies from the *Decow* dataset (Schäfer, 2015; Schäfer and Bildhauer, 2012), if a word was not present it was assigned a frequency of 0. For *Acc* $\sim$ *Dat*, $r_s(224) = -.282$, $p < 0.05$. For *Nom* $\sim$ *Dat*, $r_s(224) = -.189$, $p < 0.05$. For *Nom/Dat* $\sim$ *Acc*, $r_s(224) = -.323$, $p < 0.05$. For *Nom/Acc* $\sim$ *Dat*, $r_s(224) = -.206$, $p < 0.05$. For *Nom* $\sim$ *Acc/Dat*, $r_s(224) = -.351$, $p < 0.05$. Overall, each data configuration has a weak negative correlation between probability and frequency that is statistically significant. Given that it is weak, we conclude that frequency is not overly impactful to the classifiers probability.

H Pondering on Ergativity

An anonymous reviewer wondered how the impact of German being an nominative-aligned language affected the results in comparison to an ergative-aligned language. Here, a nominative language is one where the transitive subject and intransitive subject are morphologically[9] marked as the same. Whilst an ergative language is one where the intransitive subject and the transitive object are morphologically marked the same. Although it would be easy to remark that the experiment should just be re-run but on an ergative language, there are some particular difficulties with this: (i) we lack resources for ergative languages (see Papadimitriou et al. (2021) for a discussion on their difficulty on finding enough resources for ergative languages), (ii) ergative languages tend to have split systems (i.e. they are not always consistently ergative (Dixon, 1994)), and (iii) as ergative case is a dependent case (Baker, 2015), if you have an inherent dative object, the transitive subject should be marked as absolutive not ergative. That is by no means to say we should not be investigating language models' case representations of ergative languages (we absolutely should), it is to say it would be difficult to adapt *this* specific experiment to an ergative language, as it was designed specifically with German in mind. However, we can ponder upon how a language model may differently represent inherent cases in ergative languages.

Given that language models like *mBERT* encode intransitive subjects as objects in ergative languages (Papadimitriou et al., 2021), we may find that in an ergative language with an inherent dative on objects, the active dative and passivised dative will be classified as the same when the overt case markings are removed - obviously this depends on if a language model analyses passivised subjects as intransitive subjects (which are then analysed as 'objects'). The other interesting caveat concerns the dependent nature of ergative case; if an inherent dative object existed in this ergative language, then the absolutive case and dative case would not be in a complementary distribution like the accusative and dative in German, given that the normally ergative transitive subject would become absolutive when paired with a dative argument. However, the dative would be in a complementary distribution with the ergative case. This actually matches with the German, as the complementary distribution is between dependent and inherent cases, of which accusative and ergative are. So, we would expect some sort of encoding case split between ergative and an inherent dative. Presumably then, a language model would compute the case of the 'object' first, and if there is said X of verbs then dative would be picked over absolutive. However, unlike the German language model, there would need to be some extra information for absolutive to come about on the transitive subject, instead of predicting the ergative case. Of course, this is just a hypothesis, and an undeveloped one at that given this is not encoding the split nature of ergative systems. However, ergative case representation should absolutely be investigated with reference to this question, but with an experiment that is specifically designed for it.

[9] We are ignoring syntactic ergativity and accusativity