What do Transformers Know about Government?

Jue Hou,[†⋆] **Anisia Katinskaia,**[†⋆] **Lari Kotilainen,**[◇]
Sathianpong Trangcasanchai,[†] **Anh-Duc Vu,**[†⋆] **Roman Yangarber**[⋆]
University of Helsinki, Finland
[†]Department of Computer Science
[◇]Department of Finnish Language and Culture
[⋆]Department of Digital Humanities
`first.last@helsinki.fi`

Abstract

This paper investigates what insights about linguistic features and what knowledge about the structure of natural language can be obtained from the encodings in transformer language models. In particular, we explore how BERT encodes the government relation between constituents in a sentence. We use several probing classifiers, and data from two morphologically rich languages. Our experiments show that information about government is encoded across all transformer layers, but predominantly in the early layers of the model. We find that, for both languages, a small number of attention heads encode enough information about the government relations to enable us to train a classifier capable of discovering new, previously unknown types of government, never seen in the training data. Currently, data is lacking for the research community working on grammatical constructions, and government in particular. We release the Government Bank—a dataset defining the government relations for thousands of lemmas in the languages in our experiments.

Keywords: transformer language models, BERT, government relations, probing

1. Introduction

Many contemporary linguistic theories, such as Construction Grammar (Fillmore and Kay, 1996; Goldberg, 2006), view form-function pairs, called constructions, as fundamental elements of language. In this paper, our interest in constructions is driven by both theoretical and practical considerations. From the theoretical perspective, describing linguistic phenomena in terms of constructions provides a powerful and convenient means of capturing the complex interactions at the interface between lexical semantics and syntax.

In practical terms, constructions can provide a way of tracking language learners' progress. Constructions can be seen as describing the linguistic knowledge of a native speaker, and thus as constituting the knowledge needed to master a language—mastery of a language is mastery of its constructions. If we had a description of the constructions of a language, we could follow a learner's progress in terms of the proportion of constructions s/he has mastered so far; plan learning paths in terms of constructions of increasing complexity, etc.

The present study focuses on one of the most fundamental types of constructions: patterns of government.[1] A verb has several syntactic dependents (arguments) in a sentence; they are analyzed as *complements* vs. *adjuncts*. One may say "I listened to many songs on a trip through Europe." Here the verb *listen* has two prepositional phrase (PP) modifiers: A. "to many songs", and B. "on a trip". Phrase A is a complement—it is semantically *required* by the verb; phrase B is an adjunct—it is *optional*, in that many actions can take place while "on a trip," nothing binds this PP to this verb in particular. However, in "I went on a trip," the same PP is a complement of *went*. Government is about which complements a verb governs: about its *valency*—the patterns of complements that a verb expects in a sentence. Our linguistic competency informs us that one *listens to* a sound / person / hunch / etc., while one *goes on* a trip / rampage / quest / etc.

The high performance of pre-trained transformer Language Models (LM) on many NLP tasks has stimulated an interest in investigating the inner representations of these models to find out how linguistic knowledge is encoded inside them, and whether it agrees with linguistic theories in general, and theories of grammar in particular. Corresponding to our theoretical and practical objectives, this paper explores two Research Questions, respectively—**RQ1**: do transformer LMs encode knowledge about government, and where this knowledge is represented; and **RQ2**: can we *extract* this knowledge about government from the LM, to build government resources (or enhance existing resources), e.g., for language learning. To our knowledge, how LMs encode government has not been explored systematically to date.

To explore RQ1, we probe the BERT model using

---

[1]We expect that these methods can be applied to a broader range of constructions, in future work. We further limit ourselves to *verbal* government, in this paper; noun and other types of government to be covered in the future.

several probing classifiers, for Finnish and Russian. We find that information about government relations is encoded across all layers and heads of BERT; however, there is enough information in the initial (5 or 7) layers to detect the presence of the government relations with almost the same accuracy as by using all layers. Probing classifiers indicate that just a few attention heads encode government relations, and can be used for government prediction without a significant loss in accuracy.

To explore RQ2, we perform experiments on (a) detecting governing verbs and (b) detecting government patterns, which were held out from training data. Results show that probing classifiers perform well on both tasks, and therefore can be used to discover *new*, previously unseen patterns.

Descriptions of constructions are important for both general and computational linguistics research, evidenced, e.g., by the "*constructicon*" efforts (Janda et al., 2018; Hunston, 2019; Lyngfelt et al., 2018), although the resources resulting from this work are for human consumption, not directly usable, e.g., in CALL systems.[2] Collecting banks of constructions for any language consumes a great deal of labor; we propose a method for searching for patterns in collections of text automatically, by leveraging the information already learned by the LM and encoded in its attention heads.

The paper is organized as follows: Section 2 briefly reviews prior work on probing neural language models, and Section 3 provides some background on syntactic government. Section 4 introduces our government data and its structure. Section 5 introduces our probing classifier and the experimental setup. Section 6 discusses the results of our experiments. Section 7 concludes with a discussion of future work.

2.   Related Work

Language modeling is a fundamental task in natural language processing (NLP), on which pre-trained language models (PLMs) currently achieve the best performance. One approach to making inferences about the model "internals" is through *probing*, also known as BERTology (Rogers et al., 2021). Approaches to probing PLMs usually include a specific probing task—e.g., investigating predicate-argument agreement, or how gender is encoded in contextual representations,—data prepared for this task—e.g., minimal pairs of examples that differ only by the studied linguistic category,— and some mechanism that allows us to interact with or query the model's components. Such a mechanism can be a probing classifier (or *probe*). The behavior of a simple probe, trained on representations from the PLM, on a probing task can inform us

whether the representations include the linguistic information in question (Adi et al., 2017; Conneau et al., 2018; Hewitt and Manning, 2019; Dalvi et al., 2019; Maudslay et al., 2020; Weissweiler et al., 2022; Conia and Navigli, 2022; Arps et al., 2022).

Some researchers criticize probing classifiers and question their effectiveness, in particular, whether the probed LM in fact uses the information that is discovered by the probe (Hewitt and Liang, 2019; Tamkin et al., 2020; Ravichander et al., 2021; Voita and Titov, 2020). Later research suggests this criticism can be addressed by designing appropriate control tasks and datasets (Belinkov, 2022).

There are two research directions in probing for dependency relations: token representation or the weights of attention heads. Christopher et al. (2020) reconstruct dependency structures based on token representation. Wu et al. (2020) propose parameter-free probing based on masking tokens and measuring the impact of the masked tokens. They found that a Masked LM (MLM) such as BERT can learn the "natural" dependency structure of language. Although the dependencies learned by the MLM may differ from human annotation or linguistic theory, they consider it a good "lower bound" for *unsupervised* syntactic parsing.

Christopher et al. (2020) and Clark et al. (2019) probe attention head weights. They find that no single attention head corresponds well to dependency syntax overall, however, the combination of several attention heads can correspond substantially better than a baseline, where dependency is defined simply by a fixed offset. Furthermore, they find certain attention heads are *specialized* in certain dependency relations. Kovaleva et al. (2019) analyze different attention patterns and suggest that some attention heads could potentially be linguistically interpretable. Our work is inspired by these studies, and we likewise study how information on government is distributed among attention heads.

Weissweiler et al. (2023) analyze methodologies for probing applied to constructions, as well as probing specifically for certain constructions. They stress that probing constrictions is challenging for several reasons, in particular, the non-compositional meaning of constructions and the training objectives of PLMs.

3.   Background on Government

Government is a relation between a token and its syntactic dependents. In the context of this paper, the term *government* refers to relations between a **governor** (verb, noun, adjective) and a "*governee*," which can be a noun (phrase), infinitive verb (with its own dependents), an adpositional phrase, etc. Here are some examples from Finnish:

---

[2]Computer-aided language learning.

Lemma	(gloss)	Direct Object	Argument
ajatella	*think*	Noun + Case:partitive	Noun + Case:elative
kyllästyttää	*bore*	Noun + Case:partitive	Verb + Inf_Form:inf-A + Case:lative
taistella	*fight*	—	Postposition + Base:puolesta + Case:genitive
готовить	*prepare*	Noun + Case:accusative	Preposition + Base:к + Case:dative
давать	*give*	Noun + Case:accusative	Noun + Case:dative

Table 1: Examples of government relations: verb lemmas with their complements. The direct object is displayed as a "special" argument, for transitive verbs; other arguments follow. Third and fourth verbs show examples of a post-position "puolesta" (*for*) and a preposition "к" (*to*).

(1) **Ehdotan**   <u>teille</u>
    recommend.V.Pres.1P.Sg   *you*.Pl.Allat
    pitkää   <u>lomaa</u>.
    long.Sg.Partit   *break*.Sg.Partit

    "*I recommend to you a long break.*"

In example 1, the governor verb "ehdottaa" (*recommend*) requires a direct object—a noun phrase "pitkä loma" (*long break*), in the partitive case. The same governor requires another complement—pronoun "te" (*you*) to be in the allative case.

(2) Kuva   **helpottaa**
    picture.Sg.Nominat   *ease*.Pres.3P.Sg
    <u>ymmärtämään</u>   asiaa.
    understand.3-Infinit   *thing*.Sg.Partit

    "*A picture makes it easier to understand.*"

In example 2, the governor "helpottaa" (*ease*) requires its argument—verb "ymmärtää" (*to understand*) to be in the illative case of 3-rd infinitive.

(3) Hän   **protestoi**
    S/he.Sg.Nominat   *protest*.Past.3P.Sg
    päätöstä   <u>vastaan</u>.
    decision.Sg.Partit   *against*.Postpos

    'S/he protested the decision.'

In example 3, the governor "protestoida" (*protest*) requires the specific postposition "vastaan", which in turn governs a noun phrase in partitive case.

In the context of this paper, we focus only on verbs as governors, and use the terms "complement", "argument", or "governee" interchangeably.

4. Government Data

As of publish date, we collaborated with expert linguists, and manually collected 1184 Finnish government rules for 765 verb lemmas, and 2635 Russian governments for 1976 verb lemmas, together with several examples for each rule. For each Finnish verb, we include information about its transitivity (whether it is transitive or intransitive) and its government rules, as illustrated in Table 1.

Each object or governee indicates the part of speech governed by the verb; note, that the governee can have its own dependents. `Case` denotes the case governed by the verb, preposition, or post-position. Each verb governor can govern more than one argument. We release this extensive Government Bank, which encodes thousands of government patterns for verbs, nouns, and adjectives—for Finnish and Russian. The Bank is not "complete", but rather encodes the most frequent government patterns in each language. The patterns are stored in a structured, human-readable form, easily transformed into JSON, tsv, or other formats.[3]

4.1. Data for Probing Experiments

We build a dataset to probe the LM for government relations, using the patterns in the Government Bank. We start with sentences in the parsed Universal Dependency corpus (v2.12). We processed 37K sentences from all Finnish datasets in UD, and 116K Russian sentences from the Taiga (Shavrina and Shapovalova, 2017) and SynTagRus corpora (Droganova et al., 2018). We included 27.7K additional Finnish sentences and 102K Russian sentences from the WMT News Crawl monolingual training data (Bojar et al., 2017).

The data passes through morphological analysis, and parsers—TurkuNLP for Finnish (Kanerva et al., 2018), and DeepPavlov for Russian (Burtsev et al., 2018). Then, we filter the sentences and identify valid government instances according to the Government Bank, via rule-based pattern matching. This results in a dataset with 18K Finnish instances by 582 unique verb lemmas. For Russian, we collected 143K instances and 2805 unique verbs.

Using the syntactic parser, we label each dependent of a verb that is also a governee in the Government Bank as a **positive** instance. Each sentence can yield one or more positive instances. As a **negative** instance, we label any noun phrase or prepositional phrase that is a dependent of the governor but not matching any pattern for this governor's complements in our set of rules. This setup helps us avoid the situation where the classifier learns to identify *dependency* between two words, which is an easier task than identifying the government relation between the governor verb and its

---

[3]https://github.com/RevitaAI/GovProbing

$Dist > 3$	Class	Training		Test	
		Far	Near	Far	Near
Finnish	POS	402	381	123	144
	NEG	407	413	118	112
Russian	POS	3055	3047	809	817
	NEG	3074	3125	790	739

Table 2: The number of Finnish and Russian government instances for training and testing when "far" means the distance between governor and governee is 3 tokens away.

$Dist > 2$	Class	Training		Test	
		Far	Near	Far	Near
Finnish	POS	943	925	193	211
	NEG	999	991	137	145
Russian	POS	6905	6872	2439	1746
	NEG	7514	7598	1830	2472

Table 3: The number of Finnish and Russian government instances for training and testing when "far" means the distance between governor and governee is 2 tokens away.

arguments. We exclude adjunct dependents of the governor from the negative instances since they differ syntactically from complements and may be easier to differentiate. We make sure that the positive and negative instances are balanced in the training / test data.

4.2. Data Balancing

The government instances in our datasets are also highly skewed in two respects: in terms of linguistic features (in particular, the case of the governee), and in terms of distance between the governor and governee. We balance the instances in these respects as well.

Linguistic Features: Instances with verbs governing arguments with a certain linguistic feature may outnumber other kinds of instances—the distribution of features of the arguments is not balanced. For example, the direct object is the most frequent type of argument in Finnish government patterns, which results in a large number of such instances dominating the training data.

Distance: We also observe that in most government instances, the governees are adjacent or very near the governor. To prevent the classifier from learning to classify adjacency rather than government, we divide the instances into two types: "far" vs. "near". The governee is "far" if its distance to the governor is over $Dist$ complete words (not subword tokens); otherwise, we consider it to be "near". In this paper, we consider two distance thresholds for "far" and "near": $Dist > 3$ and $Dist > 2$. The "near" instances are much more frequent than "far".
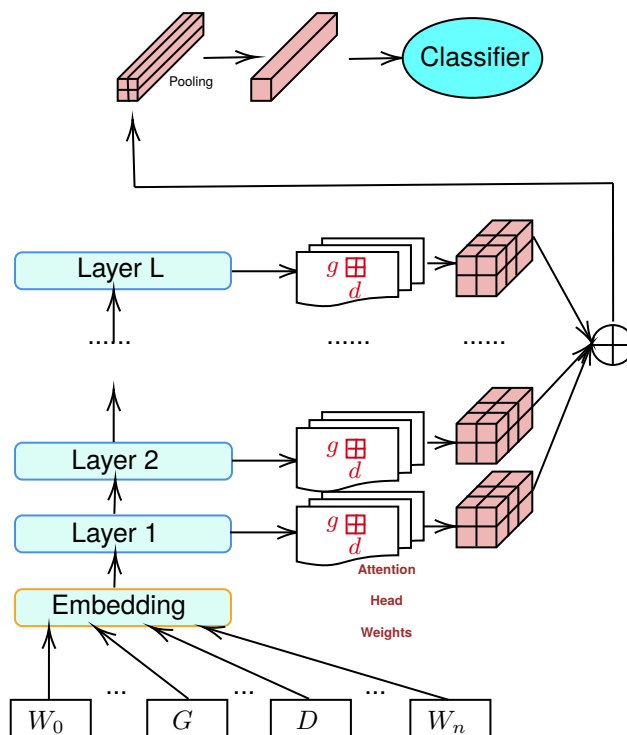

Figure 1: Weights of attention heads of transformer LM as input to probing classifier.

We balance the data by down-sampling the "near" instances so that the amount of "near" and "far" instances is about the same.

Government Patterns: To verify whether the classifier is capable of identifying patterns with verbs or arguments that it has never seen before (RQ2), we *withhold* instances with certain verb lemmas from the training data, and keep such instances only in the test data. We also withhold certain argument *types*, so that the probing classifier is not trained using all possible types of complements in the Government Bank. This allows us to evaluate the classifier's ability to discover entirely new, unseen government patterns.

Table 2 and Table 3 summarize the number of instances used in these experiments, as well as their distance distribution, for Finnish and Russian.[4]

5. Government Probing Classifier

We build the probing classifier based on the BERT model, specifically on the weights of its attention heads, from each transformer layer. Figure 1 illustrates the input to the probing classifier. From each attention head, we extract the weights for the gov-

---

[4] Detailed data statistics will be in the Appendix A.

	Model	Dist > 3				Dist > 2			
		Acc	P	R	$F1$	Acc	P	R	$F1$
Finnish	LogReg	79.20	**85.04**	75.35	79.90	77.42	85.16	72.63	78.40
	MLP-1	78.85	82.56	77.91	80.17	78.53	84.16	76.31	80.04
	MLP-2	77.13	81.32	75.70	78.41	76.54	81.85	75.04	78.30
	RF	**80.61**	82.82	**81.57**	**82.19**	**79.71**	**84.21**	**78.82**	**81.43**
Russian	LogReg	80.63	83.10	78.46	80.72	77.94	84.72	73.23	78.56
	MLP-1	**84.87**	**86.20**	84.19	**85.18**	**82.05**	**87.01**	79.32	**82.99**
	MLP-2	83.86	84.88	83.66	84.27	81.65	86.35	79.29	82.67
	RF	83.77	81.75	**88.32**	84.90	81.29	84.45	**81.01**	82.70

Table 4: Overall performance of the classifiers

ernor verb and its argument and concatenate these weights into a vector. The vector will contain $L \cdot A$ elements, where L is the number of transformer layers and A is the number of attention heads.[5] We use the vector as a one-dimensional representation for the governor-argument pair, which encodes the syntactic relations between the governor and the argument. This vector is the final input to the probing classifier.

In Figure 1, the governor word is G and its dependent is D. Any word in a sentence—including the governor and its dependents—may be segmented into multiple sub-word tokens.[6] We use byte-pair encoding (BPE) to split words into sub-word tokens before input to the LM. We denote by g a member of the list of tokens of G, and by d a segment of D. From each head, we collect the weights for *all pairs* of tokens (g, d), where $g \in G$ and $d \in D$.

First, this gives us a series of matrices (2×2 in the illustration in the figure), one for each head in each layer. These matrices are concatenated into a tensor of dimension $(L \cdot A) \times 2 \times 2$. Next, *for each head*, we max-pool all pairs of governor and dependent tokens (in the figure, 2×2 pairs), to obtain the *pooled* attention weight, which reflects the relation between the governor and its dependent, one for each head. These pooled attention weights form a one-dimensional vector of size $(L \cdot A)$. This vector is the input to the government classifier.

We experiment by training four types of probing classifiers: logistic regression, multi-layer perceptron (MLP) with one fully-connected layer, MLP with two fully-connected layers, and Random Forests. More details on the hyper-parameters of each classifier can be found in Appendix B. We implement all four classifiers using the Scikit-Learn Toolkit and mostly follow their default hyper-parameter settings.

For logistic regression, we set the maximum number of training iterations to 10000. For the MLP 1-layer classifier, we use 144 neurons. We use the same number of neurons for the first layer in the MLP 2-layer classifier, and 72 neurons for its second layer. For the Random Forest classifier, we use 300 trees.

6. Results and Discussion

We train four sets of classifier for each language and for each distance group, corresponding to four classifier types in Section 5. For each set of classifiers, we train 12 classifiers using the attention head weights only from the first N transformer layers, with $N = 1, 2, ...12$. The first classifier of each set is trained with attention head weights from only the first transformer layer, whereas the last one is trained with all attention head weights. Therefore, we expect the last model, which is trained with all weights, should be the best performing one in each set of classifiers. Since training involves some randomness, We repeat the process 5 times and average the results for more objective evaluation.

6.1. Detecting Government Relations

We first explore the overall performance of the probing classifiers. Table 4 shows the performance of the probing classifiers trained with all attention head weights, average over five repetitions. All classifiers reach generally high scores, which are—within each language—quite close to each other. The random forest classifier gives the highest score in Finnish, while the MLP-1 classifier performs the best in Russian for most metrics (except recall when $Dist > 2$). All classifiers also show a very close performance within each distance group. Except for recall in Russian, the difference among all metrics is not statistically significant, with p-values all above 0.05. The p-value for recall in Russian is 0.04. Since the test and training data contain a balanced number of positive and negative instances, we can conclude that the probing classifiers can distinguish well between government vs. non-government relations, with accuracy and F_1 over 80% for both Finnish and Russian.

Near vs. Far Arguments: To confirm the generality of the probing classifiers, we explore the *best*

---

[5]In the default BERT-base settings, $L = A = 12$.

[6]In the figure, the governor and governee are each shown with 2 tokens, but it could be more or fewer.

			Dist > 3				Dist > 2		
		Model	Acc	P	R	F1	Acc	P	R	F1
Finnish	Near	LogReg	83.20	84.83	85.42	85.12	83.15	83.86	88.63	86.18
		MLP-1	83.98	**86.52**	84.72	85.61	85.39	85.02	91.47	88.13
		MLP-2	79.69	83.82	79.17	81.43	83.71	**85.25**	87.68	86.45
		RF	**84.38**	82.10	**92.36**	**86.93**	**85.67**	84.48	**92.89**	**88.49**
	Far	LogReg	80.08	91.21	67.48	77.57	78.18	92.31	68.39	78.57
		MLP-1	83.40	86.73	**79.67**	83.05	80.30	87.21	**77.72**	82.19
		MLP-2	80.08	83.78	75.61	79.49	76.36	84.85	72.54	78.21
		RF	**85.48**	**91.51**	78.86	**84.72**	**81.52**	**93.42**	73.58	**82.32**
Russian	Near	LogReg	81.62	80.48	85.80	83.06	70.59	81.18	65.34	72.40
		MLP-1	**85.54**	**85.07**	87.88	**86.45**	**77.06**	**89.99**	68.81	**77.99**
		MLP-2	84.32	83.99	86.66	85.30	76.67	87.99	**70.06**	78.01
		RF	82.65	78.82	**91.55**	84.71	72.98	83.71	67.37	74.65
	Far	LogReg	86.74	**93.32**	79.48	85.85	82.75	92.72	75.00	82.93
		MLP-1	**89.68**	92.48	86.65	**89.47**	**85.40**	**92.81**	80.06	85.96
		MLP-2	89.06	90.96	**87.02**	88.95	85.11	91.47	80.88	85.85
		RF	87.99	90.65	85.04	87.76	88.99	92.06	**87.88**	**89.92**

Table 5: Performance on detecting *Near* vs. *Far* government between governor and complement.

performing models trained with all attention head weights from our experiments, and evaluate their performance on near vs. far governees separately. These tests check that the probes learn to predict government, rather than focusing only on the adjacency of governors to their arguments. The results are in Table 5. For Finnish, the majority of the models have higher F1 scores on detecting the government of the near than the far governees. However, the difference between far and near in $Dist > 3$ is not statistically significant. (The p-values for all metrics are above 0.05, except for recall.)

For Russian, the models score higher on detecting far governees than near ones, while for the Finnish models the scores are reversed. There is a significant difference, with all p-values below 0.05 for both $Dist > 2$ and $Dist > 3$. This confirms that the probes are not selective for the adjacency of the governors and their arguments.

Positive vs. Negative instances: To visualize the difference between the attention weights of the positive vs. negative instances, we use the t-SNE dimensionality reduction technique (van der Maaten and Hinton, 2008), implemented in Scikit learn, with default parameters. Figure 2 shows how t-SNE projects the high-dimensional input vectors of the probing classifiers onto the 2-D plane.

We plot all instances from the test data, for both languages. We can observe some strikingly well-defined clusters of the positive vs. the negative instances, particularly for Russian. This separability may help explain the very high performance of all probes on government prediction in Table 4.

6.2. Selection of Transformer Layers

We next investigate how syntactic information about government is distributed across different transformer layers. We select the best-performing model trained with all attention weights, for $Dist > 3$, and visualize it together with all the other models in its classifier set. This lets us evaluate the contribution of the first N layers to the probe's ability to predict government.

Figure 3 shows the performance in terms of the F_1 measure, for all four classifier types. Almost all curves increase monotonically as the number of selected layers increases, which is expected, as more information can be inferred with more transformer layers. Crucially, all curves also indicate that the performance of classifiers increases rapidly for the early layers–layers 1 to 4 for Finnish, to layer 5–6 for Russian—and then the performance plateaus, with much smaller subsequent increases. This suggests that the syntactic information important for government predictions is encoded in the *lowest* levels of BERT. For Russian, this information is slightly more spread out across the first 5–6 layers. This mostly aligns with the observation in other probing contexts (Hewitt and Manning, 2019; Lin et al., 2019; Liu et al., 2019; Goldberg, 2019).

We check similar plots for models with $Dist > 2$ in Figure 5, which can be found in Appendix C, and which shows a very similar pattern as in Figure 3.

6.3. Ablation of Attention Heads

We perform detailed ablation studies to see how attention heads at different layers contribute to government prediction. We use the logistic regression classifiers, trained as explained above, and explore how the weights learned in their attention heads
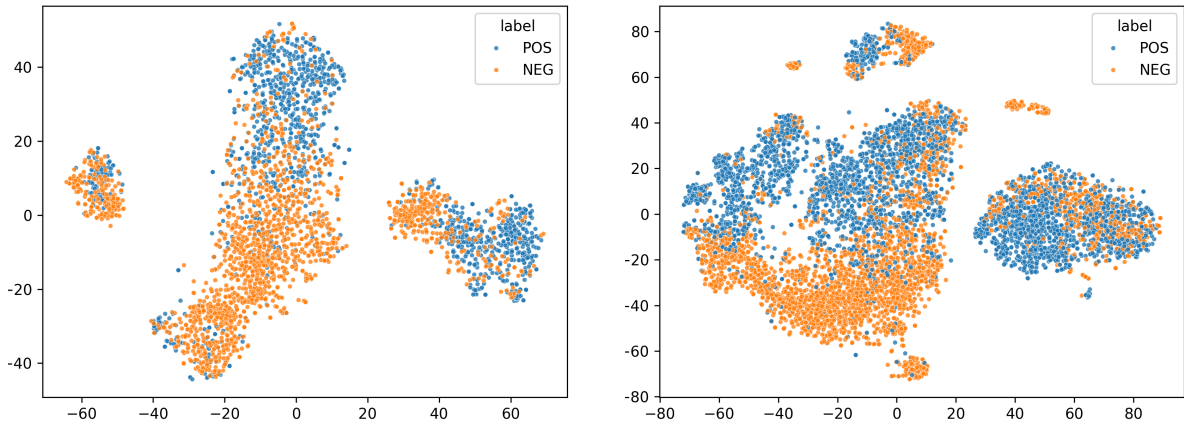
Figure 2: t-SNE visualization of positive vs. negative instances. (left: Finnish, right: Russian)
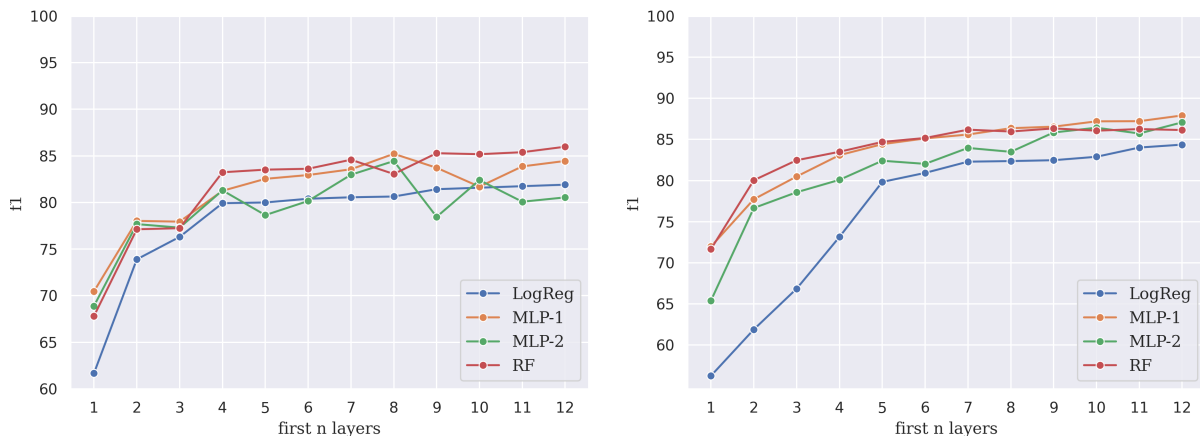

Figure 3: Probing government prediction with attention weights from the first N layers of BERT (X-axis) when $Dist > 3$. Y-axis—F_1 measure. (left: Finnish, right: Russian)

impact on each head's significance for government prediction. We fist rank all attention heads according to their logistic regression coefficients, and then ablate attention heads from two opposite perspectives: (A) training the classifier by *including* only the top-N heads, and (B) training the classifier *excluding* the top-N heads, with N ranging from 1 to $144 (= 12 \times 12)$. For comparison, we also train classifiers using a *random* subset of N heads, as a baseline.

In this experiment, we assess the relative importance and contribution of the attention heads, by "probing" our probing classifiers. We train the random forest classifier with $N = 1, 2, ...144$, for both languages. We explore models with $Dist > 3$ and plot their F_1 scores in Figure 4.

Models with $Dist > 2$ show very similar behavior—we visualize their performance in Figure 6, in Appendix C.

We observe that for Finnish, the model achieves an F_1 score of 80% either with the 3 top heads (represented by the blue curve), or by using all other heads together except the 3 top heads (orange curve). This suggests that the top 3 heads contain

most of the needed information and are reliable indicators of government. Training without the top N heads yields superior performance compared to the baseline, where the heads are chosen at random (represented by the green curve).

Similarly for Russian, 17–20 top heads are required to achieve the same performance ($F_1 = 85\%$) as when excluding these top heads. This indicates that Russian BERT requires more heads to sufficiently represent government relations. Training probing classifiers without the top N heads eventually shows a slightly worse performance for Finnish, which indicates that the Finnish BERT has less distributed government information across its heads. It also suggests that the top N heads are reliable, but *not the only* indicators of government relations between words. Government can still be inferred well from all of the remaining attention heads as well.

6.4. Error Analysis

We manually examined instances that were misclassified by the probes. Some classification errors are caused by errors in the data ("*noise*")—parsers
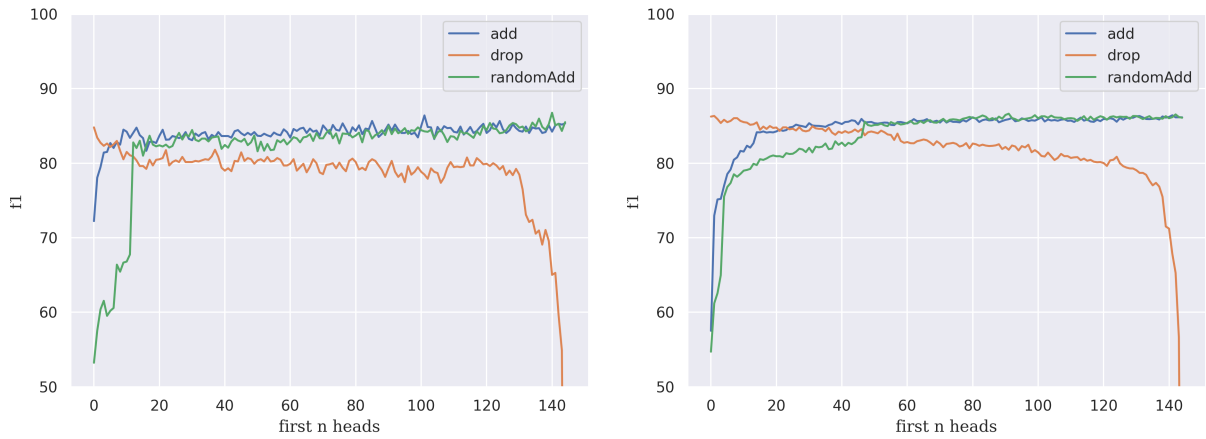
17465

Figure 4: F_1 score for Random forest classifier with selected attention heads when $Dist > 3$ (left: Finnish, right: Russian).

for both languages occasionally assign incorrect dependency relations. For example, in Finnish:

"Yritykset **investoivat** eli laajentavat toimintaa, rakentavat uutta ja hankkivat tekniikkaa." (*Companies invest or expand operations, build new ones and acquire technology.*)

The parser incorrectly labels "toimintaa" (*operations*) as the object of the governor "investoida" (*invest*). In fact, it is governed by another verb, "laajentaa" (*expand*), which is conjoined with "investoivat". All probes correctly predicted "no government," which decreases recall in our evaluation.

Similar problems occur in Russian data:

"Причем больше всего на свете они боялись, что кто-нибудь о нем узнает." (*Incidentally, more than anything, they feared that someone would find out about it.*)

The parser labels "всего" (*everything*) as an object of the governor "бояться" (*fear*). In reality, this verb's object is the following relative clause, and "больше всего" (*more than anything*) is an adverbial phrase modifier.

There are some inconsistencies between the representation of dependency relations—on one hand by the third-party neural parsers, and and on the other hand, by other components that we use to detect positive vs. negative government instances. This leads to some amount of misclassified instances. For example, in Finnish, subjects of passive verbs in the nominative case are parsed as objects. While the probing classifiers *correctly* reject government relations in such instances, this results in "false negatives" in evaluation. These examples represent "noise" in the data. They will be fixed in future work; this gives us hope that after that the results may further slightly improve.

Some proportion of the instances are true misclassifications. An analysis of these errors reveals that, occasionally, the probes fail to detect government relations with a long-distance governee

that *precedes* its governor, or incorrectly identify adjacent adverbial modifiers as arguments. In Finnish, true misclassifications are rare, mostly caused by missed arguments and misclassified adjacent words.

6.5. Discovering New Government Patterns

Research question RQ2 is: Are the probing classifiers capable of discovering new, previously unseen government patterns? To evaluate the generalization power of the probing classifiers, and check whether they can be used to enhance the Government Bank, we assess the performance of the probes on governors and government patterns that were held out from the training data ("unseen").

Unseen government patterns: For Finnish, we ran each probing classifier 8 times, while withholding two patterns from the labeled data, using these patterns only in the test data. In each run, we held out different patterns, e.g., all arguments in ablative case, or all third-infinitive arguments (e.g., "Hän **auttoi** leipomaan leipää.", "*S/he **helped** to bake bread.*"). We test analogously for Russian.

Table 6 presents the accuracy of discovering unseen government. The overall good performance implies that we can use the information encoded in the LM to discover new patterns, and extend the Government Bank. Higher accuracy for the Random Forest classifier suggests that the probes are capable of finding more unseen patterns compared to other classifiers.

Unseen governing verbs: We perform a similar assessment while withholding verbs from the training data. For each language, we run the probing classifiers 5 times. For each run, we withhold a random set of 66 verbs for Finnish (146 verbs for Russian). This lets us check the ability of the probes to discover new, unseen governor verbs.

Table 7 shows the performance, averaged across

	Model	Finnish	Russian
Dist > 3	LogReg	79.83	80.63
	MLP-1	80.17	84.87
	MLP-2	79.83	83.86
	RF	**83.79**	**83.77**
Dist > 2	LogReg	69.79	77.94
	MLP-1	73.81	82.05
	MLP-2	72.99	81.65
	RF	**77.42**	**81.29**

Table 6: Accuracy of classifiers on discovering unseen government patterns for Finnish and Russian. Reported metrics are micro-averaged.

		$Dist > 3$		$Dist > 2$	
	Model	*Acc*	*F1*	*Acc*	*F1*
Finnish	LogReg	79.12	78.15	78.86	77.53
	MLP-1	78.67	78.52	79.54	78.90
	MLP-2	76.82	76.49	77.35	76.83
	RF	**79.93**	**80.12**	**79.93**	**79.50**
Russian	LogReg	80.63	80.72	77.46	77.69
	MLP-1	**84.87**	**85.18**	**81.83**	**82.49**
	MLP-2	83.86	84.27	81.41	82.14
	RF	83.77	84.90	80.86	82.01

Table 7: Performance of classifiers on unseen governors. Reported metrics are micro-averaged.

all runs. The results indicate that probing classifiers can detect new verbs governing their complements, not previously seen during training.

These evaluations conclusively demonstrate the practical applicability of the government probing results—the information encoded in BERT's attention weights can be used for building new valuable resources—Government Banks.

7. Conclusion

The contribution of this paper is twofold:
1. We release a Government Bank (for Finnish and Russian) to the research community for work on the representation of grammatical government and constructions in neural language models.
2. We present an exploration of the representation of government constructions in transformer LMs. To the best of our knowledge, these are the first such resources to be made publicly available—not only in human-readable, but also in machine-usable form; and the first exploration of how well LMs can predict complex constructions, such as government.

We probe the syntactic information encoded in BERT's attention heads, to reveal what it knows about government relations. Our objective is to extract knowledge about government relations from inside the LM's attention mechanism.

We evaluate the performance of the classifiers on Finnish and Russian data, from several perspectives. To study RQ1, we assess the overall performance of the classifiers, and explore their probing selectivity with respect to the distance between the governor and its governees. We show that the performance of the probing classifiers is very high. Notably, the classifiers perform as well or better when the governee is far from its governor. We probed for the distribution of information about government across BERT's attention heads across different layers. We show that the probing classifiers are able to infer sufficient information from the first few layers of BERT.

We further explore the contribution of each attention head, using the coefficients learned by the logistic regression classifier. The idea is that the learned coefficients serve as a good indication of the contribution of the corresponding heads. Overall, the result shows that the most "important" attention heads (with the highest coefficients) are reliable indicators of government relations, but they are not the only ones. The probing classifiers are capable of detecting government using the remaining heads, which have lower coefficients. Comparing Finnish and Russian, we show that the information is slightly less spread out across the Finnish BERT heads, as compared to Russian BERT heads. We also performed an error analysis to understand the limitations of our probing classifiers.

For RQ2, we evaluate the ability of the probe to identify *novel* government relations, never seen during the training phase. We held out: (a) specific government patterns, and (b) governing verbs. The experiments show that the probes are able to discover novel government relations and novel governors, unseen in training.

In future work, we plan to use larger datasets, and identify large sets of government relations. This will extend the Government Bank with new patterns. We plan to work with additional languages, in particular, we are extending this approach to German and Italian government. Crucially, we will extend this work government of other parts of speech (nouns, adjectives, etc.), and to more complex types of constructions than government.

Limitations

The current work has a number of limitations to consider. (A) For now, our probing of government relations is limited to two languages. Extending to additional languages is challenging when probing this type of construction because it requires a Government Bank for each additional language, and collection of language-specific data. We are currently conducting experiments with an Italian and a German Government Bank; if the paper is accepted, we expect to be able to include those

results as well.

(B) So far, we have performed probing using only one type of transformer models—BERT-base. In future work, we plan to extend experiments to other models as well.

(C) So far, we have performed only one type of probing—using correlation probing with probing classifiers. This type of probing has received some criticism, of which we are aware. In future work, we experiment with the model representations from hidden layers and other types of probing methodologies.

Ethics Statement

We do not see any ethical issues with the current work. We use publicly available resources for all conducted experiments, and release language resources, which were created in collaboration with linguists, who were aware of how the data will be used.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks. In *International Conference on Learning Representations*.

David Arps, Younes Samih, Laura Kallmeyer, and Hassan Sajjad. 2022. Probing for constituency structure in neural language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6738–6757, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras

Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lymar, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhreva, and Marat Zaynutdinov. 2018. DeepPavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127, Melbourne, Australia. Association for Computational Linguistics.

Christopher, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Simone Conia and Roberto Navigli. 2022. Probing for predicate argument structures in pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4622–4632, Dublin, Ireland. Association for Computational Linguistics.

Alexis Conneau, German 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 (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. 2019. What is One Grain of Sand in the Desert? Analyzing Individual Neurons in Deep NLP Models. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Kira Droganova, Olga Lyashevskaya, and Daniel Zeman. 2018. Data conversion and consistency of monolingual corpora: Russian UD treebanks. In *Proceedings of TLT: the 17th international workshop on treebanks and linguistic theories*, volume 155, pages 53–66.

Charles J. Fillmore and Paul Kay. 1996. *Construction Grammar*. Manuscript, University of California at Berkeley Department of linguistics.

Adele Goldberg. 2006. *Constructions at work: The nature of generalization in language*. Oxford University Press.

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

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

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

Susan Hunston. 2019. Patterns, constructions, and applied linguistics. *International Journal of Corpus Linguistics*, 24(3):324–353.

Laura A Janda, Olga Lyashevskaya, Tore Nesset, Ekaterina Rakhilina, and Francis M Tyers. 2018. A constructicon for Russian: Filling in the gaps. In *Constructicography*, pages 165–182. John Benjamins.

Jenna Kanerva, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. 2018. Turku neural parser pipeline: An end-to-end system for the conll 2018 shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Benjamin Lyngfelt, Lars Borin, Kyoko Ohara, and Tiago Timponi Torrent, editors. 2018. *Constructicography: Constructicon development across languages*. John Benjamins.

Rowan Hall Maudslay, Josef Valvoda, Tiago Pimentel, Adina Williams, and Ryan Cotterell. 2020. A tale of a probe and a parser. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7389–7395, Online. Association for Computational Linguistics.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Tatiana Shavrina and Olga Shapovalova. 2017. To the methodology of corpus construction for machine learning: "Taiga" syntax tree corpus and parser. *Corpus Linguistics*, page 78.

Alex Tamkin, Trisha Singh, Davide Giovanardi, and Noah Goodman. 2020. Investigating transferability in pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1393–1401, Online. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description

length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Leonie Weissweiler, Taiqi He, Naoki Otani, David R. Mortensen, Lori Levin, and Hinrich Schütze. 2023. Construction grammar provides unique insight into neural language models. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, pages 85–95, Washington, D.C. Association for Computational Linguistics.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022. The better your syntax, the better your semantics? probing pretrained language models for the English comparative correlative. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10882, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176, Online. Association for Computational Linguistics.

Appendices

A. Overall instance statistics

Table 8 and Table 9 show the distribution of different types of governees—syntactic dependents of the head verb, which are governed by the verb—before balancing and sub-sampling. We should mention that we exclude certain common governees from our experiments, despite the fact that they are present in our government bank. The following direct objects governees are excluded: for Finnish—nouns in objective case and in partitive case, and for Russian noun in the accusative case. This is done because we are unable to find negative instances with these cases for data balancing, unlike for all other governee types. A large proportion of "negative" instances for these cases have been manually verified as positive instances. This suggests that our government bank is not yet complete, which is as expected. Therefore, we exclude these types of governees from our experiments.

PoS	Feature / Case	total
Noun	Inessive	2509
	Genitive	2424
	Illative	2337
	Elative	2190
	Adessive	1630
	Essive	1529
	Allative	1096
	Ablative	522
	Translative	659
Verb	Ma-Infinitive	1112
	A-Infinitive	1555
Postposition	Partitive	2
	Genitive	11
Adjective	Ablative	93
	Adessive	65
	Allative	42
	Elative	77
	Essive	230
	Genitive	312
	Illative	35
	Inessive	54
	Translative	235
total		18719

Table 8: Detailed overview of Finnish instances, before sampling

B. Hyper-parameters of classifiers

We implement the classifiers, including both MLP 1-layer and 2-layer classifiers, based on the Scikit Learn library,[7] and mostly follow their default hyper-parameter settings.

For the logistic regression classifier, we set the maximum number of training iterations to 10000.

For the MLP 1-layer classifier, we set 144 neurons. We use the same number of neurons for the first fully-connected layer in the MLP 2-layer classifier, while we set its second fully-connected layer to 72 neurons.

For the Random Forest classifier, we use 300 trees.

C. Selection of Transformer Layers and Attention Heads for $Dist > 2$

If we lower the threshold for separating "near" syntactic dependents from "far" dependents to $Dist > 2$, we observe the results in figures 5 and 6 for ablation studies in Section 6.3. These figures are included here for comparison with Figures 3 and 4, respectively.
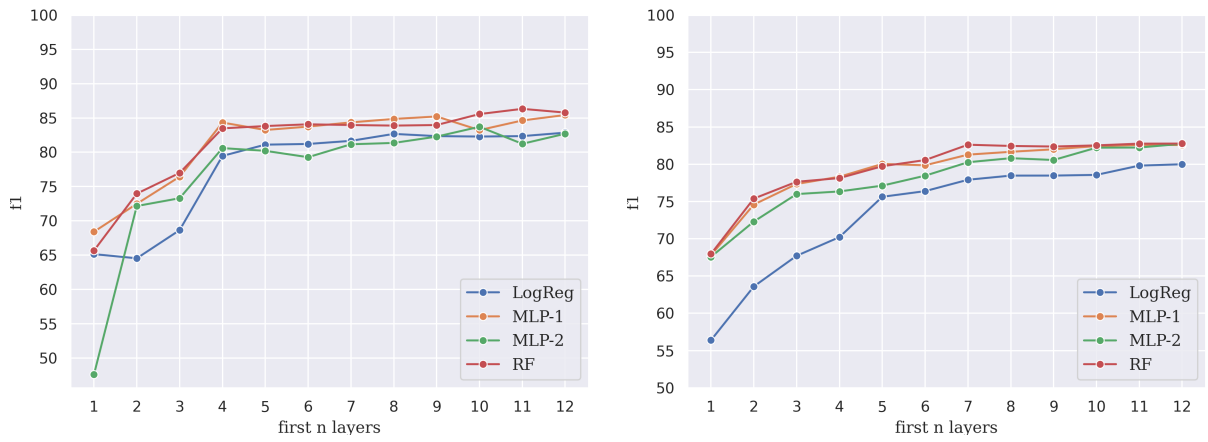
---

[7]https://scikit-learn.org/

Figure 5: Probing government prediction with attention weights from the first N layers of BERT (X-axis) when $Dist > 2$. Y-axis—F_1 measure. (left: Finnish, right: Russian)
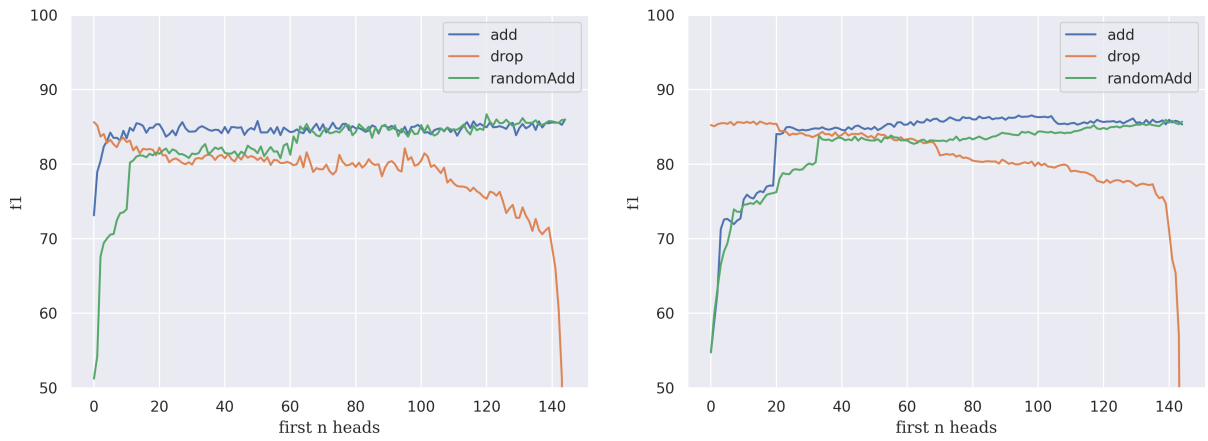

Figure 6: F_1 score for Random forest classifier with selected attention heads when $Dist > 2$ (left: Finnish, right: Russian.

PoS	Feature / Case		total
Verb	Infinitive		41090
Noun	Dative		11406
	Genitive		28926
	Instrumental		8671
Preposition	в	Accusative	6674
	за		1520
	на		8599
	о		42
	под		17
	про		236
	к	Dative	4278
	по		933
	без	Genitive	54
	для		210
	до		480
	из		2111
	из-за		25
	из-под		16
	от		2598
	против		56
	ради		6
	с		1150
	у		800
	за	Instrumental	840
	между		20
	над		234
	перед		38
	под		7
	с		3477
	в	Locative	3925
	на		2425
	о		2816
	Total		143836

Table 9: Detailed overview of Russian instances, before sampling