A linguistically-informed comparison between multilingual BERT and language-specific BERT models: The case of differential object marking in Romanian

Maria Tepei
University of Amsterdam
m.tepei@uva.nl

Jelke Bloem
University of Amsterdam
j.bloem@uva.nl

Abstract

Current linguistic challenge datasets for language models focus on phenomena that exist in English. This may lead to a lack of attention for typological features beyond English. This is particularly an issue for multilingual models, which may be biased towards English by their training data and this bias may be amplified if benchmarks are also English-centered. We present the syntactically and semantically complex language phenomenon of Differential Object Marking (DOM) in Romanian as a challenging Masked Language Modelling task and compare the performance of monolingual and multilingual models. Results indicate that Romanian-specific BERT models perform better than equivalent multilingual one in representing this phenomenon.[1]

1 Introduction

Increasingly high benchmark scores achieved by recent large language models have led to discussion as to whether models need more difficult or 'creative' testing tasks to reveal their weak points (Cífka and Liutkus, 2023). While benchmarks have become increasingly complex (MMLU, Hendrycks et al. (2020); Big-Bench HARD, Srivastava et al. (2023); GLUE, Wang et al. (2018); and many others) and state-of-the-art language models perform well on them, real-world observations, as well as recent literature (e.g. Rauh et al., 2022), reveal mismatches between benchmark and real-world performance and suggest that language models do not generalise as well as the benchmark scores might imply. This has motivated benchmarks and evaluation studies that examine the linguistic capabilities of language models at a more detailed level.

As a prominent example, the BLiMP benchmark (Warstadt et al., 2020) contains minimal pairs of very similar sentences where one is grammatical and the other is not, based on linguist-crafted grammar templates. Models are evaluated by comparing probabilities assigned to the paired sentences. In recent LLM evaluation studies, authors have focused on specific phenomena that have been extensively studied by linguists, such as the dative alternation, investigating whether models are able to accurately predict (Yao and Todd, 2024) or encode (Veenboer and Bloem, 2023) these constructions in order to gain insight into the models' linguistic capabilities. However, these efforts are typically focused on English, on phenomena that occur in various languages including English such as negative polarity items (Bylinina and Tikhonov, 2022) and the noun-preposition-noun construction (Scivetti and Schneider, 2025), or on phenomena that presumably exist in all languages such as structural priming (Jumelet et al., 2024). This is unfortunate, because English is one of the least challenging languages for multilingual language models due to its over-representation in the available training data and their Anglocentric nature. For example, although multilingual BERT (mBERT; Devlin et al., 2019) was trained on 104 languages, English is the largest one and the model performs far better on downstream tasks for English, while for the 30% least represented languages, pretrained mBERT vectors decrease performance (Wu and Dredze, 2020). Phenomena that do not exist in English should be a greater challenge to a multilingual language model.

We focus on Differential Object Marking (DOM, Bossong, 1991), a phenomenon where certain direct objects are marked differently than others. Differences are based on semantic and syntactic factors, such as animacy. For example, in Spanish, direct objects that are both human and specific are marked with the preposition *a*, while other direct objects are not:

(1) a. Elena ve a Gabriela.
Elena sees to Gabriela.

[1] Supplementary materials, including test sentences and additional results, can be found in the paper's Github repository.

Proceedings of Recent Advances in Natural Language Processing,pages 1271–1281
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-147

b. Elena ve el río.
Elena sees the river

DOM does not occur in English. It does occur in major languages spanning different language families, including Spanish, Hindi, Turkish, Persian, Tamil, Amharic and Hebrew, as well as Romanian. The Romanian version of DOM is among the more complex ones because there are two mechanisms involved. One is inherited from Romance languages (the accusative marker *pe*) and the other is drawn from Balkan languages (clitic doubling). Therefore, we use Romanian as a case study.

We perform a linguistically informed comparative evaluation of both the original mBERT model (Devlin et al., 2019) and the two language-specific Romanian BERT models (Dumitrescu et al., 2020; Masala et al., 2020) available at the time of writing that have comparable parameter sizes. We focus on bidirectional encoders as these continue to be widely used, as discussed by the authors of ModernBERT (Warner et al., 2025) and shown by the recent release of the similarly sized multilingual EuroBERT (Boizard et al., 2025), which unfortunately does not incorporate Romanian. Furthermore, bidirectionality is necessary for our experimental setup of antecedent prediction where object markers follow the antecedent that we probe for.

Multilingual models are often used for under-resourced languages as transfer learning from higher-resource languages can occur (Guarasci et al., 2022), but as there is no DOM in English, we expect there to be limited use for transfer. The language-specific training datasets of monolingual Romanian models should yield more refined linguistic representations, more accurate tokenization, and better scores on our benchmark.

2 Related work

Language models learn to represent words and their contextual meanings as multidimensional vectors in a semantic space. At present, these representations are typically learned by transformer-encoder models using a masked language modelling (MLM) training objective, such as BERT (Devlin et al., 2019), or by transformer-decoder models that use an autoregressive causal language modelling (CLM) objective, such as GPT-4 (OpenAI, 2023). CLM models have become very popular for their text generation capabilities, but are outperformed by MLM models on various fundamental NLP tasks, such as information extraction and sequence labeling (Dukić and Snajder, 2024). The bidirectional nature of MLM encoders has been argued to promote the learning of syntactic and semantic representations (Wang et al., 2022).

2.1 Testing linguistic capabilities

It remains rather unclear *how* these models achieve their performance, as well as what the nature of their underlying linguistic representations are. This has lead to the emergence of a sub-field within NLP research ('BERTology') that is focused on using linguistically-informed tasks to infer the nature of model representations based on the model's performance on these tasks (Rogers et al., 2020). As Zhou et al. (2024) point out, the lack of a strong theoretical foundation and unclear grounding of the semantic representations means that we cannot predict how well large language models will perform in various scenarios, such as in new domains or processing understudied linguistic structures, without targeted evaluation and probing. The level of syntactic and semantic granularity in BERT representations has been studied especially for English phenomena. It seems that BERT representations are hierarchical rather than linear, but syntactic structure does not appear to be explicitly encoded in the weights of BERT's self-attention heads (Htut et al., 2019). BERT represents semantic roles, but struggles with pragmatic inference (Ettinger, 2020).

2.2 Romanian language modelling

Language models reflect their training data. Therefore, many domain-specific and language-specific models have been trained, and domain-specific or language-specific BERT variants usually outperform the general or multilingual BERT model on most evaluation tasks when model size remains the same. However, this is task-dependent and there are advantages to cross-lingual models as well (Deode et al., 2023). While multilingual BERT (mBERT) can be used for Romanian text, two dedicated language-specific Transformer-based models have been developed; Table 1 summarises their parameter and training data sizes. Dumitrescu et al. (2020) introduce Romanian BERT as the first model trained exclusively on Romanian text. The training data comprises about 15.2GB of thoroughly cleaned text data from several internet corpora, such as OPUS (Tiedemann, 2012), OSCAR (Suárez et al., 2019), and Wikipedia, amounting to roughly 2.4B tokens after preprocessing. In comparison to the mBERT models (both cased

Model	TrainTokens	Parameters
mBERT	?	178M
Rom. BERT	2.42B	124M
RoBERT-small	2.07B	19M
RoBERT-base	2.07B	114M
RoBERT-large	2.07B	341M

Table 1: Language model sizes and data sizes

and uncased versions), Romanian BERT shows improved performance across a range of extrinsic tasks (simple universal dependencies, joint universal dependencies, and POS-tagging), albeit by relatively small margins. A second language-specific BERT variety was published shortly after – RoBERT ('small', 'base', and 'large', Masala et al., 2020), with small additions in training data, different training tasks, and an identical architecture to domain-general BERT models. RoBERT, just like mBERT, is trained on the masked language modelling (MLM) and next sentence predictions (NSP) tasks. The model was tested by the authors on a variety of tasks, including sentiment analysis, cross-dialect topic identification, and diacritics restoration; RoBERT-small, at 19M parameters, performs similarly to mBERT (177M parameters) across tasks, while RoBERT-base performs better than mBERT, but very similarly to Romanian BERT. Lastly, RoBERT-large outperforms all other investigated models. RoBERT is not to be confused with RoBERT*a* or XLM-RoBERTa, which we do not use in our study as there is no Romanian RoBERTa variant to compare it to.

The Romanian BERT models have been evaluated on several extrinsic tasks. Results are task-specific but favour the monolingual models. Pais et al.'s (2021) dependency parsing evaluation and Buzea et al.'s (2022) fake news detection evaluation did not include multilingual models. Dumitrescu et al. (2021) present a benchmark of ten tasks, several of which have multilingual and monolingual models on the same leaderboard. Romanian BERT is shown to outperform mBERT on a semantic textual similarity task (Dumitrescu et al., 2021), dependency parsing, tokenization and named entity recognition (Dumitrescu et al., 2020). On emotion detection, Romanian BERT outperforms the larger multilingual XLM-RoBERTa (Ciobotaru et al., 2022). On Romanian dialect identification, Ro-BERT outperformed mBERT (Zaharia et al., 2020), but on question answering, mBERT outperformed monolingual models (Nicolae et al., 2023). Marinescu and Fellbaum (2024) use Romanian BERT in comparison to human judgements as a tool to analyse the syntax-semantics interface of Romanian noun compounds, though there is no explicit evaluative perspective adopted regarding the language model's performance. We are not aware of any studies that use specific linguistic properties of Romanian for language model evaluation.

2.3 Differential Object Marking in Romanian

Differential Object Marking (DOM), a term introduced by Bossong (1991), refers to the phenomenon where certain direct objects are marked differently based on semantic and syntactic factors. This marking typically involves prepositions, particles, or case markers to signal distinctions such as animacy, specificity, definiteness, or referentiality of the noun in direct object position. DOM is cross-linguistically attested; for instance, both Spanish and Romanian prominently mark animate and specific objects, though the conditioning factors vary. Romanian exhibits a particularly intricate DOM system due to its interaction with clitic doubling (Tigău, 2010) and the interplay of semantic and syntactic constraints. While specificity and animacy largely govern DOM, syntactic constraints sometimes override these factors – for example, the presence of a definite article blocks the occurrence of DOM markers. The system has evolved from Old Romanian (OR) to Modern Romanian (MR) through a process of stabilization, incorporating both Romance-specific mechanisms (e.g., *pe*-marking, derived from Latin *per*) and Balkan influences (e.g., clitic doubling, specificity-driven marking) (Hill and Mardale, 2021).

In Modern Romanian, the marking of direct objects is realised through the grammaticalised particle *pe* (DOM-p), without (2-c), but mostly with (2-b) co-occurring clitic doubling (CD; the occurrence of a clitic pronouns, co-referent with the noun in direct object position) for direct objects. In other cases, the direct object remains unmarked (2-a):

(2) a. Am văzut un film.
have seen a movie
'I/We have seen a movie'

b. L-am chemat pe Mihai.
$CD_{3SG.MASC}$-have called DOM-p Mihai
'I called Mihai.'

c. N-am chemat pe nimeni.
not-have called DOM-p nobody
'I did not call anybody.'

As described in reference grammars (GBLR – Dragomirescu et al., 2016), as well as experimental and corpus studies (Hill and Mardale, 2019, 2021; Tigău, 2010; Montrul et al., 2015a; Montrul and Bateman, 2020), the main semantic triggers of DOM in Romanian include referentiality, animacy, specificity, and definiteness. Highly referential NPs (e.g., personal pronouns, proper names) obligatorily receive DOM-p, except for city names:

(3) Mama a văzut-o pe
mom-the has seen-CD$_{3SG.FEM}$ DOM-p
Andreea / pe ea / *(pe
Andreea / DOM-p her /*(DOM-p London)
Londra).

'Mom saw Andreea / her / *(London).'

For indefinite NPs (4), DOM-p is optional, provided that they are animate; inanimate NPs generally remain unmarked, regardless of whether they are definite specific or indefinite.

(4) Am întâlnit (pe) un student
have met (DOM-p) INDEF.ART. student
'I/We met a student.'

(5) Am vizitat (*pe) muzeul.
have visited (*DOM-p) museum-the
'I/We visited the museum.'

CD alone does not occur in DOM contexts in Modern Romanian, while *pe*-marking alone is highly restricted to bare quantifiers as in (6). Furthermore, inanimate objects usually do not undergo DOM, with some marginal exceptions, but strong pronominal objects are obligatorily *pe*-marked regardless of animacy (rather, their anaphoric or deicitic nature does not make animacy information directly accessible), as in (7).

(6) Adina nu(*-l) cunoaște pe
Adina not(*-CD$_{3SG.MASC}$) know DOM-p
nimeni de acolo.
nobody from there
'Adina does not know anyone there.'

(7) Îl cumpăr pe celălalt.
CD$_{3SG.MASC}$ buy$_{1SG}$ DOM-p the-other-one
'I'll buy the other one.'

Without being exhaustive, the number and nature of the constraints presented show the intricate semantic and syntactic triggering contexts for DOM in Romanian, as it reconciles two typological patterns in a continuous stabilisation process. Corpus studies (Mardale, 2015), as well as experiments with heritage speakers (Montrul et al., 2015b) further underscore the volatility of DOM in both Old and Modern Romanian. This complexity makes Romanian DOM a promising avenue for benchmarking language models.

3 Method

As discussed above, there is a variety of methods for probing and evaluating language models for linguistic capabilities. An option often used in benchmarks such as BLiMP (Warstadt et al., 2020) is to compare probabilities of minimal pairs that differ in grammaticality. This is quite scalable, but provides limited insight into what tokens the model would produce, as only the sentences and words provided as input are considered. An alternative that elicits tokens from the model is LAMA (LAnguage Model Analysis) proposed by Petroni et al. (2019), which involves masking specific tokens in controlled experimental sentences to get predictions for the mask. This is similar to the Cloze task (Taylor, 1953) that is widely used in psychology and linguistics, where human participants also have access to both the left and right context, and it is similar to the training objective of masked language modelling. It has been used in many investigations of linguistic capabilities of BERT-based models, such as to predict relativizers and antecedents for English relative clauses (Mosbach et al., 2020). This is also a complex phenomenon involving several factors such as animacy and definiteness, exhibiting optionality and potentially involving both the right and left context, similar to Romanian DOM. Lee and Bloem (2023) used this method to test BERT-based models for indeclinable nouns in South Slavic languages, a phenomenon that does not exist in English. To control prediction contexts on both sides of the mask, we use bidirectional transformer-encoder models.

To design such an experiment, we need to choose what to mask while giving the model enough context to make predictions. This is nicely demonstrated by Mosbach et al.'s (2020) study, where two types of masking are used:

(8) This is the dress [MASK] I saw.

(9) This is the [MASK] that I saw.

In (8), they mask the grammatical element (rela-

tivizer) and evaluate whether a correct one is predicted – *that* rather than *who*. However, although they don't mention it, it is not possible to evaluate zero predictions with this setup (*This is the dress I saw*) as the model always predicts something. In (9), the antecedent is masked and they evaluate whether it has fitting semantic properties.

3.1 Test sentences

Our setup resembles that of Mosbach et al. (2020), but the grammatical marker is more complex. Not only is it sometimes optional or ungrammatical, it can also involve multiple morphemes (and thus tokens) and they may be discontinuous, making them difficult to target with masking. Therefore, we focus on providing the model with DOM markers and masking the noun in direct object position. This approach yields more insight into the factors involved as many possible nouns can be predicted and more samples per template can be obtained. We designed a set of sentence templates containing DOM cases in such a way that each type of marking is given across sentences (DOM-p + CD, DOM-p) and the NP in object position is masked. These templates were hand-crafted by a Romanian linguistics expert on our team, based on examples from the Romanian DOM literature. This gives more control over potentially interfering factors such as sentence length. We vary the word order in our templates (SVO or OVS), so that the masked objects occur both pre- and post-verbally. All the templates and sentences we used throughout the project can be found as supplementary material in the provided Github repository. We also experiment with eliciting the DOM markers while providing the objects, following specific syntactic and semantic criteria to test the limits of the LAMA approach for evaluating linguistic knowledge. However, this strategy proved to be less robust and too sensitive to specific masking choices. We briefly discuss the outcomes below, with extended explanations provided in the supplementary materials online. Overall, we aim to probe whether the predicted noun matches the semantic (animacy, specificity) and syntactic (e.g. presence or absence of definite articles) constraints of Romanian DOM.

We passed six templates with masked objects to each of the investigated model versions, accounting for the top k = 50 predictions for each mask thus yielding a sample size of 300 sentences for each model version. We retrieved the pre-trained models using the HuggingFace API: mBERT (cased and uncased), Romanian BERT (Dumitrescu et al., 2020) (cased and uncased), and RoBERT (Masala et al., 2020) (small, base, large). The resulting sentences were assessed and manually annotated by a native Romanian speaker using four labels:

- **'incorrect'** – ungrammatical and / or nonsensical sentence: The predictions were filled in a such a way that the resulting sentence is ungrammatical, nonsensical, or both.
- **'correct non-DOM'** – grammatical and semantically valid, but not intended: The predictions yielded a sentence that is grammatical and semantically sound, but the intended structure (object marking or object itself) was not elicited.
- **'correct DOM'** – grammatical and semantically valid, as intended: The masked token(s) were filled in such a way that the sentence is grammatical and semantically sound; the intended structure was successfully elicited.
- **'ambiguous'** – ambiguous: The sentence has questionable or uncertain grammaticality or is case-sensitive, but comes from an uncased model version; the sentence potentially yields different readings, of which not all are grammatical or semantically valid.

The primary annotator was a linguistics expert who was instructed to annotate based on what is considered grammatical in the literature on Romanian DOM, rather than based on native speaker intuition, as this might be affected by regional variation. Accuracies per model version were computed as the ratio of 'correct DOM' labels to the total number of sentences. A secondary annotator annotated a batch of 140 sentences in parallel. We observed strong inter-annotator agreement (Cohen's κ: 0.801).

4 Results

Figure 1 and Table 2 summarize the results of the main experiment (the DOM markers are provided and the noun is masked), as percentages of each label per model version, as well as accuracy scores for the predictions. The multilingual model versions have much fewer 'correct' predictions compared to each of the Romanian models, with the uncased multilingual version yielding the fewest syntactically and semantically correct predictions.

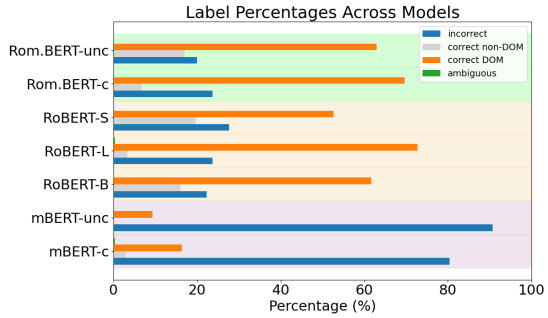


Figure 1: Percentages of each annotated label per model version, when providing DOM markers and masking the noun position.

Model	Version	Accuracy
mBERT	uncased	9.33%
	cased	16.33%
RoBERT	small	52.67%
	base	61.67%
	large	**72.67%**
Romanian BERT	uncased	63.00%
	cased	69.67%

Table 2: Accuracy per model on object prediction.

Among the language-specific model versions, the highest number of grammatical sentences with the intended morphsyntactic structure came from RoBERT-large, which brings the overall highest accuracy obtained across all models to 72.67%.

These results clearly show that the poorest-performing language-specific model (RoBERT-small) outperforms the best-performing multilingual one (mBERT-cased), despite having much fewer parameters – 19M for RoBERT-small and 130M for mBERT. Language-specific models indeed perform better at handling the uncommon or complex object marking system in Romanian, compared to their multilingual counterparts.

One potential concern is that considering the top 50 predictions may be too high of a number to consistently yield suitable candidates. Models, especially multilingual ones, may run out of fitting Romanian vocabulary to predict. In our test sentences, the pool of suitable candidates is typically expected to be larger than 50, covering all nouns referring to human entities, the full personal pronoun paradigm, and other grammatical classes. Nevertheless, to check the validity of our experimental setup, we conducted an additional analysis on the top-performing monolingual and multilingual models. Specifically, we examined (1) the distribution of grammatical predictions within these 50 samples, plotted in Figure 2, and (2) the distribution of assigned probabilities for the top 50 predicted tokens for each sentence, displayed in Figure 3.

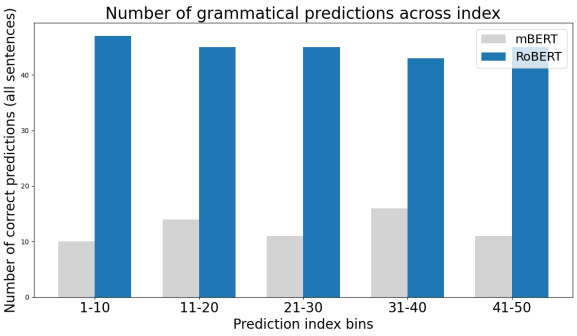
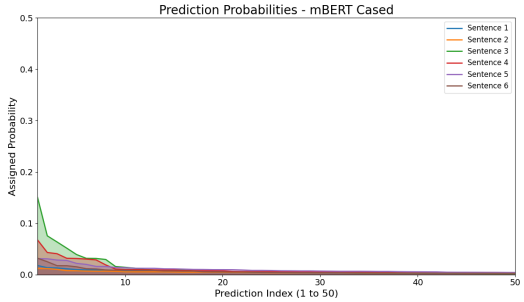
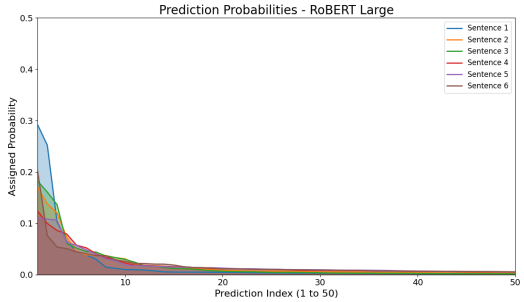


Figure 2: Number of total correct predictions across all test sentences, progressively throughout the 50 samples.



(a) mBERT-c – best performing multilingual model.



(b) RoBERT-L – best performing monolingual model.

Figure 3: Distributions of assigned probabilities for the top 50 predictions, per sentence, across the best performing models from each category.

We observe a relatively even distribution of correct predictions throughout the sample of 50 for both models in Figure 2, indicating that there is a similar amount of correct predictions for the top as for the bottom of the batch; this leads us to believe it is unlikely that the lower scores of the multilingual models are due to an insufficient number of available felicitous tokens.

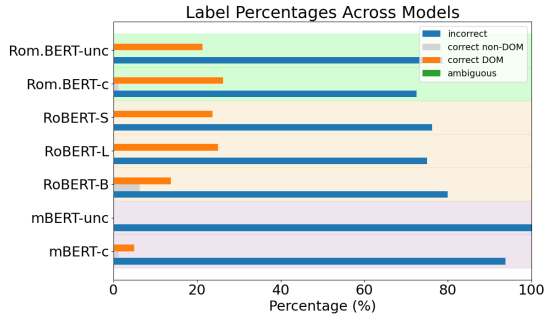


Figure 4: Percentages of each annotated label per model version, when providing nouns and masking object markers.

As for the complementary task (providing the nouns in object position and eliciting the DOM markers), the setup yielded significantly fewer correct predictions, potentially due to the limited set of valid predictions available for such masked positions (the marker 'pe' or the doubling clitic). Initial model performances were very poor, ranging between 0% and 26.2% accuracy. We discuss potential reasons for this below.

5 Discussion

Our results show that even the language-specific model versions with the poorest scores outperformed the best version of the multilingual one, mBERT-cased, despite having far fewer parameters. Language-specific training appears to make a large difference for Romanian DOM, a complex phenomenon that has no parallels in English.

As previously mentioned, we also experimented with a different elicitation strategy, by masking the DOM markers and providing different types of nouns, yet this has proven to be unsuccessful. The sentence pair below shows that the object elicitation strategy (10) and marker elicitation strategy (11) yield very different outcomes:

(10) Pe [MASK] am văzut-o azi, dar pe Ileana, nu. – A.2., template (21)

a. Pe **ea** am văzut-o azi, dar pe Ileana, nu. – RoBERT-large, top prediction, grammatical

b. Pe **Maria** am văzut - o azi , dar pe Ileana , nu . – mBERT-cased, top prediction, grammatical

(11) [MASK] Ana am văzut [MASK] azi, dar pe Ileana, nu. – A.1., template (4)

a. **pe** ana am văzut **și** azi, dar pe ileana, nu. – RoBERT-large, ungrammatical

b. **.** Ana am văzut **pe** azi, dar pe Ileana, nu. – mBERT-cased, ungrammatical

Specifically, example (10) yields grammatical sentences for both the multilingual and the monolingual model, as one predicts a personal pronoun and the other a proper name. For example (11), where the markers are masked, both models form ungrammatical sentences, except the monolingual one predicts actual words in the masked positions, while the multilingual BERT predicts both a word and a punctuation symbol. Instances such as this lead us to believe that this elicitation strategy did not necessarily reveal performance differences between monolingual and multilingual models, but rather failed to elicit grammatical sentences altogether.

Romanian object markers often involve more than one morpheme (or, token) per sentence. Numerous ungrammatical cases were annotated as such because the CD + DOM-pe marking was not predicted entirely; although DOM-p in front of the object is quite often correctly filled, the CD is not. As such, the marking as a whole was not grammatical in that context, and this may have made the task more difficult than our initial task of predicting just a single object. Furthermore, our follow-up experiment showed that the approach of eliciting markers was quite sensitive to tokenizer effects. We added an additional mask, which seems like it should make the task more difficult, but it led to drastically improved accuracy scores because the clitic was often presented as two tokens. Another issue with the approach of eliciting markers is that it is difficult for a model to predict no marking in a template where a mask for a marker is specified. This may make it easier to predict correct marking compared to an unconstrained generation scenario, resulting in inflated accuracy scores. Therefore, we find that LAMA-based probing is more reliable and informative when content words are targeted, while questions regarding markers and morphology may benefit more from BLiMP-style perplexity comparison.

Romanian is among the less frequent languages that use hyphenation ('-') to orthographically mark contraction, as opposed to the apostrophes used in Italian or English, among other languages. Furthermore, in Romanian there are numerous cases where only the contracted form is acceptable: *L-am văzut.*, but * *Îl am văzut.* for 'I/We have seen

him'. The clitic doubling often involved in object marking is one of those cases. This may impede token classification-based NLP tasks such as relation extraction for Romanian, and the observation that this type of hyphenation is challenging for tokenization has also been made by Vasiu and Potolea (2020). Injecting language-specific morphological knowledge in tokenizers for Romanian text improves their performance in that study. We observe that this is a problem not only for mBERT, but also for the Romanian-specific models. Several examples of tokenized sentences per model show that all three tokenizers separate the clitic from its corresponding hyphen, splitting it into two tokens:

(12) Te-am văzut ieri.
   a. mBERT: ['Te', '-', 'am', 'v', '##ă', '##zut', 'ie', '##ri', '.']
   b. RoBERT: ['te', '-', 'am', 'văzut', 'ieri', '.']
   c. Romanian BERT: ['Te', '-', 'am', 'văzut', 'ieri', '.']

This applies to the whole pronominal paradigm involved in clitic doubling. Therefore, surprisingly, this tokenization issue does not explain the monolingual-multilingual performance gap, but it might explain why even RoBERT-large only reaches 72.67% accuracy at object prediction. All three Romanian models could benefit from a more morphologically-informed tokenization strategy.

6 Conclusions

The primary aim of this study was to address the research gap that exists when it comes to linguistically-informed evaluations of language models on phenomena that do not occur in English, using the performance of multilingual BERT and Romanian-specific BERT models on Romanian differential object marking as a case study.

Our findings show that monolingual models outperform the multilingual model on this task. For accurate representation of language-specific grammatical phenomena, models appear to benefit greatly from language-specific datasets, which allow a more targeted representation of structures less frequently found in other languages. More morphologically informed language-specific tokenization might also benefit downstream tasks based on token labeling tasks for morphologically richer languages. This aligns with observations from other studies where monolingual Romanian BERT models outperform mBERT on most extrinsic tasks (Section 2.2). However, whether DOM knowledge specifically affects downstream task performance cannot be established without further experiments involving models trained on synthetic data.

We might expect similar benefits from language-specific data for other languages with DOM such as Spanish, though the phenomenon is less complex in most other languages. Furthermore, our findings regarding the trade-offs between eliciting grammatical markers versus the thing that they mark in Cloze-style experiments with language models will also apply to future investigations into language models' linguistic knowledge of morphosyntactic phenomena in specific languages.

It would be interesting to investigate whether fine-tuning multilingual models on more target language data could further improve performance on language-specific phenomena, although this would not solve the tokenization issues. Additionally, exploring decoder-only architectures like GPT in this context could be valuable, as their autoregressive mechanism aligns more closely with certain aspects of human language processing; however, challenges related to the cognitive plausibility of language models remain (Connell and Lynott, 2024). Furthermore, the high degree of control we get from targeting specific open slots in the MLM task would not extend to a generative setup, requiring a different experiment based on unidirectional sentence completion or minimal pair probabilities.

While our results show far better performance for language-specific models, there are also disadvantages to pre-training new models for every specific scenario as this is costly in terms of resources and environmental cost. Multilingual models may still be a better choice for these reasons.

The fact that our approach requires post-hoc annotation of model output, rather than automated comparison to a gold standard, makes it more labour-intensive and difficult to scale than benchmark-based approaches. Results using this approach are also affected by the chosen definition of grammaticality - we chose to rely on expert annotation rather than the native speaker intuitions of multiple annotators, but this approach is influenced by prescriptive norms and may not always reflect actual language use. A further limitation is that our method is not applicable to decoder-only language models such as the GPT model family, as right context is required for proper prediction on our test templates.

References

Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M Alves, André Martins, Ayoub Hammal, Caio Corro, Céline Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, et al. 2025. EuroBERT: Scaling multilingual encoders for European languages. *arXiv preprint arXiv:2503.05500*.

Georg Bossong. 1991. Differential object marking in Romance and beyond. *New analyses in Romance linguistics*, 69:143–170.

Marius Cristian Buzea, Stefan Trausan-Matu, and Traian Rebedea. 2022. Automatic fake news detection for Romanian online news. *Information*, 13(3):151.

Lisa Bylinina and Alexey Tikhonov. 2022. The driving forces of polarity-sensitivity: Experiments with multilingual pre-trained neural language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Ondřej Cífka and Antoine Liutkus. 2023. Black-box language model explanation by context length probing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1067–1079, Toronto, Canada. Association for Computational Linguistics.

Alexandra Ciobotaru, Mihai Vlad Constantinescu, Liviu P Dinu, and Stefan Dumitrescu. 2022. Red v2: Enhancing RED dataset for multi-label emotion detection. In *Proceedings of the thirteenth Language Resources and Evaluation Conference*, pages 1392–1399.

Louise Connell and Dermot Lynott. 2024. What can language models tell us about human cognition? *Current Directions in Psychological Science*, 33(3):181–189.

Samruddhi Deode, Janhavi Gadre, Aditi Kajale, Ananya Joshi, and Raviraj Joshi. 2023. L3Cube-IndicSBERT: A simple approach for learning cross-lingual sentence representations using multilingual BERT. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 154–163, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Adina Dragomirescu, Isabela Nedelcu, Alexandru Nicolae, Gabriela Pană Dindelegan, Marina Rădulescu Sala, and Rodica Zafiu. 2016. *Gramatica de bază a limbii române: Hauptband*. Univers Enciclopedic Gold.

David Dukić and Jan Snajder. 2024. Looking right is sometimes right: Investigating the capabilities of decoder-only LLMs for sequence labeling. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14168–14181, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Stefan Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020. The birth of Romanian BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4324–4328, Online. Association for Computational Linguistics.

Stefan Daniel Dumitrescu, Petru Rebeja, Beata Lorincz, Mihaela Gaman, Andrei Avram, Mihai Ilie, Andrei Pruteanu, Adriana Stan, Lorena Rosia, Cristina Iacobescu, et al. 2021. LiRo: Benchmark and leaderboard for Romanian language tasks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Raffaele Guarasci, Stefano Silvestri, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito. 2022. BERT syntactic transfer: A computational experiment on Italian, French and English languages. *Computer Speech & Language*, 71:101261.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Virginia Hill and Alexandru Mardale. 2019. Patterns for differential object marking in the history of Romanian. *Journal of Historical Syntax*, 3(5):1–47.

Virginia Hill and Alexandru Mardale. 2021. *The diachrony of differential object marking in Romanian*, volume 45. Oxford University Press.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R Bowman. 2019. Do attention heads in BERT track syntactic dependencies? *arXiv preprint arXiv:1911.12246*.

Jaap Jumelet, Willem Zuidema, and Arabella Sinclair. 2024. Do language models exhibit human-like structural priming effects? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14727–14742, Bangkok, Thailand. Association for Computational Linguistics.

Sofia Lee and Jelke Bloem. 2023. Comparing domain-specific and domain-general BERT variants for inferred real-world knowledge through rare grammatical features in Serbian. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 47–60, Dubrovnik, Croatia. Association for Computational Linguistics.

Alexandru Mardale. 2015. Differential object marking in the first original Romanian texts. In *Formal approaches to DPs in Old Romanian*, pages 200–245. Brill.

Ioana Marinescu and Christiane Fellbaum. 2024. Human and automatic interpretation of Romanian noun compounds. *arXiv preprint arXiv:2403.06360*.

Mihai Masala, Stefan Ruseti, and Mihai Dascalu. 2020. Robert–a Romanian BERT model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6626–6637.

Silvina Montrul and Nicoleta Bateman. 2020. Vulnerability and stability of differential object marking in Romanian heritage speakers. *Glossa: a journal of general linguistics*, 5(1).

Silvina Montrul, Rakesh Bhatt, and Roxana Girju. 2015a. Differential object marking in spanish, hindi, and romanian as heritage languages. *Language*, pages 564–610.

Silvina Montrul, Rakesh Bhatt, and Roxana Girju. 2015b. Differential object marking in Spanish, Hindi, and Romanian as heritage languages. *Language*, 91(3):564–610.

Marius Mosbach, Stefania Degaetano-Ortlieb, Marie-Pauline Krielke, Badr M. Abdullah, and Dietrich Klakow. 2020. A closer look at linguistic knowledge in masked language models: The case of relative clauses in American English. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 771–787, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Constantin Dragos Nicolae, Rohan Kumar Yadav, and Dan Tufiş. 2023. Evaluation of language models on Romanian XQuAD and RoITD datasets. *International Journal of Computer Communication & Control*, 18(1).

OpenAI. 2023. GPT-4 Technical Report. Retrieved April 14, 2024 from `https://doi.org/10.48550/arXiv.2303.08774`.

Vasile Pais, Radu Ion, Andrei-Marius Avram, Maria Mitrofan, and Dan Tufis. 2021. In-depth evaluation of Romanian natural language processing pipelines. *Romanian Journal of Information Science and Technology (ROMJIST)*, 24(4):384–401.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Maribeth Rauh, John Mellor, Jonathan Uesato, Po-Sen Huang, Johannes Welbl, Laura Weidinger, Sumanth Dathathri, Amelia Glaese, Geoffrey Irving, Iason Gabriel, et al. 2022. Characteristics of harmful text: Towards rigorous benchmarking of language models. *Advances in Neural Information Processing Systems*, 35:24720–24739.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Wesley Scivetti and Nathan Schneider. 2025. Construction identification and disambiguation using BERT: A case study of NPN. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 365–376, Vienna, Austria. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adri Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Wilson L. Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30(4):415–433.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the eigth Language Resources and Evaluation Conference*, pages 2214–2218.

Alina Tigău. 2010. Towards an account of differential object marking in Romanian. *Bucharest Working Papers in Linguistics*, (1):137–159.

Mihaela Alexandra Vasiu and Rodica Potolea. 2020. Enhancing tokenization by embedding Romanian language specific morphology. In *2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 243–250. IEEE.

Tim Veenboer and Jelke Bloem. 2023. Using collostructional analysis to evaluate BERT's representation of linguistic constructions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12937–12951.

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.

Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. 2022. Pre-trained language models and their applications. *Engineering*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. 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. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *5th Workshop on Representation Learning for NLP, RepL4NLP 2020 at the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 120–130. Association for Computational Linguistics (ACL).

Qing Yao and Simon Todd. 2024. BERT's insights into the English dative and genitive alternations. In *Proceedings of the Society for Computation in Linguistics 2024*, pages 52–62.

George-Eduard Zaharia, Andrei-Marius Avram, Dumitru-Clementin Cercel, and Traian Rebedea. 2020. Exploring the power of Romanian BERT for dialect identification. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 232–241.

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. 2024. A comprehensive survey on pretrained foundation models: a history from BERT to ChatGPT. *International Journal of Machine Learning and Cybernetics*.