Joint Encoding of Linguistic Information in BERT-based Language Models

Apostolos Panagiotopoulos (#14021900), **Konstantinos Papakostas** (#13914332),
Matteo Rosati (#13858149), **Giulio Starace** (#13010840)

{ apostolos.panagiotopoulos, konstantinos.papakostas,
matteo.rosati, giulio.starace } @student.uva.nl

Abstract

BERT-based Language Models (BLMs) have gained considerable popularity due to their impressive results across diverse NLP tasks. However, their representations are still not interpretable. In this work, we study the information-sharing mechanisms in BLMs and highlight the presence of joint encoding across linguistic entities, such as part-of-speech tags and dependency labels. Furthermore, we find that the way this information is shared resembles the linguistic relations between entities and is consistent across languages. Our results are validated with RoBERTa and XLM-R in English, Italian, and Greek, and take a step toward a deeper understanding of information flow in BLMs. Our code is available at https://github.com/thesofakillers/bert-infoshare

1 Introduction

In recent years, contextualized word embeddings obtained from Large Language Models (LLMs) have consistently demonstrated state-of-the-art performance across multiple NLP tasks (Devlin et al., 2019; Liu et al., 2019; Wu and Dredze, 2020). However, these representations have proven hard to interpret with regard to the information they contain, leading to a series of works (Tenney et al., 2018; Hewitt and Manning, 2019; Ravishankar et al., 2019; Libovický et al., 2020; Choenni and Shutova, 2022) that probe these models to extract insights. Choenni and Shutova (2022) showed that multilingual LLMs jointly encode information across languages, which inspired us to investigate whether they also jointly encode information across different linguistic entities, such as *nouns* and *verbs*.

We approach this problem by studying how two state-of-the-art models, RoBERTa (Liu et al., 2019) and XLM-R (Conneau et al., 2020), encode information about part-of-speech (POS) tags and dependency relations, as well as how this information may be shared between classes. We hypothesize that these BERT-based Language Models (BLMs) learn to jointly encode words from related classes in a way that mirrors linguistic relationships. We expect to see that learned representations for linguistic features are not completely independent, with information sharing between them increasing with higher co-occurrences or functional dependencies (e.g. nouns and determinants or nominal subject and clausal subject dependencies). These overlapping relationships can vary across languages, though we anticipate that lexical entities, i.e., certain parts-of-speech such as nouns, share some degree of information in a language-agnostic manner.

Our methodology leverages a cross-neutralizing method based on centroid estimation (Libovický et al., 2020), akin to the multilingual approach in Choenni and Shutova (2022). In particular, we use a pre-trained instance of RoBERTa (Liu et al., 2019) as the encoder architecture and probe various layers to perform the POS tagging and dependency labeling tasks. Then, we apply the cross-neutralizing method to uncover which entities share information. We extend our experiments to the multilingual setting using XLM-R (Conneau et al., 2020) for the English, Greek, and Italian languages.

This work bridges the gap between analyses of multilingual information sharing and attempts to localize granular semantic and syntactic information within LLMs. Tenney et al. (2019) and Clark et al. (2019) have evaluated *where* information is encoded by probing different layers within BERT. However, their work lacked an analysis of *how* this information is shared. On the other hand, papers that have explored a direction similar to ours focus on how information is shared *across languages* (Choenni and Shutova, 2022), rather than *across linguistic features*. By combining these two approaches we can arrive at a more comprehensive understanding of how information about linguistic

units is represented in BLMs.

We find that similarly to typological properties of languages (Choenni and Shutova, 2022), information is shared across different linguistic features in BLMs. We report that this phenomenon extends to the multilingual setting in the languages tested, with variations in which labels are jointly encoded that are language-specific. Additionally, cross-lingual self-neutralization (i.e., a NOUN centroid from Italian used to cross-neutralize the NOUN class in Greek) also yields a decrease in performance, suggesting that centroids can encapsulate language-agnostic features.

2 Related Work

Prior work has focused on two main approaches to inspect the data contained in contextualized embeddings; through descriptive "probing tasks" (Conneau et al., 2018) and through informative subspace representations that are used to decompose word vectors (Libovický et al., 2020).

At the sentence embedding level, a prominent curated suite of probing tasks was introduced by Conneau et al. (2018). These tasks examine whether sentence embeddings encode properties such as sentence length, tree depth, etc. This work is extended by Hewitt and Manning (2019) by proposing methods to discover linear transformations that encode word distance and tree depth in sentence parse trees using word representations. Tenney et al. (2018) goes further with "edge probing" tasks, which probe the token-level representations directly and expand to a broader range of syntactic and semantic tasks at the sub-sentence level.

In later work, Tenney et al. (2019) apply these probing methods to BERT's hidden states to quantify where linguistic information is captured within the network. They find that the model represents the steps of the traditional NLP pipeline, with lower-level, syntactic features appearing earlier than more complex semantic roles and structure. This work has been extended in the multilingual setting to probe for the previously mentioned linguistic information (Ravishankar et al., 2019) or novel information, such as determining where typological properties of certain languages are encoded (Choenni and Shutova, 2020). Additionally, Clark et al. (2019) explore BERT's attention mechanism to identify what linguistic characteristics it attends to, showing that they correspond to linguistic notions of syntax and co-reference.

Other than identifying which information is encoded in word embeddings and hidden states, substantial effort has been put into trying to identify how this encoded information is shared across linguistic categories and across languages (Şahin et al., 2020; Blevins et al., 2018), with most of the work being focused on the multilingual setting (Chi et al., 2020). In particular, Libovický et al. (2020) developed a method that successfully removes language-specific features without affecting encoded language-agnostic information (i.e., semantic meaning) by leveraging what they refer to as "language centroids". Our work draws inspiration from that of Choenni and Shutova (2022), who probe for joint encoding of typological features of different languages by extending the previous method. They find that these feature values are encoded jointly across languages and are localizable in their respective language centroids.

Compared to this existing work, we apply the cross-neutralizing method to probe BLM's hidden representations for joint encoding of syntactic classes instead of languages. In sum, we apply the multilingual methods to a multi-label paradigm. Using the centroid method introduced by Libovický et al. (2020) and extended in Choenni and Shutova (2022), we explore what class information is independent and what is jointly encoded, as well as where in the model this information emerges. We also explore how these results vary from monolingual to multi-lingual models, as well as across languages.

3 Methods

3.1 Extracting Word Representations

We use the base versions of RoBERTa (Liu et al., 2019) and XLM-R (Conneau et al., 2020) as our encoders, which produce a contextualized embedding for each sub-word token that is generated by their tokenizers. To extract a set of representations on the word level, we need to aggregate the sub-word token representations. We do so by either **(i)** taking the representation of the first sub-word token, **(ii)** taking the max-pooling over the sub-word tokens along the 768 dimensions, or **(iii)** taking the mean of the sub-word tokens.

3.2 NLP Tasks

To study information sharing in the aforementioned models we probe them for two well-studied NLP tasks: POS tagging and dependency labeling. For

POS tagging, we encode each sentence and probe its word-level representations from a given layer l using a shallow MLP. For dependency labeling, we follow a similar scheme and extract word-level representations for the child and the head of each dependency in the sentence. We concatenate them[1] to form a feature vector of size $2 \times H$, where H is the hidden state size of the encoder model, and feed it to a shallow MLP to predict their dependency relation. We include training details for the POS tagging and dependency labeling probes in Appendix A.

3.3 Information Sharing

We examine whether BLMs encode information about certain linguistic categories, such as POS tags (nouns, adjectives, etc.) or syntactic dependency relations (nominal subjects, objects, etc.) in common subspaces of their representation space.

To achieve this, we first localize the subspace of the model that corresponds to each of these categories by obtaining their mean vector representation, similarly to Libovický et al. (2020). For *POS tagging*, this corresponds to the centroid of each target POS tag t, which, given the word representations $\mathbf{v}$, is formally defined as:

$$\mathbf{u}_t^{(POS)} = \frac{1}{|V_t|} \sum_{\mathbf{v} \in V_t} \mathbf{v} \quad (1)$$

where V_t is the set of the representations of the words in our validation set that were predicted as tag t when probing the l-th layer of our encoder. For *dependency labeling*, we slightly modify the definition of the centroid, as each prediction depends on both the representation of the head $\mathbf{h}$ and the child $\mathbf{c}$ of the dependency relation. Here, the centroid of each target dependency label t is defined as:

$$\mathbf{u}_t^{(DEP)} = \frac{1}{|P_t|} \sum_{(\mathbf{h},\mathbf{c}) \in P_t} [\mathbf{h}\,;\mathbf{c}] \quad (2)$$

where P_t is the set of the (head, child) representation pairs that were predicted as dependency t when probing the l-th layer of our encoder, and $[\mathbf{h}\,;\mathbf{c}]$ is the concatenation of the two vectors.

After obtaining the mean vector $\mathbf{u}_t$ for each linguistic category t, we investigate information sharing with the cross-neutralizing method from Choenni and Shutova (2022). To cross-neutralize with $\mathbf{u}_t$, we encode each word in the sentence using our pre-trained encoder and subtract $\mathbf{u}_t$ from its representation. We pass the resulting representations through our probing network to obtain a new distribution over the linguistic categories and re-classify words, in case of POS tagging, or words with their heads, in case of dependency labeling.

To interpret our results, we group them in (`neutralizer`, `target`) pairs, meaning that we observe the change in classification accuracy for the class `target` when neutralized with the centroid of the class `neutralizer`. In the special case where the same linguistic class is both the target and the neutralizer, we refer to our method as *self-neutralizing*, and otherwise as *cross-neutralizing*. We argue that pairs of linguistic categories that result in substantial drops in performance after cross-neutralizing are jointly encoded, and hence information is shared among them.

3.4 Layer and Aggregation Selection

The choice of the layer l from which to extract the token embeddings is not arbitrary. The same non-triviality presents itself in the choice of aggregation function mentioned in Section 3.1. An ideal selection should capture the greatest amount of information for the words and the task at hand. For each aggregation function, we self-neutralize using embeddings from different layers and compare the relative drop in self-neutralizing accuracy.

We hypothesize that the configurations where the self-neutralizing percentage drop is the greatest are those closest to an ideal selection, as a higher decrease in performance signals in a higher amount of information being captured in the centroids. We also consider that this drop should be relative to a high baseline accuracy, to achieve more accurate centroid definitions. Therefore, we take the top quartile of our configurations in terms of baseline accuracy, rank them in descending order of relative drop in accuracy, and select the first to perform cross-neutralization.

4 Experiments

4.1 Datasets

We use the Universal Dependencies framework (Nivre et al., 2020) to train our probing models in POS tagging and dependency labeling, as it offers a consistent annotation style across a collection of treebanks over multiple languages. In

[1] We tested different concatenation configurations, such as including the mean vector or the absolute difference of the pair, but found the simpler approach to perform equally well.

particular, we use the GUM corpus (Zeldes, 2017) for English, which consists of 9,130 sentences with a total of 164,488 words, the VIT corpus (Delmonte et al., 2017) for Italian with 10,087 sentences and 279,723 total words, and the GDT corpus (Prokopidis and Papageorgiou, 2017) for Greek with 2,521 sentences and 63,441 total words.

All of the datasets contain token-level annotations with a total of 17 possible POS tags and 36 dependency relations. Additionally, for dependency relations, a unique head corresponds to each word, with the exception of the `root` relation, which has no head, and hence it is omitted. A detailed description of the pre-processing pipeline for all three treebanks can be found in Appendix B.

4.2 Baseline Models

We use RoBERTa and XLM-R[2]as our encoders, and we keep their weights frozen while training our probe classifiers on the aforementioned datasets for both tasks. We try various combinations of layers to probe ($l \in \{1, 3, 6, 9, 12\}$) and sub-word token aggregations (`first`, `mean`, `max-pooling`), and use the one that achieves a high baseline accuracy while resulting in the largest accuracy drop when self-neutralizing, as described in Section 3.4. We display the results we obtained per configuration in Appendix C. We report the classification accuracy for the best self-neutralizing combination for all our baseline models in Table 1, and we observe that we are in line with the current literature (Straka et al., 2019). As we only employ a pre-trained encoder without fine-tuning on our corpus, we expect to see a slightly lower accuracy compared to state-of-the-art approaches in each task.

	RoBERTa	XLM-R		
	en_gum	en_gum	it_vit	el_gdt
POS	95.6%	95.5%	97.4%	97.9%
DEP	90.9%	91.4%	93.9%	94.8%

Table 1: Classification accuracy for part-of-speech tagging and dependency labeling for our baseline models.

4.3 Cross-Neutralizing in English

To perform cross-neutralization, we subtract the centroid of a given label (`neutralizer`) from the embeddings of our encoder and compare the accuracy drop over the labels (`targets`). Although it is reasonable to expect that subtracting a fixed vector from our embeddings might always cause a performance drop, we experimented with subtracting random vectors and noticed that the performance remained unchanged. Thus, we argue that considerable changes in accuracy for a (`neutralizer`, `target`) pair suggest that the model learns some information about these two classes jointly.

While the labels for POS tagging are trivially bound to a single word, dependency labels refer to pairs of child and head words, leading to the definition of centroids being less straightforward. We experiment with considering only the embedding of the child as the centroid, versus a concatenation of both the child and the head. The latter results in more consistent self-neutralization across the task labels, which we assume suggests a better choice for centroid definition. Therefore, we use this concatenation configuration in all further dependency labeling cross-neutralization analyses.

4.4 Monolingual vs Multilingual Models

We also extend our experiments to the multilingual setting using XLM-R as the encoder and evaluating cross-neutralization in three different languages. This lets us both verify the consistency of our method across models, as we can directly compare the results between RoBERTa and XLM-R in English, but also across languages, as we can compare the effects of cross-neutralization on English, Italian, and Greek. By doing so, we can also assess the effect of cross-neutralization at a language-family level, as these are three Indo-European languages.

4.5 Cross-Lingual Cross-Neutralizing

In the multilingual setting, we test whether our findings are consistent across languages. A different direction worth exploring is whether there is a joint encoding of information between two related linguistic categories from two different languages. We test this hypothesis by cross-neutralizing every linguistic entity in one language with another entity from another language, e.g. all Italian POS tags with English nouns.

4.6 Cross-Task Cross-Neutralizing

To study joint sharing of linguistic information *across tasks*, we additionally subtract the POS centroids from the child embeddings in child-parent concatenations for dependency labelling. In this way we effectively cross-neutralise across tasks,

[2]The pre-trained models are available on Hugging Face: https://huggingface.co/roberta-base https://huggingface.co/xlm-roberta-base

with POS neutralizers and dependency label targets. Similarly, we subtract the child portion of the dependency label centroids from the POS embeddings for POS tagging, to test whether any sort of joint learning happens in both directions.

5 Results & Analysis

5.1 Monolingual Cross-Neutralization

The decrease in accuracy when cross-neutralizing English POS tags and dependency labels is presented in Figure 1. Overall, we observe a pattern that verifies our hypothesis that related classes are jointly encoded in RoBERTa, for both the POS and DEP tasks. Subtracting the centroids associated with specific classes leads to a decrease in accuracy for other classes, indicating that information is shared across them. Notably, this cross-neutralizing phenomenon is not necessarily symmetric; in several instances, the accuracy change for a label y when using neutralizer x is not similar in magnitude to the accuracy change for x when y is the neutralizer. This may be due to the structure of language, which can be represented as a directed graph in which linguistic entities may modify others, but may not themselves be modified by their object. For example, adjectives modify nouns, and yet nouns do not modify adjectives. Further work could further clarify why certain entities cross-neutralize symmetrically and why others do not. In the rest of this subsection, we kept a subset of the linguistic categories both for the POS and DEP tasks, to highlight some task-specific patterns that emerge from our cross-neutralizing experiments. The figures containing the cross-neutralizing results between all classes can be found in Appendix D.

POS Tagging The top part of Figure 1 shows the impact of cross-neutralization on a selection of POS tags. The shown POS tags have been selected to ensure a fair distribution between open and closed-class words. We observe that while there is evidence of linguistically related units being jointly encoded, this phenomenon is not ubiquitous. For instance, we see that the auxiliaries (AUX) and verbs (VERB) are jointly encoded, as indicated by the relative decrease of 53% in VERB classification accuracy. This aligns with our hypothesis since auxiliaries can be considered function words acting on verbs. However, the same is not observed when neutralizing nouns (NOUN)

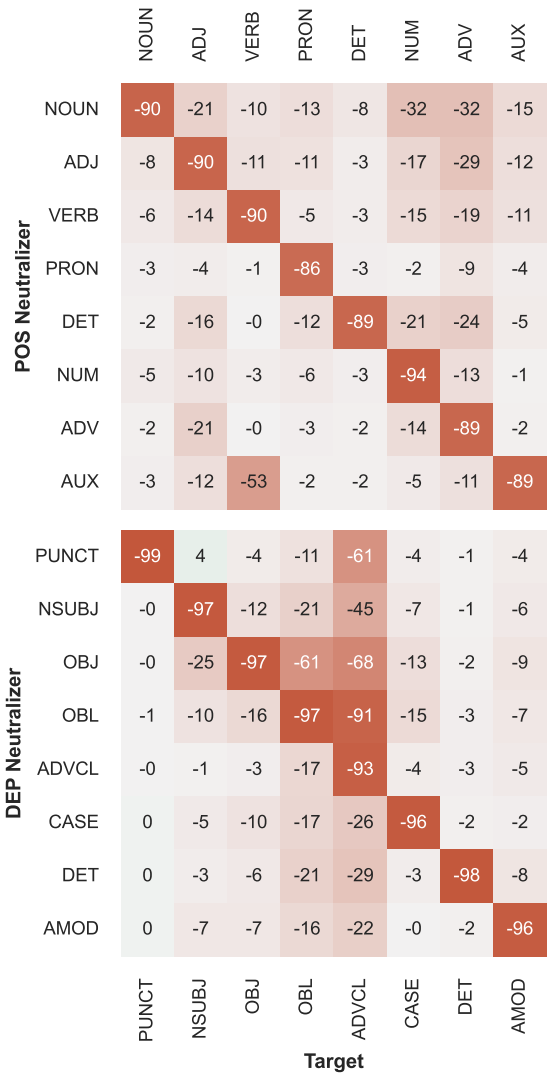


Figure 1: Relative change in accuracy when cross-neutralizing 8 sampled POS (top) and DEP (bottom) tags using embeddings from RoBERTa in English.

with determiner (DET) centroids, where the relative percentage decrease is only 2%, suggesting that these tags are learned independently, despite determiners acting as a lexical modifier for nouns. Finally, we note how certain tags are particularly susceptible to (such as numerals, NUM) or adept at (such as NOUN) cross-neutralization, regardless of their relationship to the neutralizer or target.

Dependency Labeling The lower half of Figure 1 also shows the impact of cross-neutralization on a selection of dependency labels. These labels were selected to highlight the major trends in joint encoding within this task. We observe that information sharing is most pronounced when dependency classes are (**1**) closely related by definition, or (**2**)

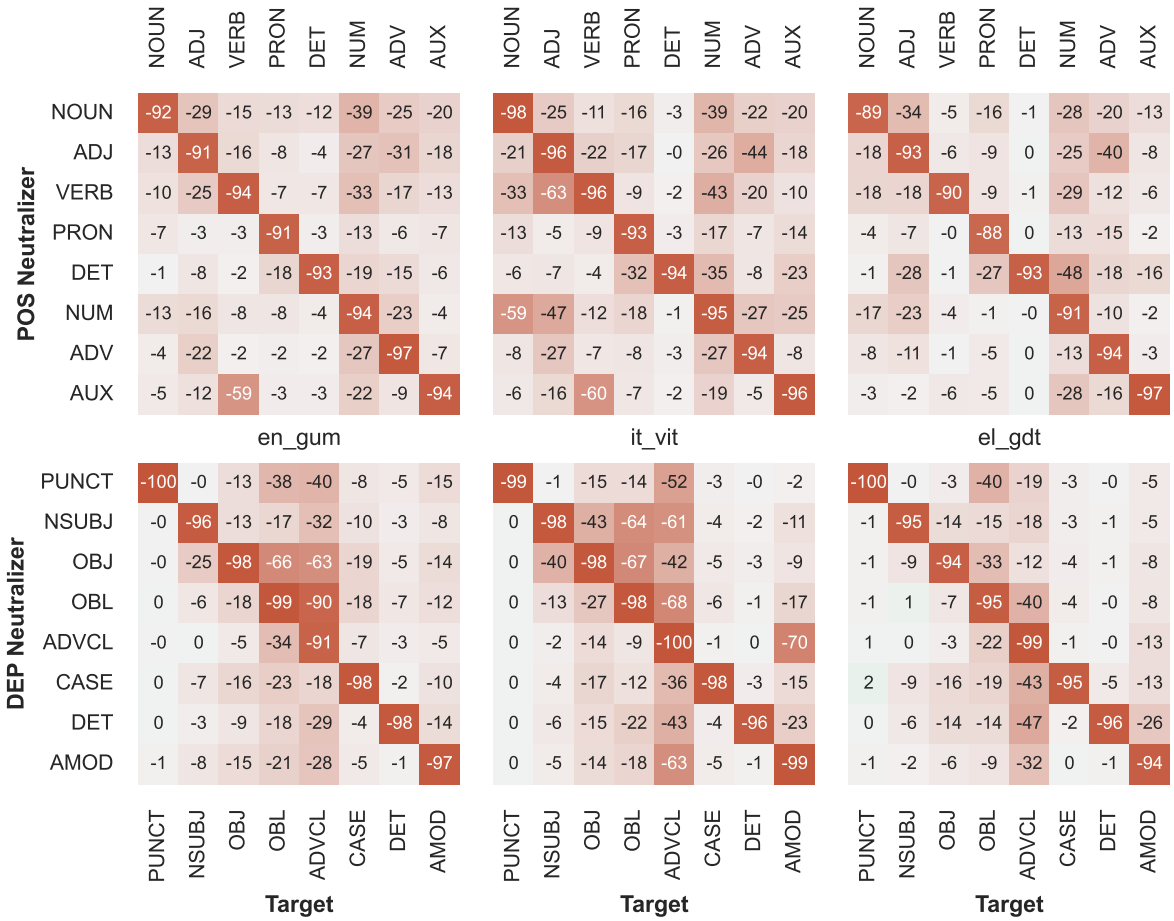


Figure 2: Relative change in accuracy when cross-neutralizing 8 sampled POS (top) and DEP (bottom) tags using embeddings from XLM-R in English (left), Italian (center) and Greek (right).

hierarchically connected. Both of these aspects support our hypothesis of joint encoding of linguistically related dependencies and elevate important details of linguistic structure. As an example of (**1**), one can consider the classes NSUBJ, OBJ, and OBL. Each is jointly encoded with the others, as they help define the basic syntactic structure of the sentence in relation to the predicate.

On the other hand, the results of PUNCT and ADVCL lend support to (**2**). ADVCL is a clausal dependency, and is therefore significantly affected when removing information that is more related to word components of clauses such as OBJ, AMOD, CASE, and PUNCT. This latter dependency is quite noteworthy; it is a strong neutralizer across several classes, and yet is not affected by the removal of any other centroid. Punctuation plays a strong role in defining the syntactic tree of a sentence and therefore removing that information will distort its structure and thus the predicted dependency labels. For example, we can consider the striking difference between the sentences "let's eat, grandma" and "let's eat grandma". On the other hand, one can read a sentence in a foreign language and still be able to point out the punctuation, indicating that, even with limited information, punctuation information is orthogonal to other linguistic features.

5.2 Multilingual Cross-Neutralization

In this section, we repeat the evaluation of Section 5.1 using embeddings from XLM-R trained on English, Italian and Greek datasets and report results in Figure 2. To facilitate comparison both across languages and with the results acquired from RoBERTa, we make the same selection of POS tags and dependency labels and present the full results in Appendix D.

Consistency Across Models When comparing RoBERTa and XLM-R on the English dataset, we observe similar patterns across linguistic categories for both POS tagging and dependency labeling. For example, numerals seem to be jointly encoded with

nouns and auxiliaries. This supports the idea that even in a multilingual training setting, the linguistic entities of each language are encoded similarly to a monolingual model.

Consistency Across Languages Similar patterns can also be observed when comparing across languages in XLM-R. For instance, we see that nouns consistently neutralize numerators by a significant factor. Likewise, in the dependency labeling setting, we note that the lack of neutralization of determiners (DET) is consistent across all three languages. This suggests that the tags involved in the cross-neutralization pairs where this observation holds are language-agnostic. On the same note, language-specific behavior is also evident. For instance, VERB POS tags cross-neutralize adjectives (ADJ) more prominently in Italian than in the other two languages. Similarly, while the neutralization of ADVCL by NSUBJ in English dependency labeling is significant, the effect is less pronounced in the other two languages. The results suggest that XLM-R discovers and leverages language-agnostic information when possible, while also learning language-specific information when necessary.

5.3 Cross-Lingual Cross-Neutralization

We extend the experiments of the previous section by cross-neutralizing using linguistic categories from different languages. Figure 3 shows the results of cross-neutralizing for Italian POS tagging and dependency labeling when using the corresponding English centroids from each linguistic category. We notice a substantial drop in performance across the diagonals of these graphs, which correspond to the self-neutralizing scenario. This indicates the presence of information sharing between languages but is limited to the same linguistic feature. We also observe a correlation between different features of different languages, such as Italian adverbs and English adjectives, or Italian nominal subjects and English objects. This behavior highlights the tremendous learning capacity of these models, as information can be shared between the representations for different languages.

Finally, we repeat the experiment by using English neutralizers on the Greek corpus, and we observed similar but milder trends. We hypothesize that this is due to English being linguistically closer to Italian than Greek, which leads to a higher degree of information sharing between these two languages. For more detailed plots on cross-lingual cross-neutralization, we refer the reader to Appendix D.

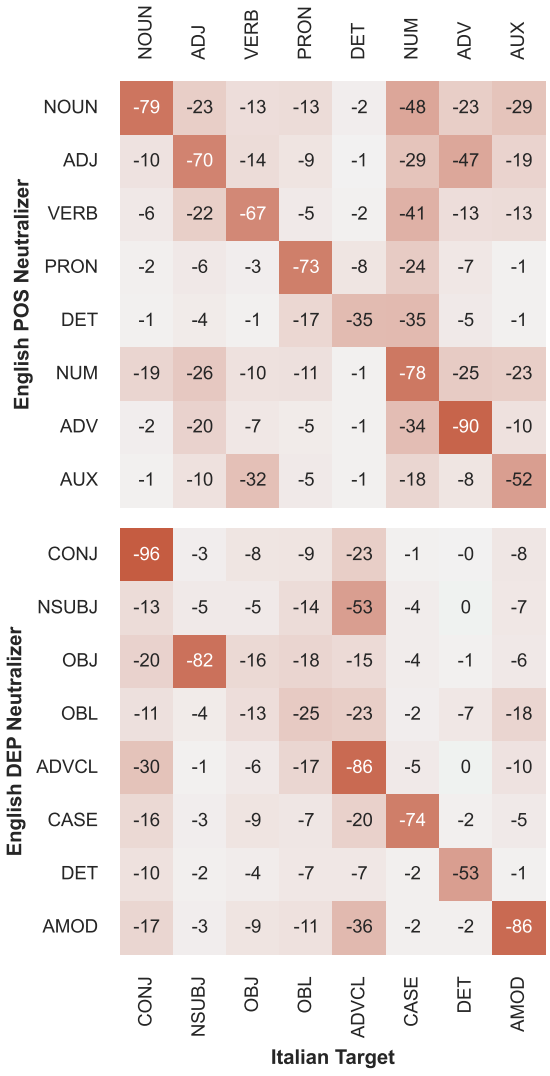


Figure 3: Relative change in accuracy when cross-neutralizing Italian with 8 sampled POS (top) and DEP (bottom) English tags using embeddings from XLM-R.

6 Conclusion

In this work, we studied the joint encoding of linguistic entities in BLMs. We approached the problem by cross-neutralizing pairs of entities in two well-studied tasks: POS tagging and dependency labeling. We first evaluated our methodology in a simple monolingual scenario using RoBERTa and found joint encoding of similar linguistic categories for both tasks. We extended our evaluation to include XLM-R and found our hypothesis to hold both across languages and to be persistent across two different BERT-based models. Finally, we noticed that related linguistic categories are jointly

encoded across different languages in XLM-R.

Our findings are consistent with recent work on information sharing in BLMs and shed light on the flow of information within them. They also give credit to the efficacy of multilingual models over the use of multiple monolingual ones. We hope that our work will aid the exploration and interpretation of BLMs and encourage further research on their information-sharing mechanisms. As a future direction, we would like to extend our experimental setting to include more downstream NLP tasks and evaluate our findings in more languages.

References

Terra Blevins, Omer Levy, and Luke Zettlemoyer. 2018. Deep RNNs Encode Soft Hierarchical Syntax. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19, Melbourne, Australia. Association for Computational Linguistics.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding Universal Grammatical Relations in Multilingual BERT. *ACL*.

Rochelle Choenni and Ekaterina Shutova. 2020. What does it mean to be language-agnostic? Probing multilingual sentence encoders for typological properties. *arXiv:2009.12862 [cs]*. ArXiv: 2009.12862.

Rochelle Choenni and Ekaterina Shutova. 2022. Investigating Language Relationships in Multilingual Sentence Encoders Through the Lens of Linguistic Typology. *Computational Linguistics*, pages 1–38.

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. 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.

Rodolfo Delmonte, Antonella Bristot, and Sara Tonelli. 2017. VIT – Venice Italian Treebank: Syntactic and Quantitative Features.

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.

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.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the Language Neutrality of Pre-trained Multilingual Representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*. ArXiv: 1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Prokopis Prokopidis and Haris Papageorgiou. 2017. Universal Dependencies for Greek. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 102–106, Gothenburg, Sweden. Association for Computational Linguistics.

Vinit Ravishankar, Memduh Gökırmak, Lilja Øvrelid, and Erik Velldal. 2019. Multilingual Probing of Deep Pre-Trained Contextual Encoders. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 37–47, Turku, Finland. Linköping University Electronic Press.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Milan Straka, Jana Straková, and Jan Hajič. 2019. Evaluating Contextualized Embeddings on 54 Languages in POS Tagging, Lemmatization and Dependency Parsing. Technical Report arXiv:1908.07448, arXiv. ArXiv:1908.07448 [cs] version: 1 type: article.

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.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2018. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Shijie Wu and Mark Dredze. 2020. Are All Languages Created Equal in Multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Gözde Gül Şahin, Clara Vania, Ilia Kuznetsov, and Iryna Gurevych. 2020. LINSPECTOR: Multilingual Probing Tasks for Word Representations. *Computational Linguistics*, 46(2):335–385.

A Training details for the probes

For both tasks, we opt for a two-layer MLP probe with a tanh activation. During training, we keep the encoder's parameters frozen and train the linear probes using the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 10^{-3} and weight decay of 10^{-2}, and employ early stopping according to the validation set accuracy.

B Dataset pre-processing

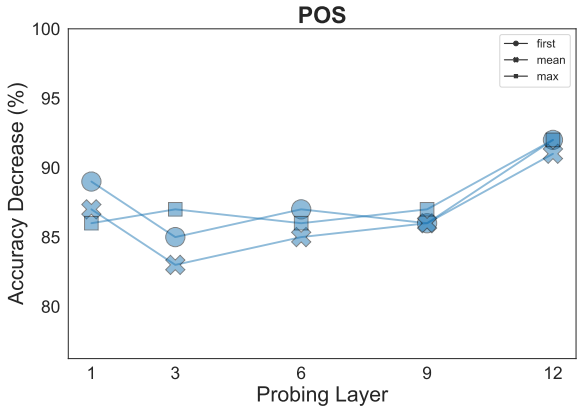
The sentences in the corpora from the Universal Dependencies framework are already tokenized to the word level and stored as lists of words in the `tokens` field. However, since we use sub-word tokenizers, namely the Byte Pair Encoding (Sennrich et al., 2016) and SentencePiece (Kudo and Richardson, 2018) tokenizer, we further split the words into their sub-word tokens. Depending on the task, we also include either the `upos` field, which is a list of integers corresponding to one of the 17 universal POS tags available, or the `head` and `deprel` fields which contain the head and one of the 36 dependence relations for dependency labeling[3]. It should be noted that we only keep the language-independent relations, as some of them appear only with a language-specific modifier, and including them would make comparisons across languages less straightforward.

Furthermore, upon inspecting the datasets we observed that the annotators had split contractions into their parts and included them next to the original contraction for the Italian and Greek corpora. However, ground-truth labels were only provided for the sub-words, with the compound words annotated as a special class "_". Hence, we filtered out the compound words from these datasets and retained their sub-parts. In addition to that, for dependency labeling, we ignored words with the root dependency label, as they have no head and their prediction is trivial.

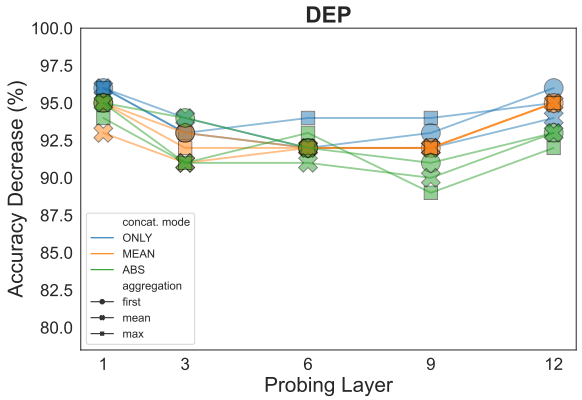
C Choosing a configuration for probing

Figures 4 and 5 showcase the decrease in accuracy when self-neutralizing in the POS tagging/dependency labeling task for RoBERTa and XLM-R accordingly. We report the best configurations for each model, language, and task combination in Table 2, and correspond to the ones we used for our cross-neutralization experiments.

[3] A full list of POS tags and dependency relations can be found on the Universal Dependencies website.



(a) Accuracy decrease for **POS tagging** when using different WordPiece aggregations.



(b) Accuracy decrease for **dependency labeling** when using different WordPiece aggregations and child-head concatenation configurations.

Figure 4: Decrease in performance for RoBERTa when **self-neutralizing** in the POS tagging (top) and dependency labeling (bottom) tasks using embeddings extracted from different layers and setup configurations.

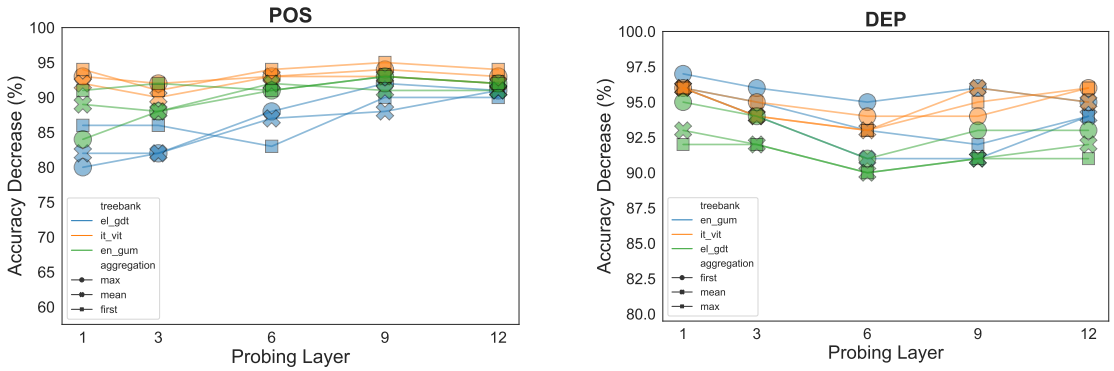
Figure 6 shows the results of cross-neutralizing in the dependency labeling task when the centroids are based exclusively on the child word's representation. As mentioned in Section 5.1, this results in less prominent self-neutralizing. We hypothesize that the inclusion of information about the head within the centroid is likely leveraged by the classifier to retain good accuracy.

D Detailed cross-neutralizing results

We display the complete results for our cross-neutralization experiments in Figures 7 – 26. More specifically, Figures 7 and 8 correspond to the monolingual setting with RoBERTa on English, Figures 9 – 14 to the multilingual with XLM-R on English, Italian and Greek. Figures 15 – 26 show the cross-lingual setting for every possible pairing of our three languages with XLM-R.

	POS		DEP		
Encoder / treebank	**Layer**	**WP Aggregation**	**Layer**	**WP Aggregation**	**Concatenation**
RoBERTa / en_gum	3	max	3	mean	[child ; head]
XLM-R / en_gum	9	max	9	first	[child ; head]
XLM-R / it_vit	9	first	9	mean	[child ; head]
XLM-R / el_gdt	12	mean	9	mean	[child ; head]

Table 2: The best probing configuration for each encoder model (RoBERTa & XLM-R), task (POS & DEP) and language (English, Italian & Greek) combination, chosen as outlined in Section 3.4. Ultimately, these are the configurations that we use for all of our cross-neutralization experiments.



(a) POS tagging Accuracy Decrease using different WordPiece aggregations.

(b) Dependency labeling Accuracy Decrease using different WordPiece aggregations.

Figure 5: Decrease in performance for XLM-R when **self-neutralizing** in the POS tagging (left) and dependency labeling (right) tasks using embeddings extracted from different layers and setup configurations, for each of for English (en_gum), Italian (it_vit) and Greek (el_gdt) treebanks. For dependency labeling, we use the best child-head concatenation mode (ONLY) based on the results we aquired with RoBERTa, as shown in Figure 4b.

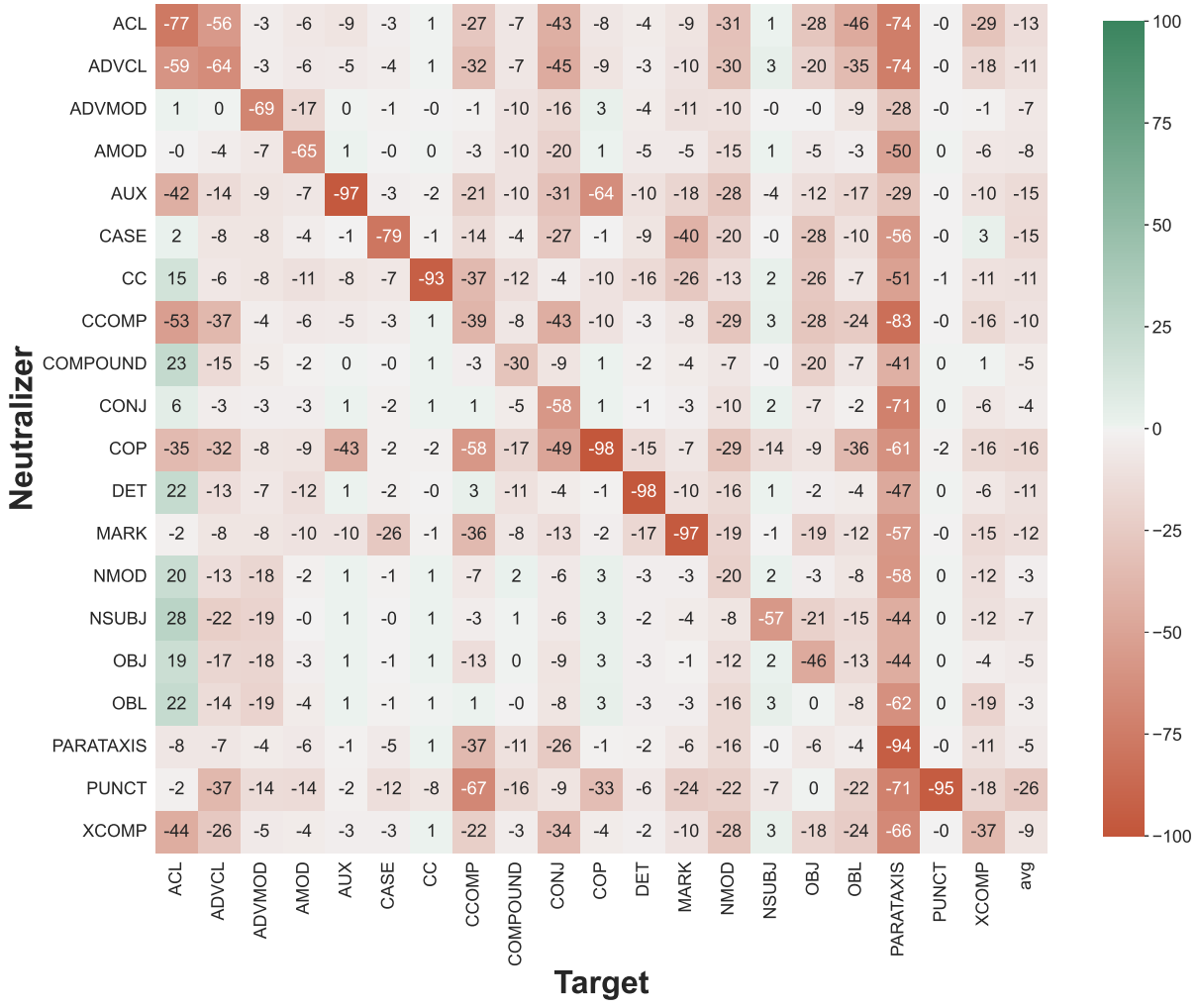


Figure 6: Relative change in accuracy when cross-neutralizing DEP tags using embeddings from RoBERTa on en_gum, in a setup configuration where we **only use the child embeddings** for the centroid calculation.

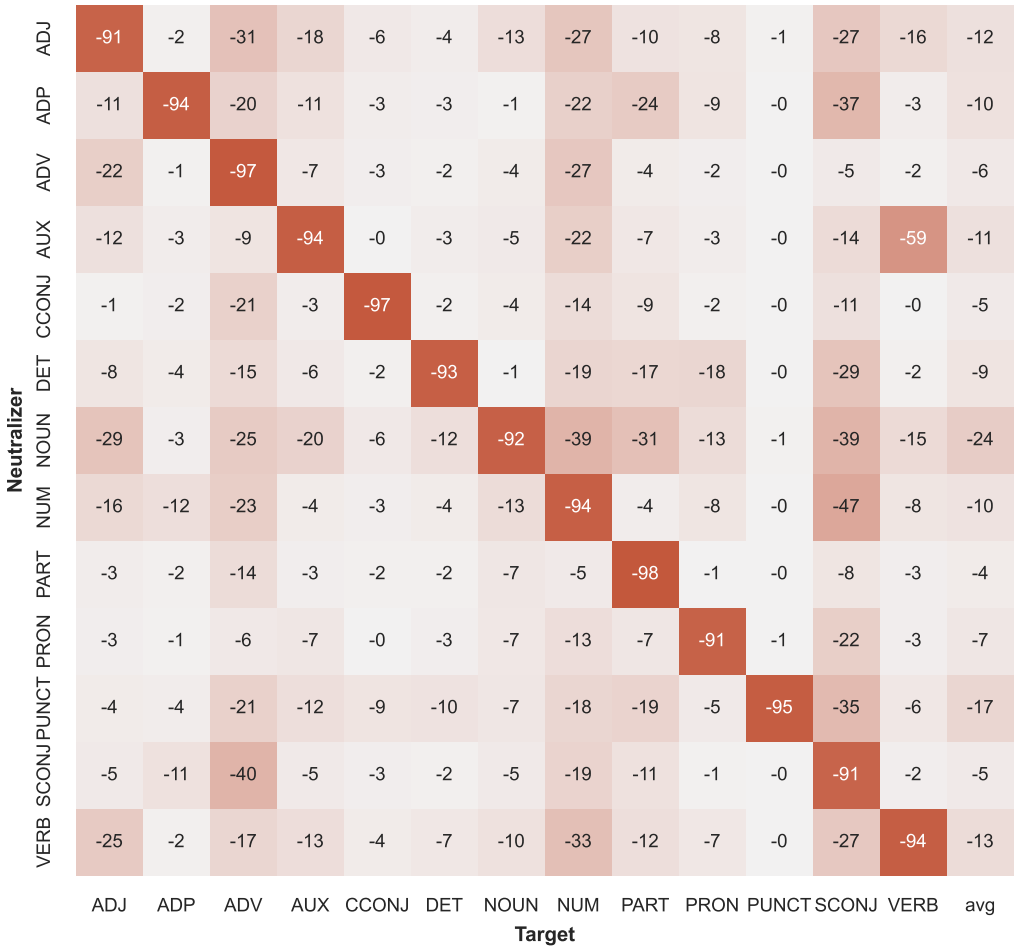


Figure 7: Relative change in accuracy when cross-neutralizing POS tags using RoBERTa embeddings on en_gum.

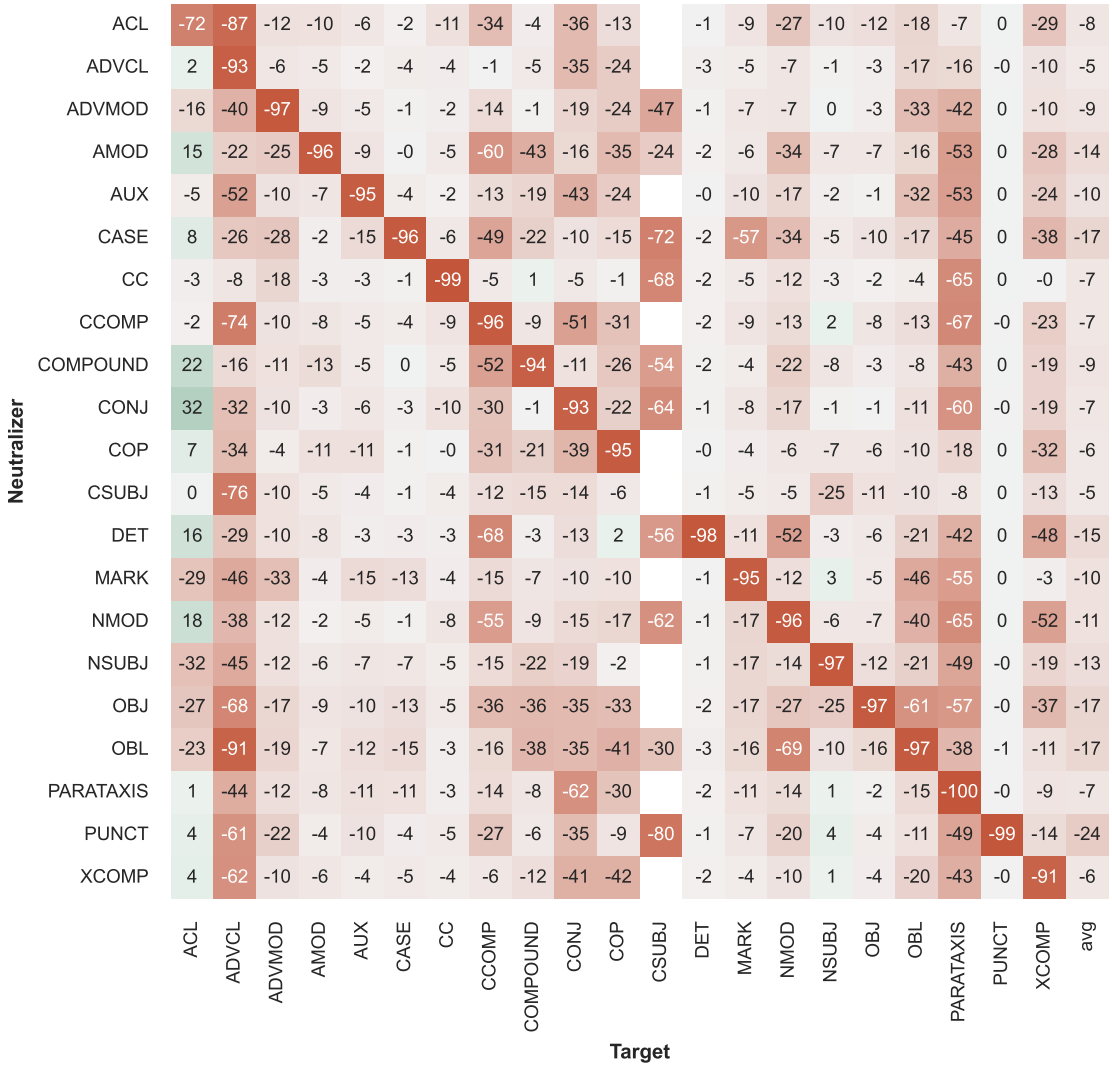


Figure 8: Relative change in accuracy when cross-neutralizing DEP tags using RoBERTa embeddings on en_gum. Note the relative increase in performance for the PARATAXIS relation; this is likely due to noise, as PARATAXIS labels make up less than 1% of the dataset.

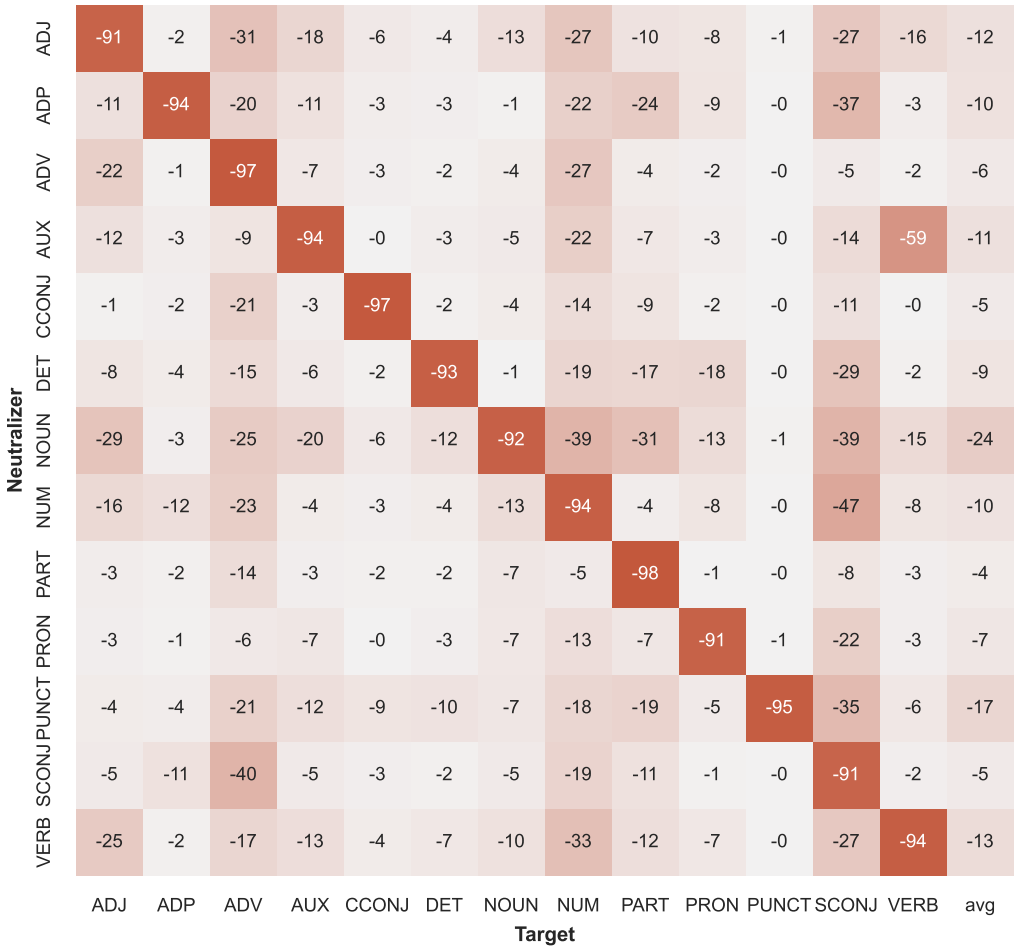


Figure 9: Relative change in accuracy when cross-neutralizing POS tags using XLM-R embeddings on en_gum.

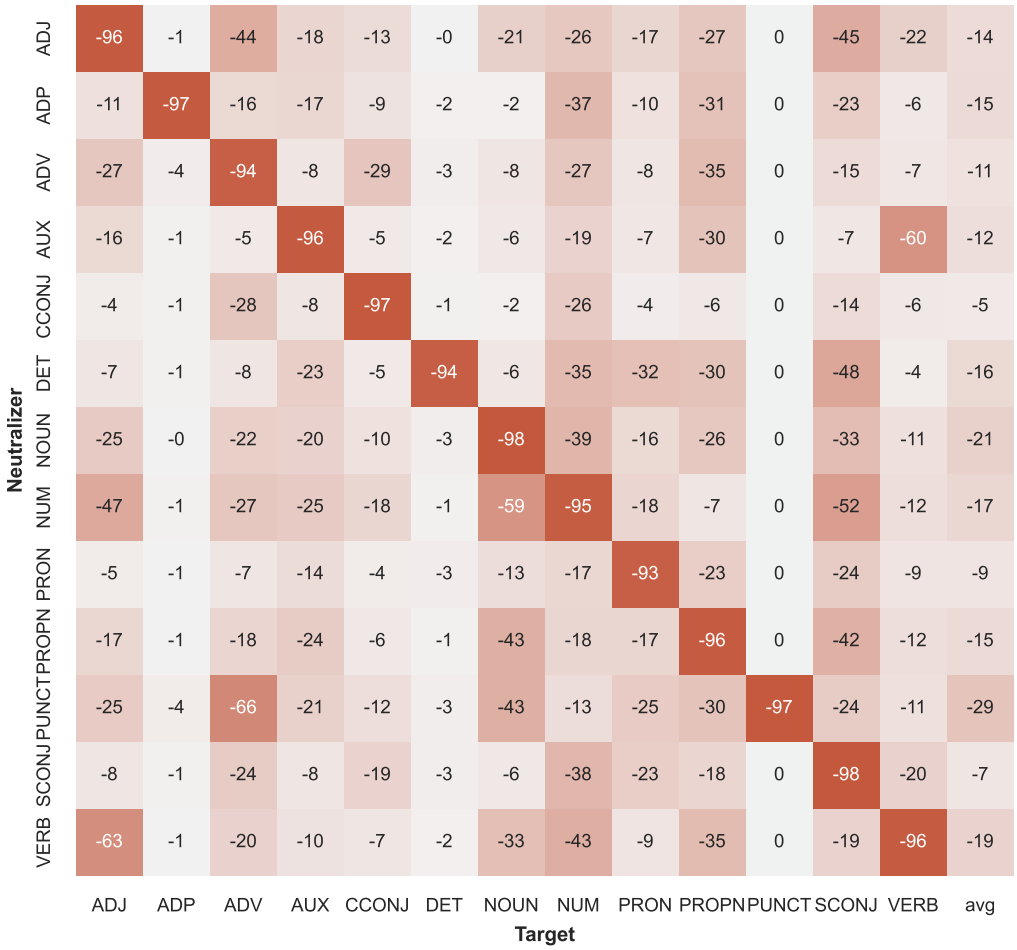


Figure 10: Relative change in accuracy when cross-neutralizing POS tags using XLM-R embeddings on it_vit.

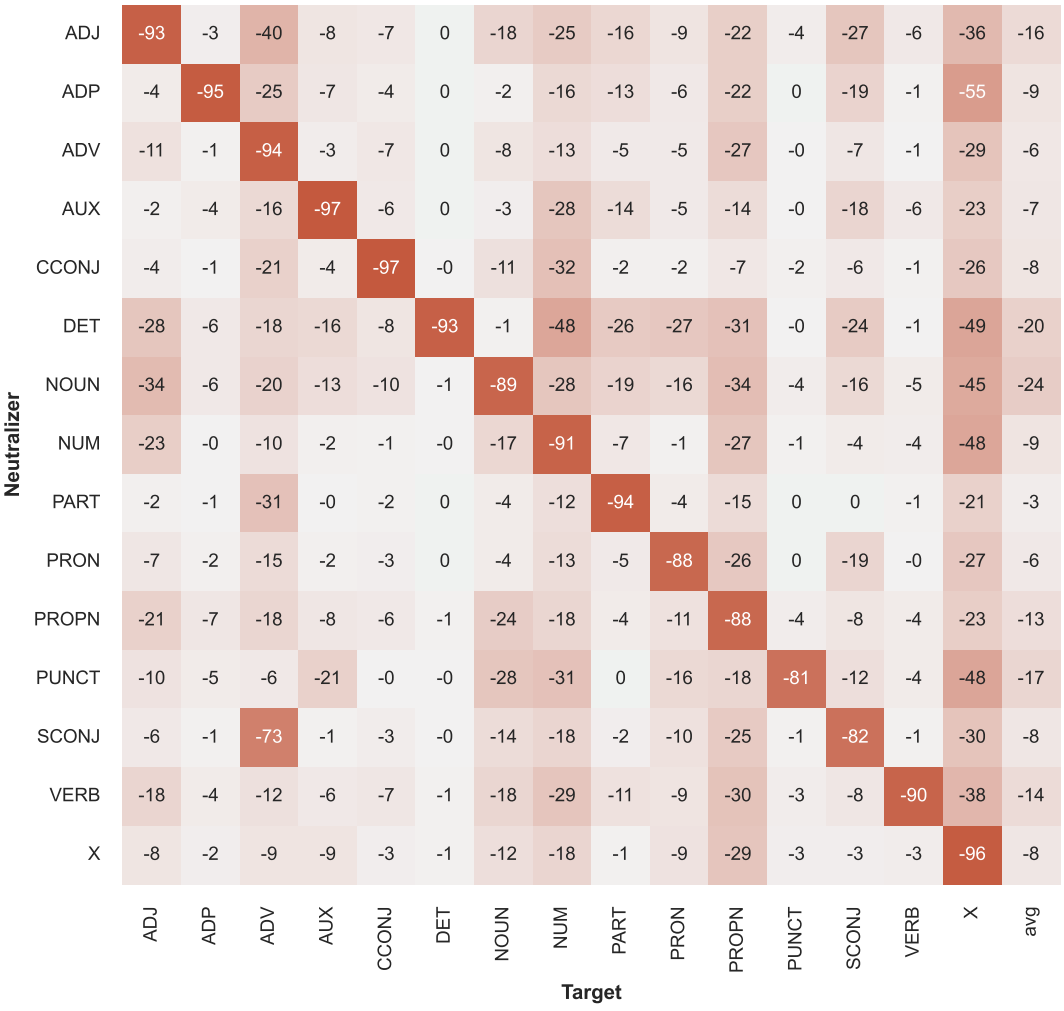


Figure 11: Relative change in accuracy when cross-neutralizing POS tags using XLM-R embeddings on el_gdt.

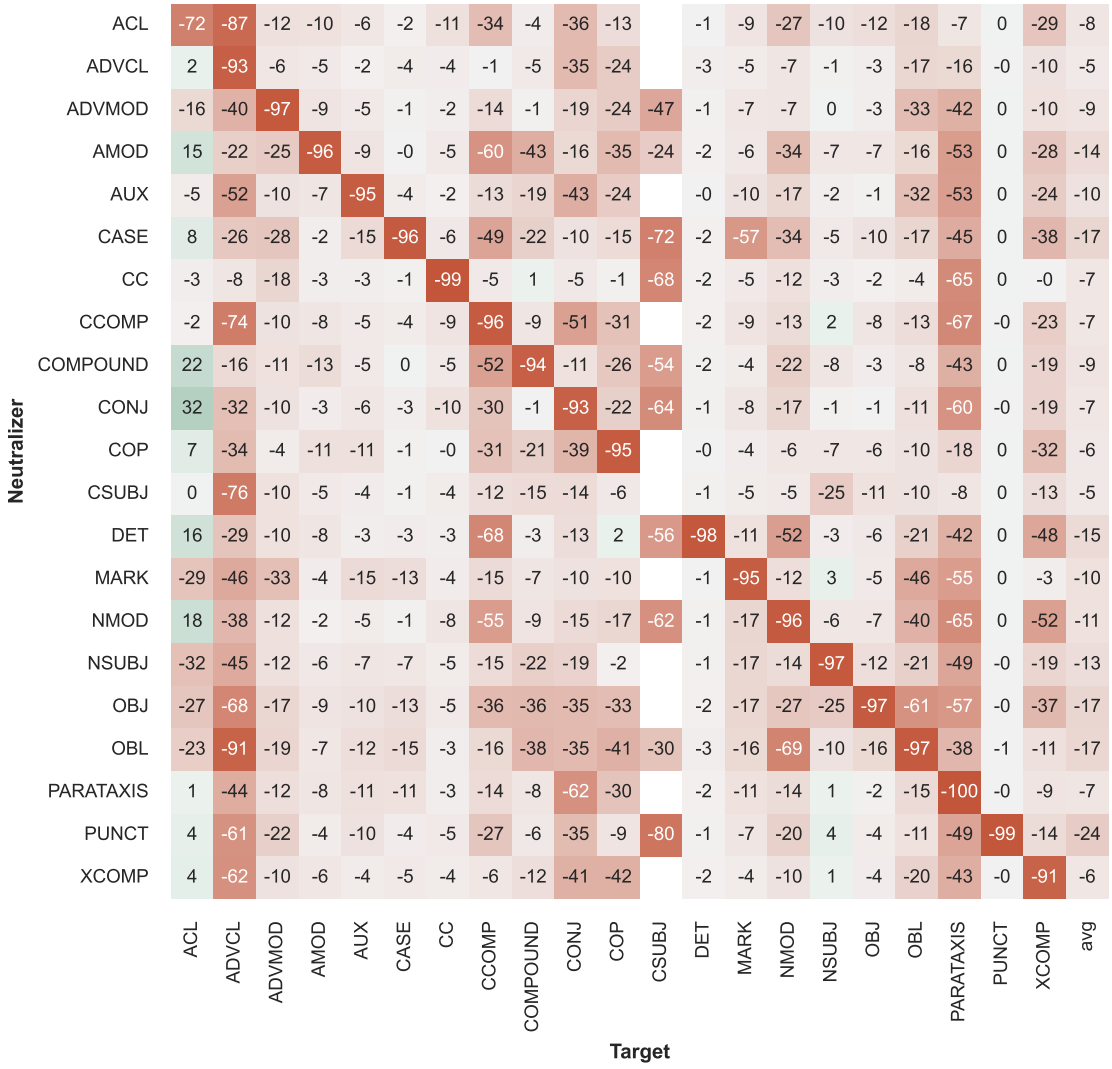


Figure 12: Relative change in accuracy when cross-neutralizing DEP tags using XLM-R embeddings on en_gum.

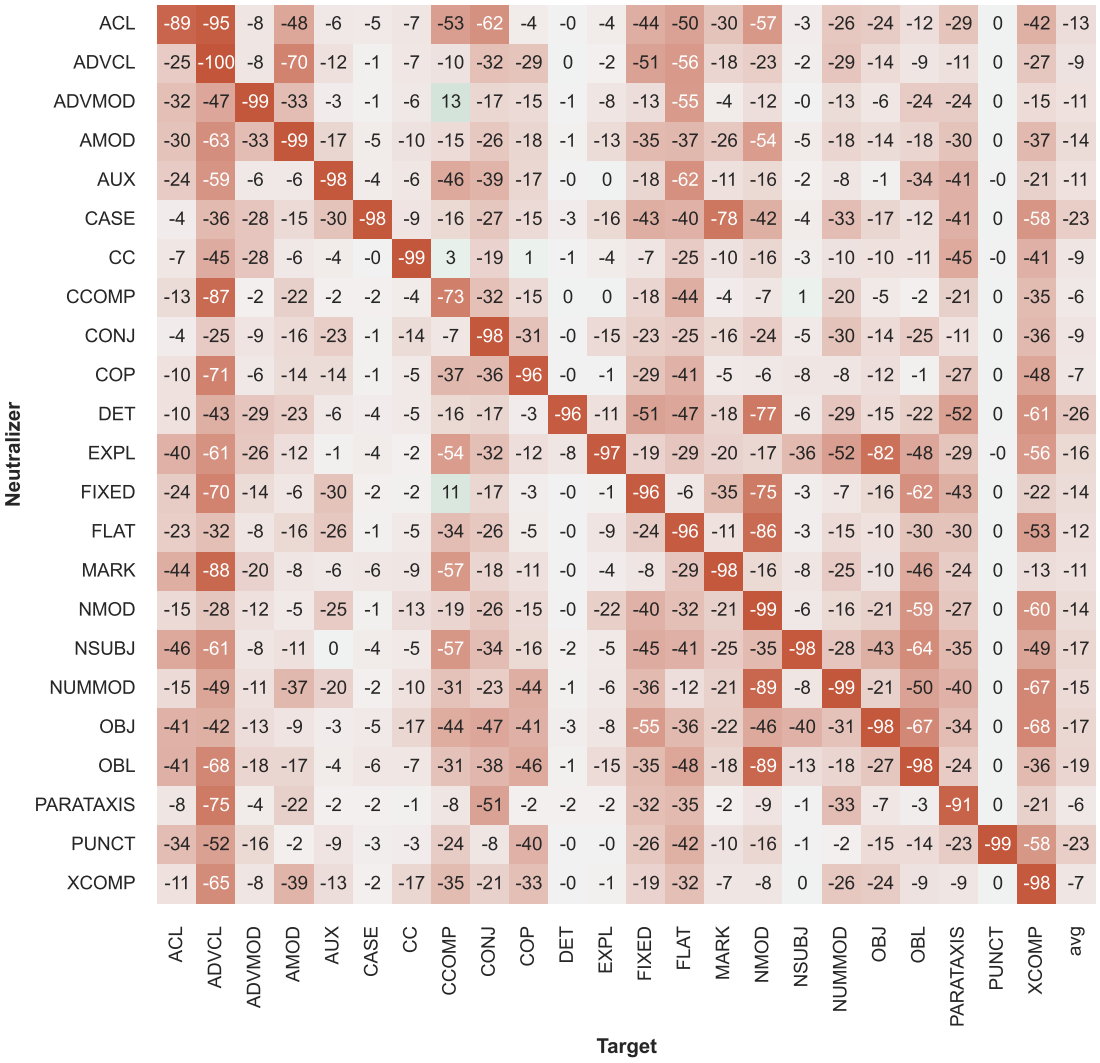


Figure 13: Relative change in accuracy when cross-neutralizing DEP tags using embeddings from XLM-R on it_vit.

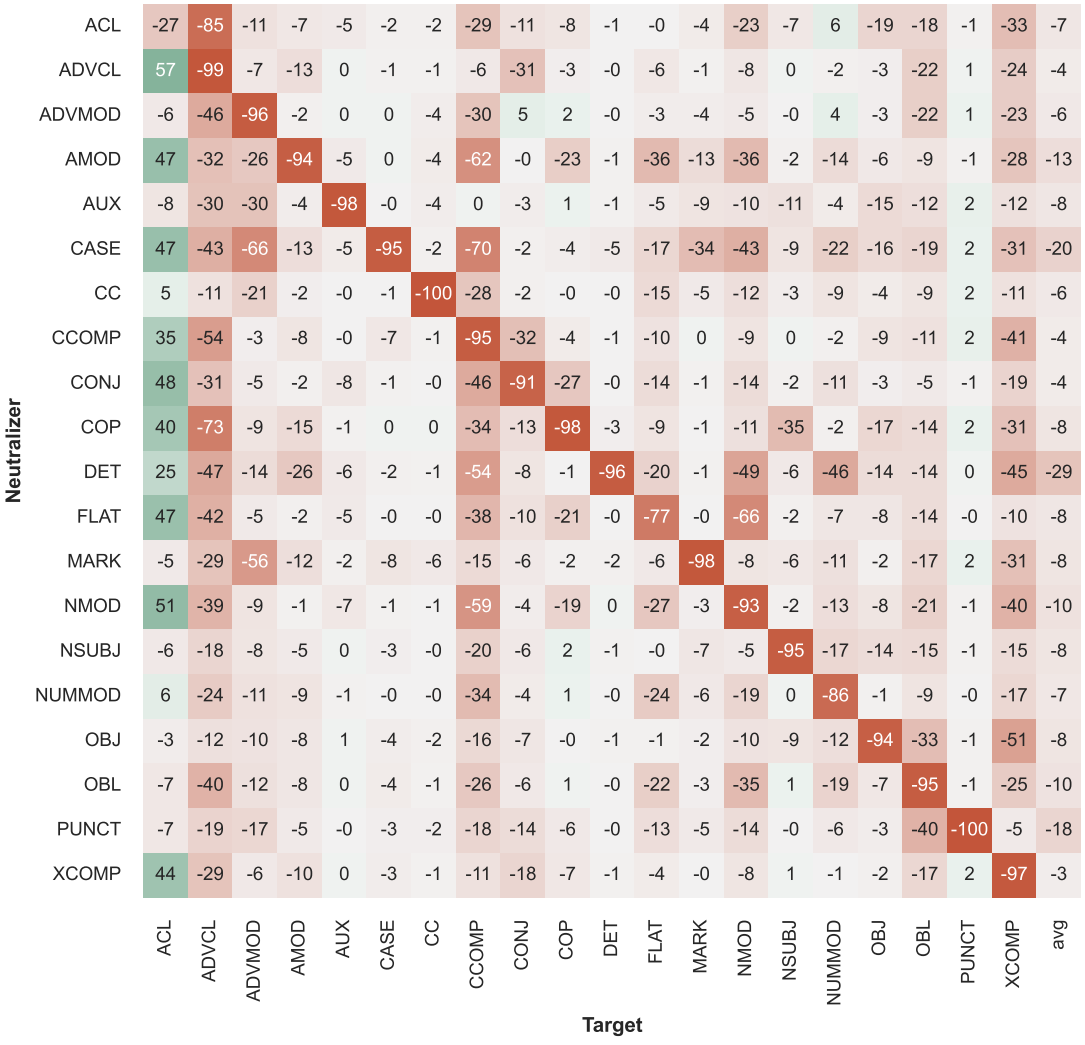


Figure 14: Relative change in accuracy when cross-neutralizing DEP tags using XLM-R embeddings on el_gdt.

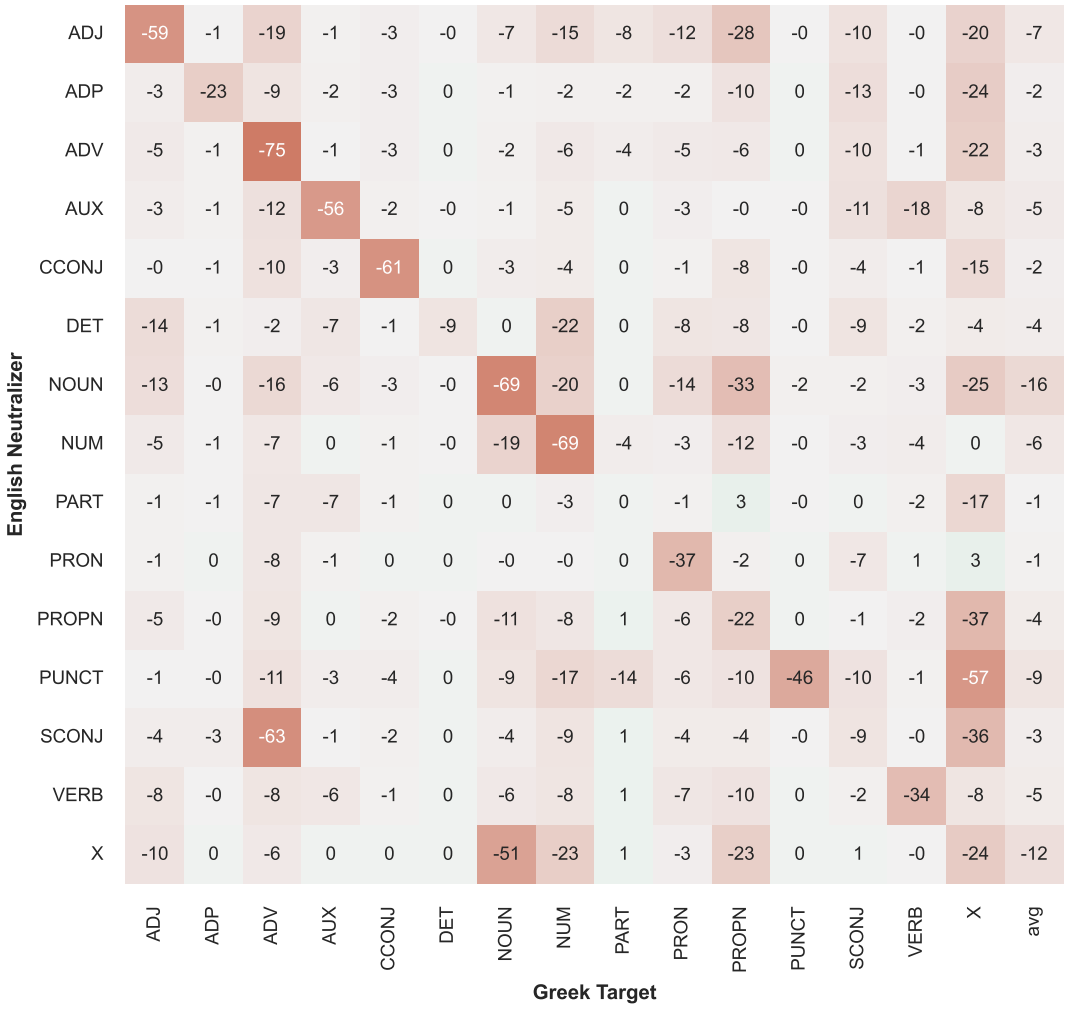


Figure 15: Relative change in accuracy when cross-neutralizing Greek POS tags using English XLM-R embeddings.

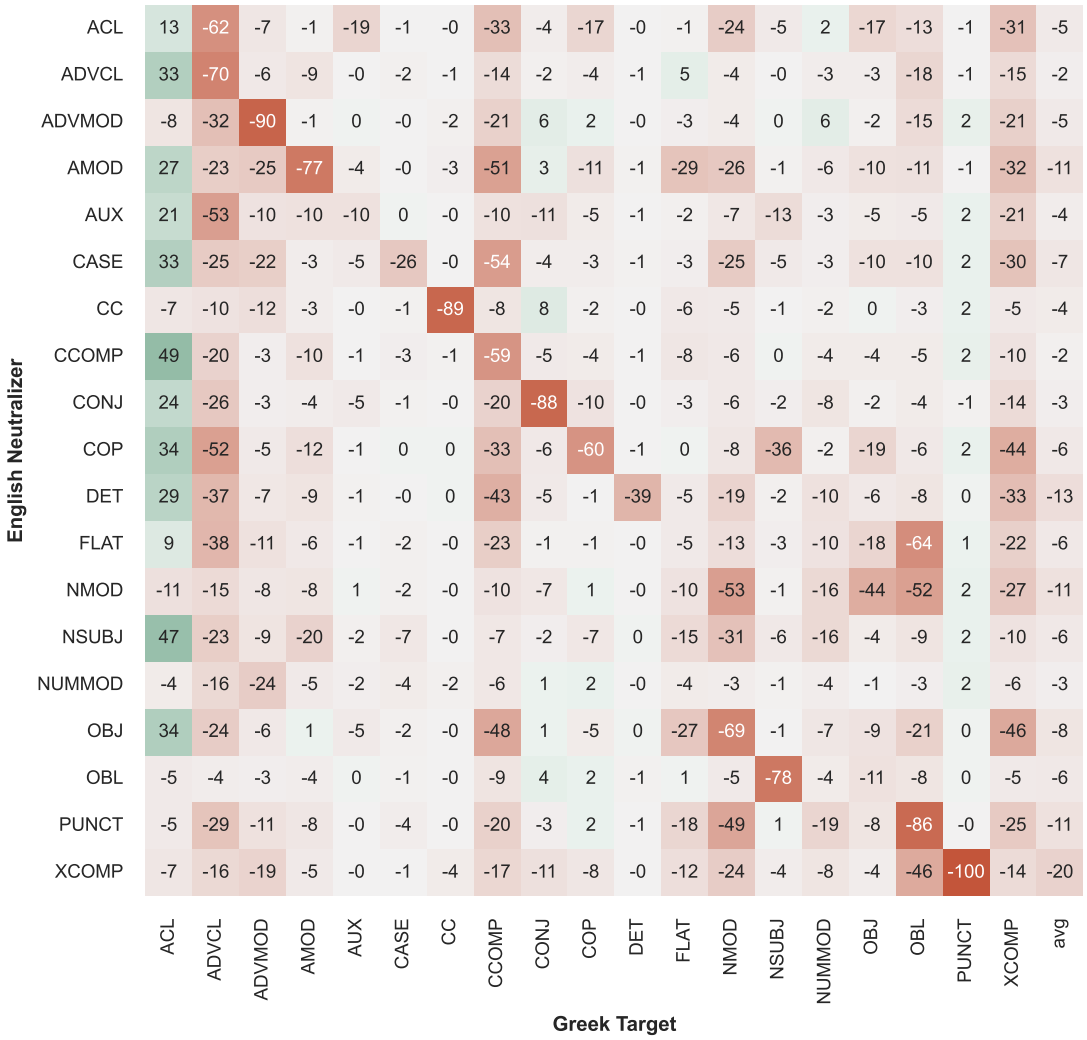


Figure 16: Relative change in accuracy when cross-neutralizing Greek DEP tags using English XLM-R embeddings.

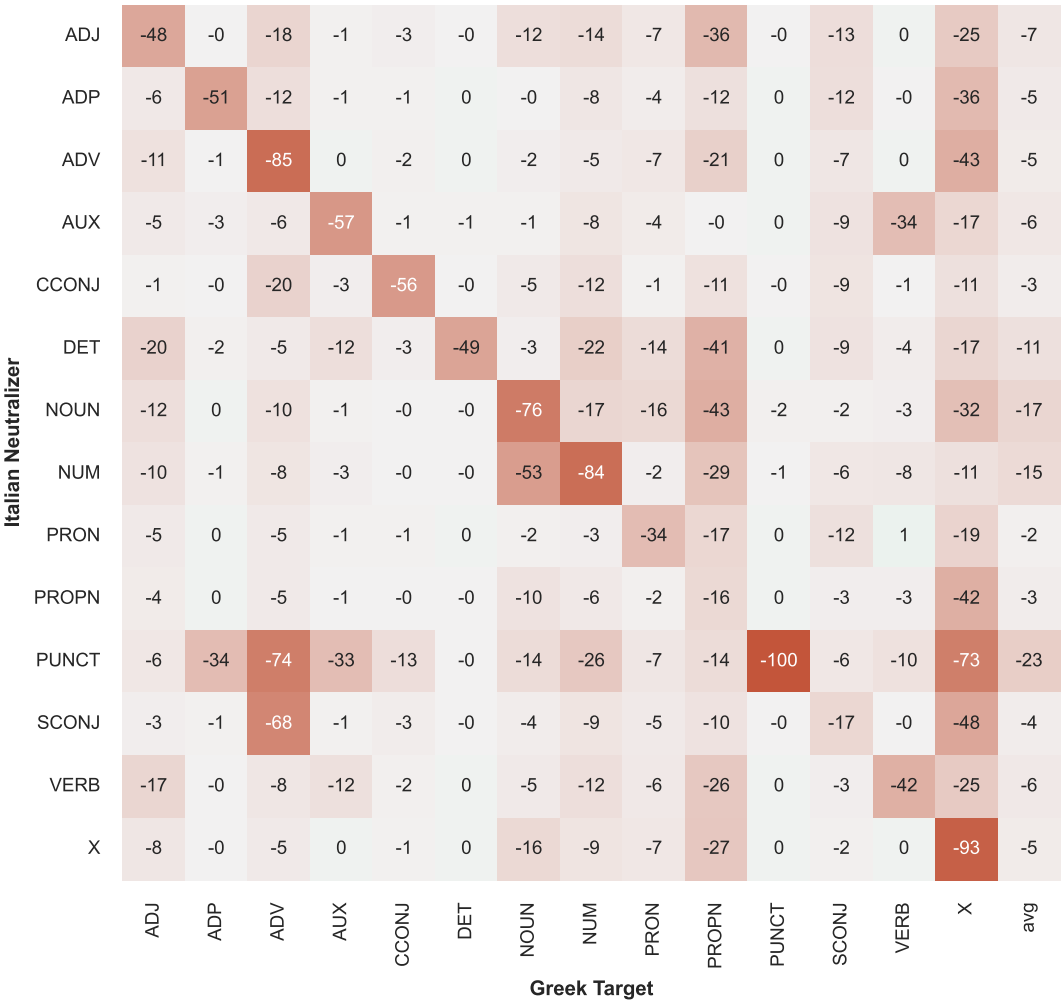


Figure 17: Relative change in accuracy when cross-neutralizing Greek POS tags using Italian XLM-R embeddings.

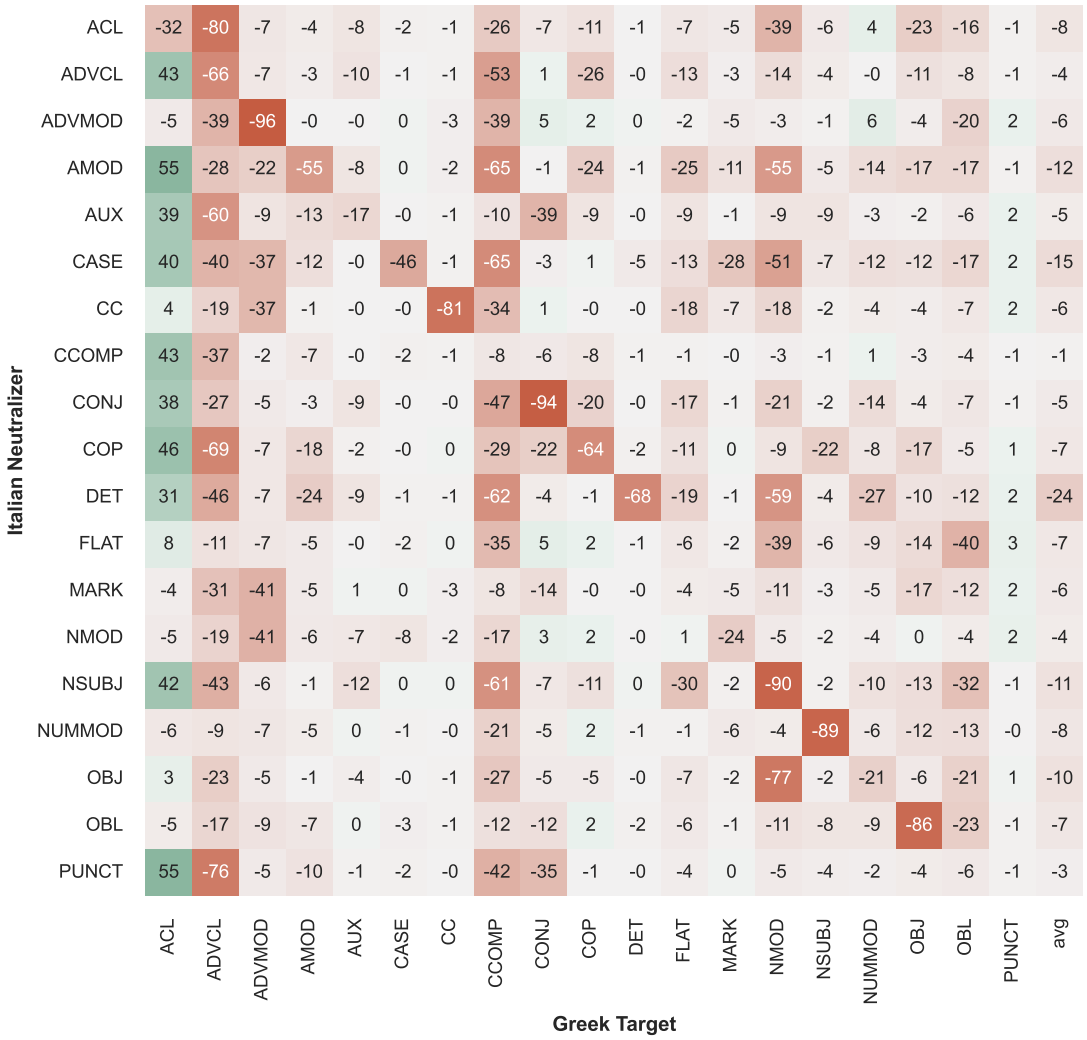


Figure 18: Relative change in accuracy when cross-neutralizing Greek DEP tags using Italian XLM-R embeddings.

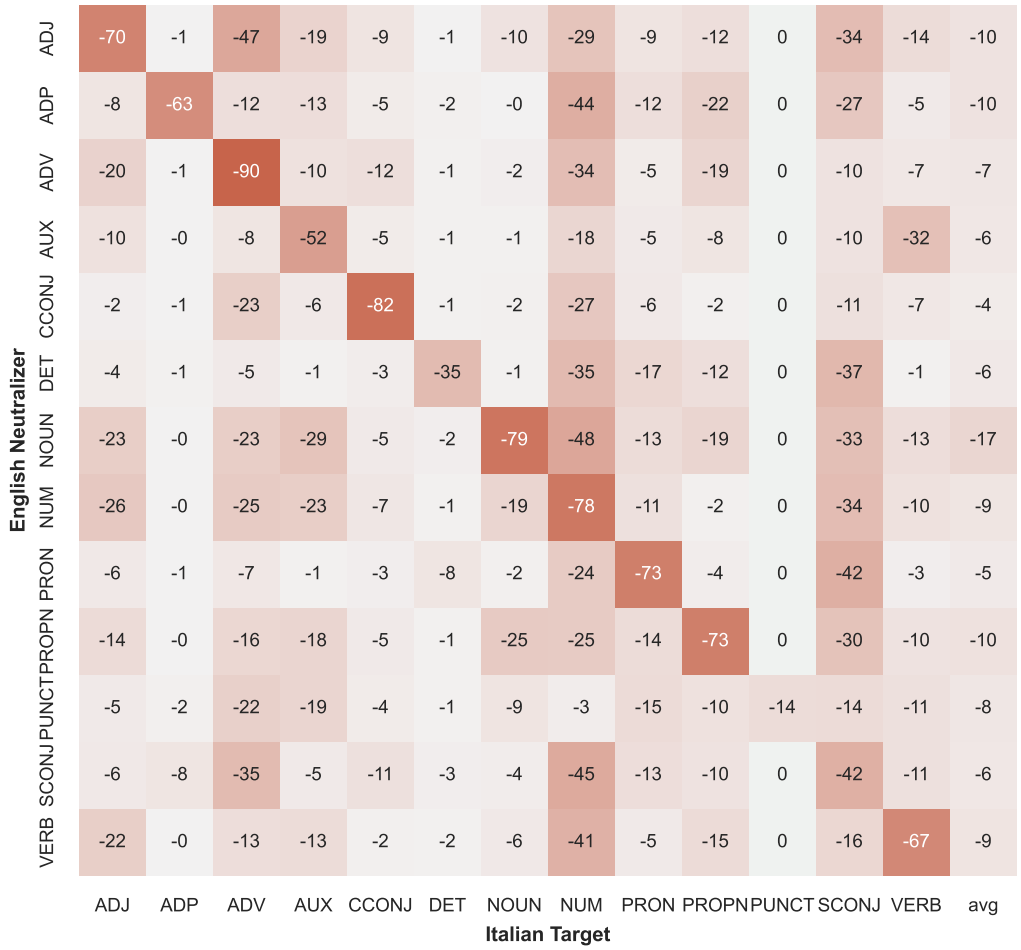


Figure 19: Relative change in accuracy when cross-neutralizing Italian POS tags using English XLM-R embeddings.

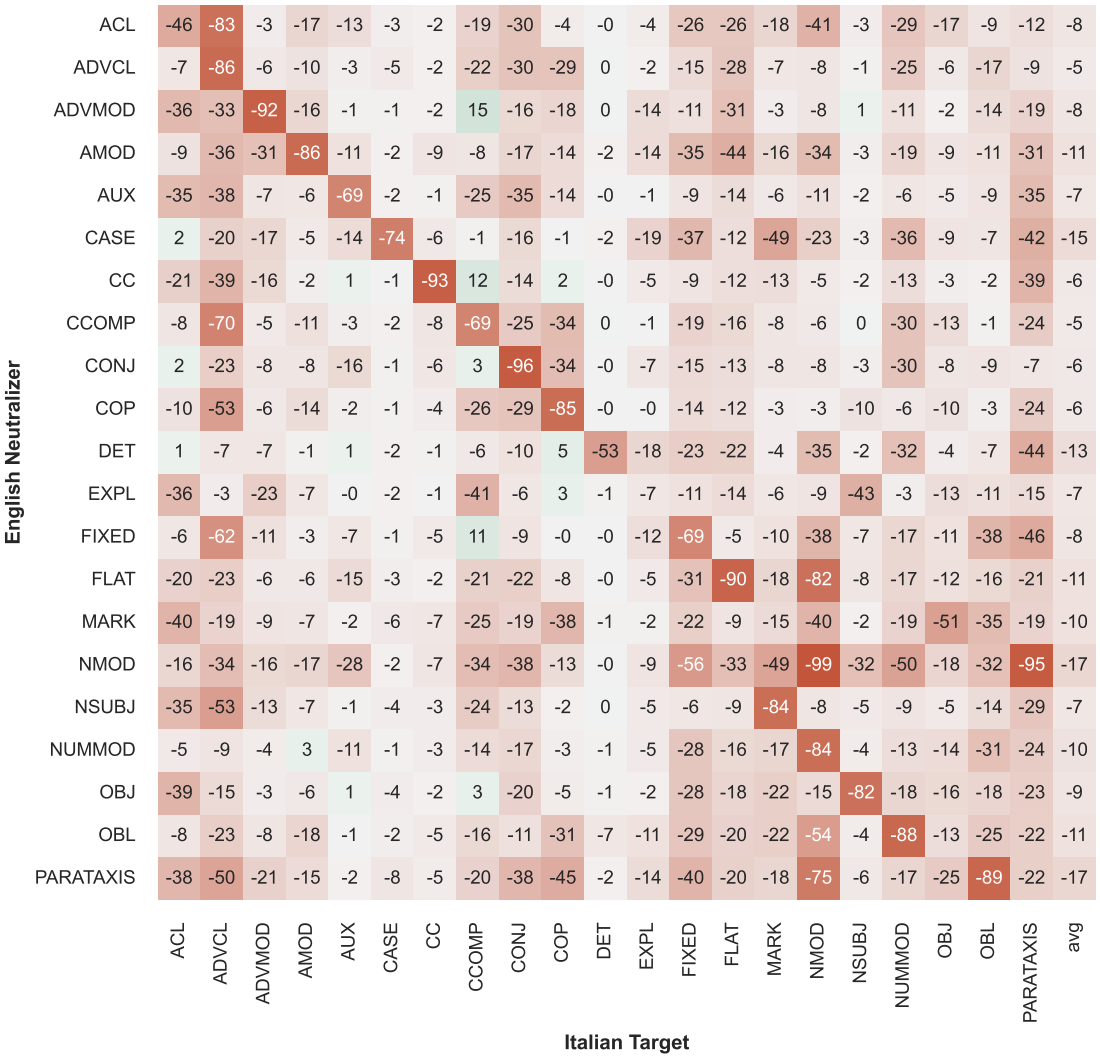


Figure 20: Relative change in accuracy when cross-neutralizing Italian DEP tags using English XLM-R embeddings.

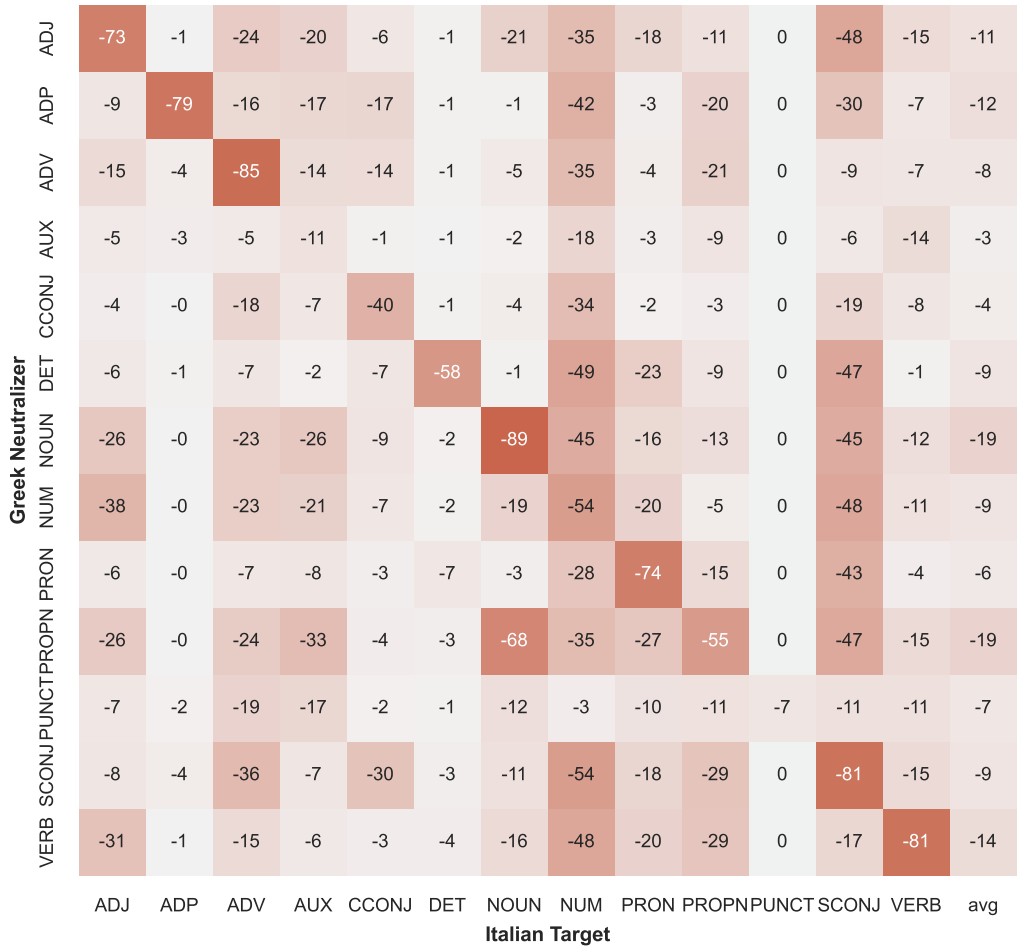


Figure 21: Relative change in accuracy when cross-neutralizing Italian POS tags using Greek XLM-R embeddings.

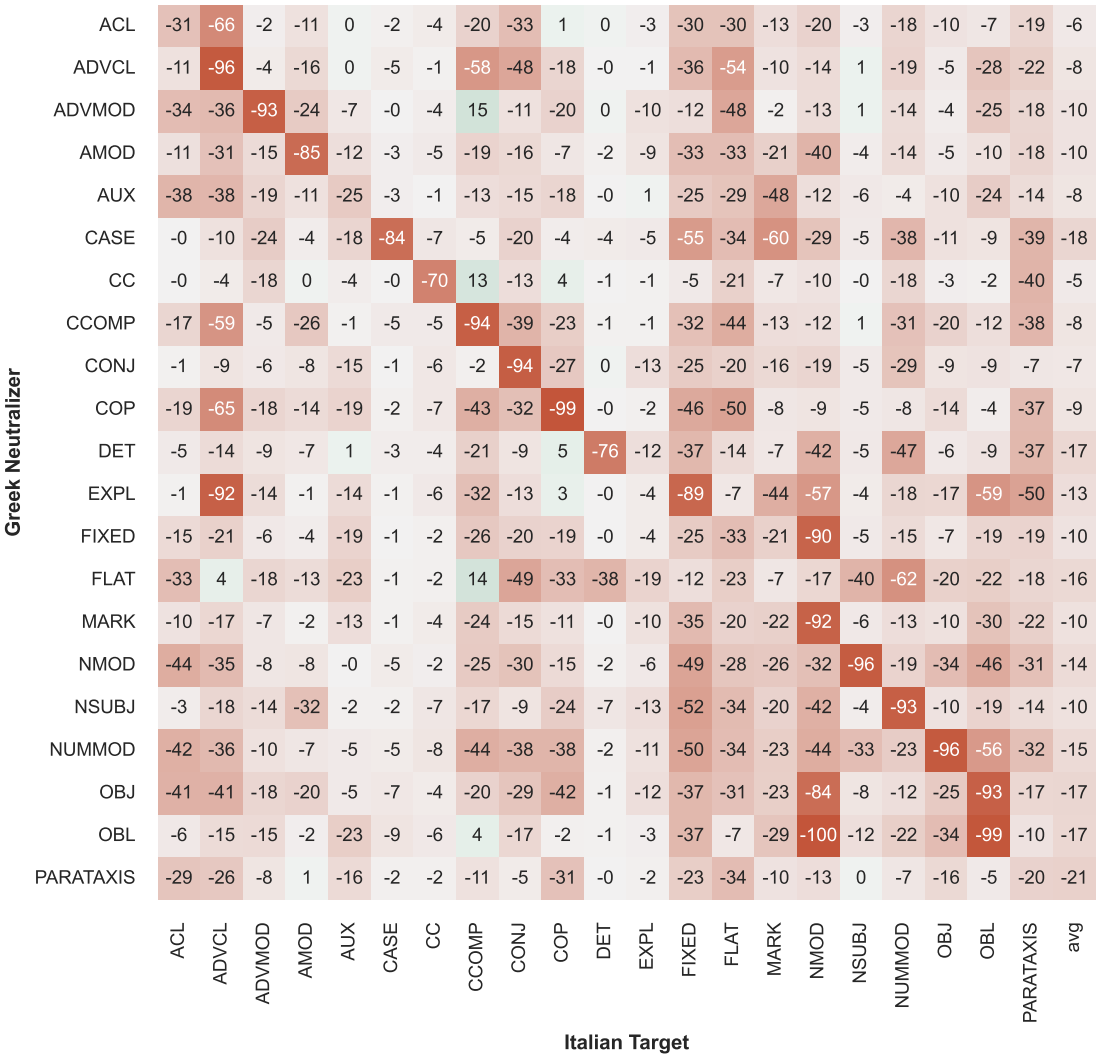


Figure 22: Relative change in accuracy when cross-neutralizing Italian DEP tags using Greek XLM-R embeddings.

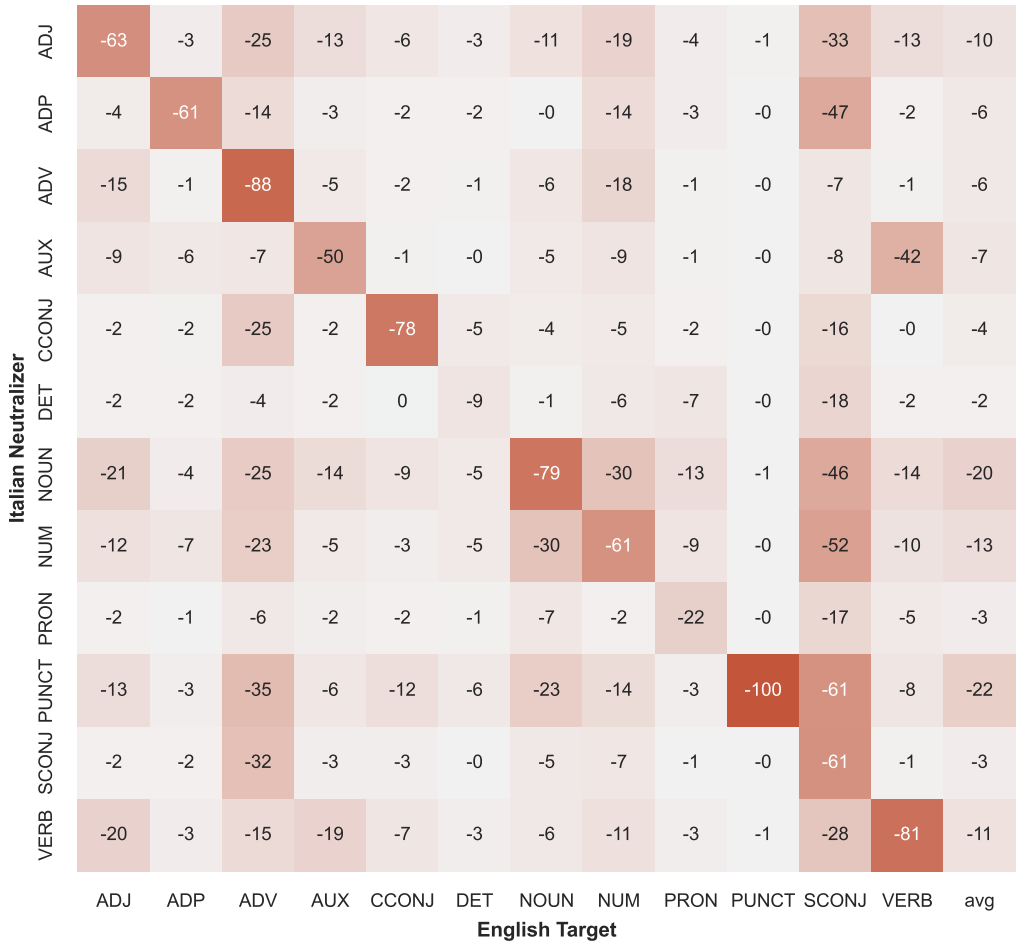


Figure 23: Relative change in accuracy when cross-neutralizing English POS tags using Italian XLM-R embeddings.

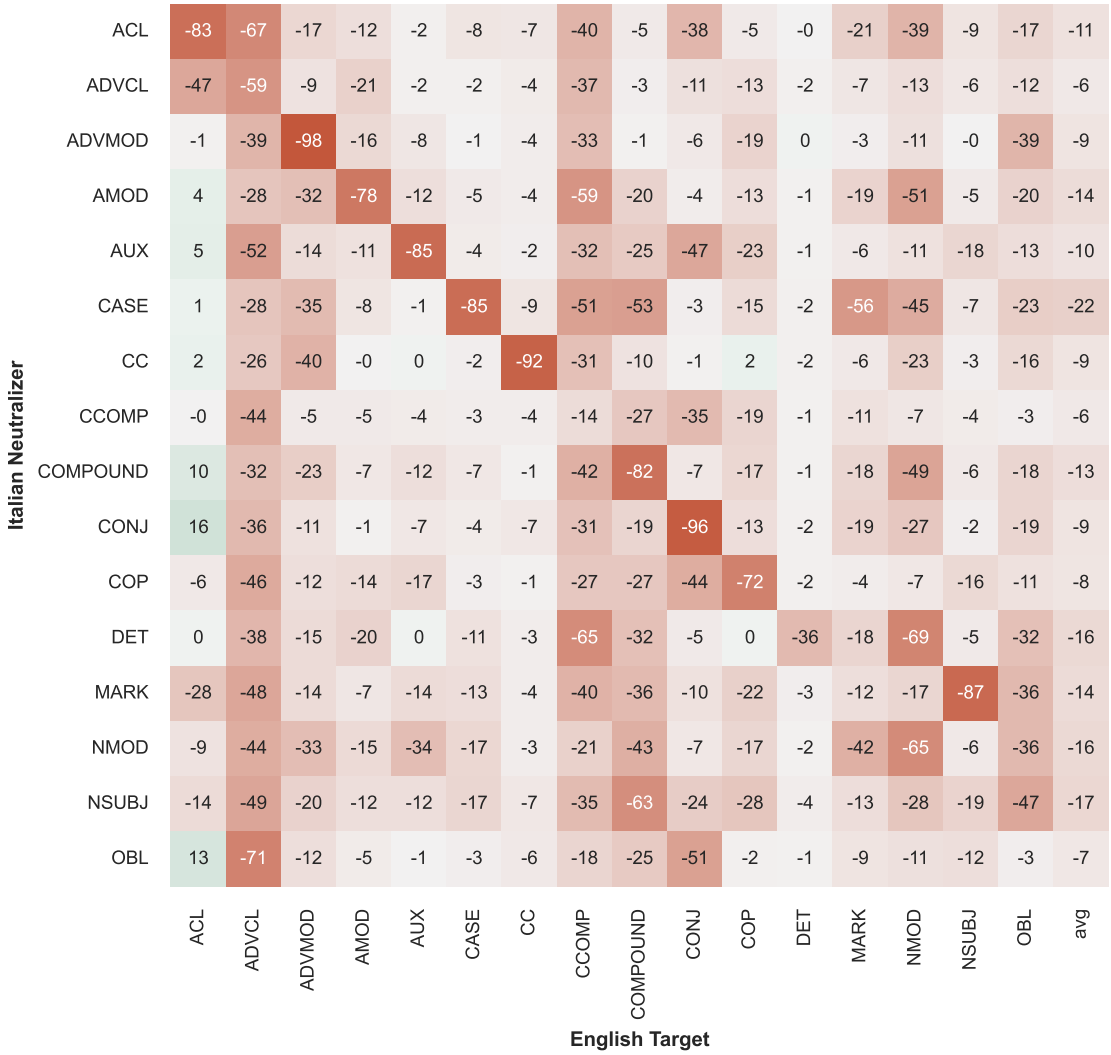


Figure 24: Relative change in accuracy when cross-neutralizing English DEP tags using Italian XLM-R embeddings.

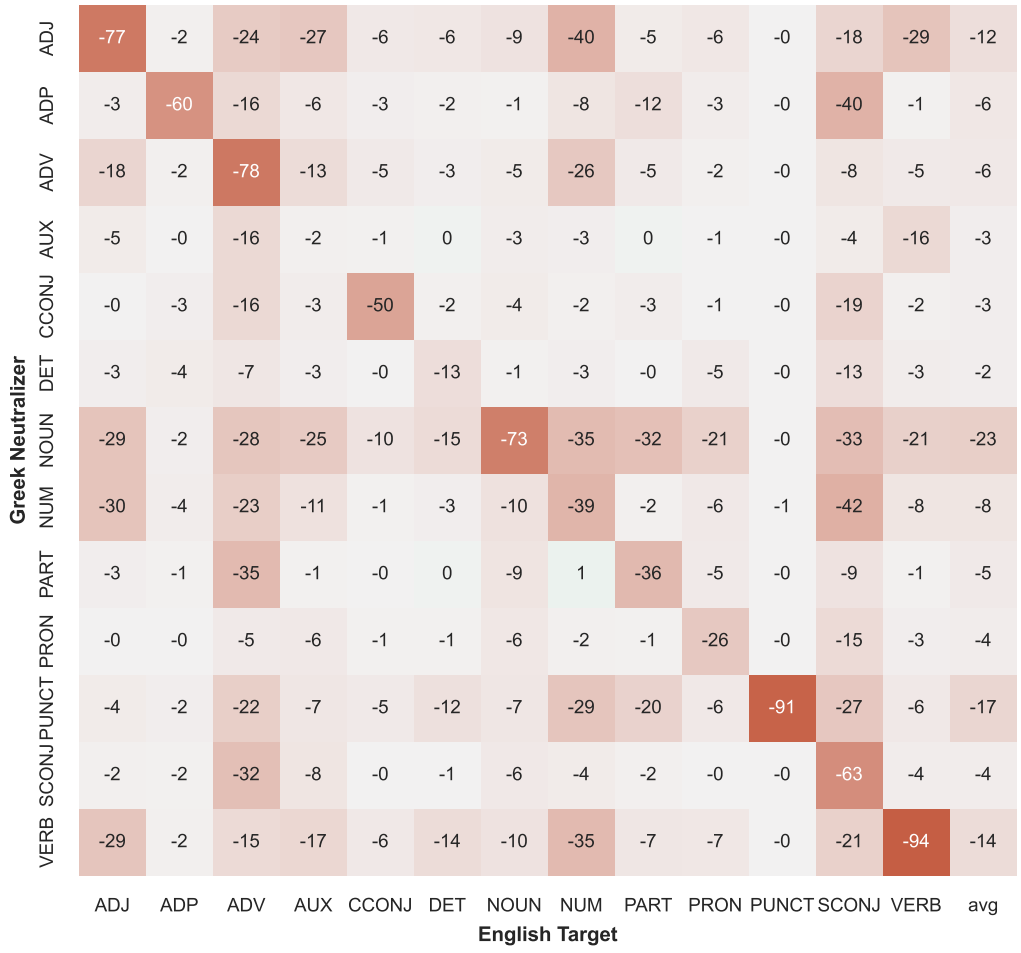


Figure 25: Relative change in accuracy when cross-neutralizing English POS tags using Greek XLM-R embeddings.

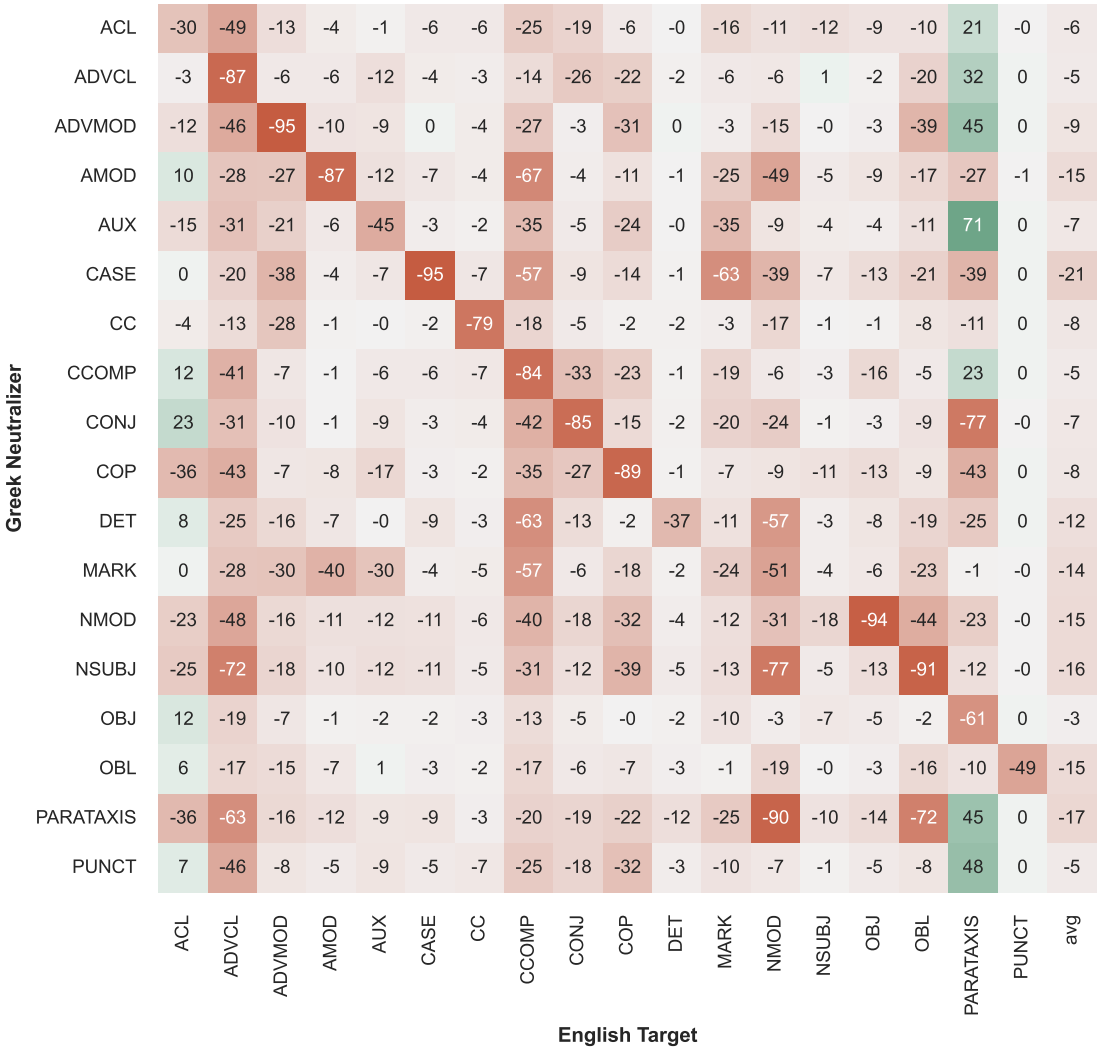


Figure 26: Relative change in accuracy when cross-neutralizing English DEP tags using Greek XLM-R embeddings.