# A cross-model study on learning Romanian parts of speech with Transformer models

**Radu Ion**
Institute for AI, "Mihai Drăgănescu"
radu@racai.ro

**Verginica Barbu Mititelu**
Institute for AI, "Mihai Drăgănescu"
vergi@racai.ro

**Vasile Păiș**
Institute for AI, "Mihai Drăgănescu"
vasile@racai.ro

**Elena Irimia**
Institute for AI, "Mihai Drăgănescu"
elena@racai.ro

**Valentin Badea**
Institute for AI, "Mihai Drăgănescu"
valentin.badea@racai.ro

## Abstract

This paper will attempt to determine experimentally if POS tagging of unseen words produces comparable performance, in terms of accuracy, as for words that were rarely seen in the training set (i.e. frequency less than 5), or more frequently seen (i.e. frequency greater than 10). To compare accuracies objectively, we will use the odds ratio statistic and its confidence interval testing to show that odds of being correct on unseen words are close to odds of being correct on rarely seen words. For the training of the POS taggers, we use different Romanian BERT models that are freely available on HuggingFace.

**Keywords:** BERT, POS tagging, Romanian, odds ratio, POS learning.

## 1 Introduction

Transformer models (Vaswani et al., 2023) and Deep Learning have changed the face of Natural Language Processing (NLP) domain, with a huge number of papers reporting superior performances of any conceivable task of NLP, including machine translation, question answering (which is now handled almost flawlessly by generative Large Language Models), and language analysis (POS tagging, dependency parsing, word sense disambiguation, etc.)

Transformer models are very good at any NLP task, provided they are pre-trained on very large corpora (billions of words) at supervised tasks such as masked language modeling or next sentence prediction (Devlin et al., 2019) and then, fine-tuned to the task at hand, e.g. POS tagging. Central to the Transformer models' remarkable ability to learn syntagmatic information about words is the attention mechanism (Vaswani et al., 2023), which encodes co-occurrence information in a large window of tokens (typically 512 tokens) for a large vocabulary of tokens (typically 50K tokens).

Comparatively, the number of papers dealing with the subject of how the Transformer model is learning a language (or multiple languages at once), which presumably makes them so good at any language processing task, is very small with respect to the number of papers presenting extensions of the model, accuracy improvements, applications, and so on.

With this paper, we want to contribute to the set of papers taping into the learning mechanisms of the Transformer models, and we present a study on if and how the BERT models (a type of Transformer models) learn the grammatical categories (e.g. noun, verb, article, determiner, etc.) of a word in its context (i.e. POS tagging with a smaller tagset). We focus on Romanian, and we use Romanian-specific BERT models for the job. We will try to experimentally prove that BERT models have about the same accuracy on unseen (during training) words as on words that were rarely seen. Furthermore, the accuracy of frequently seen words is not that much higher than the accuracy of unseen words. To quantify these comparisons, we will use the odds ratio statistic.

## 2 Related work

Experiments on how POS taggers work on words not seen during training were performed more than 20 years ago, at the time when POS taggers were actively developed using e.g. Hidden Markov Models. An example in this regard is the work by Dematas and Kokkinakis (1995), which addresses the POS tagging of unseen words with enhanced HMMs. At that time, the best tagging accuracy on these words was about 66% for English.

Kim and Smolensky (2021) investigated the ability of pre-trained Transformer models (i.e. BERT-large, Devlin et al., 2019) to perform grammatical category-based generalization of novel words, after being finetuned on limited contexts (without categorization-specific training). Inspired by an experimental design in which infants were familiarized to contexts containing novel words and then tested with new sentences that either obeyed or violated category-based co-occurrence restrictions, the authors assumed that a Masked Language Model's (MLM) ability to assign a higher probability to a word in a novel context that obeys the co-occurrence restriction for that category (over a word that does not) means the MLM makes a valid grammatical category inference about a novel word.

The two-step method involves finetuning the MLM on two signal contexts that unambiguously mark the novel words ($w_1$ and $w_2$) grammatical categories and testing the fine-tuned model by comparing the probabilities of $w_1$ and $w_2$ on multiple test contexts (higher probability to the new word in the correct test context meaning accurate category inference).

The signal and test contexts are based on MNLI corpus (Williams et al., 2018), that had different sources from the model's pre-training data. A fine-tuning set contains two signal contexts with one unseen word each ($w_1$ and $w_2$) and 400 test contexts, 200 for each grammatical category (MNLI-sampled sentences in wich words with grammatical categories of interest are masked out). Six English datasets that test for the binary classification between the four open-class grammatical categories (noun, verb, adjective and adverb) were constructed.

To use "unknown" words and make the BERT-large model to "forget" learned words, random weights were used for the unknown words' embeddings. The BERT-large model was finetuned for 70 epochs and accuracy was tested at a significance level of $p < 5\%$ with a one proportion z-test. Conclusions were as follows:

1. accuracy largely varied between category pairs, from 67.3% for noun vs. adverb to 88.1% for noun vs. verb.
2. category inference was quite slow in comparison to competent speakers' performances who often can solve the task from a single example.

In another study targeted at "what contextual representations encode that conventional embeddings do not?", Tenney et al. (2019) compare conventional word embeddings to Transformer-generated word embeddings, which they call "contextual embeddings". For this purpose, they propose to probe a contextual embeddings model based on a simple architecture employing span representations and binary classifiers. In their approach, a span corresponds to a word or a sequence of words and the classifiers are trained to predict specific labels. The authors call this approach "edge probing". For part-of-speech (POS) tagging, the OntoNotes (Weischedel et al., 2013) corpus is used (even though the authors investigate other tasks as well, making use of OntoNotes or additional corpora). The span for which a prediction is made corresponds to a single word. The classifiers are trained to predict individual part of speech tags (such as noun, verb, adjective, etc.) for the current word. The authors explore 4 contextual encoder models: CoVe (McCann et al., 2017), ELMo (Peters et al., 2018), OpenAI GPT (Radford et al., 2018), and BERT (Devlin et al., 2019). The models' weights are not fine-tuned. For BERT and GPT, contextual word vectors are obtained using two methods: concatenation of the subword embeddings with the activations of the top layer, or a linear combination of layer activations (including embeddings) using learned task-specific scalars. The authors compare the results of the entire model with so-called "lexical baselines" in which the probing model is trained only on the most closely related context-independent word representations (for example in the case of ELMo, only the activations of the context-independent character-CNN layer (layer 0) are used). Regarding POS tagging, the BERT models outperform the other models (with BERT-base, using a concatenation approach, achieving the highest F1 score). In all cases, using the full models outperform the "lexical baselines", while

ELMo "lexical baseline" outperforms the others. The authors consider that the results suggest that ELMo encodes local type information. In addition, the authors try to estimate how much information is derived from long distance tokens. Their original architecture is extended with a CNN layer of width 3 (considering one token to the left and one to the right). This addition significantly reduces the gap between a "lexical baseline" ELMo and the full model, indicating that ELMo improvements are due to the encoding of long-range information.

Metheniti et al. (2022) report on experiments in which several Transformer models (BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), ALBERT (Lan et al., 2019)) are tested with respect to their ability to capture information about telicity and duration of verbs. While these semantic features are not directly related to grammatical category learning, their study represents another very good example of analyzing how Transformer models learn the language.

An action is telic if it has an end point and atelic otherwise. Durative verbs describe an action, while stative verbs describe states. The authors work with English and French, and in one experiment, they fine-tune the transformer models for binary sequence classification of telicity and duration (separately), and of testing their accuracy on predicting these features. For fine-tuning they use a set of sentences annotated for these features.

In another experiment, no fine-tuning was performed. Instead, a logistic regression to the contextual embeddings of each layer is applied, as provided by the pre-trained models. Contextual word embeddings for the annotated verbs are extracted from each layer of the transformer model and a logistic regression model is trained to classify telicity and duration, to understand how much such information was learned by each layer.

For classifying telicity, all systems obtain an accuracy above 80% and it improved when training the models with the extra information of verb position in the sentence. BERT (both base and large) had the best results.

For classifying duration, the results are even better (higher than 93%), despite using a smaller dataset. No improvement could be noticed when training the models with the extra information of verb position in the sentence. BERT was also the best performing.

An error analysis showed that conflicting characteristics of the linguistic context prevent the correct analysis: e.g., sentences where the verb or the verbal phrase would be considered (a)telic, but part of the context defines the temporal aspect of the sentence in the opposite way.

For French, the results are not as good as for English, probably because of the characteristics of the French verbal system.

Contextual embeddings proved to be an efficient way to encode the aspectual information of a verb and its interaction with its context, and this knowledge is probably already learned in the pre-training process.

## 3 RoBERT models

RoBERT[1] is a Romanian-only, pre-trained BERT model. Masala et al. (2020) developed this model to address the gap in pre-trained language models for languages other than English. The model was designed similarly to BERT with small, base, and large variants, having the same number of layers, hidden params, and attention heads. The training time in hours for each model was 28, 77, and 255, respectively, training for 40 epochs on a v3-8 TPU on two supervised tasks: masked language modeling and next sentence prediction. The Romanian dataset that was used for training was comprised of 3 sources, totaling 2.07B words.

Without dwelling into details, the RoBERT models outperform the competition in several tasks, namely mBERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2020), and the only true Romanian BERT model at that time, BERT-base-ro (Dumitrescu et al., 2020).

## 4 Experiments

### 4.1 Preliminaries

In our experiments we use the RoRefTrees (RRT) Romanian UD corpus[2] (Barbu Mititelu, 2018), currently at version 2.13. The corpus is pre-split into the train, dev and test sets and we join the dev and test sets into a bigger test set, while only

[1] https://huggingface.co/readerbench

[2] https://github.com/UniversalDependencies/UD_Romanian-RRT

training on the train set. All accuracy results that are presented in the next sections are computed over this bigger test set.

Since, for the time being, we are interested in how BERT models learn the grammatical categories, and only the grammatical categories without any other morphosyntactic attributes such as number, gender, tense, etc., we selected the POS and its type to comprise the grammatical category label to target the training for. We ended up with 35 categories, as follows:

- Proper and common nouns and numerals.
- Main and auxiliary verbs.
- Adjectives and adverbs.
- Abbreviations (of different types, e.g. nominal, adjectival, adverbial, etc.).
- Pronouns and determiners of different types (e.g. personal, demonstrative, reflexive, indefinite, etc.)
- Articles (possessive, indefinite)
- Prepositions
- Conjunctions (coordinative, subordinative)
- Particles (infinitive, negative).

We further split the set of grammatical categories into two subsets: the *content words* set which contains all categories of "meaning bearing" words (proper and common nouns, main verbs, general adverbs, adjectives, numerals, and abbreviations) and the set of *functional words* which is the complement of the full set of categories with respect to the content words set. The reason we consider these subsets is that the rarely seen and the unseen words in the test set vastly belong to the content words set (see Table 2, below), and computing accuracies including functional words would yield an unfair advantage to the words that are seen frequently in the training set.

Table 1 presents statistics of the content and functional words in the RRT, in each split and Table 2 shows how different word types from dev plus test splits are distributed at F = 0 (do not appear at all in the train split), F = 1 (appear once in the train split) and F > 1 (appear more than once in the train split).

| | **Cont.** | **Func.** | **Punc.** |
|---|---|---|---|
| **train** | 92,694 | 68,740 | 23,691 |
| **dev** | 8,633 | 6,217 | 2,223 |
| **test** | 8,277 | 5,964 | 2,083 |

Table 1: RRT word type statistics

| | **F = 0** | **F = 1** | **F > 1** |
|---|---|---|---|
| **Content** | 3,185 | 1,702 | 12,023 |
| **Functional** | 39 | 19 | 12123 |

Table 2: Word count distribution by frequency for the dev plus test bigger test set

### 4.2 Testing methodology

The BERT models are fitted with a POS classification layer on top of the last hidden state of each token. The POS layer has 35 dimensions, one for each considered grammatical category, and it is trained with a softmax learning objective. We also update the BERT model's parameters in the backward propagation stage. The starting learning rate parameter is set at $10^{-5}$ and it is decreased by a factor of 0.9 every epoch, out of the 5 training epochs.

We will attempt to experimentally prove the following hypothesis: the POS tagging of unseen words (i.e. in the training set) is as accurate as POS training of words that were seen in the training set. We will measure the odds ratio ($OR$, Bland and Altman, 2000) of the odds of being correct vs. being incorrect when the frequency F of the targeted words is greater than 0 compared to when F is 0 in the training set. Thus, we compute the $OR$ fraction from the following contingency table:

| | **F > 0** | **F = 0** |
|---|---|---|
| **Correct** | $p_c$ | $q_c$ |
| **Incorrect** | $1 - p_c$ | $1 - q_c$ |

Table 3: $OR$ contingency table

as

$$OR = \frac{\frac{p_c}{1-p_c}}{\frac{q_c}{1-q_c}} = \frac{p_c(1-q_c)}{q_c(1-p_c)}$$

and show that it is close to 1, in a confidence interval that forbids rejecting the null hypothesis of it being different than 1. In the above equation, $p_c$ and $q_c$ are the probabilities of being correct in the chosen sample (i.e. the ratio of correctly tagged words out of all tagged words in the sample).

We will only target words that belong to the chosen BERT model vocabulary, such that the evaluated word is not split into sub-words by the WordPiece tokenizer. We enforce this constraint for two reasons:

1. We do not want to average BERT representations of sub-words to obtain a representation for the full word, because the average of embeddings is not necessarily the equivalent of producing the true representation of the full word.
2. We are interested in a study targeting the specific dimensions of a word representation that mostly decide its grammatical category.

Finally, as previously mentioned, we only compute and compare odds ratios for content words, for the reasons explained above.

### 4.3 Results with the RoBERT models

We trained the `readerbench/RoBERT-small`, `readerbench/RoBERT-base` and `readerbench/RoBERT-large` models from HuggingFace the way we described previously. Table 4 presents an overview of the accuracies we obtained on the POS tagging task, with the 35 POS labels, on all words (content and functional), at different frequency thresholds, as shown in the table's header.

| | **F ≥ 0** | **F = 0** | **F = 1** | **F > 0** |
|---|---|---|---|---|
| **small** | 96.5% | 92.6% | 90.9% | 96.6% |
| **base** | **97.9%** | 92.6% | **94.8%** | **98%** |
| **large** | 96.5% | **93.3%** | 93.1% | 96.6% |

Table 4: Accuracy on content and functional words

Table 5 below shows the same accuracy figures, but only for content words POS tagging.

| | **F ≥ 0** | **F = 0** | **F = 1** | **F > 0** |
|---|---|---|---|---|
| **small** | 93.8% | 92.9% | 89.4% | 93.9% |
| **base** | 96.4% | 92.7% | **94%** | **96.7%** |
| **large** | **96.5%** | **93.5%** | 92.4% | 95.5% |

Table 5: Accuracy on content words only

One thing we see from Tables 4 and 5 is that the large model is better at tagging unseen words while the base model is better at everything else. Comparing the values of the accuracy figure from Table 5 for F = 0 and F > 0, we see differences of at least 1%. This could suggest that the model is not able to learn the grammatical categories of unknown words, but this conclusion is going to be amended when we plot odds ratios at different frequency bands.

Figures 1 to 3 show the odds ratios plot, for each of the RoBERT models, computed as in Table 3 for content words only. Thus, we have the $OR$ on the $Y$ axis, while on the $X$ axis we have a frequency step of 1 for which the Table 3 F > 0 condition holds: 1, 1 and 2, 1 to 3, …, 1 to 10, …, 1 to 20, etc. We show the current sample (dev plus test) $OR$ variation in blue with dots, the low value of the confidence interval (CI) in orange with downward arrows and the high value of the CI in green with upward arrows, considering a 95% level of confidence.

RoBERT-small and RoBERT-large models show that the sample $OR$ statistic is close to 1 when F < 3: 1 for the small model and 1.4 for the large model. That is, being correct on unseen words happens at about the same rate as being correct on rarely seen words. Going up the frequency range, the $OR$ starts to increase in all cases: models learn to disambiguate the more frequently occurring words better, because they have seen more contexts of those words. Lastly, in all three plots we see that the sample $OR$ statistic sits comfortably within the limits of its CI, meaning that the value is very likely to be correct, and not smaller or greater than what we got.

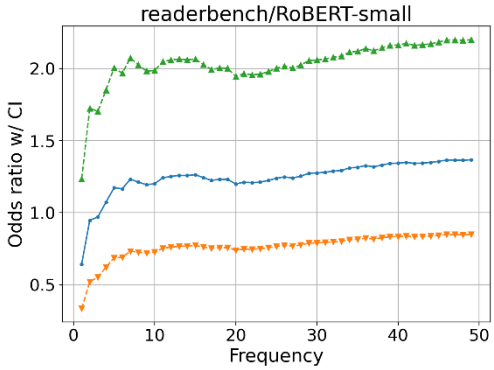


Figure 1: RoBERT-small OR variation with frequency

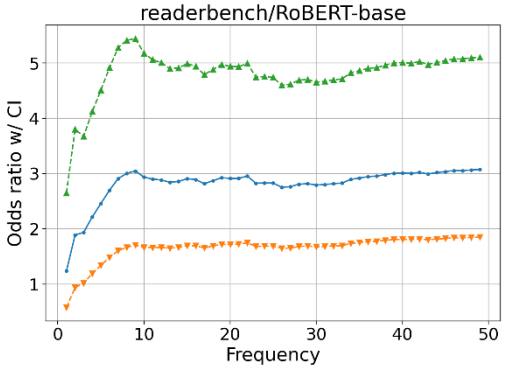


Figure 2: RoBERT-base OR variation with frequency

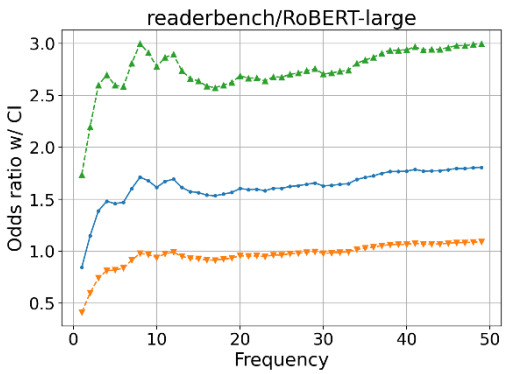


Figure 3: RoBERT-large OR variation with frequency

The RoBERT-base model is different, with the sample $OR$ settling around 2 for F < 3 indicating that this model is more likely to be correct when words have been barely seen in the training. By the time the frequency range gets to 10, the $OR$ statistic is 3, more than twice the one from the other two models.

### 4.4 Results with the CoRoLa BERT model

We previously trained a small BERT model (of approximately the same size as the RoBERT-small model) on the CoRoLA reference corpus for the contemporary Romanian language (Barbu Mititelu et al., 2019). We intended to use this model to study how a Transformer encoder learns the grammar of a language (in our case, Romanian). The model uses a vocabulary that is 13 times bigger than RoBERT's, wishing to account for the inflected nature of Romanian. The CoRoLa train set had just over 760 million words, and the CoRoLA BERT model was trained with the Masked Language Modeling training objective.

Table 6 shows accuracies at different frequencies, for all words (content plus functional) and for content words only.

| | **F ≥ 0** | **F = 0** | **F = 1** | **F > 0** |
|---|---|---|---|---|
| **All** | 93.8% | 76.9% | 89.6% | 95.8% |
| **Cont.** | 91.4% | 77.3% | 90.1% | 94.4% |

Table 6: CoRoLa BERT accuracy

We can compare these figures with the RoBERT-small's accuracies (Tables 4 and 5), as CoRoLa BERT is about the same size, parameter-wise. While RoBERT-small outperforms CoRoLa BERT at all categories, except for the accuracy on content words when the frequency F > 0, the biggest difference is when F = 0: more than 15 percents in favor of RoBERT-small. There are two explanations for this:

1. CoRoLa BERT has been under pre-trained for its massive vocabulary, which has 500K words vs. 38K words of RoBERT-small's. We pre-trained on only 760M words while RoBERT-small model was pre-trained on 2B words.
2. We only evaluate on words from the model's vocabulary, and thus, CoRoLa BERT is evaluated on many more words than RoBERT, at all frequency thresholds, because its vocabulary is much bigger. Just for the sake of comparison, RoBERT tokenizer recognizes 18K word occurrences in our test set while CoRoLa BERT tokenizer recognizes 27K word occurrences.

When we plot the variation of the $OR$ statistic with the frequency, as we did for the RoBERT models, we see the picture of an undertrained BERT model (see Figure 4, below).

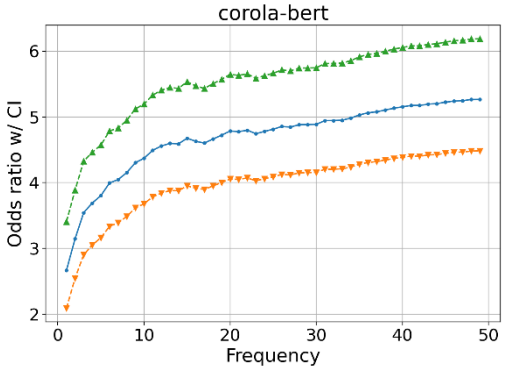


Figure 4: CoRoLa BERT OR variation with frequency

For F ≤ 3, the $OR$ statistic is already 3.5 and the function quickly increases, showing that this BERT model can do better POS tagging only on seen during training words.

### 4.5 A dimension-by-dimension hidden state analysis for POS tagging

The RoBERT-small and the CoRoLa BERT models have the same size of the hidden state vector: 256 dimensions, counted from 0 to 255. We wanted to know if we can find a common subset of dimensions that are responsible for the correct classification of each grammatical category.

To obtain the subset of dimensions that contribute the most to the output layer's highest probable grammatical category, we can use the output layer weight matrix from which we extract the row corresponding to the index of the most

probable grammatical category and multiply it, element-wise, with the hidden state of our target word. From the obtained vector, we extract the indexes of the top 10 largest elements, as the dimensions of the model hidden state that contribute the most to the correct grammatical category classification.

If we compute the most important 10 dimensions for each correctly classified word in the test set, we can derive a conditional probability distribution for each of the 35 grammatical categories, for both RoBERT-small and CoRoLa BERT. Table 7 shows which dimensions have been found as being common between RoBERT-small and CoRoLa BERT, for each grammatical category[3], with their sum of conditional probabilities.

| cat | ∑ P(d\|cat) | ∑ Q(d\|cat) | Common d |
|---|---|---|---|
| Af | 0.120 | 0.184 | 255, 28, 138, 31 |
| Cc | 0.037 | 0.066 | 77 |
| Cr | 0.068 | 0.112 | 213, 113 |
| Cs | 0.075 | 0.098 | 59, 10 |
| Dd | 0.043 | 0.012 | 52 |
| Di | 0.197 | 0.146 | 26, 65, 5, 84 |
| Ds | 0.024 | 0.013 | 22 |
| Dw | 0.137 | 0.171 | 147, 196, 104 |
| Mc | 0.152 | 0.239 | 88, 67, 218, 40, 197 |
| Mo | 0.044 | 0.048 | 101, 113 |
| Nc | 0.024 | 0.017 | 160 |
| Pd | 0.142 | 0.230 | 134, 141, 1, 9, 213, 143 |
| Pp | 0.072 | 0.160 | 234, 213, 11 |
| Pw | 0.061 | 0.102 | 234, 31 |
| Px | 0.010 | 0.012 | 103 |
| Pz | 0.110 | 0.155 | 220, 146, 187 |
| Qn | 0.105 | 0.110 | 5, 32 |
| Qs | 0.160 | 0.130 | 239, 249, 112, 128 |
| Qz | 0.038 | 0.017 | 74 |
| Rc | 0.052 | 0.101 | 212, 241 |
| Rg | 0.015 | 0.066 | 167 |
| Rp | 0.057 | 0.236 | 119, 115, 104 |
| Sp | 0.164 | 0.053 | 109, 134, 112 |
| Tf | 0.117 | 0.096 | 94, 192, 62 |
| Ti | 0.054 | 0.112 | 47, 22 |
| Ts | 0.098 | 0.113 | 36, 145 |
| Va | 0.041 | 0.139 | 49, 0 |
| Vm | 0.095 | 0.070 | 157, 193 |
| Yn | 0.084 | 0.072 | 255, 102, 179 |

Table 7: CoRoLa BERT and RoBERT-small common dimensions per grammatical category

From the cumulative probabilities of CoRoLa BERT (∑ P(d|cat)) and RoBERT-small (∑ Q(d|cat)), we see that the common dimensions do not carry a lot of the whole probability mass for a category. If the sums of the probabilities had been higher for both models (say above 0.5), that would have been an indication that the common set of categories is important for both models, but this is not the case here. Thus, we can conclude that different BERT models do not assign the same importance to the same dimensions for a given grammatical category.

## 5 Conclusions

We have presented evidence that properly trained BERT models exhibit learning words' grammatical categories, especially when the words were not seen during the training process. We drew this conclusion by measuring the odds ratio of POS tagging accuracy when the frequency of the test words (in the train set) is greater than 0 vs. when this frequency is 0. Thus, models RoBERT-small and RoBERT-large show an odds ratio that is less than 2 for the accuracy of tagging frequent words vs. tagging unseen words. We could not say that model CoRoLa BERT exhibits the same behavior due to its insufficient pre-training for its large vocabulary.

The model RoBERT-base shows a different behavior with respect to accuracy odds ratio vs. test word frequency: while the odds ratio of POS tagging accuracy is below 2 when comparing rare words ($F \leq 3$) to unseen words ($F = 0$), as in the case of the other two sibling models, when the frequency increases (e.g. $F \geq 10$), the odds ratio settles at a bit over 3 (twice as much when compared to the other two models). While it is expected that the POS tagging accuracy increases with the test word frequency (in the train set), as more contexts of those words were seen during training, RoBERT-base does much better than the

[3] For an explanation of the grammatical category codes, one can consult the MSD definitions from https://nl.ijs.si/ME/V6/msd/html/msd-ro.html

other two sibling models when test words were seen during training. This hypothesis is supported by the top POS tagging accuracy of RoBERT-base compared to any other tested model (see Tables 4 and 5). One possible explanation for this situation is that RoBERT-base has the best number of parameters (not too few, nor too many) for our POS tagging task and this enables its accuracy odds ratio curve to increase more sharply than siblings' curves, but not that sharply as the odds ratio curve of CoRoLa BERT which indicates more of an overfit of the training data than a good performance.

## References


Verginica Barbu Mititelu. 2018. Modern Syntactic Analysis of Romanian. In Ofelia Ichim, Luminiţa Botoşineanu, Daniela Butnaru, Marius-Radu Clim, Ofelia Ichim, Veronica Olariu (eds.), Clasic și modern în cercetarea filologică românească actuală, Iași, Publishing House of "Alexandru Ioan Cuza" University, 2018, pp. 67—78.

Verginica Barbu Mititelu, Dan Tufiș, Elena Irimia, Vasile Păiș, Radu Ion, Nils Diewald, Maria Mitrofan, Mihaela Onofrei. 2019. Little Strokes Fell Great Oaks. Creating CoRoLA, The Reference Corpus of Contemporary Romanian. Revue roumaine de linguistique, Issue 3. 2019

Bland, J., and Altman, D. 2000. Statistics notes: The odds ratio. British Medical Journal, 320 (7247), 1468.

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), pp. 4171–4186, Minneapolis, Minnesota.

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

Evangelos Dermatas and George Kokkinakis. 1995. Automatic stochastic tagging of natural language texts. Computational Linguistics, Volume 21, Issue 2, pp. 137—163.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. ArXiv preprint arXiv:1907.11692.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation: Contextualized word vectors. In Proceedings of NIPS, 2017.

Eleni Metheniti, Tim Van De Cruys, and Nabil Hathout. 2022. About Time: Do Transformers Learn Temporal Verbal Aspect? In Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pp. 88–101, Dublin, Ireland.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Proceedings of NAACL, 2018.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. https://blog.openai.com/language-unsupervised, 2018.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. OntoNotes release 5.0 LDC2013T19. Linguistic Data Consortium, Philadelphia, PA, 2013.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. Advances in Neural Information Processing Systems, 32:5753–5763.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 1112—1122, New Orleans, Louisiana.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. 2023. Attention is All You Need. arXiv:1706.03762v7 [cs.CL]