FrCoLA: a French Corpus of Linguistic Acceptability Judgments

Anonymous ACL submission

Abstract

Large foundation language models and Transformer-based neural language models have exhibited outstanding performance in various downstream tasks. However, there is limited understanding regarding how these models internalize linguistic knowledge, so various linguistic benchmarks have recently been proposed to facilitate syntactic evaluation of language models across languages. This paper introduces FrCoLA (French Corpus of Linguistic Acceptability Judgments), consisting of 25,153 sentences annotated with binary acceptability judgments and categorized into four linguistic phenomena. Specifically, those sentences are manually extracted from an official online resource maintained by a Québec Governments institution, and split into in-domain data splits. Moreover, we also manually extracted 2,675 from a second France-based organization source and created an out-of-domain hold-out split. We then evaluate the linguistic capabilities of three different language models for each of the seven linguistic acceptability judgment benchmarks. The results demonstrated that, for most languages, on average, fine-tuned Transformer-based neural language models are strong baselines on the binary linguistic acceptability classification tasks. However, for the FrCoLA benchmark, on average, a fine-tuned Transformer-based model outperformed other methods tested.

1 Introduction

The introduction of large foundation language models (LLM) and Transformer-based neural language model (Vaswani et al., 2017), such as GPT-3 (Brown et al., 2020) and LLaMA (Touvron et al., 2023), has led to major progress in natural language processing (NLP), substantially increasing the performance of most NLP tasks (Zhang et al., 2023). LLMs and Transformer-based neural language models were initially introduced for English (Kenton and Toutanova, 2019; Brown et al., 2020), but many other languages were later introduced, such as Norwegian (Kummervold et al., 2021), Russian (Kuratov and Arkhipov, 2019) and French (Martin et al., 2020). NLP research has approached the competencies evaluation of various natural language tasks of LLM with various benchmark corpora such as the English benchmarks GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019), and GLGE (Liu et al., 2021) to name a few. These corpora are collections of resources for training, evaluating, and analyzing natural language systems (Gao et al., 2023; Chang et al., 2023). For example, GLUE aims to benchmark an NLP system's capabilities for natural language understanding (NLU) (Wang et al., 2018). At the same time, GLGE focuses on natural language generation (NLG) tasks such as document summarization (Liu et al., 2021).

Recently, much effort has been put into creating linguistic acceptability resources to assess and benchmark LLM linguistic competency, where recent NLP research formulate linguistic competency as a binary classification acceptability judgments task (Cherniavskii et al., 2022; Proskurina et al., 2023). That is the ability, from a native speaker's perspective, to distinguish the correct form and naturalness of an acceptable sentence from an unacceptable one (Chomsky, 2014). Recently, similar non-English resources have been proposed to answer this question in typologically diverse languages such as Japanese (Someya et al., 2023), Norwegian (Jentoft and Samuel, 2023), and Chinese (Hu et al., 2023). However, the ability of LLMs to perform linguistic acceptability judgments in French remains understudied.

To this end, we introduce the **Fr**ench **C**orpus **o**f **L**inguistic **A**cceptability Judgments (FrCoLA)[1], a corpus consisting of 25,153 acceptability judgment sentences, making it the second largest linguistic acceptability resources available in the NLP

[1]Link removed for double-anonymized anonymity.

literature. Specifically, the sentences were created by French-Canadian linguists and manually extracted from the "*Banque de dépannage linguistique*[2]" (BDL), an official online resource maintained by the Québec Government.

The main contributions of this work are therefore 1) the creation and release of FrCoLA; 2) a set of experiments to assess the performance of Transformer-based models on FrCoLA; 3) a set of experiments using a monolingual and a cross-lingual Transformer-based model on English, Swedish, Italian, Russian, Chinese, Norwegian, Japanese and French, with the potential to open up novel multi-language research perspectives. It is outlined as follows: first, we study the available linguistic acceptability resources corpora and related binary classification neural language model research in section 2. Then, we propose the FrCoLA in section 3, and in section 4 and section 5 we present a set of experiments aimed at testing the performance of Transformer-based binary classifiers on all the linguistic acceptability resources corpora. Finally, in section 6, we conclude and discuss our future work.

2 Related Work

2.1 Linguistic Acceptability Judgments

Linguistic acceptability judgment constitutes a pivotal component of human linguistic competence, underlying individuals' inherent capacity to distinguish the correct form and naturalness of an acceptable sentence from an unacceptable one, even without formal grammar training. For instance, individuals can inherently distinguish between two sentences and identify the one that is more acceptable or natural-sounding. This assessment is the primary behavioural benchmark employed by generative linguists to investigate the underlying structure of human language (Chomsky, 2014). Through analyzing linguistic acceptability judgments, linguists can learn about the linguistic rules that govern languages and how these rules manifest themselves in native speakers' speech.

2.2 LLM Evaluation

Historically, evaluation of LLMs and Tansformer-based neural language models has been conducted using metrics or benchmark corpora (Chang et al., 2023; Awasthi et al., 2023). The first approach relies either on task-agnostic metrics, such as perplexity (Jelinek et al., 1977) which measures the quality of the probability distribution of words in a given corpus by a model, or alternatively on task-specific metrics, like the BLEU score that evaluates a model's performance for machine translation (Papineni et al., 2002). The second approach relies on large corpora designed for NLU or NLG downstream tasks. For example, the GLUE benchmark (Wang et al., 2018) is used to assess a model's NLU performance on tasks such as semantic similarity, linguistic acceptability judgment and sentiment analysis. In contrast, GLGE (Liu et al., 2021) evaluates language generation tasks such as summarization and question answering.

Accaptable Sentence	Not Accaptable Sentence
The cats annoy Tim.	The cats annoys Tim.

Table 1: Example of a minimal pair (Warstadt and Bowman, 2019).

2.2.1 LLM Linguistic Acceptability Judgments Evaluation

Recently, NLP researchers started using linguistic acceptability judgment tasks to assess the robustness of LLMs' and Tansformer-based neural language models against grammatical errors (Yin et al., 2020; Miaschi et al., 2023) and to probe their grammatical knowledge (Warstadt and Bowman, 2019; Zhang et al., 2021; Choshen et al., 2022; Mikhailov et al., 2022). Two approaches are used to perform this evaluation, namely minimal pairs and binary classification acceptability judgments (Wang et al., 2018; Warstadt and Bowman, 2019; Chang et al., 2023).

In the first approach, a set of minimal pairs of grammatically acceptable and unacceptable sentences, such as the pair illustrated in Table 1, is presented to an LLM that must decide which is grammatically correct. By observing which sentences the LLM assigns a higher correctness probability to, one can assess which grammatical phenomena it is sensitive to (Warstadt and Bowman, 2019) Corpus such as BLiMP in English (Warstadt and Bowman, 2019) and CLiMP in Chinese (Xiang et al., 2021) have been proposed to enable the evaluation of LLM on a wide range of linguistic phenomena.

Concurrently, in the second approach, a set of sentences that are either grammatical or ungram-

[2]https://vitrinelinguistique.oqlf.gouv.qc.ca/banque-de-depannage-linguistique

Label	Sentence
0 (Ungrammatical)	Edoardo returned to his last year city
1 (Grammatical)	This woman has impressed me

Table 2: Example sentences from the ItaCoLA dataset (Trotta et al., 2021).

matical, such as the two shown in Table 2, are provided to an LLM which must perform a binary acceptability classification task (Warstadt et al., 2019). Corpora such as CoLA in English (Warstadt et al., 2019) and CoLAC in Chinese (Hu et al., 2023) have been proposed to assess LLMs' capabilities to discriminate proper grammar from improper in their respective languages. Typically, these datasets comprise sentences collected from syntax textbooks and linguistics journals. However, as of yet no such corpus exists for French.

3 FrCoLA: French Corpus of Linguistic Acceptability Judgments

In this work, we introduce the **Fr**ench **C**orpus **o**f **L**inguistic **A**cceptability Judgments (FrCoLA), which will be the first large-scale binary linguistic acceptability judgments task dataset for the French language and the second-largest such corpus in any language. FrCoLA consists of French sentences from the "*Banque de dépannage linguistique*" (BDL), an official online resource from the "*Office québécois de la langue française*", an official provincial government public organization in Canada. The BDL is a grammatical resource index that offers 2,667 articles divided into eleven categories, such as "Spelling" (*orthographe*), "Grammar" (*grammaire*) and "Syntax" (*syntaxe*). These categories are further subdivided into sub-categories that address either a specific situation in the linguistic literature (e.g. how to use punctuation in a sentence) or a problematic linguistic situation (e.g. the use of borrowed words from English vocabulary). In these articles, the BDL explains various linguistic phenomena that are correct or incorrect and uses examples written by French linguists to illustrate both cases. For example, the "grammar" category includes the "adverbs" (*adverbes*) sub-category that includes an article about the linguistic phenomenon of proper and improper use of the adverb "surrounding" (*alentour*). A snippet of that article is shown in Figure 1. It displays two examples of well-written sentences using the adverb (marked in green) and one example of an erroneous usage (marked in red).

Figure 1: Snipped of the BDL article for the French word "*alentour*". The text is in French.

3.1 Data Collection

Sentences in FrCoLa were collected manually from the BDL online resource focusing on French grammar. Specifically, we examined all its 2,667 articles and extracted 25,153 linguistic acceptability judgment sentences. Each sentence was labelled 0 (ungrammatical) or 1 (grammatical) following the BDL green/red colour scheme as illustrated in Figure 1. Furthermore, since the BDL uses a fine-grained category structure to sort various linguistic phenomena, we collected these categories and associated them to labels according to the French linguistic literature (Fagyal et al., 2006; Chesley, 2010; Boivin and Pinsonneault, 2020; Feldhausen and Buchczyk, 2021), and labelled each extracted sentence accordingly. Our linguistic phenomena labels and their BDL-associated categories are listed below, and Table 3 present FrCoLA statistics for each one.

- **Syntax** : agreement violations, corruption of word order, misconstruction of syntactic clauses and phrases, incorrect use of appositions, violations of verb transitivity or argument structure, ellipsis, missing grammatical constituencies or words.
  BDL Categories: Editorial (*Rédaction*), Syntax (*Syntaxe*), Punctuation (*Ponctuation*), Typography (*Typographie*), and Proper nouns (*Noms propres*).

- **Morphology** : incorrect derivation or word building, non-existent words.
  BDL Categories: Ortograph (*Ortographe*), Grammar (*Grammaire*), Abbreviations and

symbols (*Abréviations et symboles*), and Pronunciation (*Prononciation*).

- **Semantic**: incorrect use of negation, violation of the verb's semantic argument structure.
  BDL Categories: Vocabulary (*Vocabulaire*).

- **Anglicism**: word and syntactical structure borrowed from English grammar.
  BDL Categories: Anglicisms (*Anglicisme*).

3.2 Analysis of FrCoLA

In this section, we compare our FrCoLA dataset to all related corpora. Table 4 present, for each corpus, the language, and the number of sentences, percentage of acceptable sentences and total vocabulary in the train, dev and test sets[3] as well as for the entire corpus. The total vocabulary sizes were computed using language-specific SpaCy tokenizers (Honnibal et al., 2020) that split each sentence into individual words or punctuation. We can see that FrCoLA is the second largest corpus behind only NoCoLA, and is approximately twice the size of all the other corpora. Moreover, FrCoLA shares a similar frequency of acceptable sentences to CoLA, CoLAC and RuCoLA datasets, and like with the other corpora all splits have a similar frequency of acceptable sentences. Finally, we can see that FrCoLA has the third-largest vocabulary size compared to the other datasets.

3.3 Out-of-Domain Hold-Out Split

Like CoLA, RuCoLA, CoLAC, and JCoLA, our FrCoLA includes an out-of-domain hold-out data split in addition to the standard train, dev, and test dataset splits to assess whether a system trained with it might suffer from overfitting. The definition of out-of-domain varies depending on the corpus. In CoLA, the out-of-domain set includes sources of varying degrees of domain specificity and time period compared to those used for the main dataset (Warstadt and Bowman, 2019). For RuCoLA, they are sentences generated by an automatic machine translation system and paraphrase generation models and annotated by a human annotator (Mikhailov et al., 2022). The JCoLA out-of-domain split comprises sentences from the Journal of East Asian Linguistics, a source with typically more complex linguistic phenomena than regular text (Someya et al., 2023). We have used a similar approach to JCoLA and CoLA; we have used an alternative and substantially different source to build our out-of-domain set. The source we picked is the *Académie française*, a France-based organization founded in 1635 as a "society of scholars" in science and literature (Académie française, 2024). This organization publishes a monthly online journal, *La langue française: Dire, Ne pas dire*[4] ("The French language: What to say, and not say"), that presents various articles on the proper and improper use of the French language, sorted into three categories, namely "neologisms and anglicisms" (*néologismes and anglicismes*), "wrongful employment" (*emplois fautifs*), and "abusive extensions of meaning" (*extensions de sens abusives*). We manually extracted all examples of grammatical and ungrammatical sentences from the 1,013 articles in the journal. In total, 2,675 sentences were extracted and binary labelled.

We present in Table 5 the number of sentences, vocabulary size and percentage of acceptable sentences of all linguistic corpora with an out-of-domain test set. However, since other corpora do not distribute their hold-out labels, we could not compute the percentage of acceptable sentences. We also note that for JCoLA, the out-of-domain hold-out split was unavailable in their official dataset GitHub repository. Once again, we can see that FrCoLA is the second largest corpus in terms of number of sentences and vocabulary size, with nearly as many sentences as RuCoLA. Compared to the main FrCoLA corpus in Table 4, we can see that the out-of-domain dataset comprises a much less diverse vocabulary, making it well distinct from the other splits. Finally, the out-of-domain hold-out split has a percentage of acceptable sentences nearly 15% lower than the overall corpus, making it more robust to highlight overfitting cases in machine learning models.

4 Experiments

We evaluate three neural-based methods for acceptability classification leveraging Transformer-based architecture.

[3]It is worth mentioning that for CoLA, RuCoLA and JCoLA, their in-domain test set labels are not publicly available to reduce the risk of overfitting. Thus, like other related work (Trotta et al., 2021; Cherniavskii et al., 2022), we use their out-of-domain dev sets as the test sets. Also, CoLAC does not offer a test set with a label or an out-of-domain dev set. Thus, as per the authors' recommendation, we have resampled the in-domain train and dev set using a 60-10-30% split using the seed 42 to create an in-domain test set.

[4]https://www.academie-francaise.fr/dire-ne-pas-dire

Category	# Sen and % Acp	Example
Syntax	5,152 77.56	*Dès son arrivée, on s'empressa de lui poser des questions à propos de son voyage.* ***Dès en arrivant, on s'empressa de lui poser des questions à propos de son voyage.***
Morphology	10,642 68.18	*La journée était calme : point de vent, point de bruit, point de mouvement.* ***La journée était calme : point de vent, poing de bruit, point de mouvement.***
Semantic	5,442 72.49	*Quand la parade est passée, le vieil homme s'est levé pour aller voir à la fenêtre.* ***Quand la parade est passée, le vieil homme s'est levé debout pour aller voir à la fenêtre.***
Anglicism	3,917 58.26	*Sauront-ils répondre aux les besoins de l'enfant?* ***Sauront-ils rencontrer les besoins de l'enfant?***

Table 3: Number of sentences (# Sen) and the percentage of acceptable sentences (% Acp) per category in FrCoLA (all three splits), and example of a positive and a negative (**bolded** with error underlined) in each category.

	Language	Train			Dev			OODD/Test			Total		
		# Sen	% Acp	Vocab	# Sen	% Acp	Vocab	# Sen	% Acp	Vocab	# Sen	% Acp	Vocab
CoLA (Warstadt et al., 2019)	English	8,551	70.44	5,778	527	69.26	1,375	516	68.60	988	9,594	70.27	6,097
DaLAJ (Volodina et al., 2021)	Swedish	7,682	50.00	6,841	890	50.00	1,799	888	50.00	1,661	9,460	50.00	7,884
ITACoLA (Trotta et al., 2021)	Italian	7,801	84.39	5,825	946	85.41	1,844	1,888	84.21	1,888	9,722	84.47	6,402
RuCoLA (Mikhailov et al., 2022)	Russian	7,869	74.52	19,057	983	74.57	4,140	1,804	63.69	9,353	10,656	72.69	26,382
CoLAC (Hu et al., 2023)	Chinese	4,134	66.09	3,835	460	66.96	1,024	1,970	67.82	2,636	6,564	66.67	4,759
NoCoLA (Jentoft and Samuel, 2023)	Norwegian	116,195	31.46	32,561	14,289	32.59	8,865	14,383	31.58	8,600	144,867	31.58	37,319
JCoLA (Someya et al., 2023)	Japanese	6,919	83.38	3,730	865	83.93	1,483	684	73.28	896	8,469	82.62	4,146
FrCoLA (This work)	French	15,846	69.49	18,350	1,761	69.51	5,369	7,546	69.49	12,690	25,153	69.49	22,131

Table 4: Comparison of FrCoLA and related corpora for number of sentences (# Sen), percentage of acceptable sentences (% Acp), and total vocabulary (Vocab). "OODD" stands for out-of-domain data split (CoLA, RuCoLA and JCoLA).

	Out-of-Domain Hold-Out		
	# Sen	Vocab	% Acp
CoLA	533	1035	N/A
RuCoLA	2,789	12,211	N/A
CoLAC	931	1,168	N/A
JCoLA	N/A	N/A	N/A
FrCoLA	2,675	1,651	53.91

Table 5: Comparison of FrCoLA with all related corpus with an out-of-domain hold-out set for the number of sentences (# Sen), the vocabulary size (Vocab) and the percentage of acceptable sentences (% Acp).

4.1 Evaluation Metrics

Following Warstadt et al. (2019), performance is measured using the accuracy score (Acc) and Matthews Correlation Coefficient (MCC) (Matthews, 1975). Accuracy on the dev set is used as the target metric for hyperparameter tuning and early stopping. We report the results averaged over ten restarts from different random seeds (i.e. $42, 43, \cdots, 50, 51$).

4.2 Models

4.2.1 Baseline

As a baseline, we evaluate a fine-tuned language-specific pre-trained monolingual and a cross-lingual neural language model using each available linguistic acceptability judgment binary classification benchmark dataset and FrCoLA. We used a different language-specific Transformed-based LLM for each language, which was optimized using different tokenization methods and training corpus. We detail the language-specific models used in our experimentation in Appendix A. We use XLM-RoBERTa-base (Conneau et al., 2020) as our cross-lingual neural language model; the model details are described in Appendix A. For each language, we name these obtained models `Monolingual FT` and `Cross-Lingual FT`, respectively.

4.2.2 State-Of-The-Art

The state-of-the-art approach to binary linguistic acceptability judgments is the topological data analysis (TDA) proposed by Cherniavskii et al. (2022). This approach extracts the attention maps of a fine-tuned Transformers-based neural language model to use as linguistic features to train a binary logistic regression. The authors report that this approach significantly outperformed previous approaches, in-

creasing by up to 0.24 Matthew's correlation coefficient score on linguistic acceptability for three languages (English, Italian and Swedish). In our case, we use the attention maps from the monolingual models we fine-tuned as baselines. For each language, we name this model `LA-TDA`.

4.3 Training Settings

Each language model is fine-tuned using the train and dev split and evaluated using the test or out-of-domain valid split (OODD) (CoLA, RuCoLA and JCoLA) following the standard procedure under the HuggingFace library (Wolf et al., 2020). Each model is fine-tuned for four epochs and uses the AdamW optimizer (Loshchilov and Hutter, 2018), with a learning rate of $3e{-}5$ and a weights decay of $1e{-}2$. Since the corpora are unbalanced, we use a weighted balanced loss based on the training percentage of acceptable sentences. We use a batch size of 32 and the other default hyperparameters. For each language model, we use the default tokenizer with a maximum sequence length of 64 tokens without lowercasing during tokenization.

5 Results and Discussion

Table 6 presents the accuracy and the MCC of our three models for each benchmark dataset on the dev and test set, with **bolded** value indicating the best score per benchmark. The table reports the average and one standard deviation over the ten restarts. We observe that, for most languages, on average `LA-TDA` outperforms other fine-tuned methods, but not on all metrics and with a smaller margin than reported by Cherniavskii et al. (2022). The two exceptions to this are CoLA and our FrCoLA. FrCoLA's performance is always slightly better when using the fine-tuned `Moolingual FT` model. Considering that `LA-TDA` is computed asymptotically in quadratic time (Cherniavskii et al., 2022), the performance gains seem marginal compared to the added computational expense. These results show that fine-tuned Transformer-based neural language models are strong baselines for the binary linguistic acceptability classification tasks.

We present in Table 7 the accuracy and the MCC of our three models trained using FrCoLA over the dataset's four categories. The table reports the average and one standard deviation over the ten restarts. We can see that the category "anglicism" has the lowest performance. For the two approaches using monolingual LLM (i.e. `Monolingual FT` and `LA-TDA`), we hypothesize that this situation is due to occurrences of anglicism in the LLM training dataset. Indeed, using word and syntactical structure borrowed from English grammar is more common over web-based (Laviosa, 2010; Planchon and Stockemer, 2019; Solano, 2021; Šukalić et al., 2022) and even official educational text (Şimon et al., 2021). Thus, fine-tuning the pre-trained LLM model can be more challenging, considering that the "anglicism" category contains the least examples. For the cross-lingual approach, since the LLM has learned word representation over English during training, we hypothesize that sentences using English words or syntax are considered more probable for the model; thus, it is more challenging for the classifier to classify these examples correctly.

Finally, we present in Table 8 the accuracy and the MCC of our three models trained using FrCoLA but evaluated using our out-of-domain hold-out set. The table reports the average and one standard deviation over the ten restarts. We can see that, once again, the `Moolingual FT` model outperforms the `LA-TDA` model. However, all three models show significant performance drops, of nearly 22% in accuracy and nearly 50% for the MCC. It shows that the fine-tuned models have overfitted over the train and dev dataset.

	Out-of-Domain Hold-Out	
	Acc (%)	MCC
`Monolingual FT`	$\mathbf{62.69 \pm 1.13}$	$\mathbf{0.286 \pm 0.020}$
`Cross-Lingual FT`	55.99 ± 4.36	0.107 ± 0.088
`LA-TDA`	61.36 ± 0.90	0.090 ± 0.019

Table 8: Acceptability binary classification result on the FrCoLA out-of-domain hold-out set. The best score per benchmark is **bolded**.

6 Conclusion and Future Works

This article introduced FrCoLA, the French Corpus of Linguistic Acceptability Judgments, a dataset comprising 15,846 sentences annotated with binary acceptability judgments. It is the first such corpus in French, and the second-biggest one in any language. The sentences it comprises were manually extracted from an official online linguistic resource maintained by a Québec Government institution, and an additional out-of-domain dataset was compiled from an equivalent French institution. We then evaluated the linguistic performances of three fine-tuned models on FrCoLA and

Model	Dev		Test/OODD	
	Acc (%)	MCC	Acc (%)	MCC
		CoLA		
Monolingual FT	83.61 ± 2.56	$\mathbf{0.639 \pm 0.030}$	$\mathbf{80.89 \pm 1.15}$	$\mathbf{0.544 \pm 0.025}$
Cross-Lingual FT	82.24 ± 1.35	0.575 ± 0.033	77.25 ± 2.42	0.452 ± 0.041
LA-TDA	$\mathbf{84.91 \pm 1.24}$	0.633 ± 0.031	80.70 ± 1.38	0.532 ± 0.034
		DaLAJ		
Monolingual FT	69.12 ± 1.53	$\mathbf{0.411 \pm 0.029}$	72.33 ± 1.40	0.467 ± 0.025
Cross-Lingual FT	55.18 ± 5.90	0.131 ± 0.144	55.21 ± 5.89	0.124 ± 0.137
LA-TDA	$\mathbf{70.08 \pm 1.24}$	$\mathbf{0.411 \pm 0.024}$	$\mathbf{73.54 \pm 1.05}$	$\mathbf{0.475 \pm 0.020}$
		ITACoLA		
Monolingual FT	83.29 ± 3.71	0.420 ± 0.051	83.45 ± 3.34	$\mathbf{0.446 \pm 0.050}$
Cross-Lingual FT	79.97 ± 6.22	0.105 ± 0.121	79.12 ± 5.99	0.117 ± 0.124
LA-TDA	$\mathbf{87.51 \pm 0.88}$	$\mathbf{0.423 \pm 0.050}$	$\mathbf{86.59 \pm 0.93}$	0.422 ± 0.054
		RuCoLA		
Monolingual FT	74.49 ± 2.56	$\mathbf{0.352 \pm 0.027}$	66.81 ± 3.56	0.379 ± 0.030
Cross-Lingual FT	71.84 ± 3.00	0.276 ± 0.038	56.81 ± 3.18	0.189 ± 0.026
LA-TDA	$\mathbf{77.56 \pm 0.61}$	0.337 ± 0.022	$\mathbf{71.09 \pm 0.92}$	$\mathbf{0.382 \pm 0.018}$
		CoLAC		
Monolingual FT	75.93 ± 1.35	0.444 ± 0.027	77.78 ± 1.43	0.482 ± 0.023
Cross-Lingual FT	73.37 ± 2.72	0.337 ± 0.022	71.09 ± 0.92	0.382 ± 0.018
LA-TDA	$\mathbf{77.33 \pm 1.79}$	$\mathbf{0.469 \pm 0.044}$	$\mathbf{79.01 \pm 0.86}$	$\mathbf{0.502 \pm 0.023}$
		NoCoLA		
Monolingual FT	77.90 ± 0.96	0.560 ± 0.009	77.90 ± 0.98	0.560 ± 0.009
Cross-Lingual FT	73.92 ± 1.40	0.504 ± 0.017	73.79 ± 1.37	0.505 ± 0.015
LA-TDA	$\mathbf{81.58 \pm 0.29}$	$\mathbf{0.582 \pm 0.007}$	$\mathbf{82.01 \pm 0.31}$	$\mathbf{0.589 \pm 0.009}$
		JCoLA		
Monolingual FT	81.34 ± 4.48	0.039 ± 0.062	73.17 ± 0.61	0.067 ± 0.111
Cross-Lingual FT	72.64 ± 8.11	$\mathbf{0.262 \pm 0.058}$	72.86 ± 4.61	$\mathbf{0.328 \pm 0.059}$
LA-TDA	$\mathbf{83.49 \pm 0.68}$	0.252 ± 0.051	$\mathbf{75.30 \pm 1.25}$	0.230 ± 0.070
		FrCoLA		
Monolingual FT	$\mathbf{84.51 \pm 0.78}$	$\mathbf{0.619 \pm 0.02}$	$\mathbf{82.92 \pm 0.61}$	$\mathbf{0.578 \pm 0.015}$
Cross-Lingual FT	70.67 ± 15.13	0.243 ± 0.263	69.91 ± 14.61	0.222 ± 0.240
LA-TDA	84.00 ± 0.48	0.606 ± 0.013	82.79 ± 0.45	0.574 ± 0.012

Table 6: Acceptability binary classification results and MCC by language. The best score per benchmark is **bolded**. "OODD" stands for out-of-domain data split (CoLA, RuCoLA and JCoLA).

Model	Category			
	Syntax	Morphology	Semantic	Anglicism
	Test Accuracy (%)			
Monolingual FT	$\mathbf{88.59 \pm 0.60}$	$\mathbf{81.76 \pm 0.74}$	$\mathbf{85.82 \pm 0.40}$	$\mathbf{74.36 \pm 1.40}$
Cross-Lingual FT	83.31 ± 4.31	74.93 ± 4.70	79.84 ± 4.88	63.79 ± 4.66
LA-TDA	88.40 ± 0.23	81.49 ± 0.51	85.39 ± 0.53	74.18 ± 1.44
	Test MCC			
Monolingual FT	$\mathbf{0.654 \pm 0.018}$	$\mathbf{0.563 \pm 0.017}$	$\mathbf{0.620 \pm 0.011}$	$\mathbf{0.506 \pm 0.028}$
Cross-Lingual FT	0.403 ± 0.279	0.327 ± 0.226	0.378 ± 0.261	0.223 ± 0.156
LA-TDA	0.649 ± 0.009	0.555 ± 0.013	0.609 ± 0.014	0.405 ± 0.026

Table 7: Acceptability binary classification results and MCC for FrCoLA per category. The best score is **bolded**.

equivalent datasets in other languages. Our results demonstrated that Transformer-based neural language models achieve high results on the binary classification task and are strong baselines. When fined-tuned on FrCoLA, a Transformer-based neural language model even outperforms the state-of-the-art TDA method proposed by Cherniavskii et al. (2022).

In future works, we plan to extend our dataset linguistic phenomena granularity and generate the complementary grammatical or ungrammatical sentence of each sentence in the dataset to create the first French minimal pair benchmark dataset. Moreover, we would also like to qualitatively explore the linguistic phenomena errors generated by the French fine-tuned Transformer-base model.

Limitations

All the sentences included in FrCoLA have been extracted from an official linguistic source on theoretical syntax. Therefore, those sentences are guaranteed to be theoretically meaningful, making FrCoLA a challenging dataset. However, the categories extracted automatically from the official source are skewed. Indeed, as shown in Table 3, nearly 42% of the dataset comprises morphological linguistic phenomena.

Ethical Considerations

FrCoLA may serve as training data for binary linguistic acceptability judgment classifiers (Batra et al., 2021), which may benefit the quality of generated texts. We acknowledge that such text generation progress could lead to misusing LLMs for malicious purposes, such as disinformation or harmful text generation and online harassment (Weidinger et al., 2021; Bender et al., 2021). Nevertheless, our corpus can be used to train adversarial defence against such misuse and to train artificial text detection models (Lewis and White, 2023; Kumar et al., 2023).

Acknowledgements

Removed for double-anonymized.

References

Académie française. 2024. L'institution et l'organisation *(The Institution and the Organization)*. Accessed: 2024-02-10.

Raghav Awasthi, Shreya Mishra, Dwarikanath Mahapatra, Ashish Khanna, Kamal Maheshwari, Jacek Cywinski, Frank Papay, and Piyush Mathur. 2023. Humanely: Human Evaluation of Llm Yield, Using a Novel Web-Based Evaluation Tool. *medRxiv*, pages 2023–12.

Soumya Batra, Shashank Jain, Peyman Heidari, Ankit Arun, Catharine Youngs, Xintong Li, Pinar Donmez, Shawn Mei, Shiunzu Kuo, Vikas Bhardwaj, Anuj Kumar, and Michael White. 2021. Building adaptive acceptability classifiers for neural NLG. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 682–697, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the ACM conference on fairness, accountability, and transparency*, pages 610–623.

Marie-Claude Boivin and Reine Pinsonneault. 2020. La catégorisation des erreurs linguistiques: une grille de codage fondée sur la grammaire moderne. *Le français aujourd'hui*, 2(209):89–116.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models Are Few-Shot Learners. *Advances in neural information processing systems*, 33:1877–1901.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*.

Wanxiang Che, Zhenghua Li, and Ting Liu. 2010. LTP: A Chinese Language Technology Platform. In *International Conference on Computational Linguistics*, page 13. Citeseer.

Daniil Cherniavskii, Eduard Tulchinskii, Vladislav Mikhailov, Irina Proskurina, Laida Kushnareva, Ekaterina Artemova, Serguei Barannikov, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. 2022. Acceptability judgements via examining the topology of attention maps. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 88–107, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Paula Chesley. 2010. Lexical Borrowings in French: Anglicisms as a Separate Phenomenon. *Journal of French Language Studies*, 20(3):231–251.

Noam Chomsky. 2014. *Aspects of the Theory of Syntax*. 11. MIT press.

Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. 2022. The grammar-learning trajectories of neural language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 8281–8297, Dublin, Ireland. Association for Computational Linguistics.

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 Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training With Whole Word Masking for Chinese Bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Zsuzsanna Fagyal, Douglas Kibbee, and Frederic Jenkins. 2006. *French: A Linguistic Introduction*. Cambridge University Press.

Ingo Feldhausen and Sebastian Buchczyk. 2021. Revisiting Subjunctive Obviation in French: A Formal Acceptability Judgment Study. *Glossa: a journal of general linguistics. 6 (1): 59.*

Philip Gage. 1994. A New Algorithm for Data Compression. *C Users Journal*, 12(2):23–38.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A Framework for Few-Shot Language Model Evaluation.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. SpaCy: Industrial-strength Natural Language Processing in Python.

Hai Hu, Ziyin Zhang, Weifang Huang, Jackie Yan-Ki Lai, Aini Li, Yina Ma, Jiahui Huang, Peng Zhang, and Rui Wang. 2023. Revisiting Acceptability Judgements. *arXiv:2305.14091*.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a Measure of the Difficulty of Speech Recognition Tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Matias Jentoft and David Samuel. 2023. NoCoLA: The Norwegian Corpus of Linguistic Acceptability. In *Proceedings of the Nordic Conference on Computational Linguistics*, pages 610–617.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Taku Kudo. 2005. MeCab: Yet Another Part-Of-Speech and Morphological Analyzer. *http://mecab. sourceforge. net/*.

Taku Kudo and John Richardson. 2018. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2023. Mitigating societal harms in large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 26–33, Singapore. Association for Computational Linguistics.

Per Kummervold, Freddy Wetjen, and Javier De La Rosa. 2022. The Norwegian Colossal Corpus: A Text Corpus for Training Large Norwegian Language Models. In *Proceedings of the Language Resources and Evaluation Conference*, pages 3852–3860.

Per E Kummervold, Javier De la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. 2021. Operationalizing a national digital library: The case for a Norwegian transformer model. In *Proceedings of the Nordic Conference on Computational Linguistics*, pages 20–29, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. *arXiv:1905.07213*.

Sara Laviosa. 2010. Corpus-Based Translation Studies 15 Years On: Theory, Findings, Applications.

Ashley Lewis and Michael White. 2023. Mitigating harms of LLMs via knowledge distillation for a virtual museum tour guide. In *Proceedings of the Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants!*, pages 31–45, Prague, Czech Republic. Association for Computational Linguistics.

Dayiheng Liu, Yu Yan, Yeyun Gong, Weizhen Qi, Hang Zhang, Jian Jiao, Weizhu Chen, Jie Fu, Linjun Shou, Ming Gong, et al. 2021. GLGE: A New General Language Generation Evaluation Benchmark. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, pages 408–420.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with Words at the National Library of Sweden – Making a Swedish BERT.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Brian W Matthews. 1975. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2023. On Robustness and Sensitivity of a Neural Language Model: A Case Study on Italian L1 Learner Errors. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:426–438.

Vladislav Mikhailov, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov, and Ekaterina Artemova. 2022. RuCoLA: Russian Corpus of Linguistic Acceptability. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 5207–5227.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora*, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the annual meeting of the Association for Computational Linguistics*, pages 311–318.

Cecile Planchon and Daniel Stockemer. 2019. Anglicisms, French Equivalents, and Language Attitudes Among Quebec Undergraduates. *British Journal of Canadian Studies*, 32(12):93–118.

Irina Proskurina, Ekaterina Artemova, and Irina Piontkovskaya. 2023. Can BERT eat RuCoLA? Topological Data Analysis to Explain. In *Proceedings of the Workshop on Slavic Natural Language Processing*, pages 123–137.

Stefan Schweter. 2020. Italian BERT and ELECTRA Models.

Ramon Marti Solano. 2021. Anglicisms and Corpus Linguistics: Corpus-Aided Research into the Influence of English on European Languages. Introduction.

Taiga Someya, Yushi Sugimoto, and Yohei Oseki. 2023. JCoLA: Japanese Corpus of Linguistic Acceptability. *arXiv:2309.12676*.

Đelaludina Šukalić, Edina Rizvić-Eminović, and Adnan Bujak. 2022. A Corpus-Based Study of Anglicisms Across Different Text Types of Online News. *Journal of French Language Studies*, 20(3):231–251.

Masatoshi Suzuki and Ryo Takahashi. 2019. BERT base Japanese (IPA dictionary). `https://huggingface.co/tohoku-nlp/bert-base-japanese`. Accessed: 2024-02-10.

Jörg Tiedemann. 2016. OPUS-Parallel Corpora for Everyone. *Baltic Journal of Modern Computing*, 4(2).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and Efficient Foundation Language Models. *arXiv:2302.13971*.

Daniela Trotta, Raffaele Guarasci, Elisa Leonardelli, and Sara Tonelli. 2021. Monolingual and Cross-Lingual Acceptability Judgments With the Italian Cola Corpus. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 2929–2940.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in neural information processing systems*, 30.

Elena Volodina, Yousuf Ali Mohammed, and Julia Klezl. 2021. Dalaj–a dataset for linguistic acceptability judgments for swedish. In *Proceedings of the Workshop on NLP for Computer Assisted Language Learning*, pages 28–37.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A Stickier Benchmark for General-Purpose Language Understanding Systems. *Advances in neural information processing systems*, 32.

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 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Warstadt and Samuel R Bowman. 2019. Linguistic Analysis of Pretrained Sentence Encoders With Acceptability Judgments. *arXiv:1901.03438*.

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

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and Social Risks of Harm From Language Models. *arXiv:2112.04359*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Édouard Grave. 2020. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In *Proceedings of the Language Resources and Evaluation Conference*, pages 4003–4012.

Wikimedia Foundation. 2024. Wikimedia downloads.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A Benchmark for Chinese Language Model Evaluation. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790.

Fan Yin, Quanyu Long, Tao Meng, and Kai-Wei Chang. 2020. On the robustness of language encoders against grammatical errors. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 3386–3403, Online. Association for Computational Linguistics.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction Tuning for Large Language Models: A Survey. *arXiv:2308.10792*.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 1112–1125, Online. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

Dmitry Zmitrovich, Alexander Abramov, Andrey Kalmykov, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergey Markov, Vladislav Mikhailov, and Alena Fenogenova. 2023. A Family of Pretrained Transformer Language Models for Russian.

Simona Șimon, Claudia E Stoian, Anca Dejica-Cartiș, and Andrea Kriston. 2021. The use of anglicisms in the field of education: A comparative analysis of Romanian, German, and French. *Sage Open*, 11(4):21582440211053241.

A Pre-Trained Monolingual and Cross-Lingual Neural Language Models Details

We selected the models based on their performance and number of downloads on the HuggingFace Hub[5].

[5] https://huggingface.co/

Dataset	Transformer Model	# Layers	# Att Hds	Hid St Dim	Tokenization	Training Dataset
CoLA	bert-base-cased (Kenton and Toutanova, 2019)	12	12	768	WordPiece (Kenton and Toutanova, 2019)	BooksCorpus (Zhu et al., 2015) and English Wikipedia (Wikimedia Foundation, 2024)
DaLAJ	bert-base-swedish-cased (Malmsten et al., 2020)	12	12	768	SentencePiece (Kudo and Richardson, 2018)	National Library of Sweden Corpus (Malmsten et al., 2020)
ITACoLA	bert-base-italian-cased (Schweter, 2020)	12	12	768	Not specified	OPUS Corpus (Tiedemann, 2016)
RuCoLA	ruBert-base (Zmitrovich et al., 2023)	12	12	768	BPE (Gage, 1994)	Various Russian corpus (i.e. Russian Wikipedia and Russian news) (Zmitrovich et al., 2023)
CoLAC	bert-base-chinese (Cui et al., 2021)	12	12	768	LTP (Che et al., 2010) and WordPiece	Chinese Wikipedia
NoCoLA	nb-bert-base (Kummervold et al., 2021)	12	12	768	WordPiece	Norwegian Colossal Corpus (Kummervold et al., 2022)
JCoLA	bert-base-japanese (Suzuki and Takahashi, 2019)	12	12	768	MeCab (Kudo, 2005) and WordPiece	Japanese Wikipedia
FrCoLA	camembert-base (Martin et al., 2020)	12	12	768	SentencePiece	OSCAR (Ortiz Suárez et al., 2019)
Cross-lingual	xlm-roberta-base (Conneau et al., 2020)	12	12	768	SentencePiece	CommonCrawl (Wenzek et al., 2020)

Table 9: Details of the pre-trained transformer models used for each linguistic acceptability corpus. For each model, it presents the number (#) of layers and attention heads (Att Hds) along with hidden states dimension (Hid St Dim), tokenization and training dataset.