Cross-linguistic comparison of linguistic feature encoding in BERT models for typologically different languages

Yulia Otmakhova[1], **Karin Verspoor**[2], **Jey Han Lau**[1]
[1]The University of Melbourne, [2]RMIT University
yotmakhova@student.unimelb.edu.au, karin.verspoor@rmit.edu.au,
jeyhan.lau@gmail.com

Abstract

Though recently there have been an increased interest in how pre-trained language models encode different linguistic features, there is still a lack of systematic comparison between languages with different morphology and syntax. In this paper, using BERT as an example of a pre-trained model, we compare how three typologically different languages (English, Korean, and Russian) encode morphology and syntax features across different layers. In particular, we contrast languages which differ in a particular aspect, such as flexibility of word order, head directionality, morphological type, presence of grammatical gender, and morphological richness, across four different tasks.

1 Introduction

Transformers (Vaswani et al., 2017) and especially pre-trained language models based on them, such as BERT (Devlin et al., 2019), had a revolutionary impact on the field of Natural Language Processing (NLP), allowing to achieve new heights in classification, retrieval, text understanding, and generation tasks. However, though major progress was made in adapting Transformers to different downstream tasks, they largely remain black-box models, especially from the linguistic point of view. In particular, though we know that they roughly follow the same pipeline as human-made natural language processing (NLP) systems when encoding the features (the lower layers of a Transformer encode part-of-speech information, the middle layers perform syntax parsing, while the top layers enable such tasks as coreference resolution) (Tenney et al., 2019), it is still unclear if there is a systematic relation between the type of morphological, syntactical and discourse features encoded and particular layers of Transformer-based models. More importantly, though there have been some attempts to examine this for languages other than English (see, for example, a study by de Vries et al. (2020) for Dutch), we still do not know if there is a consistency between models for different languages in encoding such features, especially if the languages in question are typologically different. Thus it is important to analyse how Transformers encode different linguistic features for dissimilar languages, as it would help us to better understand how they work and ultimately allow to improve their performance in such tasks as translation or multi-lingual information retrieval and summarisation.

In this study, we compare how Transformers encode particular linguistic features for the following languages: English, Russian, and Korean. The choice of languages are motivated linguistically: English, Russian and Korean are morphologically very distant languages (analytical, fusional and agglutinative respectively) which are also different in term of syntax such as word order. To examine Transformers encodings more systematically, we conduct a series of pairwise comparisons of languages which contrast in a particular linguistic feature, similarly to how it is performed in theoretical linguistics. In particular, using targeted manipulation of inputs to produce a binary correctness classification or a masked token prediction task, we examine the following research questions:

1. How sensitive are the encoders and their particular layers to correct word orders in languages with fixed and free word order?

2. Does the encoding of word order depend on the morphological type of the language (agglutinative vs inflected) and on its head directionality (head-initial vs head-final)?

3. How well is long-distance agreement encoded for languages with rich and poor morphology and agreement patterns?

4. Are gender biases encoded more strongly in languages with agreement in gender?

Proceedings of the 4th Workshop on Computational Typology and Multilingual NLP (SIGTYP 2022), pages 27 - 35

In the following sections we describe our experiments for these tasks and present our findings.

2 Sensitivity to word order in languages with fixed vs free word order

In this section we compare the ability of a BERT-based classification model to detect the corruption of word order in languages with fixed vs free word order. We hypothesise that as languages with free word order allow for more permutations in terms of token positions in the sentences, it would be more difficult for the model to detect word order corruption.

2.1 Fixed and free word order

We use English as an example of a fixed word order language and Russian as a free word order language as they are related languages which typologically differ in one aspect: while in Russian the grammatical and syntactical meaning is mostly expressed through morphological means such as suffixes and particles which allows the words to be relatively unconstrained in terms of their position in the sentence, in English due to limited morphology the grammatical and syntactical meaning is linked to the position of a word in a sentence and thus the word order is fixed. For example, to show that the word "apples" is an object rather than a subject, it should occupy the position after the predicate (verb) in English:

Emma	ate	apples.
Subject	*Verb*	*Object*

On the other hand, in Russian the word “apples” can potentially occupy both the position before and after the verb without changing the meaning[1]:

Эмма	ела	яблоки
Emma	ate	apples.
Subject	*Verb*	*Object (normal word order)*
Яблоки	ела	Эмма.
Apples	ate	Emma.
Object	*Verb*	*Subject*
(inverted word order, more focus on “Emma”)		

Despite this flexibility, the word order in Russian is not random or arbitrary: there is still a strong tendency for constituents (such as noun phrases and predicates) to occupy a particular position, and the movement of words across the constituent borders is very limited. However, as there is still more word movement allowed compared to English, we postulate that it would be more difficult to automatically detect ungrammatical word order changes in Russian than in English.

[1]Though, as we explain in Section 3, there is a strong preference for the position after the verb.

2.2 Dataset

For the experiments in this section we use UMC 0.1, a Czech-Russian-English corpus of news articles automatically aligned at sentence level (Klyueva and Bojar, 2008). We chose to use a parallel corpus for this task to ensure that the difficulty of classification is not affected by the syntactic complexity or the length of the sentences. We use the train/test data split provided by the authors of the corpus. For both the train and test data we remove the pairs where either the English or Russian sentence contains less than two tokens, since otherwise it is impossible to swap tokens. We also remove pairs where either of the sentences has over 100 tokens, as the task’s difficulty would increase in case of very long inputs. The statistics of the resulting dataset are provided in Table 1.

2.2.1 Model and experiments

The *bigram shift* (BShift) probing task introduced by Conneau et al. (2018) allows to check the capability of a model to distinguish between sentences with correct and incorrect word orders. Specifically, it is a binary classifier which has to distinguish between intact sentences and sentences where some two random adjacent tokens were swapped. For this task, we randomly sample half of the sentences in the training and test datasets and corrupt the token order in them at a random position.

We use BERT-Base Cased (Devlin et al., 2019)[2] as a pre-trained model for English and RuBERT (Kuratov and Arkhipov, 2019)[3], which was initialized with the multilingual version of BERT-Base and trained on the Russian Wikipedia and news, for Russian. We choose these particular models as the most closely matching in terms of their architecture and parameters (12-layer, 768-hidden, 12-heads for both models, 110M parameters for BERT-Base Cased and 180M Parameters for RuBERT). Unlike most studies which use the uncased version of BERT, we choose the cased one, as only

[2]https://huggingface.co/bert-base-cased

[3]https://huggingface.co/DeepPavlov/rubert-base-cased

cased models are available for Russian.

To ensure that the classification results reflect the performance of the pre-trained model itself on the task, we add a only single linear layer on top of the pre-trained model, freeze the BERT layers, and train the model only for 1 epoch. For the training optimization we use Adam (Kingma and Ba, 2014) with the learning rate of $2e^{-5}$. As usual for the classification tasks, we use the embedding of the first token [CLS] as the input to the linear layer. However, to explore how well the word order information is encoded in the different layers of the pre-trained model, we do this not only for the last layer, but for all layers in the pre-trained model.

2.3 Results

The results of the BShift classification tasks for all layers of BERT-Base Cased and RuBERT models are shown in Table 2. We report classification accuracy averaged over 5 runs with various random seeds. For all layers, the difference between the accuracy of the models was statistically significant, while the variation among the different runs was minimal, which shows that there is a visible difference in the ability of the Russian and English models to encode the correct word order.

In particular, though at lower layers the Russian model underperforms in terms of accuracy, it performs better than the English one at middle and higher layers. The English model also achieves its maximum performance at an earlier layer (5) than the Russian one (11). Both of these phenomena can be largely explained by the fact that the models' layers follow the so-called classic NLP pipeline (Tenney et al., 2019), where the lower layers specialize in lower-level language features such as parts of speech and other morphological information, the middle layers are responsible for more complex syntactic relations, while the higher layers deal with even more high-level language phenomena such as anaphora and coreference. Therefore, we might conclude that though it takes more layers for the Russian model to encode more complex morphology and syntax relations which are necessary to detect if the word order was corrupted, once it does that, it performs better on the task since the morphology of the inflected language binds the words together by the means of suffixes showing their gender, number, aspect, or tense. Thus, in contrast to our expectations, the free word order is not that free, as moving a word arbitrarily has a high probability of breaking such rich morphological and syntactical ties.

	Train		Test	
	EN	RU	EN	RU
Sentences	85663		2753	
Tokens	1798267	1599786	49642	44006

Table 1: Dataset statistics for the English vs Russian BShift task.

	EN	RU
Layer 1	0.786	0.903
Layer 2	0.902	0.867
Layer 3	0.893	0.824
Layer 4	0.926	0.855
Layer 5	0.931	0.937
Layer 6	0.903	0.944
Layer 7	0.895	0.945
Layer 8	0.893	0.935
Layer 9	0.875	0.935
Layer 10	0.869	0.944
Layer 11	0.873	0.948
Layer 12	0.863	0.911

Table 2: The accuracy of detecting word order corruption for the languages with fixed and free word order.

3 Sensitivity to word order corruption in agglutinative and inflected languages with different head directionality

In this section we compare the ability of pre-trained models to recognize word order corruption in such languages as Russian and Korean. We originally chose to analyse these languages as they differ in terms of head directionality (Haider, 2015) (see below) while both having a free word order. We hypothesized that since the attention mechanism in Transformer models (Vaswani et al., 2017) is able to capture the context to the both sides of a focus token, the performance on the word order corruption task should be comparable between these two languages. However, our experiments showed that some other aspects of these languages are affecting the task, namely the type of their morphology (inflected for Russian vs agglutinative for Korean).

3.1 Head directionality

Head directionality refers to the position of a head (main) word in a phrase relative to its subordinate word (Haider, 2015), and languages can be roughly categorized into head-initial and head-final. In head-initial languages such as Russian the verb (predicate) normally precedes the object (VO

word order), while in head-final languages such as Korean they follow the object (OV word order) (Lehmann, 1973)[4]:

Эмма	ела	яблоки.
Emma	ate	apples.
Subject	*Verb*	*Object*

엠마는	사과를	먹었다
Emma	apples	ate.
Subject	*Object*	*Verb*

As head directionality essentially refers to the expected position of subordinates relative to head words, it affects the importance of right-hand and left-hand context of a focus (head) word in representing the input text. However, as the attention mechanism (Vaswani et al., 2017) allows to capture both the right-hand and left-hand context equally well, we hypothesised that there should not be a remarkable difference in word corruption detection between these two languages, i.e. neither VO nor OV syntax should make it more difficult for a direction-agnostic model.

3.2 Inflection vs agglutination

Both Korean and Russian are morphologically rich synthetical languages, that is, grammatical meaning is expressed by adding a diverse variety of morphemes to the lexical root. However, while in Russian morphemes expressing tense, aspect, gender, person, number etc are fused together and one suffix can thus carry several grammatical meanings, in Korean morphemes can be stacked on top of each other in various combinations. For example, while in Russian the verb иду ("I am going") is a fusion of a stem ид- and a morpheme у which simultaneously signifies present tense, inperfect aspect, single number, and 1st person, in Korean the verb 가고있어요 with the same meaning can be split into a stem 가 and a stack of morphemes: 고 (continuous aspect), 있어 (present tense), 요 (politeness marker). Such difference in morphology is reflected in approaches to tokenization: while in Russian texts are normally tokenized by space, i.e. each token represents a lexical item together with its grammatical meanings, in Korean words are usually split into stems and particles, each representing a distinct lexical or grammatical meaning. Thus, in this experiment we also compare how the tokenization before word order corruption affects the model's ability to recognize the latter.

3.3 Dataset

For this task, similar to Section 2, we use a parallel corpus of Russian and Korean sentences, this time based on TED talk subtitles[5]. We apply the same preprocessing steps, and randomly sample 10% of the sentences to create the training/test split. The statistics of the dataset are presented in Table 3.

3.4 Model and experiments

For this set of experiments we use the same RuBERT model for Russian as in Section 2; for Korean we use a similar BERT-Kor-base model[6]. We follow the same method for corrupting the word order, training and evaluation as in Section 2. However, for this task we restrict BShift to VO chunks for Russian and OV chunks for Korean. To do that, we apply part-of-speech tagging using morphological analysers for Russian (Korobov, 2015)[7] and Korean[8], and then restrict the application of BShift only to spans with a verb and the following (for Russian) or preceding (for Korean) noun.

We also experiment with two types of tokenization for BShift in Korean. For the first experiment, we tokenize the text using Kkma parser[8] and then swap the resulting tokens which can be lexical stems or grammatical particles; for the second one, we apply the usual whitespace tokenization and thus swap the whole words with grammatical particles attached to them. The reason for such setup is that we intend to compare the effect of swapping the entire lexical units vs swapping subunits, potentially including grammatical ones.

3.5 Results and discussion

Table 4 reports the accuracy of detecting the sentences which underwent BShift in Korean and Russian. For Korean, we report the results for two approaches to word order corruption described above.

The most striking finding is probably an almost perfect accuracy even at the lowest layers achieved by the Korean model with the native (morphology-based) tokenization, where morphological particals/subunits could potentially be switched. This

[4] As both languages are relatively free in their word order, VO structures are possible in Korean while OV structures are legal in Russian, but such word order is inverted or emphatic.

[5] https://github.com/ajinkyakulkarni14/TED-Multilingual-Parallel-Corpus

[6] https://huggingface.co/kykim/bert-kor-base

[7] https://pypi.org/project/pymorphy2/

[8] https://konlpy.org/en/v0.4.4/api/konlpy.tag

	Train			Test		
	KO1	KO2	RU	KO1	KO2	RU
Sentences		299769			33308	
Tokens	7477635	3597639	4283921	826887	397301	473650

Table 3: Dataset statistics for the Korean vs Russian BShift task. KO1 refers to BShift using morphology-based tokenization; KO2 refers to BShift based on whitespace tokenizer.

shows that the order of particles that are attached to the stem is strongly encoded even at the lowest layers of the model, i.e. the pre-trained model is aware of agglutination and learned the correct slots for suffixes with a particular meaning. On the other hand, compared to the Korean model with the native tokenization, the performance of the model where the BShift occurred after whitespace tokenization is considerably lower and worsens even more at higher levels. This shows that whitespace tokenization makes it much harder to recognize word swaps, as in that case the whole word moves together with its suffixes and thus the basic morphology is preserved.

Interestingly, the performance of the Russian model and the Korean model with whitespace-based corruption is very similar at the lowest (morphology) layers, which supports our claim that the attention mechanism is agnostic to head directionality at lower levels and can recognize incorrect word order equally well for both OV and VO languages. However, at higher layers, with more syntactic information taken into account, it becomes easier for the Russian model to recognize corruption, while the performance of the Korean model falls significantly.

Another thing to note is that the accuracy of classification for the Russian model is higher here than reported in Section 2, and the best results are achieved at lower RuBERT layers. This can be explained by the fact that we restricted possible movements to one type of structure, so the task is inherently easier and potentially requires less information about the syntactic structure; another reason for such discrepancy can simply be a much larger training dataset available for this task. Thus, though these experiments provide some intuition into the abilities of the models to encode word order information and detect different types of word order shift, more rigorous experiments across different datasets and with more exact and diverse chunking strategies are in order.

	KO1	KO2	RU
Layer 1	0.988	0.940	0.948
Layer 2	0.989	0.928	0.928
Layer 3	0.995	0.942	0.886
Layer 4	0.996	0.921	0.899
Layer 5	0.994	0.896	0.981
Layer 6	0.991	0.895	0.983
Layer 7	0.990	0.902	0.982
Layer 8	0.987	0.888	0.977
Layer 9	0.985	0.863	0.976
Layer 10	0.989	0.844	0.979
Layer 11	0.990	0.888	0.979
Layer 12	0.990	0.875	0.921

Table 4: The accuracy of word order corruption for an agglutinative SOV language vs inflected SVO language. KO1 refers to BShift using morphology-based tokenization; KO2 refers to BShift based on whitespace tokenization.

4 Long-distance agreement in morphologically rich and poor languages

In this task we test the ability of attention-based models to encode long-distance agreement, in particular the agreement in number (plural vs singular). We choose this task as it tests the model's ability to encode hierarchical syntactic structure, for example to determine the number of a verb based on the form of the noun related to it in the syntactic tree, rather than on the form of the closest noun.

4.1 Long-distance agreement

Agreement refers to a linguistic phenomenon where the grammatical form of one word depends on the form of another word. While in English agreement is restricted only to nouns and verbs in the present tense, in Russian nouns agree with verbs in all tenses and also with adjectives and some pronouns. Though in Russian words agree in several different grammatical aspects such as gender, person, case, and number, in this task we focus only on the number agreement as it is the only agreement type present in English. In particular, we focus on long-distance agreement, which occurs when a head word (cue) that

determines the form of a dependent (target) is separated from it by intervening words (context). As such agreement requires the model to consider not only the nearby context but often far removed tokens that are nevertheless closely connected with the cue in terms of their position in the syntactic tree, it should be able to encode at least some hierarchical syntactic structure. We hypothesise that a higher performance can be achieved on this task for Russian, since the majority of parts of speech are marked for number there. Thus it is highly likely that there are some words in the context between the cue and target words that provide some clues for the correct long-distance agreement. Consider the following example:

Выросли	любимые	мамой	розы
Have grown	loved	by mom	roses
pl verb	*pl adj*	*sing noun*	*pl noun*

In this sentence, the words выросли (*have grown*) and розы (*roses*) should both have the plural form as they agree in number, but there is a noun in singular (*mom*) between them which can potentially interfere with the ability of the model to assign the correct plural number. However, unlike English, in Russian there is another word in the context between the cue and the target (любимые, *loved*) which has plural number and thus allows to infer the correct form.

4.2 Dataset

For this task we use the long-distance agreement test set created by Gulordava et al. (2018)[9]. We choose this dataset rather than a more popular agreement dataset by Linzen et al. (2016) as in addition to regular sentences with long-distance agreement the authors generate nonce sentences which retain the same syntactic structure but have no meaning. They do so by replacing all content words in a sentence by random words with the same grammatical properties; 9 nonce sentences are generated for each normal sentence in this manner. Gulordava et al. (2018) do so in attempt to disentangle the abilities of the model to capture syntactic and semantic information, as they notice that models tend to rely on semantic and lexical features such as frequency of co-occurance when resolving long-distance relationships. Thus, in the example above the model can choose the plural form of "have grown" for "roses" not because it learned to abstract such features as number and detect the long-distance relationships, but simply because "have grown" occurs more frequently around "roses" than "has grown" in the corpus it was trained on. To see how much of the performance on the task is due to such effect, the nonce sentences contain random words which are unlikely to frequently co-occur. Overall, the dataset contains 41 original vs 369 generated sentences for English and 442 original vs 3978 generated sentences for Russian.

[9]https://github.com/facebookresearch/colorlessgreenRNNs

4.3 Model and experiments

For this set of experiments we use BERT-Base Cased and RuBERT models introduced above, and BERT-Base Uncased model in addition to them. We compare the cased and uncased variants of the model to estimate the effect of capitalization on the encoding and detection of long-distance agreement.

We adapt the evaluation protocol proposed by Goldberg (2019) for our task. Namely, we cast it as a masked token prediction task where we replace the word which form we are trying to predict by [MASK]. To predict the token, we use a masked language model which is essentially a BERT model with a feed-forward network projecting onto the vocabulary. Then for each masked token we compare the probability of the word in the correct form (plural or singular) with the probability of the incorrect form (opposite in number). We consider the prediction to be correct if the probability of the expected token is strictly higher than then probability of the alternative form. As in the tasks above, we perform the experiments for all layers of the pre-trained models.

4.4 Results and discussion

	EN uncased		EN cased		RU cased	
	orig.	gen.	orig.	gen.	orig.	gen.
L 1	0.683	0.477	0.683	0.423	0.464	0.471
L 2	0.659	0.477	0.756	0.439	0.489	0.481
L 3	0.707	0.485	0.707	0.472	0.502	0.485
L 4	0.707	0.458	0.659	0.496	0.500	0.501
L 5	0.659	0.466	0.683	0.520	0.523	0.518
L 6	0.732	0.499	0.757	0.537	0.539	0.543
L 7	0.805	0.623	0.780	0.602	0.559	0.538
L 8	0.780	0.612	0.854	0.664	0.520	0.515
L 9	0.878	0.737	0.951	0.734	0.568	0.531
L 10	0.927	0.770	0.976	0.797	0.586	0.558
L 11	0.951	0.816	0.976	0.824	0.618	0.569
L 12	0.951	0.810	0.976	0.821	0.991	0.919

Table 5: The accuracy of long-distance agreement.

Table 5 shows the accuracy of grammatical form

prediction related to long-distance agreement for both semantically correct and nonce sentences. First, we can observe a clear gap in performance between the original and generated (nonce) test sets for all three models, which shows that the ability to assign the correct number is largely due to the co-occurrence and frequency effects rather than recognizing syntactic structure. However, it can be noticed that as the number of layers grow, the gap in accuracy for normal and nonce sentences diminishes, which means that at higher layers the models do learn to abstract from the lexical information and encode long-distance syntactic relations even when they cannot rely on co-occurrence.

When comparing the three models, it can be observed that, as expected, the Russian model performs better at the last layer than both English ones, which shows that rich morphology helps to encode long-distance relationships. Interestingly, the cased variant of BERT-Base had a higher accuracy than the uncased one, especially at some intermediate layers, which can mean that capitalization helps to encode morphological and syntactic relationships. Another thing to note is a remarkable difference in the progression of accuracy across layers between the Russian and English models: while in BERT-Base Cased and Uncased there is a consistent improvement with moving to higher layers, in RuBERT the accuracy grows very slowly at all but the last layer, where there is a huge jump in performance. It can be explained by the fact that due to complex morphology of the Russian language it takes more layers to encode some lower-level morphology and syntactic features before the model is ready to handle long-distance agreement. Lastly, compared to lower-level morphology tasks in Sections 2 and 3, where the performance actually downgraded at the last layer, here the last layers are important, especially for Russian.

5 Gender bias encoding in languages with and without gender marking

Though gender bias is a wide-known issue affecting such down-stream tasks as machine translation or text generation, it has been mostly studied only through such phenomena as co-reference and pronoun resolution (Rudinger et al., 2018). In this task we aim to explore if the gender bias is more prominent in languages such as Russian where nouns, adjectives and verbs can be marked for gender, i.e. have masculine, neutral or feminine gender. We hypothesise that the gender bias would be even more pronounced in such languages.

5.1 Dataset

For this task we construct the test set using the dataset provided by Stanovsky et al. (2019)[10] as a starting point. In particular, we extract the mentions of professions (triggers) and the relevant sentences from their English dataset and modify them as follows:

- We remove triggers that have a strong feminine gender (i.e. feminine ending) since they can only be used with feminine forms of words according to grammar rules. For example, referring to a nurse (медсестра) as "her" is the only correct way in Russian as the feminine ending of the word requires such agreement. Therefore, agreement with such words is due to grammatical conventions rather than bias. For the same reason, we remove triggers which can be translated both in a masculine and feminine form, as that would pre-determine the agreement.[11] As the result we selected 30 triggers (see Appendix A).

- We simplify the sentences so that there is only one trigger and it is referred to unambiguously. We do this to ensure that any discrepancy in gender usage is due to the model attending to the trigger noun rather than other nouns.

- For English we mask the pronoun referring to the trigger to test the assumed gender of co-reference resolution.

- For Russian, we modify the sentence to create three variants: with a masked pronoun, as in English, with a masked adjective referring to the trigger, and with a masked verb referring to it, to compare the degree of bias for these parts of speech. While doing so we ensure that other words in the sentence do not reveal the assumed gender; for instance, we change the past tense verbs (marked for gender) into their present tense forms (which are the same for both genders). On the other hand, we try to ensure that the masked word is predicted in a form marked for gender, such as past tense, by adding adverbs such as "yesterday".

[10] https://github.com/gabrielStanovsky/

[11] Some of nouns can have both neutral-style masculine forms and derogative feminine forms; we included them as we expect neutral forms also to be used for women.

5.2 Model and experiments

We use the same approach to evaluation as in Section 4, but here instead of comparing probabilities of two tokens we extract the list of 50 most prominent candidates and compare the probabilities assigned to the top tokens with masculine and feminine gender. If either a masculine or feminine form did not occur in the list of top 50 tokens, we record its probability as 0. We examine both the percentage of cases where either masculine or feminine genders were the winning ones (winning rate), and the average probabilities assigned to masculine and feminine forms.

5.3 Results and discussion

Table 6 shows the winning rate and the average probability for masculine and feminine forms appearing in the pronoun, verb or adjective slot.

	Winning rate		Avg. prob.	
	M	F	M	F
EN pronouns	50%	50%	0.268	0.296
RU pronouns	93%	7%	0.460	0.03
RU verbs	100%	0%	0.299	0.047
RU adjectives	100%	0%	0.091	0

Table 6: Winning rate of one gender over the other and average probabilities for genders in particular position.

As it can be seen from the table, in Russian there is a large skew towards masculine forms for all analysed parts of speech, while in English the gender labels for pronouns were distributed almost equally. It does not in any way imply the absence of bias: we observed the well-known phenomenon of assigning the masculine gender to both "manly" professions such as *mechanic* or *guard* and high-status jobs such as *CEO*, *manager* or *lawyer*, while the feminine gender was mainly assigned either to assisting roles such as *clerk* or *secretary* or to creative professions such as *editor* or *designer*. However, in Russian even such professions as *hairdresser* or *assistant*, which are more likely to be marked as feminine in English, had a higher probability of masculine forms than that of feminine ones. This is even more so for verbs and adjectives, all of which had masculine forms. Thus we can conclude that the strong gender bias in Russian which we observed is rather a grammatical phenomenon than encoded connotations of professions of particular type, as in English. In particular, though the words we studied are gender-neutral in terms of their applicability to people of both genders, grammatically they have a masculine form, and unlike native speakers who would choose feminine forms when referring to female professionals, the model is unable to do that and selects the most probable form based on the grammatical form only.

6 Conclusions and future work

In this study we used linguistic probes and masked language models to explore several aspects of morphology and syntax representation in Transformer-based models. In particular, we examined the ability of the model and its particular layers to encode the correct word order in languages with contrasting morphology and syntax, their ability to capture hierarchical structure represented by long-distance agreement, and the degree of bias encoding in languages with and without gender morphology. In doing so, we once again showed that the number of layers in the model roughly corresponds to the complexity of the encoded features, but also discovered that languages differ in layers where such encoding happens.

One of the most important takeaways of analysing the pre-trained models' performance layer by layer is that the best accuracy is not necessarily achieved at the last layer, which leads us to question the practice of using the complete model for all downstream tasks. Therefore, a potential extension of this work would be to explore the performance of such tasks as classification or generation when using only some layers of the pre-trained model. Another observation is that in general the Russian model needed more layers to achieve its optimal performance, while both Korean and English ones showed their best results at much earlier layers. It can be explained by more complicated morphology and syntax of Russian language which potentially can require more layers to be properly encoded. Thus it leads to a question whether adding more layers to pre-trained models for inflected languages with rich morphology and syntax (for example, Spanish or German) can help to improve performance of downstream tasks. That said, one of the limitation of the present study is that we focused only on one type of Transformer-based models and compared only two languages at a time; to ensure the general applicability of our experiments, they should be expanded to more languages with similar typological characteristics to those analysed above, and to attention-based models with different training approaches.

Acknowledgements

This initiative was funded by the Department of Defence and the Office of National Intelligence under the AI for Decision Making Program, delivered in partnership with the Defence Science Institute in Victoria, Australia.

References

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Wietse de Vries, Andreas van Cranenburgh, and Malvina Nissim. 2020. What's so special about BERT's layers? A closer look at the NLP pipeline in monolingual and multilingual models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138*.

Hubert Haider. 2015. Head directionality. *Contemporary Linguistic Parameters: Contemporary Studies in Linguistics*, pages 73–97.

Diederik P Kingma and Jimmy Ba. 2014. ADAM: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Natalia Klyueva and Ondřej Bojar. 2008. UMC 0.1: Czech-Russian-English multilingual corpus. In *Proceedings of International Conference in Corpus Linguistics*.

Mikhail Korobov. 2015. Morphological analyzer and generator for Russian and Ukrainian languages. In *International conference on analysis of images, social networks and texts*, pages 320–332. Springer.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for Russian language. *arXiv preprint arXiv:1905.07213*.

Winfred P Lehmann. 1973. A structural principle of language and its implications. *Language*, pages 47–66.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *ACL*, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*.

A Selected triggers

For the gender bias experiments we selected the following English names of professions and their direct translations into Russian:

English: developer, mechanic, clerk, mover, analyst, assistant, salesperson, librarian, lawyer, hairdresser, cook, teacher, physician, baker, farmer, CEO, manager, guard, editor, auditor, secretary, designer, supervisor, cashier, driver, construction worker, counselor, carpenter, janitor

Russian: разработчик, механик, клерк, грузчик, аналитик, ассистент, продавец, библиотекарь, адвокат, парикмахер, повар, учитель, врач, пекарь, фермер, CEO, менеджер, охранник, редактор, аудитор, секретарь, дизайнер, супервизор, кассир, водитель, строитель, психолог, плотник, дворник