Targeted Syntactic Evaluation of Language Models on Georgian Case Alignment

Daniel Gallagher and **Gerhard Heyer**
Institute for Applied Informatics (InfAI), Leipzig
gallagher@infai.org, heyer@infai.org

Abstract

This paper evaluates the performance of transformer-based language models on split-ergative case alignment in Georgian, a particularly rare system for assigning grammatical cases to mark argument roles. We focus on *subject* and *object* marking determined through various permutations of *nominative*, *ergative*, and *dative* noun forms. A treebank-based approach for the generation of minimal pairs using the Grew query language is implemented. We create a dataset of 370 syntactic tests made up of seven tasks containing 50–70 samples each, where three noun forms are tested in any given sample. Five encoder- and two decoder-only models are evaluated with word- and/or sentence-level accuracy metrics. Regardless of the specific syntactic makeup, models performed worst in assigning the ergative case correctly and strongest in assigning the nominative case correctly. Performance correlated with the overall frequency distribution of the three forms (NOM > DAT > ERG). Though data scarcity is a known issue for low-resource languages, we show that the highly specific role of the ergative along with a lack of available training data likely contributes to poor performance on this case. The dataset is made publicly available and the methodology provides an interesting avenue for future syntactic evaluations of languages where benchmarks are limited.

1 Introduction

Small and Large Language models (LMs) have led to a paradigm shift in how we consider the learning of language by a machine. A natural question to ask has been: *what understanding have they of the structure of the language they're trained on?* A standard approach to tackling this question has been the use of *minimal pairs*, that is, grammatical/ungrammatical sentence pairs differing by a single syntactic feature. A significant effort has been put into compiling benchmarks for evaluating models on such pairs, however much of this has focused on a small subset of the world's approximately 7,000 languages. This has exacerbated a recognised dearth of research on *low-resource languages* (LRLs), typically defined as languages that are under-studied, less commonly taught, or for which relatively few online resources exist (Singh, 2008; Cieri et al., 2016). We consider Georgian one such language due to a significant lack of relevant research in this domain. This presents an opportunity however to analyse the rarer aspects of its syntax.

In this work we evaluate one such aspect by looking at its split-ergative case-alignment system, exhibited in the same manner by only three other languages according to the *World Atlas of Language Structures* (WALS; Comrie (2013)). This revolves around three combinations of subject–object marking depending on verb tense/aspect/mood: nominative–dative, ergative–nominative, and dative–nominative. We use the Grew query language and Georgian Language Corpus Universal Dependencies treebank (GLC UD; Lobzhanidze et al. (2024)) to generate minimal sets of syntactic tests. The resulting 370 case-alignment tests mark one of the first such datasets for Georgian[1] and an evaluation of seven LMs reveals that models particularly struggle with the ergative case. Many LRLs suffer from a lack of available syntactic benchmarks and the approach used here may be generalised to other languages and phenomena where there is a treebank available.

---

[1]The full dataset is available on Hugging Face at DanielGallagherIRE/georgian-case-alignment.

2 Related Work

Transformer-based architectures (Vaswani et al., 2017) have become the standard for the training of high-performing language models and introduced a new wave of research in the field of Natural Language Processing (NLP). It appears that elements of syntax have been encoded by *some* means, however our understanding of their sense of grammaticality is still limited (Kulmizev and Nivre, 2022). There have been a variety of methods used to evaluate linguistic knowledge in LMs such as probing classifiers (Conneau et al., 2018; Hewitt and Manning, 2019; Giulianelli et al., 2018) or psycholinguistic tests (Futrell et al., 2019). Targeted Syntactic Evaluation (TSE) was introduced by Marvin and Linzen (2018) and brought a focus on minimal-pair acceptability judgements into the LM domain, where a minimal pair is defined as a set of two items with a single differing feature. In this context, the term is used to distinguish between a grammatical and ungrammatical sentence. Example 1 shows agreement across a subject-relative clause with test *run/runs*.

(1)   The officers that chased the thief __.

$p(run \mid \text{context}) > p(runs \mid \text{context})$ ✓
$p(run \mid \text{context}) \leq p(runs \mid \text{context})$ ✗

The *Corpus of Linguistic Acceptability* (CoLA) was released soon after containing 10,657 sets of grammaticality judgements (Warstadt et al., 2019) and incorporated into the General Language Understanding Evaluation benchmark (GLUE; Wang et al. (2018)). Newman et al. (2021) refined the goals of TSE further by focusing not just on how LMs fared with a single pair for a given sentence, but also expanded to include a wider variety of pairs beyond the most likely. It was observed that models could typically assign the correct conjugation to the expected verb, such as the distinction in Example 1, but struggled when it came to verbs that were less frequent despite the same grammatical/ungrammatical distinction e.g. *rest/rests*. This provided evidence that models can have high syntactic performance on some tests without having generalised for that phenomenon across the language. Further work has been carried out on evaluating whether models have generalised syntactic phenomena across languages (Someya et al., 2024; Hu et al., 2020).

2.1 Linguistic Minimal Pairs

Minimal-pair syntactic tests have become the standard for evaluating model performance on syntactic phenomena and this ethos has resulted in a concerted effort towards the creation of standardised minimal-pair benchmarks. The *Benchmark of Linguistic Minimal Pairs* (BLiMP; Warstadt et al. (2020)) compiled and published 67 datasets at 1000 sentences each for evaluating model performance on English syntax, remaining one of the most important datasets in the TSE domain. Similar benchmarks have been compiled for other languages such as German (Zaczynska et al., 2020), Spanish (Bel et al., 2024), Chinese (Song et al., 2022; Wang et al., 2021), Russian (Taktasheva et al., 2024), Japanese (Someya and Oseki, 2023), and Dutch (Suijkerbuijk et al., 2025).

2.2 Low-Resource Languages

Work in the syntactic evaluation of LRLs has been limited. There has however been a notable buildup of momentum in the creation of BLiMP-style datasets, though with some being much smaller and covering fewer syntactic categories. These include datasets for Turkish (Baar et al., 2025), Swedish (Volodina et al., 2021), Icelandic (Zhang et al., 2024), and more recently Irish (McGiff et al., 2025). Kryvosheieva and Levy (2025) evaluated multilingual models on phenomena in three different languages: Swahili noun-class agreement, Hindi split ergativity, and Basque verb agreement. Performance correlated with how well a language was represented in the training data, with models performing strongest on Hindi. Leong et al. (2023) focused on Southeast Asian languages and introduced a test suite for the evaluation of LMs on Tamil and Indonesian. A significant recent contribution is that of MultiBLiMP (Jumelet et al., 2025), covering 101 languages with all data stemming from UD treebanks. They include Georgian as one of the languages including subject-verb and subject-participle agreement for both `number` and `person`, however they do not examine case alignment. Furthermore, they do not use a

Alignment	Screeve	Subject	Object
Accusative	PRES/FUT	NOM	DAT
Ergative	PAST	ERG	NOM
Inverted	PERF	DAT	NOM

Table 1: Outline of the 3 Georgian alignments, their correspondings tense–aspect–mood category (i.e. screeve) triggers, and the resulting cases assigned to the subject and object in both transitive and intransitive constructions.

query language to match specific syntactic constructions but rather implement this programmatically. In this work we will take a query-based approach using Grew, a graph-based query language that can identify syntactic patterns in treebanks (Guillaume, 2021, 2019).

3 Georgian Syntax

Georgian is a Kartvelian language of the southern Caucasus syntactically characterised by its derivational morphology as well as its split-ergative case alignment system (Bolkvadze and Kiziria, 2023), the latter being discussed in detail in Section 3.2. Modern Georgian uses the *Mkhedruli* script written left-to-right and without distinction between upper and lower case (Tuite, 1996, p. 4). An important aspect of this language's verb paradigm is that it does not distinguish clearly between *tense*, *aspect*, and *mood*, instead categorising them into distinct properties known as *screeves* (Amiridze, 2006). Henceforth we will refer to these tense–aspect–mood combinations using this term.

The typical properties of an agglutinative language are exhibited with words built from an inventory of largely invariable morphemes (Plungian, 2001). Unique suffixes represent any of seven noun cases as well as singular/plural marking that are appended to corresponding root forms (Tuite, 1996, p. 33). A large variety of word forms are thus able to be produced. Only the nominative (NOM), ergative (ERG), and dative (DAT) cases pertain to argument structure, therefore they will be our focus. How these cases are mapped to the argument roles determines the case-alignment system. We focus here solely on subject–object relations and will sometimes refer to particular alignments using X–Y notation, where X represents the subject's case and Y the object's.

3.1 Accusativity & Ergativity

The most common case-alignment system is *accusative* (Comrie, 2013), where the nominative marks the subject of both transitive and intransitive verbs (e.g. *he* runs, *he* helps the boy) and the accusative marks the object of transitive verbs (e.g. he helps *him*). Georgian however has no true accusative case and so the dative acts in its stead. In the screeves corresponding closely to the *present* and *future* tense we observe a nominative subject and dative object. Uniquely, in the *perfective* screeve e.g. "have done something" (though requiring evidence of completion), these undergo inversion (Harris, 1981, ch. 8) and the subject is marked as dative and the object as nominative. On the other hand, in the screeve corresponding to the *past* tense, ergativity is observed. The key feature of this alignment is that a special case is assigned to the subject of transitive verbs. The subject of intransitive verbs and the object of transitive verbs thus take the same case, often referred to as the *absolutive* case (Dixon, 1994) though we label it simply as the nominative. Table 1 shows an overview of the various alignments along with their corresponding screeves and assigned cases.

3.2 Split Ergativity

Ergative languages are often only partially ergative; under certain conditions they follow an ergative system and under others an accusative system (Coon, 2013). If a language exhibits a mixture of alignments including at least partial ergativity, as is the case for Georgian, then this language typically falls under the category of *split-ergative* (DeLancey, 1981). Screeves and their alignments are typically grouped into sets of 'series' that are labelled I, II, and III (Harris, 1981, ch. 1), however we will refer to the exact subject–object case alignments instead of individual series for simplicity. We leave the incorporation of further screeves as valuable future work. The full case-alignment system with corresponding examples is shown in Figure 1. The noun ბავშვ 'bavshv' meaning 'child' takes, for instance, three different cases when acting as the subject depending on the screeve: NOM: -ი '-i' , ERG: -მა '-ma' , or DAT: -ს '-s'. Example glosses are provided in Example 2.
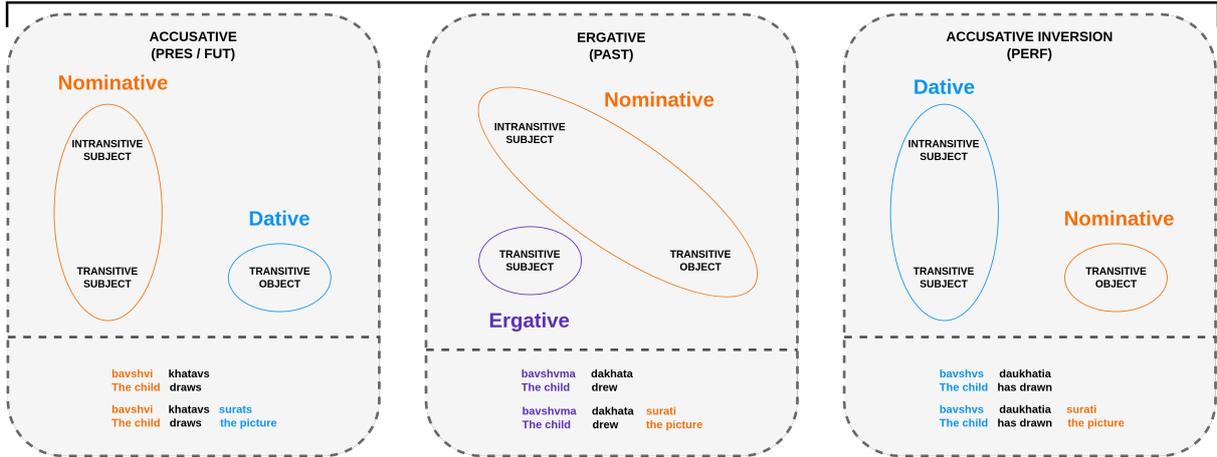
Figure 1: The Georgian split-ergative case system, made up of accusative, ergative, and inverted alignments. Examples are provided top-to-bottom for intransitive and transitive constructions. In most languages, including English, there is solely accusative alignment.

(2) a. **Present**   (NOM)
   *Bavshv-**i**   tchams*
   child-NOM   eat.3SG.PRES

   "The child eats."

  b. **Present**   (NOM–DAT)
   *Mama   sakhl-**s***
   father.NOM   house-DAT

     ashenebs
     build.3SG.PRES

   "Father builds a house."

  c. **Past**   (ERG–NOM)
   *Mama-**m**   sakhl-**i***
   father-ERG   house-NOM

     aashena
     build.3SG.PST

   "Father built a house."

  d. **Perfective**   (DAT–NOM)
   *Mama-**s**   sakhl-**i***
   father-DAT   house-NOM

     aushenebia
     build.3SG.PERF

   "Father has built a house (evidentially)."

We can see from Figure 1 that the same case may be used for a wide variety of syntactic or semantic roles and some naturally tend to occur more frequently than others. The Georgian nominative case, for instance, appears 11,438 times in the GLC UD treebank, while the dative 10,034 times and the ergative only 475 times. The nominative case thus occurs over 24 times as often as the ergative. This trend may persist in Georgian text generally.

4   Data

We use Grew, a graph-based query language, for the creation of minimal-pair syntactic tests from the GLC UD treebank which contains 3,013 annotated sentences. Example 2 shows the queries used for a sample of the datasets. We search the treebank for four different forms of subject–object constructions: (1) intransitive nominative (2) transitive nominative–dative, (3) transitive ergative–nominative, and (4) transitive dative–nominative. These correspond respectively to accusative (for both (1) and (2)), ergative, and inverted alignments as discussed in Section 3.

All data is in the Mkhedruli Georgian script as this is the expectation of model tokenisers. Table 2 shows an outline of the datasets, with separate sets for testing the subject and object in transitive constructions resulting in a total of seven datasets of 50–70 samples each totalling 370 case-alignment tests. Each test sentence is accompanied with a target noun provided in the nominative, ergative, and dative form. Additional ergative forms were created and validated by a human annotator in order to reach at least 50 in any given set.

Test Set	Subject	Object	Relation Tested	# Tests
`intransitive-nom-subj`	Nominative		Subject	70
`transitive-nom-dat-subj`	Nominative	Dative	Subject	50
`transitive-nom-dat-obj`	Nominative	Dative	Object	50
`transitive-erg-nom-subj`	Ergative	Nominative	Subject	50
`transitive-erg-nom-obj`	Ergative	Nominative	Object	50
`transitive-dat-nom-subj`	Dative	Nominative	Subject	50
`transitive-dat-nom-obj`	Dative	Nominative	Object	50

Table 2: Overview of the 7 Georgian case-alignment datasets from the GLC UD treebank totalling 370 tests. Table contains the dataset name, subject–object relations, relation tested, and number of tests.

```
// Transitive Erg-Nom
pattern {
  V [upos="VERB"];
  SUBJ [Case="Erg"];
  OBJ [Case="Nom"];
  V -[nsubj]-> SUBJ;
  V -[obj]-> OBJ;
}

// Intransitive Nom
pattern {
    V [upos="VERB"];
    SUBJ [Case="Nom"];
    V -[nsubj]-> SUBJ;
}
without {
    V [upos="VERB"];
    V -[nsubj]-> SUBJ;
    V -[obj]-> OBJ;
}
```

Figure 2: Grew queries to match sentences in the GLC UD treebank that are transitive ERG-NOM and intransitive NOM constructions.

5 Methodology

We propose a query-based approach for generating syntactic tests from UD treebanks through the use of the Grew query language [2]. The usage of treebanks allows us to take advantage of (1) dependency relations between words for syntax specification, and (2) syntactic annotations. The latter allows us to find specific word forms and adjust them by a single morphosyntactic feature to create minimal pairs e.g. NOM *bavshvi* (lit. 'child')→ERG "*bavshvma*". Furthermore, the usage of a query language allows the intuitive specification of syntactic constructions. This approach relies heavily on the richness of the UD treebank in question, with some constructions and word

forms naturally occurring less often than others, thus placing a limit on the number of syntactic tests that can be generated. In the case of Georgian, nominative and dative nouns occur frequently and ergative forms less so, therefore a native speaker provided assistance in creating additional ergative noun forms where necessary. Each test contains a masked sentence along with a nominative, dative, and ergative form of the same noun. Example tests with the three forms for transitive accusative and ergative alignments are shown in Example 3.

(3) a. *bavshvi*✓/*bavshvma*✗/*bavshvs*✗
       child.NOM/child.ERG/child.DAT

    khatavs   surats
    draws   picture

    'The child draws a picture.'

  b. *bavshvi*✗/*bavshvma*✓/*bavshvs*✗
       child.NOM/child.ERG/child.DAT

    dakhata   surati
    drew   picture

    "The child drew a picture."

5.1 Minimal Sets

The creation of a minimal pair using a treebank-generated lexicon requires (1) a base word form as it appears in text and (2) a single feature adjustment. In a typical TSE experiment configuration, two word forms are evaluated against each other, however we create a minimal set of three noun forms to account for all cases, one valid and two invalid in a given context. We thus make two feature adjustments, e.g. NOM→{DAT,ERG}.

---

[2]Dataset creation and evaluation code is available on GitHub at DanielGall500/georgian-case-alignment.

5.2 Metrics

The standard unit of measuring syntactic performance is *accuracy*, defined as the proportion of answers that a model gets correct. Typically, accuracy relates to discrete or categorical values. The value with the highest probability is taken to be the final prediction $\hat{y}_i$ and compared with the ground truth y_i. We are not however comparing one discrete label with another but rather probabilities. For a given grammatical–ungrammatical pair (w_G, w_{UG}), a model is deemed correct if it assigns a higher probability to w_G than to w_{UG} given a context C (Marvin and Linzen, 2018). Accuracy is defined in Equation 1 as applied across all tests.

$$\frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\Big[P(w_G \mid C) > P(w_{UG} \mid C) \Big] \quad (1)$$

Measuring accuracy in this manner may not account for the highest-probability token from a model inference as is otherwise typical. Rather, it compares any target words that have been pre-defined for the minimal pair. In our case we extrapolate this to the minimal *set* of three forms where the grammatical item w_G must be assigned a higher probability than the two ungrammatical items w_{UG_1} and w_{UG_2}.

5.2.1 Word- versus Sentence-level

Comparing the likelihoods of words or sentences involves calculating the joint probabilities of their sequences of tokens. We implement two distinct approaches to the calculation of joint probabilities: *sentence-level* (SL) and *word-level* (WL). To calculate SL we take the joint probability over entire tokenised sentences auto-regressively. This can result in extremely low probabilities for long sentences, thus we convert this metric to a logarithmic scale to alleviate this issue. An additional problem with this metric is that the ungrammatical item may be present in the chain of probability multiplications long after the target word, causing unknown downstream effects on the results. WL evaluations, on the other hand, measure certainty at the exact moment that a grammatical/ungrammatical item is chosen by taking the joint probability of the tokenised target word only, given the full context.

5.2.2 Why use one or the other?

WL can be more representative of performance as it measures certainty at the exact moment that a model chooses an item in the minimal pair, whereas SL measures this across an entire tokenised sentence. However, in models trained on next-token prediction (i.e. decoder-only), WL only make sense if the grammaticality of a minimal pair is determined linearly *earlier* in the sentence. Decoder-only models *cannot* see the full context if the target is not the last word. For instance, in Georgian the verb often appears at the end of a sentence but the noun case is marked linearly earlier. SL metrics do not suffer from this as the full context is taken into account and thus ambiguities are limited. If we include decoder-only models, we should include SL evaluations to some extent despite WL being more representative of performance.

6 Experiments

We evaluate seven LMs, with encoder-only models evaluated on both WL and SL accuracy and decoder-only solely on SL accuracy for the reasons discussed in Section 5.2.2. The language marker for Georgian is 'ka' or 'geo'. The encoder-only models selected are XLM-RoBERTa base and large (Conneau et al., 2020), BERT multilingual base (Devlin et al., 2018), RemBERT (Chung et al., 2020), and HPLT-BERT-ka (Samuel et al., 2023; de Gibert et al., 2024). All are multilingual except for the HPLT model which is trained solely on Georgian. The decoder-only models are GPT2-geo (Kuduxaaa, 2025) and mGPT-ka (Shliazhko et al., 2024; AI-Forever, 2025), both models fine-tuned on Georgian though with the latter additionally fine-tuned on English and Russian. Models range from 110M–1.3B parameters and 1–104 languages. They use a variety of tokenisation algorithms consisting of WordPiece (Schuster and Nakajima, 2012) for mBERT and HPLT-BERT-ka, SentencePiece (Kudo and Richardson, 2018) for RemBERT and the XLM-R models, and Byte-Pair Encoding (BPE; Sennrich et al. (2016)) for the GPT models.

| | | | NOM | NOM−DAT | | ERG−NOM | | DAT−NOM | |
Model	Size	#langs	SUBJ	SUBJ	OBJ	SUBJ	OBJ	SUBJ	OBJ
				Word-Level Accuracy (%)					
XLM-RoBERTa(bs)	270M	94	83	96	52	26	84	56	96
XLM-RoBERTa(lg)	550M	94	86	94	56	34	94	54	98
HPLT-BERT-ka	110M	1	97	100	70	40	96	74	100
mBERT	110M	104	100	92	42	8	98	36	98
RemBERT	559M	104	59	72	20	14	72	28	56
				Sentence-Level Accuracy (%)					
XLM-RoBERTa(bs)	270M	94	50	54	18	14	52	24	62
XLM-RoBERTa(lg)	550M	94	56	48	18	18	50	20	62
HPLT-BERT-ka	110M	1	66	54	16	18	54	28	52
mBERT	110M	104	71	68	10	14	84	14	80
RemBERT	559M	104	61	62	30	26	70	24	60
GPT2-geo	124M	1	69	68	38	16	70	34	66
mGPT-ka	1.3B	3	97	100	100	84	98	78	100

Table 3: Accuracy scores for all models across tasks. X−Y indicates the subject–object makeup and the tested role is underneath. Model performance is highest on the nominative regardless of the task and most models perform particularly poorly on assigning the ergative case.

7 Results

The resulting WL and SL accuracy scores are shown in Table 3. The cases assigned are indicated in the X–Y format and the SUBJ or OBJ label indicates which role was tested. Figure 3 shows box plots with the average probability assigned to each case for each dataset, listed in the same order top-to-bottom as in Table 3. Note that this shows the results for the word-level only as they are more comparable due to fewer probability multiplications. The results show the strongest overall performance on assigning the nominative correctly, followed by notably poorer performance on dative and ergative. Across all tasks that tested the correct assignment of the nominative, the average word- and sentence-level accuracy was 88.6% and 67.3%, respectively. This was followed by 48.8% and 32.3% for the dative, and 24.4% and 27.1% for the ergative. The majority of models perform poorest on the ergative. Similarly, Figure 3 shows a dramatic reduction in model certainty for WL evaluations when evaluating the grammatical ergative form. On the other hand, models assign a much higher average probability to the correct form for tasks that test the nominative and dative. This effect persists regardless of the syntactic task.

7.1 Discussion

The strongest performance across tasks is observed from the mGPT-ka decoder-only model, notably the largest tested. High model accuracy alone however tells us little of the generalisation of a syntactic feature due to issues with testing on unseen data. Poor performance, on the other hand, *does* indicate a meaningful gap in knowledge. We thus see the almost universally poor performance assigning the ergative, discussed further in Section 7.2, as of particular interest in these results. We additionally observe a notable nominative bias that extends across *all* tasks, indicating an effect that is beyond simple memorisation. We will explore this in Section 7.3.

The frequency distribution of Georgian case approximately follows NOM > DAT > ERG (see Section 3). There is an interesting correlation between this general frequency of cases and the word-level accuracy, sentence-level accuracy, and probability results: NOM (WL: 88.6%, SL: 67.3%, WL P(X): 0.03) > DAT (WL: 32.3%, SL: 36.6%, WL P(X): 0.029) > ERG (WL: 24.4%, SL: 27.1%, WL P(X): 0.008). These results indicate a possible relationship between the general frequency of syntactic constructions and their learnability in data-scarce contexts.
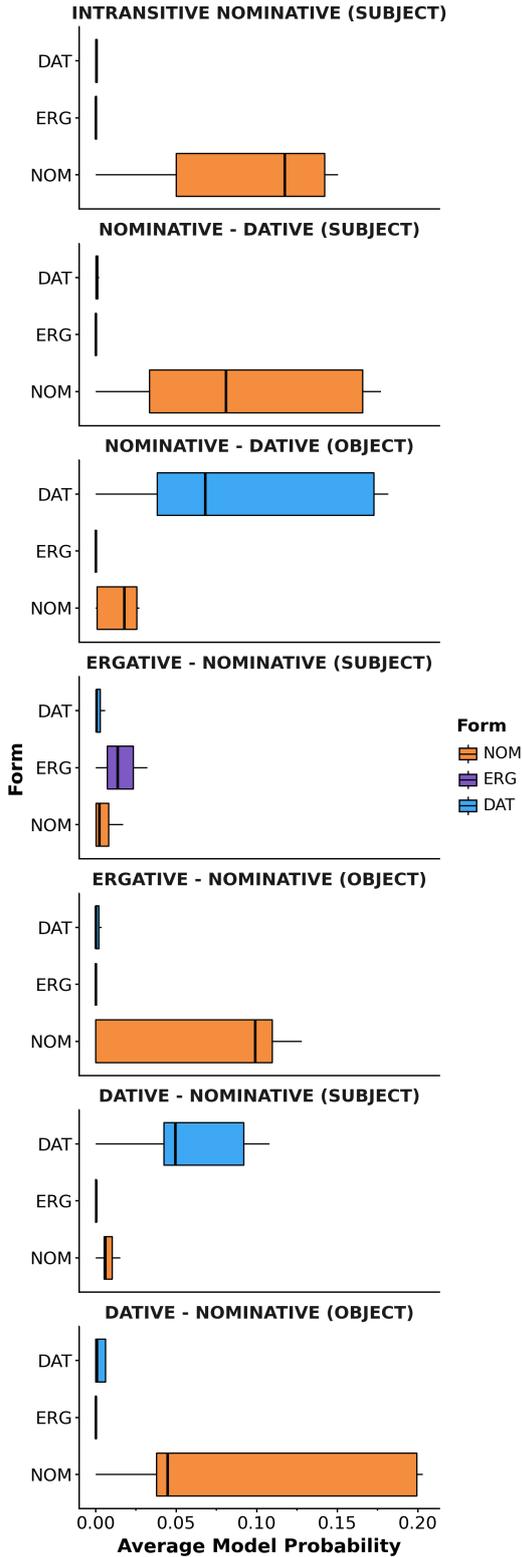
Figure 3: The average word-level probability assigned to each case represented as box plots. In all cases, the highest average probability is assigned to the correct grammatical case. Most significantly however, the average probability is significantly lower where the model must assign the ergative case correctly.

TSE results can be difficult to interpret (see Section 2) and we have included multiple accuracy metrics to alleviate this. We will however focus primarily on word-level metrics to identify patterns in the results due to their more direct measurement of model preference (see Section 5.2.1). We will additionally assess the possible impact of tokenised word or sentence length on performance in Section 7.4.

7.2 Models Struggle with Ergativity

A key finding is that models tend to struggle most with correctly assigning the ergative case. The nominative and dative are used for a wider variety of purposes and naturally occur more frequently, while the ergative is used in very specific cases (sometimes defined as a more *marked* feature (Andrews, 1990)). There is thus a natural reduction in instances of the ergative case in the training data, strained further by limited data for this language generally. This leads us to a situation where models may easily over-generalise and thus fail to learn the correct triggers for the ergative case, namely, the usage of the perfective screeve. Furthermore, the rarity of this case-alignment system across languages likely means that there is limited scope for improvement through transfer learning. Data scarcity is a well-understood issue for training models on LRLs (Chang et al., 2024; McGiff and Nikolov, 2025), however here we have isolated its effects on the learning of syntax to a single syntactic feature.

7.3 Nominative Bias

There is conversely a clear default bias towards the nominative case regardless of syntactic makeup. In cases where the dative was the correct choice but models chose wrongly, the nominative was preferred 83.2% of the time with only minor deviation among datasets (σ=2%). In failed tests for the ergative case, the nominative was preferred an average of 74.4% of the time. Lastly, where the nominative was the correct choice but the model chose wrongly, the dative was preferred over the ergative an average of 79% of the time with limited deviation across the four datasets (σ=3%). Even where models did not choose the nominative as they should have, the dative is a secondary default and the ergative is rarely preferred.

7.4 Assessing Tokenisation Impact

Agglutinative languages such as Turkish or Georgian provide particular challenges for tokenisation algorithms due to the building of particularly complex word forms (Altnta et al., 2025). However, we found that their effect on the end result was in this case limited and, furthermore, we did not find a qualitative word or sentence type that correlated with poor performance. Models used a variety of tokenisation approaches (see Section 6) and the average token length of the target word for the WL results differed significantly: mBERT – 4.1 > RemBERT – 2.3 > XLM-R(lg) – 2.2 > XLM-R(bs) – 2.1, > HPLT-BERT-ka – 1.4. Upon performing a Pearson correlation analysis between model performance and both tokenised (a) target word length, and (b) sentence length, we found no statistically significant correlations for the nominative (a. $r=-0.34$, $p=0.21$; b. $r=0.04$, $p=0.45$), the dative (a. $r=-0.24$, $p=0.2$; b. $r=-0.04$, $p=0.53$), or the ergative (a. $r=0.03$, p=0.34; b. $r=0.07$, $p=0.42$) tasks. These findings partially alleviate the limitation that we did not control for sentence length and point further toward the rarity of the ergative and general data scarcity for Georgian as a key issue.

8 Conclusion

In this work we created a dataset for the syntactic evaluation of Georgian case alignment and evaluated seven LMs trained on Georgian text. Performance correlated with the distribution of case seen within the language: NOM > DAT > ERG. The low frequency of the ergative case, its highly specific use case, as well as an overall lack of available training data make this a particularly difficult syntactic phenomenon for models to learn. We found that models tended to default to using the nominative regardless of the syntactic construction as shown by an error analysis of model preferences. Furthermore, a Pearson correlation analysis revealed that tokenised word and sentence length did not appear to be strong indicators of performance. The dataset, primarily generated from the GLC UD treebank, contributes towards a syntactic benchmark for Georgian. The approach used may also be generalised to other languages for which a treebank exists. The incorporation into the dataset of additional screeves as well as a broader range of intransitive tests would be a fruitful avenue for future work.

Limitations

A treebank-based approach to the creation of minimal pairs comes with the caveat that any generated tests have possibly been seen in the training data, leading to difficulties in properly evaluating the generalisation of syntactic features or performance on unseen data. Furthermore, an upper limit is placed on the number of syntactic tests that can be generated for a given feature that is determined by the treebank size. Dataset size was thus a limiting factor in this work. Due to these limitations we additionally did not control for sentence length, though this was shown not to have major implications in Section 7.4. There are additional difficulties in the interpretation of the results due to tokenisation biases and context ambiguities.

Acknowledgments

References

AI-Forever. 2025. mGPT-1.3B-Georgian: Georgian Language Model. https://huggingface.co/ai-forever/mGPT-1.3B-georgian. Accessed: 2025-12-26.

Gül Sena Altnta, Malikeh Ehghaghi, Brian Lester, Fengyuan Liu, Wanru Zhao, Marco Ciccone, and Colin Raffel. 2025. Toksuite: Measuring the impact of tokenizer choice on language model behavior. *Preprint*, arXiv:2512.20757.

Nino Amiridze. 2006. *Reflexivization Strategies in Georgian*. Ph.D. thesis, Utrecht University, Utrecht, The Netherlands.

Edna Andrews. 1990. *Markedness theory*. Duke University Press.

Ezgi Baar, F.P Padovani, Jaap Jumelet, and Arianna Bisazza. 2025. Turblimp: A turkish benchmark of linguistic minimal pairs.

Núria Bel, Marta Punsola, and Valle Ruiz-Fernández. 2024. EsCoLA: Spanish corpus of linguistic acceptability. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6268–6277, Torino, Italia. ELRA and ICCL.

Tinatin Bolkvadze and Dodona Kiziria. 2023. *Georgian: A comprehensive grammar*. Routledge.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2024. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096, Miami, Florida, USA. Association for Computational Linguistics.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. *Preprint*, arXiv:2010.12821.

Christopher Cieri, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. 2016. Selection criteria for low resource language programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549, Portorož, Slovenia. European Language Resources Association (ELRA).

Bernard Comrie. 2013. Alignment of case marking of full noun phrases (v2020.4). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jessica Coon. 2013. *Aspects of split ergativity*. Oxford University Press.

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128, Torino, Italia. ELRA and ICCL.

Scott DeLancey. 1981. An interpretation of split ergativity and related patterns. *Language*, 57(3):626–657.

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.

R.M.W. Dixon. 1994. *Ergativity*. Cambridge Studies in Linguistics. 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.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium. Association for Computational Linguistics.

Bruno Guillaume. 2019. Graph Matching for Corpora Exploration. In *JLC 2019 - 10èmes Journées Internationales de la Linguistique de corpus*, Grenoble, France.

Bruno Guillaume. 2021. Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion. In *EACL 2021 - 16th conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, Kiev/Online, Ukraine.

A.C. Harris. 1981. *Georgian Syntax: A Study in Relational Grammar*. Cambridge Studies in Linguistics. Cambridge University Press.

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

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2025. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Preprint*, arXiv:2504.02768.

Daria Kryvosheieva and Roger Levy. 2025. Controlled evaluation of syntactic knowledge in multilingual language models. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 402–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Kuduxaaa. 2025. GPT-2Geo: Georgian Language Model. https://huggingface.co/Kuduxaaa/gpt2-geo. Accessed: 2025-12-26.

Artur Kulmizev and Joakim Nivre. 2022. Schrödinger's treeon syntax and neural language models. *Frontiers in Artificial Intelligence*, Volume 5 - 2022.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. Bhasa: A holistic southeast asian linguistic and cultural evaluation suite for large language models. *Preprint*, arXiv:2309.06085.

Irina Lobzhanidze, Erekle Magradze, Svetlana Berikashvili, Anzor Gozalishvili, and Tamar Jalaghonia. 2024. Building a Universal Dependencies treebank for Georgian. In *Proceedings of the 22nd Workshop on Treebanks and Linguistic Theories (TLT 2024)*, pages 40–45, Hamburg,Germany. Association for Computational Linguistics.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Josh McGiff and Nikola S. Nikolov. 2025. Overcoming data scarcity in generative language modelling for low-resource languages: A systematic review. *Preprint*, arXiv:2505.04531.

Josh McGiff, Khanh-Tung Tran, William Mulcahy, Dáibhidh Ó Luinín, Jake Dalzell, Róisín Ní Bhroin, Adam Burke, Barry O'Sullivan, Hoang D. Nguyen, and Nikola S. Nikolov. 2025. Irish-blimp: A linguistic benchmark for evaluating human and language model performance in a low-resource setting. *Preprint*, arXiv:2510.20957.

Benjamin Newman, Kai-Siang Ang, Julia Gong, and John Hewitt. 2021. Refining targeted syntactic evaluation of language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3710–3723, Online. Association for Computational Linguistics.

Vladimir Plungian. 2001. Agglutination and flection. In Martin Haspelmath and 1 others, editors, *Language Typology and Language Universals: An International Handbook*, volume 1, pages 669–678. Mouton de Gruyter, Berlin.

David Samuel, Andrey Kutuzov, Lilja Øvrelid, and Erik Velldal. 2023. Trained on 100 million words and still in shape: BERT meets British National Corpus. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1954–1974, Dubrovnik, Croatia. Association for Computational Linguistics.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mGPT: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

Anil Kumar Singh. 2008. Natural language processing for less privileged languages: Where do we come from? where are we going? In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

Taiga Someya and Yohei Oseki. 2023. JBLiMP: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, Dubrovnik, Croatia. Association for Computational Linguistics.

Taiga Someya, Ryo Yoshida, and Yohei Oseki. 2024. Targeted syntactic evaluation on the Chomsky hierarchy. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15595–15605, Torino, Italia. ELRA and ICCL.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. 2022. SLING: Sino linguistic evaluation of large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4606–4634, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L. Frank. 2025. Blimp-nl: A corpus of dutch minimal pairs and acceptability judgments for language model evaluation.

Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. 2024. RuBLiMP: Russian benchmark of linguistic minimal pairs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9268–9299. Association for Computational Linguistics.

Kevin Tuite. 1996. B. g. hewitt. georgian: A structural reference grammar. *Functions of Language*, 3.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 60006010, Red Hook, NY, USA. Curran Associates Inc.

Elena Volodina, Yousuf Ali Mohammed, and Julia Klezl. 2021. DaLAJ – a dataset for linguistic acceptability judgments for Swedish. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 28–37, Online. LiU Electronic Press.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman.

2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Yiwen Wang, Jennifer Hu, Roger Levy, and Peng Qian. 2021. Controlled evaluation of grammatical knowledge in Mandarin Chinese language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5604–5620, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Karolina Zaczynska, Nils Feldhus, Robert Schwarzenberg, Aleksandra Gabryszak, and Sebastian Möller. 2020. Evaluating german transformer language models with syntactic agreement tests. *Preprint*, arXiv:2007.03765.

Ziyin Zhang, Yikang Liu, Weifang Huang, Junyu Mao, Rui Wang, and Hai Hu. 2024. MELA: Multilingual evaluation of linguistic acceptability. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2658–2674, Bangkok, Thailand. Association for Computational Linguistics.