The argument–adjunct distinction in BERT: A FrameNet-based investigation

Dmitry Nikolaev **Sebastian Padó**
Institute for Natural Language Processing, University of Stuttgart
dnikolaev@fastmail.com pado@ims.uni-stuttgart.de

Abstract

The distinction between arguments and adjuncts is a fundamental assumption of several linguistic theories. In this study, we investigate to what extent this distinction is picked up by a Transformer-based language model. We use BERT as a case study, operationalizing arguments and adjuncts as core and non-core FrameNet frame elements, respectively, and tying them to activations of particular BERT neurons. We present evidence, from English and Korean, that BERT learns more dedicated representations for arguments than for adjuncts when fine-tuned on the FrameNet frame-identification task. We also show that this distinction is already present in a weaker form in the vanilla pre-trained model.

1 Introduction

The widely used Transformer-based contextualized language model BERT (Devlin et al., 2019) has been extensively studied regarding its capability to uncover linguistic patterns from raw text, with analyses focused mostly on syntax. Both constituency and dependency trees were either found encoded inside the model or were used to probe for syntactic rules such as agreement (Jawahar et al., 2019; Rogers et al., 2020).

In this paper, we shift the focus of BERT analysis to the syntax-semantics interface, considering the foundational distinction between arguments and adjuncts. According to Koenig et al. (2003), arguments and adjuncts differ in two crucial ways: arguments describe necessary participants in the event described by the verb and are therefore both *obligatory*, i.e. they have to be realized by default, and *specific*, i.e. they express idiosyncratic properties of the event or the event class. In contrast, neither is necessarily true for adjuncts. For example, in the sentence *Peter praised his colleague repeatedly*, the praising event is accompanied by two necessary, specific participants, namely a communicator, *Peter*, and an evaluee, *the colleague*; in contrast, the adverb *repeatedly*, which specifies the frequency, could be left out and applies to a very broad range of events. The argument–adjunct distinction has played a major role in linguistic theory (Chomsky 1981; Pollard and Sag 1994, but see Przepiórkowski 2016) and has implications for human language processing (Tutunjian and Boland, 2008) and semantic NLP (Zhang et al., 2020).

We empirically assess the status of the argument–adjunct disinction in BERT by making use of FrameNet (Baker et al., 1998) – an implementation of frame semantics (Fillmore, 1982), a theory of predicate-argument structure, which describes predicate meaning in terms of frames (prototypical situations) and frame elements (the situations' participants). FrameNet maintains a distinction between *core elements* and *non-core elements*, which maps onto the argument–adjunct distinction (see Section 2 for details).

We use a modification of the method of model analysis proposed by Rethmeier et al. (2020) for associating neurons inside neural-network models with features they are particularly attuned to. In our main analysis, we use FrameNet annotations to fine-tune BERT for a task – frame identification, – for which frame elements are informative, without exposing the frame-element labels to the model, and then correlate the learned model representations with the presence of these labels. We also repeat the correlational analysis on the vanilla (pre-trained) BERT model.

Our contribution is twofold: (a) we extend Rethmeier et al.'s methodology, which targeted LSTMs, to BERT and, instead of constructing a probability distribution of features a given neuron is attuned with, we extract tight neuron–feature combinations using correlation analysis, reminiscent of the larger neuroscience literature on input-specific neural activations (Dayan and Abbott, 2001); (b) we use this method for an analysis of the representation of arguments vs. adjuncts in Multilingual BERT

Proceedings of the 15th International Conference on Computational Semantics pages 233–239

(mBERT) based on English and Korean data. We find that even though BERT representations are dominated by frequency effects, with common input patterns more robustly tracked by individual neurons, arguments and adjuncts differ in their activation patterns (arguments produce relatively more robust activations while adjuncts generally lack highly specialized neurons that track them) and that this distinction is already present, to a lesser extent, in a vanilla pre-trained model.[1]

2 Frame Semantics and FrameNet

Frame semantics (Fillmore, 1982) posits that a key element of the understanding of an utterance is knowledge about the situations that the predicates in it evoke. This knowledge is captured through *frames*, schemas that associate predicates (*frame-evoking elements* / FEEs) with situations, their inferences, and their relevant participants, which are realized in language as so-called *frame elements*. Frame-semantic resources were first developed for English (FrameNet; Baker et al., 1998) but have been extended to other languages (Baker et al., 2018).

The example given in the introduction, *Peter praised his colleague repeatedly*, evokes the JUDGMENT_COMMUNICATION frame where a COMMUNICATOR expresses an evaluation of an EVALUEE. These are two of the *core elements* (CEs) of this frame, which generally meet both of Koenig et al.'s criteria for argumenthood: they are obligatory (unless they are null-instantiated, cf. Fillmore 1986) and they are specific to frames (or groups of closely related frames, cf. Fillmore et al. 2004). In contrast, the JUDGMENT_COMMUNICATION frame contains a number of *non-core elements* (NCEs), which do not meet at least one of the two criteria and thus show adjunct behavior: they are either not specific (MANNER, FREQUENCY) or not obligatory (GROUNDS: the basis for the judgment; ROLE: the capacity of the evaluee). A similar situation obtains with many other frequently found frames, and we assume that the core vs. non-core distinction largely mirrors the argument/adjunct dichotomy.

Data For our experiments, we use FrameNet corpora in English and Korean. For English, we use the FrameNet 1.7 lexical unit annotations, which cover over 1.2k frames and 13k unique predicates. The Korean FrameNet was created around a set of about 4k sentences translated from English, which were then added to using crowd sourcing. It aims for full compatibility with the English FrameNet (Hahm et al., 2020). We select 50 most frequent frames in both languages for analysis; the full list is given in the Appendix. There are 34,373 sentences in the English train set and 3,819 sentences in the test set. We use the Korean dataset only as a test set in a zero-shot setting. It contains 4,591 sentences.

3 Experimental Setup

Fine-tuning BERT We start from a pre-trained BERT model and fine-tune it to assign a single frame to each sentence (Hermann et al., 2014) in line with the FrameNet annotation (cf. Section 2).

We experiment with two variants of the task. In the *FEE present* setting, the model is shown complete sentences, including the FEEs, but no frame-element annotation. This task aims at encouraging the model to connect FEEs with arguments, which are known to be relevant for frame identification (Yang and Mitchell, 2017). Adjuncts are expected to be less relevant (as they are unspecific) or less reliable (as they are optional). To select the frame, we feed the first subword of the first FEE token to a fully-connected 50-neuron layer (corresponding to the 50 frames) and obtain a prediction by applying the usual softmax.[2]

In the *FEE masked* setting, all FEE tokens are replaced with the [MASK] token, so that the model has to rely on the sentential context to identify the frame. Our hypothesis is that this version of the task incentivizes the model to more actively focus on extracting arguments. In this case, we feed the embedding of the first masked token into the frame classification head as above.

In both variants, the model is trained end-to-end using cross-entropy loss for twenty epochs with early stopping when the performance on the test set decreases. We use the pre-trained mBERT model provided by HuggingFace (Wolf et al., 2020). For English, we report results for the test set. For Korean, we adopt a zero-shot setting and, after checking that mBERT fine-tuned on English has some success in identifying Korean frames, analyze the activations that Korean sentences produce in it.

[1]The code used for the analyses in this paper is available at https://github.com/macleginn/argument-adjunct-framenet

[2]We opt for a simplistic classifier head to keep more information in the embeddings.

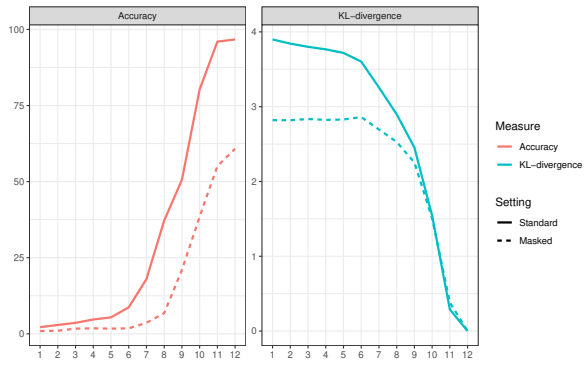

Figure 1: Left: Accuracy of predictions based on the output of different layers (development set). Right: the Kullback–Leibler divergence between the probability distribution of frame labels induced by intermediate layers and by the final layer.

Probing analysis Once the BERT model has been fine-tuned, we can analyze the activation patterns of different layers of the model (Rethmeier et al., 2020). On the data side, we cast input sentences as a binary matrix whose columns correspond to the presence or absence of each CE or NCE in these sentences, i.e. to their indicator functions. On the model side, we associate each input sentence with a $d = 768$-dimensional embedding of the first subword of the FEE token or the embedding of the first [MASK] token, depending on the setting, for a selected subset of BERT layers. We then carry out correlational analysis to identify, for each CE or NCE indicator function and for each layer of interest, the neuron whose activations are most strongly correlated with these functions.

To choose layers for the analysis, we evaluated English model predictions based on the representations in each layer. The results are shown in Figure 1. For both variants of the task, we find similar results: the outputs of the 11th layer are close to the final layer, and there is a swift increase in prediction accuracy from the 7–8th layer onward. On this basis, we probe the activation patterns of layer 11 (near-convergence) and layer 9 (start of competitive performance).

Analysis of neural activity was performed in a similar fashion by Durrani et al. (2020). They, however, extract activations in the context of specific tasks, such as POS tagging and syntactic chunking, instead of feeding sentences to a headless embedding model in an unsupervised setting.

Language	FEE present	FEE masked	MBL	RBL
EN	96	55	15	2
KO	40	21	12	2

Table 1: Frame ID accuracy in % on test set (layer 11). MBL: majority class baseline, RBL: random baseline

4 Results and Discussion

English Table 1 shows the test-set performance of layer 11 in the fine-tuned model.[3] As expected, the FEE-present setting is much easier than the FEE-masked one, where the model still substantially outperforms the baselines.

The results of the correlation analysis are presented in the scatterplots in Figure 2. Individual points show, for a frame element with a given frequency, how large the correlation with the most attuned neuron activation vector in the respective model is. The left plot shows core elements, the right plot non-core elements.

The plots show that frequency is the dominating factor: high-frequency frame elements tend to have (more or less) dedicated neurons tracking them, with correlations of 0.4 and above, while this is not true for low-frequency frame elements. This is to be expected given the maximum-likelihood training objective.

However, there still is a clear difference between CEs and NCEs: even the most frequent NCEs do not attain correlations above 0.3, and only a handful show correlations above 0.2, in both the standard and masked settings. In contrast, the correlations for CEs with frequencies above 100 are all higher than 0.2. This shows the model's low reliance on NCEs for frame identification.

Comparing the behaviors at layers 9 (red) and 11 (turquoise), we do not see major differences: in particular for NCEs, the plots are extremely similar. Comparing the two variants of the task (solid vs. dashed), we see that the masked-task model learns less dedicated representations for the CEs but spends some more effort on representing high-frequency NCEs – contrary to the expectation we formulated in Section 3. The global advantage of CEs over NCEs in all settings leads us to believe that the model simply relies on arguments in either case, and that in the masked setting the model just struggles more to identify where they are.

[3] Results for layer 12: 96.5/60.8 (EN), 41.3/23.9 (KO).

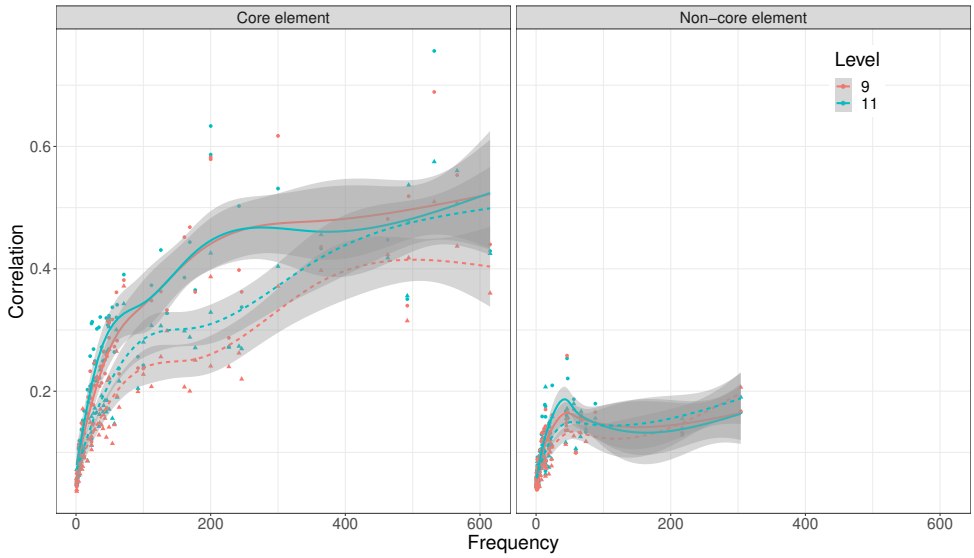

Figure 2: English fine-tuned setting: Averages and 95% confidence intervals for maximal correlations between BERT neurons and CEs (left) / NCEs (right), by frequency. Solid/dashed lines: FEE present/masked task. The curves show GAM-smoothed averages with 95% confidence intervals.

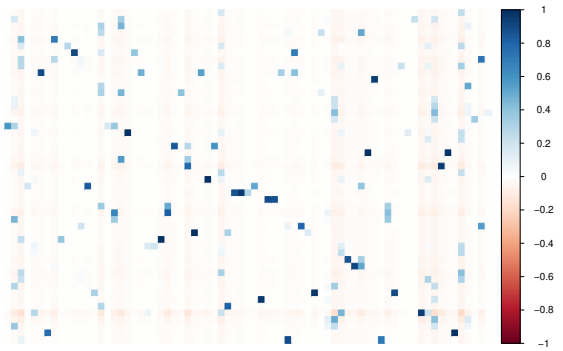

Figure 3: Correlations between frames (rows) and their core elements (columns).

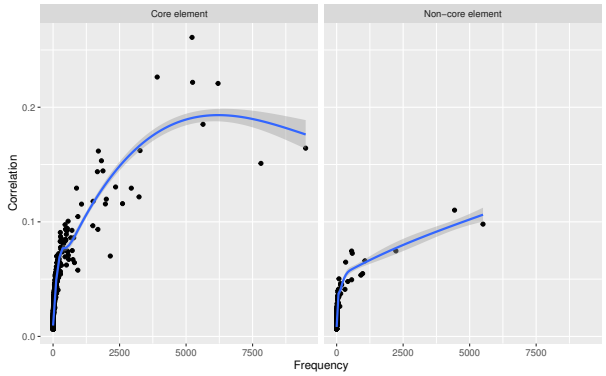

Figure 4: English vanilla BERT: GAM-smoothed averages and 95% confidence intervals for maximal correlations between neurons and CEs (left) / NCEs (right), by frequency. Solid/dashed lines: FEE present/masked task.

This interpretation is corroborated by an analysis of CE information content. Figure 3 shows a matrix of correlations between frames with non-masked FEEs (rows) and their CEs (columns). Some frames are in a nearly one-to-one correspondence with their CEs, but other CEs can be found with several frames. Arguably, when FEEs are present, they form a strong signal together with the CEs pointing towards particular frames. When FEEs are masked, however, frequent CEs – precisely those that are found with many different frames – become less informative, and the model shifts some of the weight towards NCEs.

Korean The accuracy results for the zero-shot application to Korean in Table 1 show similar tendencies to English, but with much lower accuracies. We attribute this to the simplistic linear classifier we use (cf. the observations on multilingual zero-shot transfer by Lauscher et al. 2020). However, the results of the correlation analysis shown in Figure 5 are strikingly similar to English: (a) top correlations of neural activations with CEs are much higher than those with NCEs; (b) strong frequency effects are evident; (c) the masked variant moves some focus from CEs to high-frequency NCEs. We take these observations as evidence that mBERT represents arguments and adjuncts in a remarkably similar way across languages as different as English and Korean, with the latter's rich morphology and SOV word order.

Without fine-tuning The above analysis uses a fine-tuned model. This begs the question of

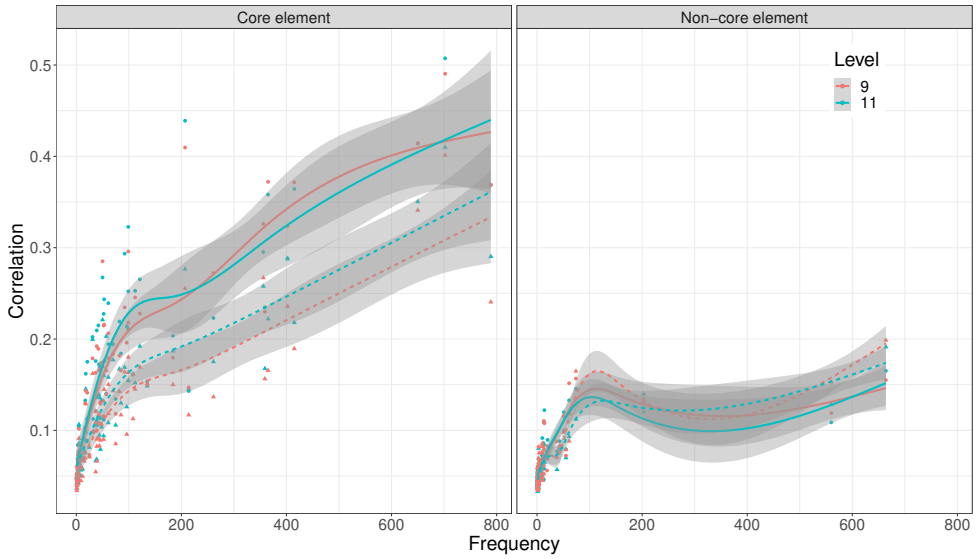

Figure 5: Korean zero-shot setup (model fine-tuned for English): GAM-smoothed averages and 95% confidence intervals for maximal correlations between BERT neurons and CEs (left) / NCEs (right), by frequency. Solid/dashed lines: FEE present/masked task.

whether the distinction between arguments and adjuncts is a side-effect of the fine-tuning task, as opposed to mBERT's acquiring it in an unsupervised way in pre-training (Tenney et al., 2019). To test this, we repeat the experiment using a vanilla English pre-trained model and the complete Berkeley FrameNet 1.7 release instead of the sentences with most-frequent frames. The results for layer 11, shown in in Figure 4, are remarkably similar in terms of the general pattern but with significantly weaker correlations: for CEs, correlations exceed 0.1 reliably for $N > 1000$, with maximum values approaching 0.3.[4] For NCEs, correlations are almost always < 0.1, reaching this value only for the most frequent NCEs, with $N \approx 5000$. This indicates that after pre-training BERT already has some notion of the distinction between arguments and adjuncts, but that this distinction becomes substantially more pronounced after fine-tuning on a task for which it is relevant.

5 Conclusion

Our study asked whether BERT can distinguish between arguments and adjuncts and operationalized these concepts via FrameNet's core vs. non-core frame-element distinction. For both English and Korean, our analysis of the presence of dedicated neurons that track individual frame elements found that this is the case, with frequency as a major covariate. The picture is clearer for a fine-tuned model, but the main patterns emerge already after pre-training.

On the neural-language-model side, our study confirms the ability of such models to recover 'deep' linguistic categories in an unsupervised manner. On the FrameNet side, our results have bearing on the status of borderline-core frame elements (Ruppenhofer et al., 2006), for which the behaviour of the model may serve as a heuristic. A promising avenue for future work would be to turn around our setup and to explore BERT representations in order to identify a set of properties that differentiate arguments and adjuncts from the model's point of view, *à la* Geva et al. (2021).

This work has focused on FrameNet. Other frameworks giving access to semantic-role information, such as the PropBank annotation scheme (Palmer et al., 2005), AMR (Banarescu et al., 2013), and UCCA (Abend and Rappoport, 2013), also may be fruitful for this type of analysis.

[4] Two most-frequent frames, AGENT and THEME, are very general and unsurprisingly display weaker correlations. By comparison, the next three most-frequent frames, SPEAKER, GOAL, and TIME, are much richer semantically and have more dedicated representations.

References

Omri Abend and Ari Rappoport. 2013. Universal Conceptual Cognitive Annotation (UCCA). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 228–238, Sofia, Bulgaria. Association for Computational Linguistics.

Collin F. Baker, Michael Ellsworth, Miriam R. L. Petruck, and Swabha Swayamdipta. 2018. Frame semantics across languages: Towards a multilingual FrameNet. In *Proceedings of the 27th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 9–12, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Noam Chomsky. 1981. *Lectures on government and binding*. Forus.

Peter Dayan and Laurence F Abbott. 2001. *Theoretical neuroscience*. MIT Press.

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.

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. 2020. Analyzing individual neurons in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4865–4880, Online. Association for Computational Linguistics.

Charles J. Fillmore. 1982. Frame semantics. In *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Co., Seoul, South Korea.

Charles J Fillmore. 1986. Pragmatically controlled zero anaphora. In *Proceedings of the Berkeley Linguistics Society*, volume 12, pages 95–107. Berkeley Linguistic Society, BLS.

Charles J. Fillmore, Collin F. Baker, and Hiroaki Sato. 2004. FrameNet as a "net". In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Younggyun Hahm, Youngbin Noh, Ji Yoon Han, Tae Hwan Oh, Hyonsu Choe, Hansaem Kim, and Key-Sun Choi. 2020. Crowdsourcing in the development of a multilingual FrameNet: A case study of Korean FrameNet. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 236–244, Marseille, France. European Language Resources Association.

Karl Moritz Hermann, Dipanjan Das, Jason Weston, and Kuzman Ganchev. 2014. Semantic frame identification with distributed word representations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1448–1458, Baltimore, Maryland. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Jean-Pierre Koenig, Gail Mauner, and Breton Bienvenue. 2003. Arguments for adjuncts. *Cognition*, 89(2):67–103.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Carl Pollard and Ivan Sag. 1994. *Head-driven phrase-structure grammar*. Chicago University Press.

Adam Przepiórkowski. 2016. How not to distinguish arguments from adjuncts in LFG. In *Proceedings of the Joint 2016 Conference on Head-driven Phrase Structure Grammar and Lexical Functional Grammar*, pages 560–580.

Nils Rethmeier, Vageesh Kumar Saxena, and Isabelle Augenstein. 2020. TX-Ray: Quantifying and explaining model-knowledge transfer in (un-)supervised NLP. volume 124 of *Proceedings of Machine Learning Research*, pages 440–449, Virtual.

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

Josef Ruppenhofer, Michael Ellsworth, Miriam R L Petruck, Christopher R Johnson, and Jan Scheffczyk. 2006. FrameNet II: Extended Theory and Practice.

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*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Damon Tutunjian and Julie E. Boland. 2008. Do we need a distinction between arguments and adjuncts? evidence from psycholinguistic studies of comprehension. *Language and Linguistics Compass*, 2(4):631–646.

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 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Bishan Yang and Tom Mitchell. 2017. A joint sequential and relational model for frame-semantic parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1247–1256, Copenhagen, Denmark. Association for Computational Linguistics.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware BERT for language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9628–9635.

A Appendix

Most frequent frames used in the analysis

ARRIVING, ATTEMPT SUASION, AWARENESS, BECOMING AWARE, BODY MOVEMENT, BRINGING, CATEGORIZATION, CAUSE HARM, CAUSE MOTION, CHANGE POSITION ON A SCALE, CHANGE POSTURE, COGITATION, COMING TO BELIEVE, COMMITMENT, COMMUNICATION MANNER, COMMUNICATION NOISE, COMMUNICATION RESPONSE, CONTACTING, COTHEME, DEPARTING, DESIRING, EVIDENCE, EXPERIENCER FOCUS, EXPERIENCER OBJ, FILLING, FLUIDIC MOTION, GIVE IMPRESSION, IMPACT, INGESTION, JUDGMENT, JUDGMENT COMMUNICATION, JUDGMENT DIRECT ADDRESS, KILLING, LOCATION OF LIGHT, MANIPULATION, MOTION, MOTION NOISE, PERCEPTION ACTIVE, PERCEPTION EXPERIENCE, PLACING, REMOVING, REQUEST, RESIDENCE, REVEAL SECRET, SCRUTINY, SELF MOTION, STATEMENT, TELLING, TEXT CREATION, USING.