Exploring Typology with mBERT: A Cross-Lingual Comparison of the Features behind Morphosyntactic Alignment

Nicholas Hankins
Montclair State University
hankinsn1@montclair.edu

Abstract

It has previously been proven that there is evidence of the transfer of grammatical information in pretrained languages in Multilingual BERT (mBERT). These findings can be replicated and observed by the process of extracting contextualized word embeddings that allow us to classify what is constituted as a Subject, otherwise known as Morphosyntactic Alignment, among languages with a Nominative-Accusative alignment or Ergative-Absolutive/Split Ergative alignment. We constructed experiments where language groups with unique linguistic features, such as pronoun dropping and distinct genealogical trees, are deliberately compared with a balanced set of languages without those features. The objective of this is to see if there would be a clear variation when, given an intransitive subject, the trained classifiers have a greater or lesser influence on Morphosyntactic Alignment after abstract grammatical features are ostensibly isolated. Contingent on our results, we can further speculate that a more profound understanding of the intricacies occurring in the grammatical transfer of multilingual information in Large Language Models (LLMs) is possible, and could potentially lead to the facilitation of future work regarding low-resource languages with under-examined grammars.

1 Introduction

We've historically been able to complete tasks such as Part-of-Speech tagging and Named Entity Recognition even before the emergence of Transformer architecture through Naïve methods, notably with Hidden Markov Models. However, what we are still beginning to uncover is how we can analyze more abstract grammatical features of language within Large Language Models. An example of an abstract grammatical feature, and the one that we analyze within this paper, is Morphosyntactic Alignment. Morphosyntactic Alignment can be considered abstract because it is not something that can be parsed in sequences or at a word level. It is a feature of the entire grammar of a language that can be observed through the analysis of language data. The contextual word embeddings in Large Language Models such as mBERT hold valuable grammatical information related to abstract grammatical features such as this one (Devlin et al, 2019). To further extrapolate on this information, we can look towards literature where it has been proven that there is evidence of cross-lingual transfer within and across languages (Papadimitriou et al, 2021).

This means that, based on the case marking of a given language, we can use a classifier to predict the chances that a given intransitive subject will be classified according to its expected output, otherwise adhering to its language's typical alignment. To expand on these findings, this paper endeavors to form scenarios where this use of transfer learning is done to observe how different languages react to typological changes. The experiments and following discussion go into detail on the process and findings[1].

2 Background

Syntax and Typology are examples of areas of Linguistics that have numerous different theories related to what the most efficient representation of Linguistic information is. For example, a commonly accepted theory of Syntax is the Generative Grammar Syntax theory (Closs, 1965) which was originally established and revised over the years. Another prominent theory of Syntax is the Dependency Grammar theory (Debusmann, 2000). For this assignment, the research involved using data that had been annotated using a dependency grammar schema to denote the syntactic and morphological information included in the relevant language data. It is important to also mention that the initial

[1] My GitHub for reference: https://github.com/nhankins

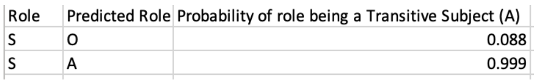
purpose of the paper inspiring this project was to observe the impact that contextual word embedding had in languages with varying grammatical features, specifically related to Morphosyntactic Alignment. The code that was used as a starting point for this project (Papadimitriou et al, 2021) was able to extract the grammatical roles necessary to fine tune and train mBERT for a classification task. As a result of these breakthroughs, it is possible to continue cross-lingual analysis of abstract grammatical features and truly understand what similarities influence grammatical similarities and differences as a whole sample of a language, not as a particular sequence.

3 Methodology

The starting point to this research involved adjusting preexisting code from prior literature for a more modernized virtual environment[2]. This was necessary to ensure that only specific parameters were selected during the initial baseline training within the classifier. Among the changes included: Correcting the batch shell files which were responsible for the zero-shot transfer learning so there were no Type or Syntax Errors, creating folders for the output data to be stored, verifying that the version of the transformers library is compatible and is not flagged as an Attribute error, adjusted list indices to optimize ranges for different languages, and set all random seeds to 0.

Once the starting code had been prepared and the data was assembled, the logic used here was that even though the classifier is not specifically being trained on features beyond whether a token's grammatical role is A, O, or S, the mBERT contextual embedding may carry valuable alignment information in this synthetically balance 3x3 language differentiation for each experiment. The reason for this being to observe if the classifiers are able to infer and learn alignment information at the prediction stage without the need to be explicitly trained for all the grammatical features presented.

When filtering and parsing the output data, the piece of information that was found to be most pertinent to this goal of isolating specific typological features to pinpoint the level of influence on morphosyntactic alignment, was the probability of A information. Each role that was found to truly be an Intransitive Subject (S) was filtered at this stage. Then, this information was separated at each layer and averaged. Therefore, if the actual embedded role (S) had a softmax predicted role of A, than the probability of A would be extremely high i.e. close to 1. If the predicted role was output as O, it would be an extremely low probability of A. These calculated averages were then plotted for further analysis. Each experiment has 6 line plots; one for each language that had its own data as the training reference, and the 5 other languages passed through in an effort to notice grammatical reactions. The x-axis correlated to each layer of the mBERT classifier (13 total) and the y-axis correlating to the accuracy of predicting the grammatical role as being SA or SO. These line plots did not show performance, but rather if a line was higher, than it indicated that the language was more likely to be associated with SA or nominative accusative alignment, whereas if it was lower then it would be more likely classified as SO or Ergative-Absolutive or Split-Ergative alignment. It is important to mention that the SA and SO averages for the probability of A outputs were not separated, but rather combined at the visualization stage.



Role	Predicted Role	Probability of role being a Transitive Subject (A)
S	O	0.088
S	A	0.999

Figure 1: Example of SO and SA roles with probabilities of a transitive subject being A given S taken from two English Test Sequences

4 Data

All of the data gathered for training and testing was from the Universal Dependency Treebank[3]. As mentioned previously, it contains dependency parsed information relevant for the classification tasks. It was a heuristic process in searching for data that existed with a train and test split and had a sufficient amount of usable tokens i.e. enough marked for the alignment roles with which we are interested. This could not be easily observed in preprocessing, since the dependency Treebank data is not explicitly marked for case alignment. So, when the initial baseline classifier for each language was trained individually, it was uncovered whether or not the specific dataset could be used in cross-lingual comparison. In addition, the languages chosen had to have been one of the pre-

[2] Paper where code was found as a starting point: https://github.com/toizzy/deep-subjecthood

[3] Link to Universal Dependency Treebank where data was collected: https://universaldependencies.org/

trained mBERT languages so that a contextual representation could be made. There was no modification to the data from its .conllu format and the selection criteria for what tokens the classifier is trained to recognize is not altered either from the original code.

5 Experiment 1: Areal and Split-Ergativity Comparison

As a way to segue into novel research, the approach for this first experiment was to include a total of 6 languages for zero-shot language transfer. The languages selected for this experiment were Western Armenian, Persian, Turkish, Hindi, Indonesian, and Urdu. Three languages that have split-ergative morphosyntactic alignment (Sinha, 2017) or share a language family with a language that has split-ergative morphosyntactic alignment (Aldridge, 2016), and then three languages that are not known to have non-ergative morphosyntactic alignment in any way, that is, strictly a Nominative-Accusative case alignment system. At the same time, we looked to concurrently perceive if the areal typological differences of the 3 non-ergative languages hold any influence over the final results. Despite this particular feature being hard to identify by itself, it is curious to speculate on the geographic implications of language differences.

During the baseline training for each classifer, it was observed that these languages output a significant amount of data related to each relevant sequence containing an A (Transitive Agent), O (Transitive Object), or S (Intransitive Subject) token. Some examples of what the reference code was organized to output are features such as animacy, case, and the probability of the alignment role being A.

5.1 Experiment 1 Results

Given that this experiment occurred first and acted as a test of the methodology, there was much to learn and absorb with the line plot results. There was plenty of variation between the 6 languages, so it was a challenge to confine the attribute to a consistent trend. Nevertheless, there were indeed trends to be seen within the line plots. For example, if the line plot for the Indonesian language outcomes are distinguished between all of the 6 total lanugages involved with this Experiment, it was aligned most frequently with SA or what we typically associate with Nominative-Accusative alignment. It is also curious that with the classifier trained on Indonesian data, there were not many convergences in the languages SA/SO average values.

It is likewise noteworthy to say that there is a U-curve that does not appear elsewhere in this Experiment, nor the following two. The significance of the curve is that as the classifiers are trained throughout the 13 layers, there is an learned trend that the alignment of SO can exist more than a language strictly known as aligning more with SA can expect. This can most likely be attributed to the split-ergative nature of the languages. Given that the cross-lingual transfer of contextual mBERT word embeddings across languages has been proven (Papadimitriou et al, 2021), it is supported here in a replicated, yet distinct environment.

Some takeaways from this first Experiment is that we can use contextual embeddings to, at the very least, test traditional notions of alignment that a language may have. For example, Western Armenian is a language that is usually known in present day as an SA language (Meyer, 2020), yet nearly every classification matrix, besides its own, contains a lower line for Armenian, indicating a more likely association with SO. This intuition of testing linguistic features can also be extended towards typological aspects such as agglutination and polysyntheticism in future work.

6 Experiment 2: Pronoun Dropping

After discerning the results of the first experiment and reviewing more literature on the subject, it appeared that the utility of the pronoun in relation to subjecthood is quite important (Fox, 1987). Therefore, now that we had the first experiment to reference as a template, including the functions created for filtering and averaging the data, the language data was replaced with three languages where you can drop the pronoun and three languages where the pronoun cannot be dropped. These languages are English, French, German, Portuguese, Korean, and Hebrew. Purposely, there are no languages that exhibit any ergativity among these 6, so we are able to ideally examine and analyze typological differences, in this case pronoun dropping, without the influence of languages that sway to a lower SO average when predicting an Intransitive subject's role.

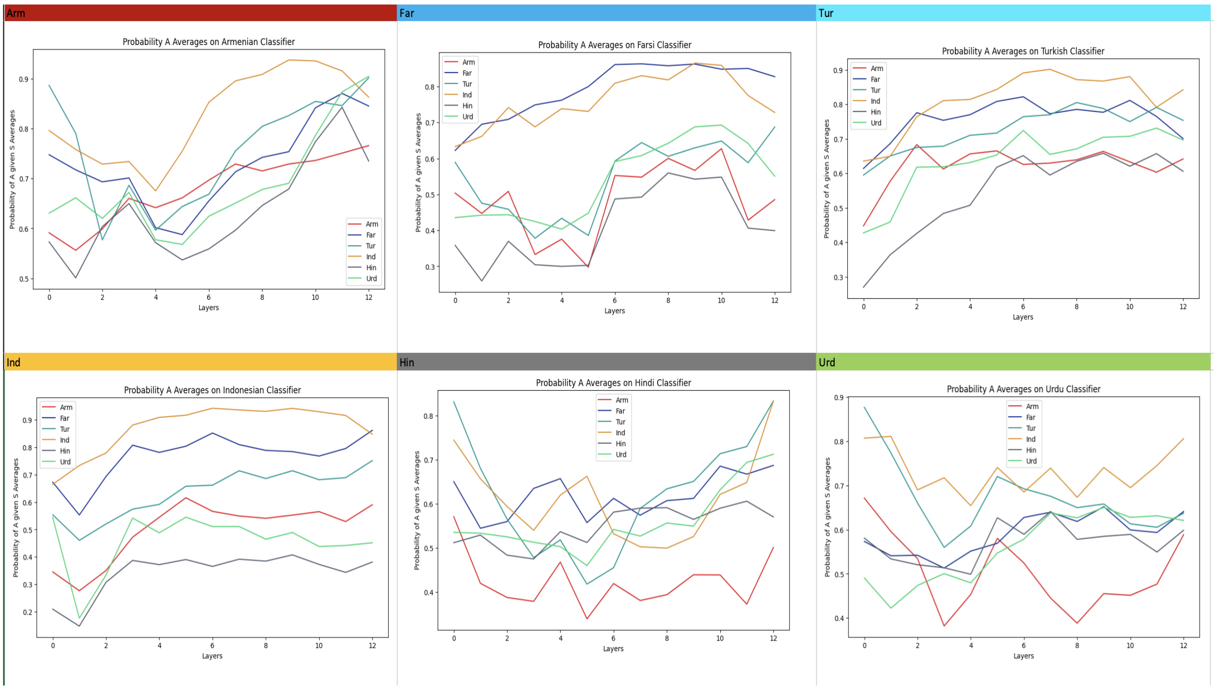


Figure 2: Experiment 1 Results (Areal and Split-Ergativity Comparison)

6.1 Experiment 2 Results

There was not any research that specifically covered how the typological feature of pronoun dropping can be analyzed through a lens of morphological alignment so there was uncertainty of what to expect. That being said, the groups behaved mostly as one might expect. That is, in language classifiers trained on grammatical data where pronoun dropping is not allowed (English, French, and German), we can notice that they all associate more closely with an SA alignment. The same three languages seem to drop closer to SO when predicted on pro-drop language classifiers. On the other hand, the other three language classifiers (Portuguese, Korean, and Hebrew) generally perform closer to SO, both when tested on the non pro-drop triplet and pro-drop triplet. In other words, it can be stated that if a model is to be fine tuned or trained on subjecthood, using a language model that allows and/or is familiar with pronoun dropping may be a wise decision since there are more robust results that allow for uniform comparability.

We can assume other positional factors are learned since Korean is the only SOV language among the languages used in this particular experiment(Jung and Lee, 2018). Additionally, Korean is the only language of this set where object dropping is allowed as well (Kim, 2000). Bearing this in mind, it explains why Korean emerges to have an irregular line plot amongst all the other examples in this Experiment. This supports the notion that other typological features play a role in determining how contextual word embeddings are transferred and classified.

7 Experiment 3: Genealogical Comparison

Upon inspecting the results of the preceding Experiment 2 regarding Pronoun dropping, there was supplemental speculation that genealogical or language family may be a strong contributing factor in the LLM contextual embedding information. This became evident specifically when, while analyzing the Portuguese trained classifier, French, a non pro-drop language (Kim, 2000), performed more closely with SA than the other two non pro-drop languages. This is intriguing since French and Portuguese are both Romance languages and clearly have similar learning curves in the Portuguese classifier.

Building on this intuition, the structure for the third and final Experiment are the same as the previous two, except for the languages trained and analyzed. The languages chosen for this Experiment were Bulgarian, Czech, Slovenian, Catalan, Italian, and Romanian. The first three languages are all in the Slavic language family, while the last three are all in the Romance language family.

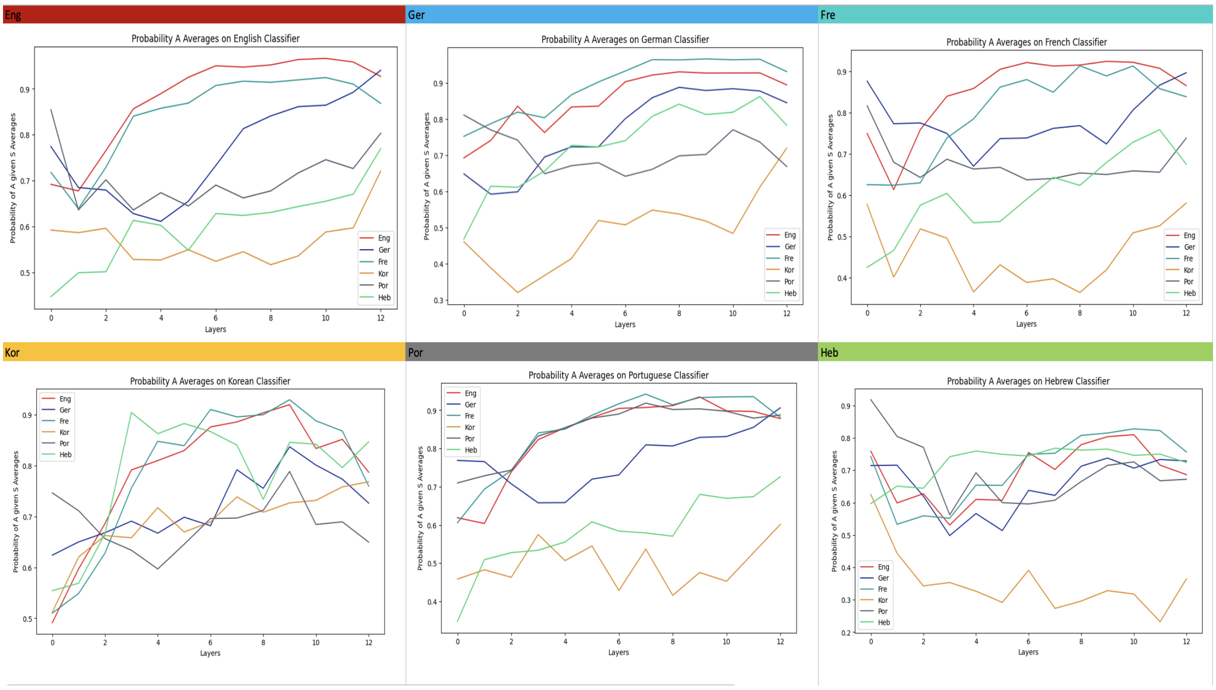


Figure 3: Experiment 2 Results (Pronoun Dropping Comparison)

7.1 Experiment 3 Results

For this last experiment, it was potentially the most difficult to distinguish an overarching trend among the 6 different line plots denoting the 6 trained languages. It was clear that many of the languages converge around the final layer, e.g. Romanian and Italian. It has been stated in prior literature that we can assume grammatical details in mBERT are held between higher layers of 7-10 (Papadimitriou et al, 2021), so this is fairly consistent with those findings. One can argue that the typological similarities existing in their respective language families could be impacting the results, such as the fact that Slavic languages do not usually have articles (Pereltsvaig, 2013), therefore guiding an SA or SO line plot average decision. An example of an unexpected outcome, though, would be Catalan aligning more highly with SA than Bulgarian on a Bulgarian trained classifier, especially when there is no known existence of ergativity in Bulgarian.

A certain big mystery thus appeared in the all 3 Experiments with how the source languages did not always score as closely to what their language's structure might imply. It is crucial to reiterate that a language's line plot average being higher does not necessarily mean it is better, conversely a line plots average being lower does not necessarily mean it is worse; those are simply how the grammars of those languages average out to be given a zero-shot training scenario of a different language alongside their own language's data. With that being said, it was noticed that during the prediction step of all the experiments, there were sequences of tokens not only classified as A, O, or S. This revelation creates conjectures that are expanded on in the limitations section. Be that as it may, it is still helpful to mention that a clear tendency of this third experiment was that Romance language's seem to adhere towards SA alignment regardless of a classifier's family.

8 Limitations

As introduced at the end of the Experiment 3 results, there may be some adjustments to be done in truly getting desired results for these experiments as a whole. More concisely, at the prediction transfer learning phase of each experiment, when test data is passed through the fine tuned classifiers to generate the inferences, each languages was set to select exactly 2,000 tokens that appear in sequences. The expectation was that the classifier selects only A, O, and S as the relevant tokens for predictions. Instead of that being the case, the testing code was not set to filter out other grammatical roles besides purely A, O, and S. Nevertheless, the experiments yielded a majority of A, O, S roles, while likewise including roles such as passives, expletives and auxiliaries associated with the roles.

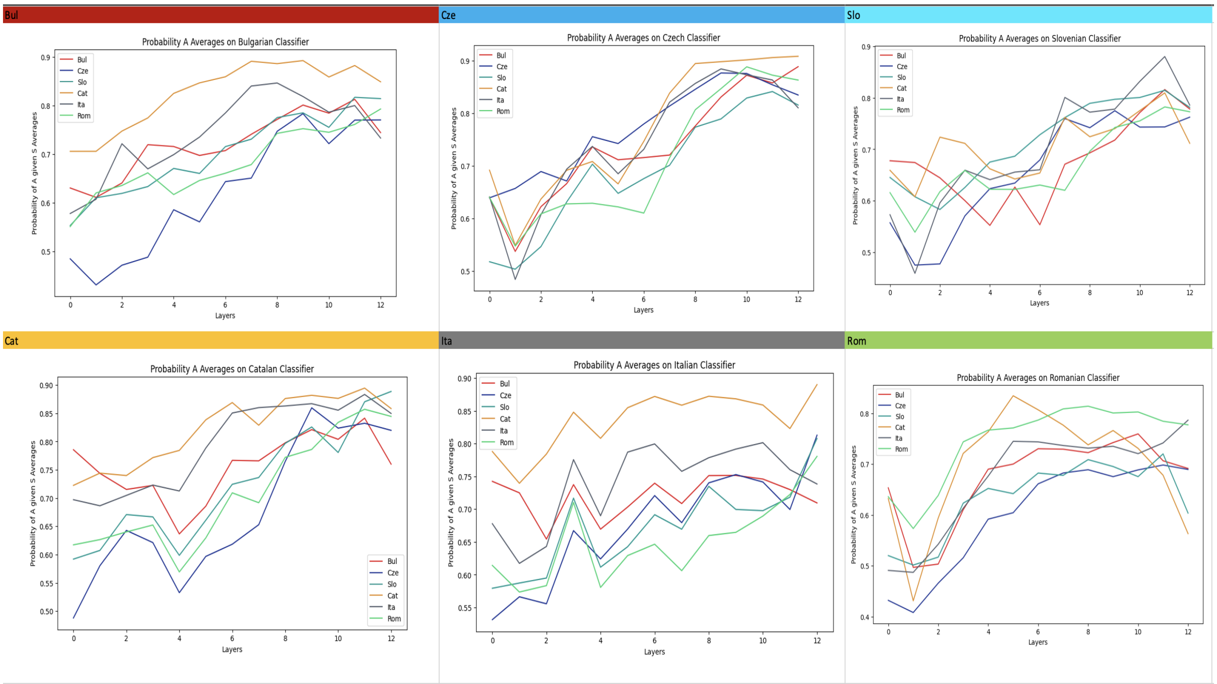


Figure 4: Experiment 3 Results (Genealogical Comparison)

This could have potentially confounded the classifier in some situations and given a role that might have otherwise included another S role prediction to augment the plotted averages.

This was not, by and large, an extreme limitation, since not every language experienced this counting selection. For example, Korean had exactly 2,000 tokens that we solely A, O, S extracted in the testing stage, when languages such as Italian had a combination that can be seen below:

All Italian Tokens Tested: ’O’: 995, ’S’: 408, ’A’: 308, ’S-passive’: 158, ’S-expletive’: 125, ’A-passive’: 3, ’S-expletive-passive’: 3

Hence, roles that could have been included as more SA or SO data were looked over by other roles such as the examples listed. Correspondingly, the source languages, also known as the test language data trained on languages that it should recognize and conform with, performed differently than they might have been without confounding the model of what the definition of subjecthood is.

Other limitations that should be addressed aside for human error of overlooking parameter settings, one can say that these experiments are quite limited in the availability of data. It is purely by chance that the selected languages have the sufficient quantity of tokens required for Morphosyntactic Alignment comparisons, so one could assume that it may be easier to compile a corpora of sequences only consisting of applicable token roles. Moreover, these experiments beg the question of whether or not the typological features such as genealogical classification or geographic location are too generalized to divulge accuracies evident in contextual embedding analysis at this time.

9 Discussion

In an attempt to string a conclusive narrative with this information, it can be said that the most obvious clues that were detected involved languages that behaved as one would expect according to their historically documented morphological alignment. More precisely, we are concerned with how languages features are impacted when they occur in the similar environments to similar and dissimilar typological precedence. Examples of languages occurring as one might expect are, in Experiment 2, when looking at the classifier that had been trained on English data and tested on 6 languages (including itself), and the non pro-drop languages align higher, indicating an SA preference.

Some surprises took place in the analysis of this data in all the experiments that may or may not be able to clearly define. Reconciling word order could have the ability to elucidate the averages such as the performance of Korean in the second experiment or of Turkish in the first experiment. The noisiness of the data could also enlighten us as to the peculiarities of these data plots.

10 Conclusion

As far as future work is concerned, there would need to first be verification that the batch zero-shot transfer learning of test data step is selecting the appropriate roles at runtime. More specifically, that only the roles of A, O, and S are considered. It may be that each language's dataset does not have enough purely A, O, S tokens so the algorithm has to pad with passives and expletives as a consequence. This would need to be tested and confirmed.

The implications of this research are that we can begin to gain a greater understanding about grammatical features, including but not limited to, Morphosyntactic alignment. This lends us the possibility of extending this analysis to more under-examined low-resource languages as they are made available on platforms such as the Universal Dependency Treebank. It would be curious to analyze the results with a larger dataset of languages that have at least split-ergativity. More grammatical features can likewise be taken into consideration such as what specific syntactic organization patterns provide insight into how LLM contextual word embeddings inform predictions.

Acknowledgements

Advised by Dr. Anna Feldman and received valuable input from the Montclair State University NLP lab participants, as well as Linguistics and CS faculty.

11 References

Papadimitriou, I., Chi, E. A., Futrell, R., and Mahowald, K. (2021). Deep Subjecthood: Higher-Order Grammatical Features in Multilingual BERT. ArXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2101.11043

Fox, B. A. (1987). The Noun Phrase Accessibility Hierarchy Reinterpreted: Subject Primacy or the Absolutive Hypothesis? Language, 63(4), 856–870. https://doi.org/10.2307/415720

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/1810.04805

Pereltsvaig, A. (2013), Noun Phrase Structure in Article-less Slavic Languages: DP or not DP?. Language and Linguistics Compass, 7: 201-219. https://doi.org/10.1111/lnc3.12014

Aldridge, E. (2016). Ergativity from Subjunctive in Austronesian Languages. Language and Linguistics, 17(1), 27–62. https://doi.org/10.1177/1606822X15613499

Meyer, R. (2020). Armenian Morphosyntactic Alignment in Diachrony. Historische Sprachforschung, 109(1), 40-44.

Closs, E. (1965). Diachronic Syntax and Generative Grammar. Language, 41(3), 402–415. https://doi.org/10.2307/411783

Debusmann, R. (2000). An introduction to dependency grammar. Hausarbeit fur das Hauptseminar Dependenzgrammatik SoSe, 99(1), 16.

Sinha, Y. (2017). Ergative case assignment in Hindi-Urdu: Evidence from light verb compounds. Proceedings of the Linguistic Society of America, 2, 32:1–14. https://doi.org/10.3765/plsa.v2i0.4079

Jung, S. and Lee, C. (2018), Deep Neural Architecture for Recovering Dropped Pronouns in Korean. ETRI Journal, 40: 257-265. https://doi.org/10.4218/etrij.2017-0085

Kim, YJ. Subject/Object Drop in the Acquisition of Korean: A Cross-Linguistic Comparison. Journal of East Asian Linguistics 9, 325–351 (2000). https://doi.org/10.1023/A:1008304903779